ISSN 0280-526X: Working Papers
ISSN 1018-4554: ESCA/ETWR

Lund University
Department of Linguistics

General Linguistics
Phonetics

PROSODY
WORKSHOP

European Speech Communication Association

Working Papers
41. 1993

Edited by
David House and Paul Touati

Proceedings of an
ESCA WORKSHOP ON PROSODY
September 27-29, 1993
LUND, SWEDEN

Lund University
Department of Linguistics

General Linguistics
Phonetics

PROSODY
WORKSHOP

European Speech Communication Association

Working Papers
41. 1993

Edited by
David House and Paul Touati

Proceedings of an
ESCA WORKSHOP ON PROSODY
September 27-29, 1993
LUND, SWEDEN

ESCA WORKSHOP ON PROSODY
September 27-29, 1993
LUND, SWEDEN

Working Papers
Department of Linguistics and Phonetics
Lund University
Helgonabacken 12
S-223 62 LUND
Sweden
Fax +46 46 104210
email: lingf@lingf.lu.se

This issue was edited by David House and Paul Touati

ISSN 0280-526X: Working Papers
Dept. of Linguistics and Phonetics, Lund University
ISSN 1018-4554: ESCA/ETRW
European Speech Communication Association

Printed in Sweden, Reprocentralen, Lund University, 1993

Introduction

On the initiative of the European Speech Communication Association (ESCA) the area of *Prosody* has been selected for a Tutorial and Research Workshop (ETRW). The Department of Linguistics and Phonetics, Lund University, Sweden, agreed to organize an ESCA Workshop on Prosody in Lund, September 27-29, 1993. The ESCA Workshop on Prosody is related to two other events during 1993: thematically to the ELSNET Summer School on Prosody, University College London, and temporally as a satellite event to the 3rd European Conference on Speech Communication and Technology (EUROSPEECH '93), Technical University of Berlin.

Researchers working with prosody represent a large variety of disciplines and interests, ranging across theoretical linguistics and phonology, phonetic based models and descriptions, and applications such as text-to-speech and automatic recognition. The aim of the Workshop is to encourage an international scientific exchange of current work in prosody in much the same way as the Nordic Prosody Meetings function within Scandinavia. The response to our call for papers to the ESCA Workshop on Prosody has been overwhelming. Approximately 90 abstracts were submitted for review. In order to maintain the plenary nature of the three day Workshop we decided to restrict the number of contributions. The review of the abstracts was shared by the International Scientific Committee and the Organizing Committee. The final decision taken by the Organizing Committee in Lund resulted in the acceptance of around 70 contributions (oral and poster presentations). We would like to thank all of the contributors, and we regret that we were not able to make room for everyone who had intended submitting a paper to the Workshop.

The Proceedings of the ESCA Workshop on Prosody comprise the present volume of the Working Papers from the Department of Linguistics and Phonetics at Lund, which over the years have included a fair number of papers on prosody by a variety of authors. The most influential of these authors has been Professor Eva Gårding, who is the keynote speaker of the Workshop. The structure of this volume closely mirrors the structure of the Workshop. The Workshop has been organized into five oral sessions devoted to the following topics: Prominence, Grouping, Discourse, Perception of prosody, and Production, models and synthesis of prosody. Each of these sessions is opened with a tutorial paper by an invited speaker. In addition, three plenary poster sessions contain papers covering a variety of subtopics within prosody. Participants from more than 15 European countries, as well as from Canada, Japan and the USA, dealing with more than 20 different languages, contribute to the present volume.

We would like to thank the members of the International Scientific Committee: Mary Beckman, Dafydd Gibbon, Björn Granström, Nina Grønnum, Carlos Gussenhoven, Julia Hirschberg, Daniel Hirst, Robert Ladd, Jacques Terken, and Jacqueline Vaissière for their help in reviewing the submitted abstracts and for other advice concerning the Workshop. We would also like to acknowledge the generous support from the following organizations: COST 233, The European Speech Communication Association, The Einar Hansen Research Foundation, Lund University Faculty of the Humanities and Theology, The Elisabeth Rausing Memorial Foundation, The Wenner-Gren Center Foundation. Special thanks are due to the following persons connected to the Organizing Committee in Lund: Duncan Markham, Mechtild Tronnier, Lena Hörnlund and Yasuko Nagano Madsen, for their assistance in preparing the Proceedings.

Lund, August 1993

The Organizing Committee

Gösta Bruce, David House, Paul Touati and Britt Nordbeck

ESCA WORKSHOP ON PROSODY
September 27-29, 1993
LUND, SWEDEN

Organizing committee

G. Bruce, D. House, P. Touati, B. Nordbeck
Dept of Linguistics and Phonetics, Lund University (Sweden)

International scientific committee

M. Beckman, Dept of Linguistics, Ohio State University, Columbus, Ohio (USA)
D. Gibbon, Faculty of Linguistics and Literature, Univ of Bielefeld, Bielefeld (Germany)
B. Granström, Dept of Speech Comm and Music Acoustics, KTH, Stockholm (Sweden)
N. Grønnum, Dept of General and Applied Linguistics, Copenhagen (Denmark)
C. Gussenhoven, Dept of English, University of Nijmegen, Nijmegen (Netherlands)
J. Hirschberg, AT&T Bell Laboratories, Murray Hill, New Jersey (USA)
D. Hirst, Dept of Phonetics, University of Provence, Aix-en-Provence (France)
R. Ladd, Dept of Linguistics, Edinburgh University, Edinburgh (UK)
J. Terken, IPO, Eindhoven (Netherlands)
J. Vaissière, Dept of Phonetics, Sorbonne-Nouvelle-Paris III, Paris (France)

Sponsoring organizations

COST 233
The European Speech Communication Association
The Einar Hansen Research Foundation, Lund (Sweden)
Lund University Faculty of the Humanities and Theology, Lund (Sweden)
The Elisabeth Rausing Memorial Foundation, Lund (Sweden)
The Wenner-Gren Center Foundation, Stockholm (Sweden)

CONTENTS

ORAL SESSION ON PROMINENCE

ORAL SESSION ON GROUPING

POSTER SESSION ON SYNTHESIS & TTS DEMO

ORAL SESSION ON DISCOURSE

POSTER SESSION ON PROMINENCE, GROUPING & DISCOURSE

POSTER SESSION ON MODELS, PERCEPTION & APPLICATIONS

ORAL SESSION ON PERCEPTION

ORAL SESSION ON PRODUCTION, MODELS & SYNTHESIS

* Papers marked with an asterisk were not available for inclusion in this volume.

Notes on the Phonology of Prominence

D. Robert Ladd
Department of Linguistics
University of Edinburgh, Scotland

ABSTRACT

An explicitly phonological approach to prosody, of the sort provided by autosegmental and metrical theory, is necessary for understanding the relationships among different prosodic phenomena. Prominence is not a simple phonetic property of syllables but a complex phenomenon reflecting the metrical structure of a phrase or utterance, the location of (intonational) pitch accents, and the utterance's paralinguistic aspects.

INTRODUCTION

It is difficult to give a coherent brief overview of the topic "prominence", because the topic itself is not one but several. A quick glance at the six workshop papers grouped together under this rubric shows that we are dealing with a many-headed beast: one (Cruttenden) is about the pragmatics of sentence-stress, one (Gronnum) deals with the relation between stress and rhythm, one (Grabe et al.) studies the use of certain English-specific prosodic cues in human sentence-processing, and one (Hermes & Rump) explores the contribution of pitch range to perceived prominence. Only two - Campbell's and Fant & Kruckenberg's - deal with something that is clearly the same topic, namely the role of a variety of acoustic cues, especially duration cues, in signalling both prominence and phrase boundaries. Given this variety, it is pointless to try to discuss each of the six papers and relate them to a single core of theoretical issues and ideas. Instead I would like use them as points of reference in an argument for approaching the study of prominence - and indeed, prosody in general - in explicitly phonological terms.

Specifically, I wish to argue that the theoretical framework provided by autosegmental and metrical phonology (Liberman & Prince 1977, Pierrehumbert 1980, and much work since then; for a review relevant to intonation see Ladd 1992) is essential to reconciling the diversity of methodologies and points of view of the papers in this section. I hasten to add that I am more or less agnostic about many specific issues within autosegmental and metrical phonology, and those are not my topic here. My point in invoking the general autosegmental/metrical approach is to suggest that it provides an appropriate way of thinking about the *relationships* among prosodic phenomena, and between prosodic form and function.

By this I mean two things. First, observable acoustic properties like F0, duration, and intensity are not direct correlates of functional categories like focus, nor the direct realisation of morphosyntactic structures. They are rather the correlates of phonological categories and phonological structures, and as such may only indirectly reflect focus, phrase boundaries, and so on. I shall return to this point below. Second and more specifically, what I find important about the autosegmental/metrical point of view is the idea that stress, duration, rhythm, and prosodic grouping form one coherent cluster of phenomena, and intonation forms another. This is what is implied by a representation of the sort in Fig. 1. The prominence relationships and surface constituency - stress and phrasing, in other words - are represented in the metrical tree, and the intonation is represented in the tonal string. At the risk of being accused of Procrustean misrepresentation, I think that many of the observations in Gronnum's paper are based on just such an implicit distinction between duration, stress, and foot structure on the one hand, and pitch features and rules for their alignment with prominent syllables on the other.

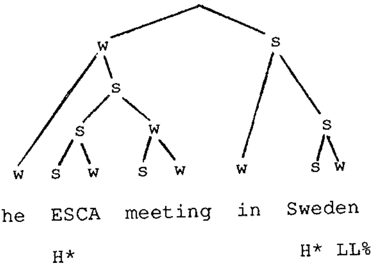


Fig. 1. *An autosegmental/metrical representation of the phrasing and prominence structure (metrical tree) and intonation contour (tonal string) of a simple utterance.*

I suspect that many participants in the symposium will be rather skeptical about such phonological representations. Metrical trees and autosegmental tone strings are neither hard acoustic data nor clear functional correlates of acoustic data. Moreover, they seem subject to lots of changing fashions, and the empirical constraints on the changes of fashion are often obscure to people whose methodological biases are experimental rather than theoretical and whose primary concern is, say, producing the undcrstanding that will lead to better speech technology. With this audience in mind, then, let me proceed to some concrete illustrations of my general theme.

PHONOLOGY AND PHONETICS

First, consider the question of cues to prominence. Fry's classic experiments of the 1950's (e.g. Fry 1955) gave rise to a received view that the acoustic correlates of stress are F0, duration, and intensity - generally in that order, but not always, to the long-standing confusion and frustration of many phoneticians. As long as stress was taken to be a fairly uncomplicated phonetic category applying to individual syllables, then it made sense to try to find measurable acoustic properties of syllables that relate directly to that category. Unfortunately, the received view based on Fry is about the best we can do if we approach the question that way. By accepting that we are dealing with a complex phonological structure, rather than a straightforward phonetic property of syllables, we begin to unravel some of the remaining confusion.

Specifically, once we accept the autosegmental/metrical notion that intonation is essentially separate from prominence and rhythmic structure, we see that the reason F0 is curiously unreliable as an "acoustic correlate of stress" is that it is in the first instance an aspect of *intonation*. Pitch accents are intonational features; they are anchored to prominent syllables, but prominent syllables need not have a pitch accent. When a pitch accent is present, it is an unmistakable signal of prominence, but prominence is quite detectable without the help of pitch accents. Once we see beyond the statistically overwhelming effects of pitch accent on stress judgments, we realise that duration and intensity are considerably more important than Fry's work made it appear. This has been a consistent finding of more recent work, including the papers by Campbell and by Fant & Kruckenberg.

Campbell's paper illustrates a different sense in which it is useful to get beyond a simple phonetic taxonomy of prosodic phenomena. Campbell shows that it is possible to distinguish increased duration due to prominence from the added duration due to pre-boundary lengthening. It has already been shown (Edwards & Beckman 1988) that this distinction is observable in the kinematics of speech production, but Campbell shows that the acoustic correlates are distinct as well. The key to Campbell's findings is not to state the acoustic data simply as raw durations of segments or syllables. It would admittedly be stretching things to claim that Campbell's notion of "duration contour" is inspired by metrical phonology, but metrical phonology is at least consistent with Campbell's leading idea that duration has a complex structure, and that by discovering that structure we make it possible to identify prominence acoustically without reference to F0.

PHONOLOGY AND FUNCTION

A more general consequence of a phonological approach to prosody is that it makes us less likely to look for direct links between phonetic form and communicative function. Everybody knows that prosodic features have a bewildering variety of "functions" - signalling focus, emphasis, phrasing distinctions, lexical distinctions, speaker attitude, and many more. To my mind, this variety of functions is bewildering only if we assume that the link between sound and meaning is direct.

By way of illustration, consider the "delimitative" cues to dividing up the stream of speech into chunks corresponding to words and phrases. There are many delimitative cues in segmental phonology, but these are essentially accidental. For example, in American English one can clearly distinguish the utterances *can't race* and *can trace* by the allophonic variation in the three segments /n/, /t/, and /r/, but it is not very revealing to say that the "function" of segmental allophonic variation is delimitative. Segmental allophony exists (for whatever reason), and it may be exploited by listeners when it happens to provide "delimitative" information.

In the same way, I believe that if we find prosodic features playing a delimitative or otherwise disambiguating role in sentence processing, we should always assume that this role is a useful accident, and only later consider the possibility that delimitation is the central function of a given prosodic feature. This is relevant to the paper by Grabe et al. I find it unremarkable that listeners can distinguish the unshifted stress pattern of *Chinese teacher* 'teacher of Chinese' from the shifted pattern of *Chinese teacher* 'teacher who is Chinese', even when they have heard only the first syllable, or that listeners should make syntactic use of that information as soon as they get it. These findings tell us about human sentence processing, not about the function of prominence. That is, Grabe et al.'s results tell us how quickly listeners can make use of information from a variety of sources, not that the "function" of stress shift is syntactic disambiguation. Like segmental allophony, stress shift exists - apparently for reasons of maintaining as regular an alternating speech rhythm as possible. In the specific case of *Chinese teacher,* stress shift happens to have an effect that we can put to use in syntactic processing, just like the allophony of /n/, /t/, and /r/. As far as I am aware, Grabe et al. interpret their findings in more or less the way suggested here, but I have discussed this example at length because it is precisely the sort of finding that might be taken as evidence for the essentially delimitative function of prominence, or for a "functional" explanation of stress shift. I do not think that such interpretations are worth pursuing.

If the linguistic function of prominence is not delimitative, what is it? In my opinion the phenomena dealt with in Cruttenden's paper are closer to the essence of what prominence is for. Prominence is like a grammatical category, similar to number or case. As with number or case, there are broad similarities across languages in the way prominence is distributed, but also language-specific pragmatic, syntactic, and phonological principles. Cruttenden's paper discusses a couple of specific points on which languages differ. I find this work very useful as a corrective to the idea that, as a universal of intonation systems, pitch prominences go on focused or otherwise emphasised words. Focus may help to govern the distribution of pitch accents, but focus can apply to whole constituents, and accents, by definition, must be associated with individual words. We must therefore assume that phonological and other rules come into play here in building the prominence structure of an utterance on the basis of the intended focus (cf. e.g. Gussenhoven 1983, von Stechow & Uhmann 1986). The simple idea of a direct correspondence between pitch accent and some intuitive notion of "focus" on individual words does explain a substantial percentage of cases, but it ignores the language-specific differences, and gets in the way of understanding the remaining percentage of cases that don't fit the simple pattern.

PHONOLOGY AND PARALINGUISTICS

So far I have suggested that a phonological perspective on prominence will keep us from looking for too direct a link between simple phonetic properties and communicative functions. In the remainder of the paper I wish to explore a third puzzling area of prosody where I believe we can make progress by considering the phonological implications of experimental findings. This is the problem of "gradience".

Prosody appears to differ from the rest of phonology in the way it treats continuous acoustic dimensions such as pitch range and duration. Instead of such acoustic continua being divided up into discrete categories, as is usual in phonological organisation, at least in some cases gradual acoustic changes can result in steadily perceptible non-categorical shades of meaning. This is clearly the way a lot of paralinguistic signalling works - the broader the smile, the happier the smiler - but it is not very common in language proper, except in prosody. Perhaps the most common example is one directly relevant to prominence, and well-documented in the papers by Fant & Kruckenberg and by Hermes & Rump: the gradual increase in emphasis or contrastiveness associated with a gradual increase in overall pitch range or with gradual increases in intensity or duration. Such gradient prominence is allowed for even in rigorously phonological descriptions. Within the autosegmental/mctrical approach to prosody, both the degree of emphasis of individual accents within phrases and the overall pitch range of individual phrases within utterances are normally said to be freely variable.

I have argued elsewhere (Ladd forthcoming) that the concept of free gradient variability of prominence poses serious theoretical and empirical problems. Here I wish to go beyond my earlier arguments and propose that much of what is taken as evidence for gradient prominence of individual words actually reflects paralinguistic differences of *overall* prominence, emphasis, interest, etc., which affect the perception of the most prominent word. That is, when listeners are presented with a short test utterance under experimental conditions and asked to judge the degree of prominence of the accented word, what they are really doing is judging the degree of emphasis *of the utterance as a whole* and applying that judgement to the accented word. This is exactly analogous to something that Fant & Kruckenberg report for word prominence and syllable prominence: they found that judgements of a word's degree of prominence correlate highly with judgements of the prominence of the word's stressed syllable. I therefore claim that for the most part prominence really is a largely categorical, phonological matter: this is what is implied by the presence or absence of pitch accent in the tonal string. Gradient prominence on an accented word is simply gradient overall prominence or emphasis, which the speaker, under experimental conditions, interprets as affecting the accented word. I disagree with the notion that the prominence of every individual accent can vary freely; I take quite literally the categorical nature of relative prominence implied by the autosegmental/metrical representation.

Of course, in some cases individual words really *do* have extra emphasis or prominence within a given utterance, like the word *do* in a colourful reading of this sentence. In this case the presence of extra emphasis is signalled categorically by the presence of extra intensity, extreme pitch range, and so on. This claim is also supported by Fant & Kruckenberg: they found that intensity comes into play as a correlate of perceived degree of prominence only for "emphatic" accents. Implicitly, that is, they draw a categorical rather than a gradient distinction between those accents for which intensity is relevant and those for which it is not. I believe this is correct: I think there is a fairly discrete boundary between a neutral reading like *He's Ukrainian* (interpreted as providing new information about the subject), and a paralinguistically marked emphatic reading *He's UKRAINIAN* (interpreted as emphasising the adjective, e.g. to contrast with erroneously presupposed *Russian)*. More generally, I think that many paralinguistic dimensions - certainly including local emphasis - have a "neutral" range, and it is only beyond this neutral range that classic gradience sets in.

If this general interpretation of gradient prominence is to hold up, we need to distinguish paralinguistic effects on an utterance as a whole from paralinguistic effects on individual words - and to distinguish both from phonological distinctions of relative prominence. I believe this can be done. To begin with, we have the results of several studies on pitch range in at least three different languages (Liberman & Pierrehumbert 1984 on English, Bruce 1982 on Swedish, Pierrehumbert & Beckman 1988 on Japanese, and others). These studies all find that the relative pitch range of individual accents in a phrase and of individual phrases in an utterance remains *virtually constant* when the overall pitch range of the utterance is experimentally varied for paralinguistic reasons (e.g. by getting subjects to "speak up" or to talk as if they were "more involved"). That is, gradient variability generally seems to affect the pitch range of utterances or larger chunks of discourse as a whole without affecting the pitch range relations within the utterance. This is very difficult to explain if the pitch range of each accent and each phrase is freely and independently variable, but makes sense if the relative prominence of accents within an utterance is tightly constrained by the phonology and not controlled paralinguistically.

More concretely, Ladd, Verhoeven and Jacobs (forthcoming; henceforth LVJ) have provided experimental evidence that accentual prominence is perceived in a way consistent with the proposal just made. Their study replicates and extends an earlier discovery by Gussenhoven & Rietveld (1988). Gussenhoven & Rietveld found that, in an utterance with two accent peaks (e.g. a sentence like *Her mother's a lawyer)*, a decrease in pitch range on the first accent causes a *decrease* in the perceived prominence of the *second* accent. If the degree of emphasis on accents were independently variable, one would expect the opposite effect - i.e. one would expect a decrease on one accent to enhance the prominence of the other. LVJ explain Gussenhoven & Rietveld's finding by suggesting that, for moderate peak heights, gradient variability applies to the pitch range *of the utterance as a whole*, not to each accent individually. This means that lowering one accent lowers the perceived degree of overall emphasis of the utterance, and hence the degree of emphasis on all the accents of the utterance. This is again the utterance level analogue of the word/syllable effect found by Fant & Kruckenberg.

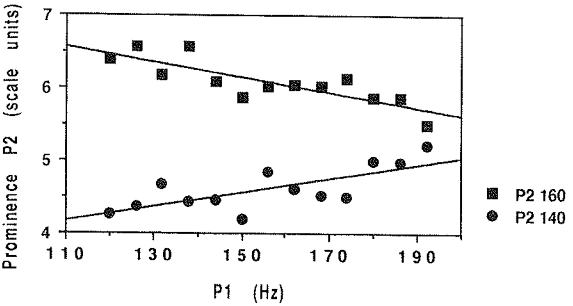


Fig. 2 *Perceived prominence of the second accent peak of a two-accent utterance, as a function of the F0 of the first accent peak. The two F0 levels of the second peak (140 and 160 Hz) show different effects. From Ladd, Verhoeven and Jacobs (forthcoming).*

However, LVJ also found that Gussenhoven & Rietveld's effect is reversed (and the more expected effect obtained) when the second accent peak is *relatively high*. This is shown in Fig. 2. LVJ suggest that this reversal reflects a categorical difference between "normal" or "neutral" accent peak height - with which pitch range is evaluated globally - and "emphatic" peak height - which overrides global pitch range and treats every accent in its own right. If LVJ are correct, gradient variability applies accent-by-accent only beyond the threshold of the emphatic. This, as noted earlier, is quite consistent with Fant & Kruckenberg's implicitly categorical distinction between emphatic and non-emphatic accents.

Obviously, the most rigorous test of the idea that there is a categorical distinction between neutral and emphatic accents will come from something akin to a categorical perception experiment. A promising pilot study along these lines has just been done under my supervision by Rachel Morton, as the basis of her undergraduate honours dissertation in Edinburgh. Under certain conditions Morton's results seem to show a stepwise increase in perceived prominence as overall pitch range increases, suggesting a categorical shift from neutral to emphatic. But further work needs to be done before this result can be regarded as established.

REFERENCES

G. Bruce (1982), "Developing the Swedish intonation model", *Working Papers* (Lund University Dept. of Linguistics) no. 22, pp. 51-116.

J. Edwards & M. E. Beckman (1988), "Articulatory timing and the prosodic interpretation of syllable duration", *Phonetica* 45, pp. 156-174.

D. B. Fry (1955), "Duration and intensity as physical correlates of linguistic stress", *Journal of the Acoustical Society of America* 27, pp. 765-768.

C. Gussenhoven (1983), "Focus, mode, and the nucleus", *Journal of Linguistics* 19, pp. 377-417.

C. Gussenhoven & T. Rietveld (1988), "Fundamental frequency declination in Dutch: testing three hypotheses", *Journal of Phonetics* 16, pp. 355-369.

D. R. Ladd (1992), "An introduction to intonational phonology", in *Papers in Laboratory Phonology II: Segment, Gesture, Prosody,* ed. by G. J. Docherty & D. R. Ladd (Cambridge University Press, Cambridge), pp. 321-334.

D. R. Ladd (1993), "In defense of a metrical theory of intonational downstep", in *The Phonology of Tone: The Representation of Tonal Register,* ed. by H. v.d.Hulst and K. Snider (Mouton DeGruyter, Berlin), pp. 109-132.

D. R. Ladd (forthcoming), "Constraints on the gradient variability of pitch range (or) Pitch Level 4 Lives!", to appear in *Papers in Laboratory Phonology III,* ed. by P. Keating (Cambridge University Press, Cambridge).

D. R. Ladd, J. Verhoeven, & K. Jacobs (forthcoming), "Influence of adjacent pitch accents on each other's perceived prominence: two contradictory effects", to appear in *Journal of Phonetics.*

M. Liberman & J. Pierrehumbert (1984), "Intonational invariance under changes in pitch range and length", in *Language Sound Structure,* ed. by M. Aronoff and R. Oerhle (MIT Press, Cambridge MA), pp. 157-233.

M. Liberman & A. Prince (1977), "On stress and linguistic rhythm", *Linguistic Inquiry* 8, pp. 249-336.

J. Pierrehumbert (1980), *The Phonology and Phonetics of English Intonation,* PhD Dissertation, MIT.

J. Pierrehumbert & M. Beckman (1988), *Japanese Tone Structure* (MIT Press, Cambridge MA).

A. von Stechow & S. Uhmann (1986), "Some remarks on focus projection", in *Topic, Focus, and Configurationality,* ed. by W. Abraham & S. de Meij (Benjamins, Amsterdam), pp. 295-320.

The De-Accenting and Re-Accenting of Repeated Lexical Items

Alan Cruttenden
Department of Linguistics
University of Manchester
Manchester M13 9PL

ABSTRACT

The phenomenon of the de-accenting of old or 'given' information has often been taken for granted as an intonational universal, despite hints to the contrary in the literature. An attempt is made to construct a test of such de-accenting and apply it to eight languages. It is found that de-accenting and re-accenting vary according to discourse structure and to language.

INTRODUCTION

It is regularly assumed that old or repeated information is not accented in discourse. One of the most quoted statements to this effect is in Halliday (1967: 23): 'Marked tonicity occurs, in general, under either (or both) of two conditions. Either some element other than the one just specified [= the last lexical item in the tone-group] is "contrastive"; or the element just specified (and possibly others before it) is "given"—has been mentioned before or is present in the situation.' The archetypal case of "given-ness" involves a repeated lexical item, e.g. *The stadium where Manchester United play is in the EAST of Manchester* In this example *Manchester*, as a repeated item, is de-accented and the accent is thrown back onto the word *east*.

Such de-accenting is certainly the norm for the type of English which Halliday was describing (roughly RP) and for most other dialects of English. And it is easy to think of de-accenting as some sort of cognitive universal: we do not wish to re-accent repeated information because, in Chafe's (1974) terms, it is already in the consciousness of the speaker. Yet obligatory de-accenting of this sort (for it is obligatory in English) may not be as universal as first thought. There have been hints in the literature to this effect, e.g.

(a) Crystal (1975: 44) on Brazilian Portuguese: 'This tendency [to keep the tonic syllable on the last item in the tone-unit] applies even when one has repeated items in co-ordinate constructions', e.g.

Esti livro custa cinco dolares e esti aqui tres DOlares.

(b) Ladd (1990) gives examples from Roumanian, e.g.

[...o să vedem] ce aveţi şi ce nu AVEŢI
SUBJ we.see what you.have and what not YOU.HAVE

(c) Vanderslice & Pierson (1967) give examples from Hawaian English, e.g.

Forty t'ree per cent is gavment owned, and fifty seven per cent is privately OWNed.

In more phonetically- rather than informationally-oriented research Gårding (1981) elicited Greek, Swedish and French versions of the sentence *Madame Marianne Mallarmé has a mandolin from Madrid* in a context which called for a focus on *Mallarmé*. She reported that 'In Greek and French the pitch contour is flattened after focus. For Swedish, on the other hand, the situation is different. Here the accents still have their pitch configurations after focus'. The reason for this was suggested to lie in the lexical status of the accents in Swedish. The hints in the literature above suggest that it might not be only in languages with lexical accents that post-focal accents are unflattened.

With such comments in mind I constructed an intonational test of 13 setting-response pairs where the response involved a lexical item repeated from the setting. The pairs have been loosely translated into 8 languages, and pairs of native speakers of the languages (varying in number from 1 to 14) have read and recorded the setting-responses. I have analysed all the responses auditorily and a selected sample instrumentally. The languages and number of speakers are as follows: French 15, Italian 12, Tunisian Arabic 4, Russian 2, Greek 1, Macedonian 1, Swedish 1, Spanish 1. I have further, as yet unanalyzed, data on Albanian 4, German 1, and Lithuanian 1. It is of course obvious that this is in the nature of a prototype study of intonational typology and no statistical validity can yet be claimed for the results. I present here a sample of the responses in a number of languages and then some of the general conclusions.

SETTING-RESPONSE 1

The English version of the first example is as follows:

A: If you don't hurry up, you'll be late.
B: I don't care if we are late.

In this example *late* , a repeated item, and *we,* with identical, though shifted, reference to *you,* are treated as old information in English and are hence de-accented. There are two alternatives for the placement of the accent, depending on the reduction or non-reduction of *we're* or *we are.* With reduction the accent will be on *care* hence *I don't CARE if we're late;* without reduction the accent may be on *are,* hence *I don't care if we ARE late.* The latter accentuation emphasizes the positive polarity of the outcome. When we look at other languages, we find that most languages behave similarly to English. Indeed almost all my varying numbers of informants de-accented. In French where I had 14 informants for the sentence *Je m'en fiche d'être en retard,* 11 put the last accent on *fiche* and only 3 re-accented *retard.* In Spanish I had only one informant who, in the sentence *No me importa si llegamos tarde,* re-accented *tarde.* However two further facts make the Spanish informant important. One is that my Spanish informant re-accented in every response pair that I tested. The second is that a Ph.D. thesis (Ortiz-Lira, 1993) is about to be presented in Manchester which finds re-accenting as the norm for other Spanish informants who read this sentence. One speaker of Russian and Tunisian Arabic re-accented but no significance can be attached to this. The overall conclusion for this setting-response pair is that it produces de-accenting in all the languages tested with the almost certain exception of Spanish.

SETTING-RESPONSE 2

The next example is as follows:

A: I make the answer sixteen point one.
B: Well, I make it twenty-six point one.

This example is different from the preceding one in that it might be said to involve a positive accentuation for contrast as well as the de-accenting of old information. In this example the *point one* is de-accented in English and the accent in the response is on *six.* In French the sentence *Eh bien ma réponse est vingt six virgule un* produced de-accenting with the accent on *six* from 12 of the 15 informants and re-accenting with the accent on *un* from 3 informants. The Italian sentence *A me invece ventisei virgola uno* produced de-accenting on *ventisei* from 8 informants and re-accenting with the accent on *uno* from 4 informants. Otherwise there was one case of re-accenting in the Tunisian Arabic data and the single Macedonian and Spanish speakers both re-accented. As said in the previous discussion the Spanish data is the only one of these last three to which importance should be attached. Overall this setting-response again generally produces de-accenting. But the cases of re-accenting have grown because, while the situation in French and Spanish remains the same as in setting-response 1, there is now a substantial minority in Italian using re-accenting.

SETTING-RESPONSE 3

The next example concerns the reading of football results (for a detailed exposition of this topic in English see Cruttenden, 1974). The setting-response is as follows:

A: What was the score?
B: Liverpool 1 (one), Manchester United 1 (one).

This example is clearly different from the preceding ones: the old information is now completely within the response. In English the repetition of the word *one* produces de-accenting and hence the last accent is thrown back onto the preceding word *United.* Now it might be objected that this is a very specialised type of intonational context but in fact it is not. The same sort of correlated construction produces sentences like *John has two and MARY has two.* and many similar. What happens in other languages? Firstly it has to be said that the test does not work (as I found out too late) in all languages, because a diffferent way of reading results on radio and television is often used corresponding to *Liverpool-Manchester United 1-1.* This applied in Greek, Macedonian and Russian where the informants said that they could not possibly read results in my way. So these languages do not enter the present comparison. (Incidentally my lone Swedish informant sent me the data and fro some reason chose to alter the scores to 3-2 so that data was not relevant either!) For those remaining languages where the test was legitimate, the findings were quite unambiguous: only one informant (Italian) out of the whole total used de-accenting. What this means is that de-accenting appears to be impossible with this setting-response for French, Italian, Spanish, and Tunisian Arabic, e.g. the Tunisian Arabic production was always:

l-ʔafrīqī	waḥd	u-t-tarājī	WAḤD
Afriqi	one	Taraji	one

and in Italian: *Inter uno Roma UNO*

SETTING-RESPONSE 4

The last type of setting-response to be discussed here involves pairs like:

A: That reply is correct.
B: You mean that reply is incorrect.

and

A: I think the locals are very friendly.
B: I think they are very unfriendly

I call this type of de-accenting (where it applies) morphological de-accenting. In English in such cases obligatory de-accenting takes place, e.g. *I think they are very UNfriendly.* Greek, Macedonian, and Swedish follow the English pattern (but there was only one informant in each of these languages) where all the other languages favour re-accenting (Italian 9-3, French 12-3, Tunisian Arabic 4-0, Russian 2-0 plus the solitary but presumed highly significant Spanish speaker). So, for example the French informants preferred *Moi, je les trouve très impoLIS*, the Italians preferred *A me invece sembrano molto disSImili,* and the Russians preferred *A mne kažetsja,̌cto oni nepriVETlivy.*

CONCLUSIONS

To sum up, there appear to be at least two dimensions of variation operating: (1) There is a scale of structural likelihood of de-accenting: firstly, where it arises only from old, repeated, information, as in the first setting-response pair discussed, de-accenting appears to be universal in so far as the sample of languages here is concerned. Secondly and paradoxically, where the presence of a repeated item is bolstered by the presence of a contrast earlier in the sentence as in the second setting-response discussed, the likelihood of de-accenting appears to be somewhat less (although of course one sentence is not enough from which to draw any strong conclusions and in any case the difference from the first example is largely in the Italian data). Thirdly, and continuing the paradox, in the third setting-response, where a repeated item and a contrast are again involved, but where the repetition and the contrast are within one speaker, there is almost no posssibility of de-accenting in a number of languages. Lastly the special type of morphological contrast tested in the fourth setting-response also strongly disfavours de-accenting. A general conclusion from the four examples is that while the presence of old repeated information may encourage de-accenting, the presence of contrast may actually inhibit de-accenting in some languages.

(2) Some languages clearly favour de-accenting or re-accenting more than others. It will by now be obvious that re-accenting is much more common in the Romance languages, Spanish, Italian, and French, than in the other languages tested. But it is by no means a clear-cut difference: while de-accenting is obligatory in English (and perhaps some other non-Romance languages) both de-accenting and re-accenting are sometimes options in the Romance languages and there may be variation both between speakers, and indeed even within one speaker on grounds not yet known and of course not tested here.

Finally. to return to the quotation from Gårding (1981) quoted at the outset of this article, even the re-accenting of repeated material may involve an accent which is never particularly prominent and indeed never as prominent as an earlier accent. So we may have a sort of semi-flattening which approaches that of the accenting of post-focal lexical material in Swedish. First inspections of pitch traces of a sample of relevant data from Spanish do indeed suggest that falling pitch patterns on re-accented syllables always consist of a relatively narrow fall.

REFERENCES

W. L. Chafe (1974), "Language and consciousness", *Language*, Vol. 50, pp.111-133.

A. Cruttenden (1974), "An experiment involving comprehension of intonation in children from 7 to 10", *Journal of Child Language*, Vol.1, pp.221-232.

D. Crystal (1975), "Prosodic features and linguistic theory", In *The English Tone of Voice* (Edward Arnold, London).

E. Gårding (1981), "Contrastive prosody: a model and its applications", *Studia Linguistica*, Vol. 35, pp.146-165

M. A. K. Halliday (1967), *Intonation and grammar in British English.* (Mouton,The Hague).

D. R. Ladd (1990), Review article: "Intonation vs. grammar"(review of D. L. Bolinger (1989), *Intonation and its uses*), *Language*, Vol. 66, pp.806-816.

R. Vanderslice and L.S. Pearson (1967), "Prosodic features of Hawaiian English", *Quarterly Journal of Speech*, Vol. 53, pp.156-166.

Rhythm in regional variants of Standard Danish

Nina Grønnum
Institute of General and Applied Linguistics
University of Copenhagen
80 Njalsgade, DK-2300 Copenhagen, Denmark

ABSTRACT

The acoustic analysis is an attempt to establish the durational correlates of differences in perceived rhythm in six regional variants of Standard Danish. Durational relations between vowels and consonants in disyllabic sequences vary considerably across the six regions, but the correlation with differences in perceived rhythm is nearly nil. Fo patterns seem to be the main determinants of perceived rhythm.

INTRODUCTION

Space limitations make this a very sketchy and rather glossed-over presentation of the results. A complete account is forthcoming.

The analysis is based on a material recorded for intonation investigation purposes (Grønnum 1991, 1992), and is motivated by my clear auditory impression from the recordings that the six regions concerned differ with respect to the internal rhythmical organization of the stress group (or the foot, i.e. the succession of a stressed and all succeeding unstressed syllables). They are: Copenhagen, Næstved (South Zealand), Aalborg (North Jutland), Tønder (West South Jutland), Sønderborg (East South Jutland) and Bornholm (an island south of Sweden).

Note that I am not studying the effect of *systematically* varying parameters that may influence the perception of rhythm (duration, intensity, pitch, phonological structure). I am rationalizing after the fact: looking for acoustic differences in recordings that appear auditorily to be different rhythmically at the level of the prosodic stress group.

TIMING

Insofar as the stress-timed versus syllable-timed dichotomy is operative at all (Dauer 1983, Eriksson 1991, Miller 1984), all of the varieties under investigation here are stress-timed, and the rhythmic groups all fall under the heading *falling* or *trochaic/ dactylic*, i.e. their durational, pitch and other characteristics are such that the accent unmistakably falls on the first syllable in the sequence.

The search for durational parameters that would differentiate the speech samples from the six regions was successful only for the 'V(:)/C(C) and 'V(:)/$_{\circ}$V̌ ratios in 16 stress groups of rather varying segmental and syllabic structure. The material in the analysis presented here consists of 5 immediately comparable 'V̌N$_{\circ}$V̌ sequences (V̌ = low front or mid, N = [m/n]), extracted from non-final position in three short utterances, recorded six times by two speakers from each region. Initial consonants in the stressed syllable are excluded from calculation, because vowel onset seems to determine the onset of the unit within which temporal compensations may take place (Fant and Kruckenberg 1989, Fischer-Jørgensen 1982, Strangert 1985), and stressed vowel onset also constitutes the boundary between successive Fo patterns (Thorsen 1984).

The results are presented in Fig. 1, which represents an average over items and speakers. The correlation between the two parameters is high ($r=0.96$, $p<0.001$). Næstved (N) appears with the relatively longest and Sønderborg (S) with the relatively shortest stressed vowels, respectively. Aalborg (A), Bornholm (B) and Copenhagen (C) rather cluster in the middle of the continuum, and Tønder (T) is located somewhat more towards the lower end.

AUDITORY EVALUATION

Four colleagues in the department were supplied with tapes which contained 2x9 different complete utterances by each of two speakers from each region. (On the assumption that perceived rhythmical differences between different regions will cut across any type of stress group in any position. And, further, that measurable differences in any type of directly comparable structure, thus also $'\check{V}N_{\circ}\check{V}$, will be representative of differences under other segmental conditions.) They were asked to evaluate the foot internal rhythm of each region in terms of 11 pairs of antonymous adjectives. They found the task extremely difficult (and two adjective pairs meaningless), and the averages (over 4 listeners and 4 groups of adjectives) as presented in Fig. 2 must be taken cum grano salis. Certain aspects nevertheless stand out: nowhere do Næstved and Sønderborg constitute the extremes, as one would have expected from Fig. 1, on the contrary; and regions that cluster in Fig. 1 do not invariably do so in Fig. 2. Note specifically Tønder and Copenhagen. I would have expected, ceteris paribus, that the longer the 'beat' (the stressed vowel), the higher the score on 'regular' etc., and the lower the score on the remaining three groups of adjectives. But, of course, everything is *not* equal and listeners are apparently not reacting to durational relations per se to any (great) extent. (However, they were *not* asked to judge *durations*, and I cannot really know how they understood 'foot internal rhythm'; nor how, exactly, they interpreted the adjectives.) The other obvious candidate then is Fo/pitch:

STRESS GROUP PATTERNING IN THE SIX VARIETIES

Fig. 3. depicts stylized models of the very different Fo patterns of the six variants.

Perceived rhythm may well be the result of an interplay between segment duration and its associated Fo movement, granted that Fo movement may make segments appear different in duration, ceteris paribus (Lehiste 1976, Hombert 1977, Sundberg 1973). But perceived rhythm may not even be any direct correlate of 'pitch-corrected' durations (even if we knew exactly how to quantify them in this kind of material), pitch may play its own, more direct role, although how I cannot venture to speculate at present. (I also suspect that much less easily quantifiable idiosyncratic factors, like voice quality, speech rate and over-all fluency, come into play.)

On the above assumptions, let me attempt a very tentative outline of an explanation of the more pertinent facts as they appear in the figures. But note that no single aspect can unambiguously account for the scores in Fig. 2. Durations *and* the *total* Fo pattern are probably evaluated holistically, and thus there will be apparent contradictions in the statements to follow: Tønder is irregular, vivacious (and inciting) due both to its shortish and rising 'V and to the very large and steep fall. Sønderborg will be comparatively more regular but less vivacious due to its less extensive and steep fall (in spite of its even shorter 'beat'). Copenhagen may be regular and energetic due to the fact that the two significant events, the Fo turning points, are located comparatively early in the sequence, but dull for its rather limited range and the modest slope of the fall.

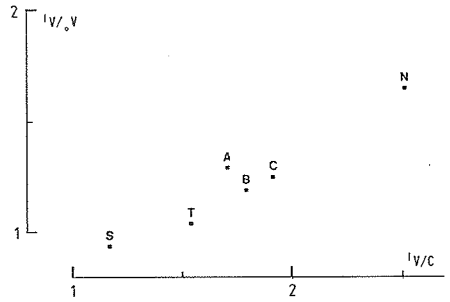


Figure 1. *Ratio of stressed vowel to post-tonic consonant duration plotted against the ratio of stressed to post-tonic vowel duration; see further the text.*

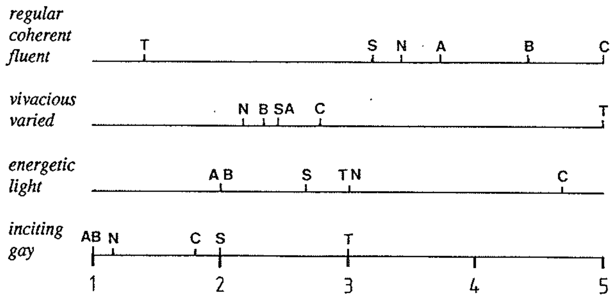


Figure 2. *Auditory evaluation of the stress group internal rhythm in the six varieties investigated; see further the text.*

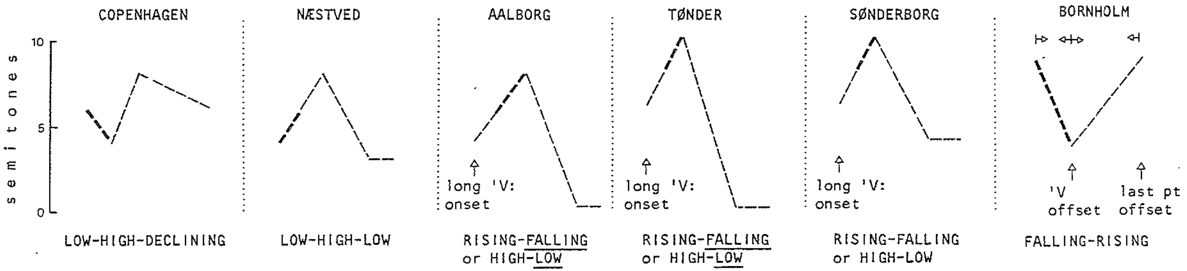


Figure 3. *Model stress group patterns from the six varieties investigated. Heavy lines denote stressed vowels. Horizontal arrows delimit movements which may be expanded or compressed in time. Vertical arrows depict turning points which anchor segmental events. The patterns are characterized auditorily beneath each frame.*

Bornholm may be regular due to the frequency constant 'highs' and 'low' but heavy and dull by virtue of its not very large range, and particularly its pronouncedly *falling*-rising pattern. Aalborg may be rather regular but heavy and dull for its slow movements, and perhaps also because the fall lands at the bottom of the range (but cf. Tønder!). Næstved is perhaps monotonous and dull both for its long 'beat', the limited range and the modest slopes but, curiously, not very regular.

FINAL REMARKS

In and of itself this little experiment naturally does not tell us very much besides the fact that perceived rhythm is not directly correlated to durational relations in the foot. Nor does it pretend, of course, to exhaust questions of timing in the six varieties. Obviously, other structures besides the simple 'V̌N$_\circ$V̌ sequences would have to be included. But the results raise a host of questions - with obvious implications for speech synthesis - which can only be answered through a multitude of experiments with synthetic manipulation of duration and Fo. They would demonstrate the significance of and the requirements for appropriate realizations of stress group patterns (in terms of magnitude, slopes and timing of Fo events relative to the segmental structure) in conjunction with the proper durational relations, for any given variant.

REFERENCES

Dauer, R.M. (1983), "Stress-timing and syllable-timing reanalyzed", *Journal of Phonetics,* Vol. 11, pp. 51-62.

Eriksson, A. (1991), *Aspects of Swedish speech rhythm,* (Gothenburg Monographs in Linguistics, Vol. 9, Department of Linguistics, University of Göteborg, Göteborg).

Fant, G. and A. Kruckenberg (1989), "Preliminaries to the study of Swedish prose reading and reading style", *Speech Transmission Laboratory - Quarterly Progress and Status Report,* Vol. 2, pp. 1-83.

Fischer-Jørgensen, E. (1982), "Segment duration in Danish words in dependency on higher level phonological units", *Annual Report, Institute of Phonetics, University of Copenhagen,* Vol. 16, pp. 137-189.

Grønnum, N. (1991), "Prosodic parameters in a variety of regional Danish standard languages, with a view towards Swedish and German", *Phonetica*, Vol. 47, pp. 182-214.

Grønnum, N. (1992), *The Groundworks of Danish Intonation* (Museum Tusculanum Press, Copenhagen).

Hombert, J.-M. (1977), "Difficulty of producing different Fo in speech", *Working Papers, University of California at Los Angeles,* Vol. 36, pp. 12-19.

Lehiste, I. (1976), "Influence of fundamental frequency pattern on the perception of duration", *Journal of Phonetics*, Vol. 4, pp. 113-117.

Miller, M. (1984), "On the perception of rhythm", *Journal of Phonetics*, Vol. 12, pp. 75-83.

Strangert, E. (1985), *Swedish Speech Rhythm in a Cross-Language Perspective* (Umeå Studies in the Humanities, Vol. 69, Almqvist & Wiksell, Stockholm).

Sundberg, J. (1973), "Data on maximum speed of pitch changes", *Speech Transmission Laboratory, Quarterly Progress and Status Report*, Vol. 4, pp. 39-47.

Thorsen, N. (1984), "Fo timing in Danish word perception", *Phonetica*, Vol. 41, 17-30.

Resolving Category Ambiguities - Evidence from Stress Shift

Esther Grabe, Paul Warren & Francis Nolan
Department of Linguistics
University of Cambridge, Cambridge CB3 9DA, Great Britain

ABSTRACT

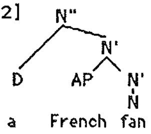
This paper presents an experimental study of stress shift in category-ambiguous material. Sequences such as Chinese fan exhibit phonological evidence for two structural analyses. If Chinese is an adjective, fan is stressed; the sequence is a syntactic phrase. If Chinese is a noun, fan is deaccented and the sequence a compound. Additionally, as Chinese is a stress shift item, stress shift may apply in the phrasal interpretation. Thus, category-ambiguous sequences with a potential for stress shift might contain earlier cues to syntactic category than sequences without such a potential. Production data show that stress shift patterns do indeed map onto syntactic categories, but only if the second element in the sequence is not right-branching. A comprehension experiment suggests that stress shift may facilitate category assignment.

1. INTRODUCTION AND BACKGROUND

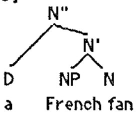
Syntactic, semantic, morphological and phonological information supply evidence for grammatical categories. So far, little research has been carried out to establish the nature of the phonological correlates of grammatical class (Kelly 1992). In the category-ambiguous potential stress shift sequence *Chinese fan*, *Chinese* may be either an adjective or a noun, i.e. the sequence is [A+N] or [N+N]. This paper investigates whether the incidence of stress shift (e.g. Giegerich 1985) varies with the different structural analyses, and whether the (non-)occurrence of the process can cue the correct structural interpretation in comprehension.

A category-ambiguous (CA) sequence such as [1] *a French fan* may be interpreted as (a) a fan from France or (b) a fan of the French language. Structurally, (a) is commonly described as a syntactic phrase and (b) as a compound (Liberman & Prince 1977, Bauer 1983, Radford 1988). At first glance, the difference seems clear; however, as Matthews (1974) points out, the distinction between compounds and phrases in English is problematic. Relevant criteria may be drawn from various sources and these do not yield the same results. Radford (1988) correlates different syntactic categories with different stress patterns. Items such as *increase* carry primary stress on the second syllable when verbs and on the first syllable when nouns. Similarly, the two possible interpretations of *a French fan* are said to correlate with two different stress patterns found in phrases and compounds. The phrase has greater prominence on the second element, and the compound has greater prominence on the first. This corresponds to a difference in the syntactic category of *French*: when this word is an adjective in a phrase [2], then *fan* has greater relative prominence; if *French* is a noun in a compound [3], then it is more prominent than *fan.*

[2]



[3]



Selkirk (1984) suggests that the incidence of stress shift varies with respect to syntactic category and attributes this to the prominence relations at the supra-word level: when *fan* is more prominent, as in the phrase, then stress shift will apply to an immediately preceding stress shift item. This is of interest to speech comprehension. If stress is shifted reliably in the phrasal interpretation, then this might signal that an accented word is likely to follow and a [A+N] interpretation might be cued early. Thus, CA sequences containing stress shift items [5] might contain an additional cue to syntactic category that non shifting CA sequences [4] do not have.

[4] *French fans are popular.*
[5] *Chinese fans are popular.*

In [4], the category ambiguity is likely to be resolved when listeners hear *fans*. If the information in [5] is processed as it is heard, then the relative stress pattern of *Chinese* may resolve the ambiguity earlier.

First, we need to determine how well the incidence of stress shift correlates with differences in syntactic category. Liberman and Prince (1977) state that relative prominence is defined on syntactic constituents and define prominence relations on each pair of sister nodes in a syntactic structure. According to the nuclear stress rule (a) and compound stress rule (b) the following should apply:

(a) if a sequence such as *Chinese fans* is a phrasal category, *fans* is strong;
(b) if a sequence such as *Chinese fans* is a compound, the second element is strong if it branches.

(b) means that sequences such as [6] *French entrance exams* and [7] *Chinese entrance exams* will not be distinguished in their phrasal and compound forms because the branching right element (*entrance exams*) will be strong. [7] will exhibit stress shift in both phrasal and compound interpretations, unless speakers produce stress clashes in order to map rhythmic patterns of stress shift items onto syntactic structure systematically. The next section examines the reliability of stress shift as a marker of syntactic category.

2. INCIDENCE

A production experiment has been carried out testing the incidence of stress shift in non-right-branching and right-branching CA sequences. We predict (i) that in non-right-branching sequences speakers produce phrasal stress patterns in [A+N] sequences and compound stress patterns in [N+N] sequences; and that stress shift applies in phrases, but not in compounds. Furthermore (ii) we expect phrasal stress patterns and stress shift in both [A+N] and [N+N] right-branching sequences.

Twelve native speakers of Southern British English read 24 sentences with the following category-ambiguous sequences:

- non-right-branching CA sequences with a non-stress shifting (-SI) first element, e.g. *French fans*
- non-right-branching CA sequence with a stress shifting (+SI) first element, e.g. *Chinese fans*
- right-branching sequences with a -SI CA first element, e.g. *French entrance exam*
- right-branching sequences with a +SI CA first element, e.g. *Chinese entrance exam*

The sentences were read with over 70% fillers in two sessions, approximately one week apart. A trained phonetician performed an auditory analysis of the data, assigning phrasal or compound stress patterns to each sentence. Table 1 below presents the percentage of phrasal stress patterns for *French fans* and *Chinese fans*. In the -SI sequence *French fans*, speakers produced only phrasal stress patterns for [A+N] and only compound stress patterns for [N+N] sequences. In the +SI sequence *Chinese fans*, the [A+N] sequence also always had phrasal stress, and a small number of [N+N] sequences showed phrasal stress. Each incidence of phrasal stress in the +SI items was accompanied by stress shift on Chinese. The results confirm hypothesis (i). A clear preference for the phrasal stress pattern emerged in sequences with branching second elements (Table 2). Again, phrasal stress on +SI items was accompanied by stress shift. This confirms our second hypothesis (ii).

Table 1. *Distribution (%) of phrasal stress patterns in CA sequences without (-SI) and with (+SI) stress shift items (French fans and Chinese fans respectively).*

	-SI	+SI
A+N	100	100
N+N	0	17

Table 2. *Distribution (%) of phrasal stress patterns in right-branching CA sequences without (-SI) and with (+SI) stress shift items (French entrance exams and Chinese entrance exams respectively).*

	-SI	+SI
A+N	100	75
N+N	100	100

3. COMPREHENSION

On the basis of the production data, we ask whether (i) listeners are sensitive to a phonological distinction between compounds and phrases, and (ii) stress shift provides additional information which facilitates syntactic category assignment. To address these questions, we carried out a comprehension experiment with non-right-branching test sequences only, since the production data suggests that right-branching sequences do not exhibit a correlation between stress shift and syntactic category.

In a cross-modal naming task, subjects heard sentence fragments ending with category-ambiguous -SI and +SI tokens. All -SI tokens had two syllables and late stress, thus matching +SI tokens in number of syllables but not stress shift. Examples of fragments with -SI and +SI tokens are given in [8a] and [8b] respectively.

[8a] *I'm trying to find reliable Malay*
[8b] *I'm trying to find reliable Chinese*

The subjects' task was to name a visual probe presented at the end of each sentence fragment. For the sentence pair in [8] the probe word was SERVANTS. This probe is appropriate if the preceding CA word (*Malay* or *Chinese*) is an adjective, but not when it is a noun. The sentence materials from which the fragments were taken were recorded with nuclear accent either on the CA word or on the following word, giving four context conditions as in Table 3. We predict that when a fragment is heard with nuclear stress on the CA word, this signals a [N+N] compound with prominence on the first element, as no further accented syllables should follow. In this case, the probe SERVANTS will be inappropriate. By contrast, if the CA item is prenuclear, it is likely to be interpreted as the first element of a phrase, and the probe will be appropriate.

Table 3. *Test conditions and average response times (in msec) in comprehension experiment.*

condition	structure	item type	accent	CA item	probe	average RT
A	N+N	-SI	+nucleus	Ma'**lay**	inappropriate	619.6
B	A+N	-SI	-nucleus	Ma'**lay**	appropriate	586.3
C	N+N	+SI	+nucleus	Chi'**nese**	inappropriate	638.6
D	A+N	+SI	-nucleus	'**Chi**nese	appropriate	538.3

Thus we predict (i) that for -SI and +SI items, the reaction time (RT) will be longer if the probe does not agree with a [N+N] interpretation (i.e. conditions A and C should be slower than B and D respectively) and (ii) that the difference in RT for appropriate and inappropriate probes will be larger for +SI items (i.e. the difference between C and D will be greater than that between A and B). The average reaction time data are given in the final column of Table 3, and confirm both these hypotheses. For both -SI and +SI items, subjects react significantly faster to probes in the A+N condition. Furthermore, the effect is stronger for the +SI item *Chinese* (where the difference between appropriate and inappropriate probes is 101 ms) than for the -SI item *Malay* (where it is 34ms). That these effects result from the interpretation of the CA item and the consequent integration of the probe word, and not from some other property of the context fragments, is made clear in corresponding control conditions where the same fragments were followed by the probe word TEACHERS. This word is appropriate in all conditions (i.e. as the second element of a compound and as the head of a noun phrase modified by the CA word), and the response times did not differ across conditions A to D.

The pattern of results for SERVANTS would seem to confirm both hypothesis (i), that nuclear stress placement will determine the category interpretation of the CA item, and hypothesis (ii) that stress shift provides additional information which facilitates parsing. However, we found that stress shift does not cue syntactic category reliably for other category-ambiguous items such as *dark-blue* or *grown-up*. This is presumably attributable to morphological and semantic differences between such items and the nationality/language items. For example, *dark-blue* has two free morphemes, either of which may be contrastively stressed. This means that greater prominence on *blue* does not necessarily exclude the adjectival interpretation, and thus the following probe word may have been appropriate in all conditions.

4. DISCUSSION AND CONCLUSION

In Grabe & Warren (1993) we show that in connected speech the application of stress shift is highly likely, but can rarely be predicted with any degree of certainty for a particular sequence. The current production data suggest that stress shift occurs reliably in CA stress shift sequences with a branching right element. Furthermore, the data suggest that in non-right branching CA sequences the application of stress shift correlates with syntactic category. It would appear that category-ambiguous stress shift items allow us to predict the application or non-application of the process with a higher degree of certainty than our previously studied materials.

The comprehension data show that in CA sequences of the type *Malay servant*, listeners are sensitive to stress signalling a distinction between compounds and phrases, and that they are able to supplement this with stress shift information on CA items such as *Chinese* in *Chinese servant*. This suggests that the phonological information in stress shift sequences affects syntactic parsing as follows: the application of the process is likely to indicate that an accented item will follow - this cues a phrasal interpretation. Non-application suggests that the following item (if N) should be compounded. However, it appears that this applies only to sequences with nationality/language items such as *French* or *Chinese* followed by non-right branching items.

5. ACKNOWLEDGEMENTS

The research reported in this paper has been funded by the Joint Research Councils' HCI initiative under grant SPG 9030657 to P. Warren, F. Nolan and E. Briscoe and by SERC Fellowship B/ITF/226 to P. Warren.

5. REFERENCES

L. Bauer (1983), *English word formation* . Cambridge: CUP

H. Giegerich (1985), *Metrical phonology and phonological structure*. Cambridge: CUP.

E. Grabe & P. Warren (1993), Stress shift - do speakers do it or do listeners hear it? Presentation at LabPhon 4, Oxford, 11 - 14 August 1993.

M.H. Kelly (1992), Using sound to solve syntactic problems: the role of phonology in grammatical category assignment. *Psychological Review, 99, No 2.*

M. Liberman, & A. Prince (1977), On stress and linguistic rhythm. *Linguistic Inquiry*, 8, 249-336.

P. Matthews (1974), *Morphology*. Cambridge: CUP

A. Radford (1988), *Transformational Grammar*. Cambridge: CUP.

E. O. Selkirk (1984), *Phonology and Syntax: The relation between sound and structure.* Cambridge, MA: MIT Press.

The Role of Pitch in Lending Prominence to Syllables

Dik J. Hermes and H.H. Rump
Institute for Perception Research, P.O. Box 513
NL-5600 MB Eindhoven, The Netherlands

ABSTRACT

Pitch is one of the prosodic variables that play an important role in lending prominence to an accented syllable. Little is known, however, about which characteristics of a pitch contour are important in this respect. A model will be described in which it is assumed that the contribution of pitch to prominence depends linearly on the difference between the average pitch level in the nucleus of the accented syllable and the average pitch level in the nucleus of the preceding syllable. When excursion size is defined as the difference between upper and lower declination line, a consequence of this model is that, excursion sizes being equal, a fall lends more prominence to a syllable than a rise. An experiment will be described, in which this difference in prominence between a rise and a fall is measured for various declination slopes. The results are in good agreement with the predictions of the model.

INTRODUCTION

It is well known that pitch plays an important role in lending prominence to accented syllables. In this study we investigated which properties of the pitch contour contribute to the prominence of an accented syllable. Hermes and Van Gestel (1991) showed that subjects are capable of comparing the prominence of accented syllables in two utterances presented in different registers, when the prominence was lent by the same accent-lending pitch movement, a rise, a rise-fall or a fall. In another study, Hermes (1991) showed that subjects are also capable of comparing the prominence of the accented syllables in two utterances when prominence was lent by different types of pitch movements. It was found that, when lending equal prominence, the rise and the rise-fall had a larger excursion size than the fall. It could be excluded that this discrepancy between the rise and the rise-fall on the one hand and the fall on the other was due to a difference in timing. A model was proposed, in which the contribution of pitch to prominence depends linearly on the difference between the average pitch in the syllabic nucleus of the accented syllable and the average pitch in the syllabic nucleus of the preceding syllable. As shown in Fig. 1, this difference is larger for the fall than for the rise, when the excursion sizes are equal. Excursion size is defined as the distance between the upper and the lower declination line. To be more precise, a calculation shows that this difference S equals

$$S = 2dT, \tag{1}$$

in which d is the declination slope, and T is the time interval between the two points subjects use in estimating the pitch-level differences, the pitch-level estimation points. Note that S is independent of the excursion size of the pitch movements. The model predicts that this difference will increase linearly with increasing declination. In this contribution an experiment will be described to test this prediction.

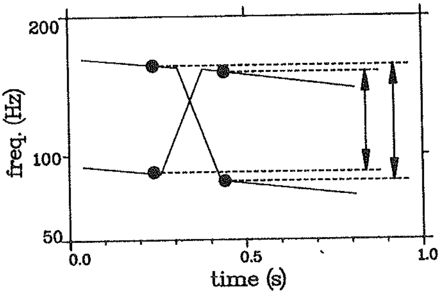


Figure 1: *Diagram showing that for equal excursion sizes the distance between pitch levels in a fall is greater than in a rise. The filled circles indicate pitch levels which might be used for the perceptual determination of prominence. Note that the difference in length between the arrows is due to the presence of declination.*

METHODS

The experiment was an adjustment experiment. Its design is shown in Fig. 2. It shows the pitch contours of two PSOLA-resynthesized utterances, /mamáma/, a Dutch nonsense word, with an accent on the second syllable. The utterances are identical, except that the second syllable in the first stimulus is accented by a fall, while the second syllable in the second stimulus is accented by a rise. The excursion size of the pitch movement in the first stimulus, the test stimulus, was fixed. The excursion size of the pitch movement of the second stimulus, the comparison stimulus, was under the subject's control. Subjects were asked to adjust the excursion size of a pitch movement in a comparison stimulus until the prominence of the accented syllable was equal to the prominence of the accented syllable in the test stimulus. When the subject had done this, the procedure was repeated for a test stimulus with another excursion size. This was done for a total of six different excursion sizes in the test stimulus, shown in the left part of the figure. When this was completed, the experiment was repeated with the rise in the test stimulus, and the fall in the comparison stimulus. In this way an accurate estimate could be obtained of the difference in excursion size between a rise and a fall, when lending equal prominence (for details see Hermes, 1991). This excursion size is expressed in E, the number of ERBs, since this psycho-acoustically defined frequency scale (Patterson, 1976) appears to be most adequate for expressing the frequency of speech intonation (Hermes and Van Gestel, 1991).

The experiment was carried out for five different declination slopes: 0.0, 0.33, 0.70, 1.09 and 1.50 E/s, which, in this register, correspond to slopes of 0 and about 2.4, 4.9, 7.3 and 9.7 semitones/s, respectively.

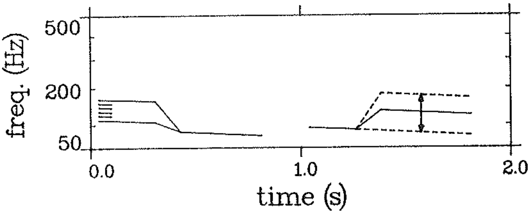


Figure 2: *Set-up of an adjustment run. The subject was first presented with one of the six test stimuli displayed on the left-hand side. Then the subject heard the comparison stimulus with an adjustable excursion size shown on the right-hand side. The run stopped when the subject indicated that the same prominence was perceived in the accented syllables of both stimuli.*

RESULTS

Twelve subjects took part in this experiment. They were students and research associates of the Institute for Perception Research. All of them reported normal hearing and most of them had experience in auditory perception experiments. In order to determine whether subjects performed consistently, the criterion for consistency described in Hermes and Van Gestel (1991) was applied. As a consequence, the results of three of the twelve subjects were excluded from the analysis. The results of the nine consistent subjects are presented in Fig. 3. The data points show the difference in excursion size between the rise and the fall as found in the experiment. The straight line shows the predictions for S as obtained from Eq. 1. This figure shows that the predictions are quite good.

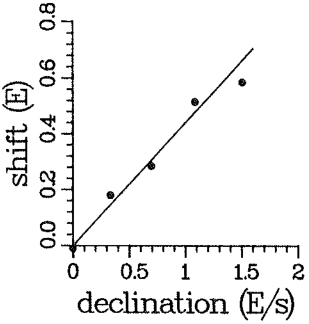


Figure 3: *The data points show the average difference S in the excursion size of the rise and the fall for the five different declination slopes. The line represents Eq. 1, the predicted relation between the difference S and declination.*

The value chosen for T in Eq. 1 was 0.22 s, the interval between the vowel onsets of the first and the second syllables. The vowel onsets were taken, since they are quite accurately defined in time, much more accurately than the nucleus which actually comprises a time interval.

DISCUSSION

We tested the effect of declination on the difference in prominence lent to a syllable by a rise and by a fall. More specifically, we tested the assumption that this prominence depended linearly on a difference in pitch level at some point before the pitch movement and at some point after the pitch movement. For the interval between these pitch-level estimation points we chose the interval between the vowel onsets, which in this utterance was equal to the distance between the syllabic nuclei.

It follows from the model that S, the difference in excursion size of the rise and the fall when lending equal prominence, is independent of the excursion size of the rise and the fall. This has a strange consequence when the data points are extrapolated to excursion size zero: a fall of excursion size zero will be adjusted to a rise with a positive excursion size. Although this was not tested experimentally, this agrees well with informal listening to stimuli of this kind. In the presence of no other pitch movement than is due to declination, a falling pitch accent is perceived on the second syllable. A similar observation can be made when only inclination is present, in which case a rising pitch accent is perceived. A prerequisite is probably that the accent on the second syllable is clear due to other cues such as loudness and vowel length. In the utterance used in this experiment, the speaker had pronounced the utterance in such a way that both loudness and vowel length in the accented syllable were larger than in the preceding syllable.

Another related consequence of this model is that, on the short time scale of these stimuli, there is no compensation for declination. One might have expected that listeners would be able to predict the discrepancy between the rise and the fall on the basis of the declination preceding and following the pitch movement, and would compensate for this. The results of the experiment described here show that listeners do not compensate.

It is not yet clear how long the time interval should be between the two pitch-level estimation points which the listeners use in comparing the difference in pitch level. In the model presented here it is implicitly assumed that the pitch-level estimation points are in the syllabic nuclei of the accented and the preceding syllables. There may, however, very well be a fixed distance which accidentally corresponds quite well with the distance between the nuclei of succeeding syllables in normal speech. In the stimuli used in these experiments the time interval between succeeding nuclei was 0.22 s. In fluent speech it has the same order of magnitude. Future experiments will have to clarify whether the distance between the pitch-level estimation points is fixed, or whether it varies with the actually realised distance between the syllabic nuclei in fluent speech.

ACKNOWLEDGEMENTS

We would like to thank Jacques Terken and Jan Roelof de Pijper for their constructive comments on the manuscript.

REFERENCES

Hermes, D.J., and Van Gestel, J.C. (1991) "The frequency scale of speech intonation", *Journal of the Acoustical Society of America,* Vol. 90, pp. 97-102.

Hermes D.J. (1991) "Prominence caused by rising and falling pitch movements with different positions in the syllable", *IPO Annual Progress Report,* Vol. 26, pp. 17-28.

Patterson R.D. (1976) "Auditory filter shapes derived with noise stimuli", *Journal of the Acoustical Society of America,* Vol. 59, pp. 640-654.

Peak, boundary and cohesion characteristics of prosodic grouping

Daniel Hirst
Institut de Phonétique, CNRS - Parole et Langage, Université de Provence
29 avenue Schuman, 13621 AIX (France)
e-mail : Hirst @ fraix11.bitnet

INTRODUCTION

Prosodic grouping can be defined as the way in which phonological units are assumed to join together to form larger units. Such grouping has, at various times and by various authors, been accounted for by appealing to one or more of three distinct but inter-related characteristics which I refer to here as *peaks* (= prominence, headedness), *boundaries* (= disjuncture, discontinuity) and *cohesion.* (= juncture, continuity).

There is a certain redundancy in these three characteristics. Thus boundaries and cohesion, for example, are mutually exclusive, together defining prosodic goups (=units, constituents or domains). The relationship between domains and groups is more subtle and also more theory dependent. Halle & Vergnaud (1987) argue that the pure representation of heads (= peaks) and the pure representation of domains (= groups) constitute "conjugate representations" which together define the notion of government :*"the property of being a head is the same as that of being a governing element in a constituent, and the property of being in some domain is the same as that of being governed by some head."* (p 16)

Despite this redundancy, I argue that these three characteristics are worth distinguishing when we look at the way prosodic constituents are defined. Before looking at the way they apply to spoken language, as an exercise it may be interesting to see how these concepts can be applied to written language where the notion of constituents seems more easily defined.

Grouping in written language.

Note that we are interested here in looking at the formal graphic constituents of written language, not the linguistic constituents. At the lowest level we may distinguish the *letter.* Letters combine to form *words.* Words combine to form *punctuation groups* (i.e. groups delimited by punctuation marks). Punctuation groups combine to form *paragraphs.* Above the level of the paragraph, constituents vary depending on the type of document. A book may be further organised into *chapters* and *volumes* : other documents may have a much more highly organised and specific structure.

Assuming that we are concerned with printed texts using the roman alphabet, letters, as constituents are defined by the characteristic of cohesion : a letter is a graphic symbol presenting a horizontal continuity (some letters such as i, j and accented letters are discontinuous but the discontinuity is always vertical). To take into account cursive handwriting (or other writing systems such as Arabic or Devanagari where individual letters are not necessarily disjoint) we should need to appeal to a notion of graphic peak to define the way in which we separate one letter from another. The graphic word can also be thought of as being defined by its cohesive properties : a word consisting of a sequence of contiguous letters. The punctuation-group is a clearcut case of a boundary-defined constituent marked. Paragraphs, like words, are units which can be defined in terms of cohesion. Chapters are defined by boundaries (chapter headings) and usually by cohesion (new page for new chapter). Volumes too are defined both by cohesion (binding) and boundaries (covers).

Besides these formal constituents, graphic texts can also be divided into lines and pages which are not constituents although as I have mentioned they are used as part of the cohesive properties defining paragraphs and chapters respectively. If we set up a hierarchy among punctuation boundary symbols defining minor symbols (, ; :) and major symbols (. ? !) then we can define a further hierarchical level : the sentence, intermediate between the punctuation group and the paragraph. Unlike the other graphic constituents we have defined, this distinction is not based on formal properties of the boundary symbols themselves. Sentences are also marked by a boundary symbol in the form of an initial capital letter but initial capitals are not sufficient defining characteristics (this is even more true in German).

It is controversial whether reading appeals to intermediate constituents between the letter and the word. Against the classic idea that reading can make use of both direct (orthographic) and indirect (phonological) lexical access, recent research (Treiman & Chafetz 1987) suggests that intermediate units between the letter and the word may play a greater role in reading than has hitherto been thought. Interestingly, if we look at characterisations of the *sonority hierarchy* of phonemes such as the following (adapted from Goldsmith 1990, p112) :

`{ a>e;o>i;u>r>l>m;n>s>v;z;D;f;Q>b;d;g>p;t;k }`

it is a rather striking fact, which to my knowledge has not been pointed out before, that the shape of the letters partially encodes this hierarchy. All of the symbols on the lefthand (strong) side of the hierarchy (except for /l/) are small letters whereas all of those on the right are ascending or descending letters. It seems difficult to imagine that such a fact is a mere coincidence. Small letters seem, then, to act as potential syllable peaks, which explains both why lower-case letters are generally considered more legible than upper case (Tinker 1963), and how the shape of a word can function as a cue to its lexical identity (Walker 1987).

We can conclude from this discussion that even in the case of printed language where a document is entirely composed of a sequence of discrete symbols, there are a number of formal constituents (line, page) whose role is only marginal while other constituents (syllable, sentence) are only indirectly related to or derivable from the printed form.

GROUPING IN SPEECH

As is to be expected, the relation between the observable data and the phonological units which are assumed to structure this data is even more indirect in the case of speech than in the case of written language. In speech, unlike in writing, there are no unequivocable units on which an analysis can be based. The only direct formal constituents of a speech-signal are portions delimited by silences and/or by breathing. Both of these have in the past been taken as boundary signals for phonological constituents, ie 'breath-groups' (Sweet 1890; Jones 1949; Lieberman 1980) or 'pause-defined units' (Brown, Currie & Kenworthy 1980). Lieberman's categorical affirmation : *"the breath-group is the primary element that people use to segment the flow of speech into sentence-like units"* (p 240) is worth contrasting with Sweet's far more cautious :*"These breath-groups correspond partially to the logical division into sentences"* . The parallel with line ends and page breaks in reading is instructive since these, as we saw above, just like breaths and silences, are neither necessary nor sufficient criteria for isolating linguistic units even though higher order linguistic units will often make use of such direct physical characteristics to reinforce the internal cohesion of the units.

Since the observation of directly delimited prosodic units does not prove very rewarding, the next step is to look at those indirect units which have been proposed in various analyses. Between the phoneme and the utterance, the following hierarchy is a fairly traditional one in prosodic analysis :

syllable < accent group < intonation unit

None of these prosodic units are uncontroversial. I propose to review each in turn together with alternative proposals that have been made for units of approximately the same level. There have been a number of proposals concerning prosodic units above the intonation unit - paratones - the prosodic equivalent of paragraphs. Research in this area, in particular into the prosodic structure of discourse, is a world unto itself and I shall not attempt to deal with it at all in this paper.

AROUND THE SYLLABLE

It could be claimed that the syllable is the prosodic unit for which there is the greatest consensus among linguists and phoneticians. Speakers' intuitions concerning the number of syllables for example are far more consistent than those conerning phonemes. Thus de Cornulier (1982) has shown that French speakers can easily spot an odd-man-out in sequences of words consisting of three syllables (quantité, apéro, Nicolas,...) whereas they are totally incapable of performing the same sort of operation on words made up of three phonemes (nuit, avis, tard, train...)(p 59.). The syllable, it has been claimed is a good contender for the role of basic unit of speech perception (Mehler et al. 1981). The demonstration of a syllabic effect in French (i.e. shorter reaction time for a detection task when targets were syllables than when they were smaller or larger than syllables) was not however replicated for English (Cutler, Mehler, Norris & Segui 1986). This was taken to imply the existence of a language-specific comprehension strategy and seemed to support the claim that certain consonants in English are ambisyllabic, that is belong to both adjacent syllables. (Anderson & Jones 1974, Kahn 1976). Other phonologists (Kiparsky 1979, Selkirk 1982) have claimed that ambisyllabicity is not necessary within a framework making use of the notion of 'foot'.

Most of the arguments in favour of the syllable as a unit are really arguments in favour of syllabicity rather than syllabic constituency : syllables, in other words (in languages like English) are identified by their peak characteristics rather than by their boundaries or their internal coherence. Recent work on temporal patterning has brought up an alternative candidate for modelling segmental durations. This is the V-to-V unit used by Fant & Kruckenberg 1988, Fant, Kruckenberg & Nord 1991 consisting of a sequence beginning with the vowel onset and continuing up to but not including the following vowel onset. There is considerable evidence that the vowel onset is a particularly crucial part of the speech signal (Dogil 1986). Barbosa & Bailly 1992 suggest that normalised duration of V-to-V units gives a better image of prosodic boundaries than does syllable duration. My own analysis of a reading of a one-minute continuous text in French reveals that V-to-V units have smaller variance than the corresponding syllables in the same text. This is to be put into relationship with the finding of Wightman et al (1992) that final lengthening only applies to the segments in the final rime, not the final syllable. The picture which emerges from these findings is that the syllable is not perhaps so strongly entrenched as a phonological unit as has often been believed. Instead of syllables, it could turn out to be their immediate constituents (onsets and rimes) which are the basic building blocks for speech.

AROUND THE ACCENT GROUP

Above the syllable, the accent group, or foot, is probably the most respectable phonological unit in common use. This unit, like the syllable, is obviously essentially a peak phenomenon. It is not evident that foot boundaries are themselves in any way well defined. The term 'accent group' or 'foot' is often used indiscriminately to refer to units which should perhaps be more carefully distinguished. Thus the foot can be taken to be a strictly phonological unit (Selkirk 1978) a purely rhythmic unit (Abercrombie 1964), a pitch accent or tonal unit, or even a mixed syntactic/phonological unit. The relationship of the accent group to the word has always been controversial. Thus Pike (1945) and Jassem (1951) both distinguished rhythmic units from tonal units. For both authors rhythm units were essentially groups of words. Unlike Pike's 'contour', however Jassem's Tonal Unit does not take into account word boundaries but groups unstressed syllables with the preceding stressed syllable, irrespective of word boundaries. This is precisely the same unit which Abercrombie (1964), followed by Halliday (1968) later promoted as the Foot. Most of the evidence on final lengthening seems to suggest that any boundary effects are associated with the word (Beckman & Edwards 1990) or with higher order constituents (Wightman et al. 1992) rather than with the foot itself. The controversy as to the relevance of the Foot to prosodic analysis is, however, far from over as can be seen from the discussion between Nooteboom 1991, Fant 1991 and Kohler 1991 for example. It also remains to be seen whether the Foot can be generalised to account for language-specific characteristics. Hirst & Di Cristo (in press) suggest that Germanic languages in general are characterised by left-headed accent groups (initial prominence) whereas similar structures in Romance languages are right-headed (final prominence) (cf the analysis of French in Wenk & Wioland 1982).

A second question relates to the position of the boundary of the foot. Although most authors have defined the foot as a group of syllables, others, following Lea (1980), have preferred to work with the 'interstress interval' defined as the sequence running from the onset of one stressed vowel to the next (Fant Kruckenberg & Nord 1991, Wightman et al. 1992). Systematic comparisons of the approporiateness of different prosodic constituents in different languages remain to be made.

AROUND THE INTONATION UNIT

Feet are generally held to be grouped into higher order constituents for which there are a great number of different proposals in the literature. Most work within the British tradition of intonation analysis has assumed, with a few exceptions, a single hierarchical level variously referred to as the "sense-group", "tone-group" or "intonation unit" (see Hirst & Di Cristo in press for a review). Recent work on the TOBI transcription system (Silverman et al 1992, Hirschberg & Beckman 1993) has followed Beckman & Pierrehumbert (1986) in assuming that at least in English and Japanese there are two levels of intonation units : called 'intermediate phrase' and 'intonational phrase' respectively. Although it is claimed that these units are distinguished by tonal boundary signals ('phrase accent' and 'boundary tone' respectively, this distinction is as yet extremely theory internal, since there is no simple one-to-one correspondance between acoustic data and such boundaries. A valley in the F0 curve for example may be interpreted as a L% boundary tone, a L phrase accent, the -L or *L of a pitch accent or simply an unspecified sagging transition between two high pitch accents.

Recent corpus-based studies of durational cues for prosodic constituency have suggested that more than two levels of intonation units can be systematically distinguished. Thus both Campbell & Ladd (1991) and Wightman & al. (1992) found that at least four degrees could be established on the basis of final lengthening. This has been interpreted by Ladd & Campbell as evidence in support of the claim (Ladd 1986) that there is no principled limit to the depth of prosodic structure. A similar claim has been made by Martin (1975) and is implicit in the 'performance structures' of Gee & Grosjean (1983), Monnin & Grosjean (in press). It still remains to be shown whether there are categorical differences between different types of intonation units and how far such distinctions hold across different languages.

One of the most promising lines of research into cues for prosodic grouping is that of tonal cohesion. Downstepping or downdrifting sequences have long been known to provide important cohesion cues in many languages. Thus in Bambara (Mali) a downdrifted sequence : /sìsé nà ꜝkúmá/ will be interpreted as a single clause : "Sissé is going to talk!" whereas the same sequence without downdrift /sìsé nà kúmá/ will be interpreted as two clauses : "Sissé! Come and talk!" (Hirst 1979). An upward resetting of pitch has a distinct boundary effect even in the absence of a nuclear pitch movement in the preceding intonation unit (Kingdon 1958 p62). Research into prosodic parsing on the basis of predictive models of tonal cohesion only just begun (Schuetze-Coburn, Shapley & Weber 1991: Bruce, Granström, Gustafson & House in press) but promises to be one of the most rewarding fields in prosody in years to come.

The investigation of the relationship between Intonation Units and syntactic, semantic and pragmatic representations has a long and varied history. Prosodic structure theory (Selkirk 1978, Nespor & Vogel 1986) presents a rich articulated theory of the relation between syntax and phonology. This is further developped in Selkirk (1986, 1990) where it is claimed that each lower-order constituent in a given language (phonological word, phonological phrase) is characterised by two parameters : a designated syntactic category (maximal projection/ lexical word) and a designated edge (left/right). Given these two parameters the mapping from syntax to phonology is defined by ensuring that the designated edge of the designated syntactic constituents aligns with the corresponding edges of the prosodic constituents. Selkirk has claimed several times that Intonational phrases seem be governed by semantic rather than by syntactic constraints. I have argued (Hirst 1993) that a non-deterministic version of Selkirk's rule will in fact account for the data which she found troublesome for a syntactic account. One way of formulating this Mapping Rule is as follows :

Map a syntactic structure exhaustively onto a linear sequence of intonational phrases such that :

a.) the left end of each intonational phrase corresponds to the left end of a major syntactic constituent *b.) the intonational phrase is no longer than the corresponding syntactic constituent.*

The Mapping Rule also gives an interesting explanation for the fact that certain Intonation Units boundaries are optional while others appear to be compulsory. I claim that the difference stems from the non-linear nature of certain syntactic structures (parentheticals, vocatives, non-restrictive relatives etc) and that linear order is only imposed after the (optional) choice of Intonation Unit boundaries.

CONCLUSION

There is clearly a need for considerably more empirical data concerning the appropriateness of prosodic constituents at different levels for modelling durational and tonal characteristics of utterances. In particular very little is known concerning the way in which such characteristics vary from one language to another and this is an obvious priority task for the years to come.

ACKNOWLEDGEMENTS

This research has been partially supported by the French CNRS Cogniscience programme : *Les unités de traitement dans la perception de la parole et la lecture.* (Responsable : J. Pynte).

REFERENCES

Abercrombie, D. (1964) "Syllable quantity and enclitics." in Abercrombie, Fry, MacCarthy, Scott & Trim (eds) *In Honour of Daniel Jones* (Longman), 216-222.

Anderson, J. & Jones, C. 1974, 'Three theses concerning phonological representation.' *Journal of Linguistics* 10, 1-26.

Barbosa, P. & Bailly, G. (1992) "Prédiction de la durée segmentale : le paradigme des groupes inter-P-Centers". in *Actes du Séminaire Prosodie* (La Baume-lès-Aix, October 1992), 73-79.

Beckman, M. & Edwards, J. (1990) "Lengthenings and shortenings and the nature of prosodic constituency." in Kingston & Beckman (eds) Papers in Laboratory Phonology

Beckman, M. & Pierrehumbert, J. (1986) "Intonational structure in English and Japanese" *Phonology Yearbook* 3, 255-309

Brown, G.; Currie, K.L. & Kenworthy, J. (1980) *Questions of Intonation.* (Croom Helm; London).

Cornulier, Benoît de (1982) *Théorie du vers : Rimbaud, Verlaine, Mallarmé.* (Seuil, Paris).

Cutler, A; Mehler, J; Norris; Segui, J; (1986) "The syllable's differing role in the segmentation of French and English." *Journal of Memory and Language* 25, 385-400

Dogil, Grzegorz (1986) *The pivot theory of speech parsing.* (University of Vienna Press, Vienna)

Fant, G. (1991) "Units of temporal organization. Stress groups versus syllables and words." *Proc. I.C.Ph.S. XII* 1, 247-250.

Fant, G.; Kruckenberg, A. & Nord, L. (1991) "Durational correlates of stress in Swedish, French and English." *Journal of Phonetics* 19, 351-363.

Gee, J.-P., & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology* 15, 411-458.

Goldsmith, J. (1990) *Autosegmental and Metrical Phonology* (Blackwell; Oxford)

Halle, M. & J.-R. Vergnaud *An Essay on Stress* (MIT Press; Cambridge, Mass.)

Halliday, M.A.K. (1968) *Intonation and Grammar in British English.* (Mouton; the Hague)

Hirschberg, J. & Beckman, M. (1992) "Report on proposed transcription system and some recommendations." unpublished ms.

Hirst, D.J. (1979) 'Pitch features for tone and intonation.' *Travaux de l'Institut de Phonétique d'Aix* 6, 177-191.

Hirst, D.J. (1993) "Detaching intonational phrases from syntactic structure." *Linguistic Inquiry* 24 (4).

Hirst, D.J. & Di Cristo, A. (in press) "A survey of intonation systems" in Hirst & Di Cristo (eds) (in press).

Hirst, D.J. & Di Cristo, A. (eds) (in press) *Intonation Systems : a Survey of Twenty Languages* (Cambridge University Press; Cambridge)

Jassem, W. (1952) *Intonation of colloquial English.* (Panstowe Wydawnictwo Naukowe; Warswawa).

Jones, D. (1949) *An Outline of English.Phonetics* (Heffer; Oxford)

Kahn, D. (1976) *Syllable based generalisations in phonology.* PhD Thesis, MIT.

Kingdon, R. (1958) *The Groundwork of English Intonation.* (Longmans, London)

Kiparsky, P. (1979) "Metrical structure assignment is cyclic." Linguistic Inquiry 10 (3), 521-541

Kohler, K.J. (1991) "Isochrony, units of rhythmic organisation and speech rate. *Proc. I.C.Ph.S. XII* 1, 257-261.

Ladd, D.R. (1986) "Intonational phrasing : the case for recursive prosodic structure." *Phonology Yearbook* 3, 311-340.

Ladd, D.R. & Campbell, W.N. (1991) "Theories of prosodic structure : evidence from syllable duration." *Proceedings I.C.Ph.S. XII* 2, 290-293.

Lieberman, P. (1980) "The acquisition of intonation by infants : physiology and normal control." in C. Johns-Lewis (ed) *Intonation in Discourse.* (Croom Helm; London), 239-257.

Martin, P. (1975) "Eléments pour une théorie de l'intonation".*Rapport d'Activités de l'Institut de Phonétique de Bruxelles* 9(1), 97-126

Mehler, J; Segui, J; Frauenfelder, U. (1981) "The role of the syllable in language acquisition and perception." in Myers, Laver & Anderson (eds) *The Cognitive Representation of Speech.* (Amsterdam, North Holland).

Monnin, P., & Grosjean, F. (in press). "Les structures de performance en francais: caracterisation et prediction." L'Année Psychologique.

Nespor, M. & Vogel, I. (1986) *Prosodic Phonology* (Foris, Dordrecht)

Nooteboom, S.G. (1991) "Some obsservations on the temporal organisation and rhythm of speech." *Proc. I.C.Ph.S. XII* 1, 228-237.

Pierrehumbert, J. & Beckman, M. (1986) "Intonational structure in Japanese and English." *Phonology Yearbook* 3, 255-310.

Schuetze-Coburn, S.; Shapley, M. & Weber, E.G. (1991) "Units of intonation in discourse: a comparison of acoustic and auditory analyses." *Language and Speech* 34(3), 207-234.

Selkirk, E.O. (1978) "On prosodic structure and its relation to syntactic structure." *Indiana University Linguistics Club.* (in *Nordic Prosody* II 1981)

Selkirk, E.O. (1982) "The Syllable." in Van der Hulst & Smith (eds) The Structure of Phonological Representations. vol II (Foris, Dordrecht),

Selkirk, E.O. (1986) "On derived domains in sentence prosody." *Phonology Yearbook* 3, 371-405

Selkirk, E.O. (1990) "On the nature of prosodic constituency : comments on Beckman and Edward's paper." in Kingston & Beckman (eds) Papers in Laboratory Phonology, 179-200

Silverman K. et al., (1992) "TOBI: A Standard for Labeling English Prosody," Proc. of the Intern. Conf. on Spoken Language Processing, 867-870.

Sweet, H. (1890) *A Primer of Phonetics* (Clarendon; Oxford)

Tinker, M.A. (1963) *The Legibility of Print* (Iowa State University Press)

Treiman, R. & Chafetz, J. 1987 "Are there onset- and rime-like units in printed words." in M. Colheart (ed.) Attention and Performance XII. The Psychology of Reading. (Erlbaum; Hillsdale, NJ).

Walker, P. (1987) "Word shape as a cue to the identity of a word : an analysis of the Kucera and Francis (1967) word list." *Quart.Jour. of Exper. Psy.* 39A, 675-700.

Wenk, B. & Wioland, F. (1982) "Is French really syllable-timed?" *Journal of Phonetics* 10, 193-216.

Wightman, C.; Shattuck-Hufnagel, S.; Ostendorf, M. & Price, P. (1992), "Segmental Durations in the Vicinity of Prosodic Phrase Boundaries." *J.A.S.A.*, 1707-1717

Durational Cues to Prominence and Grouping

W. N. Campbell
ATR$^{\pi}$ Interpreting Telecommunications Research Labs
Hikari-dai 2-2, Seika-cho, Kyoto 619-02, Japan.
e-mail: nick@itl.atr.co.jp

ABSTRACT

This paper discusses the analysis of prosody in a speech-to-speech-through-machine environment and describes the types of processing that can be used to facilitate transfer of extra-textual information in automatic speech translation. In particular, it details the use and contribution of duration-based sub-routines in such a system.

INTRODUCTION

In order to improve the efficiency of automatic translation, we need to make use of prosodic information. A speech interface to machine translation typically only processes a textual representation of the utterance, and in the three main stages of the task (speech recognition, language processing, and speech synthesis) prosody is only considered in the third. However, speech encodes more than just segmental information, and to make use of this extra channel to enrich the communication, we need to code the meaning contained in the prosody for processing alongside the orthographic representation. In order to study the flow of prosodic information, we need a database of utterances coded for speech-act and analysed for corresponding acoustic correlates. This paper describes the first steps in building such a database.

Prosodic Information Processing

Decoding not just the sentence, but the issuing of an utterance in a speech situation, we need detailed annotation of prosodic events and their correlates at several levels of representation. Leaving aside the interesting area of language-specific prosodic realisations and the translation of prosodic gestures between different cultural environments, we can simplify the task initially by assuming monolingual domains, analysing Japanese and English aspects of prosody separately, and coding them into a language-neutral message structure[1]. We are then left with the question of what constitutes the basic data for analysis. This can be answered by consideration of a) the needs, and b) the resources available.

Needs:

There is not space here for an exhaustive list of the uses of prosodic information in translation, but necessary in the long term, and perhaps most difficult to achieve, is the recovery of the speaker's intention; the pragmatic content or illocutionary force of each utterance. More immediately realisable are attitudinal characteristics (Hirschberg & Ward, 1992; Murray & Arnott, 1993), clues to the discourse structure and role (Hirschberg & Pierrehumbert, 1986; Swertz *et. al.*, 1992), and to the focus and prosodic boundary relations (Veilleux & Ostendorf, 1993; Wightman & Ostendorf, 1993). A further requirement for translating into and out of Japanese is an understanding of the degree of politeness (Ogino, 1986); this too is largely carried in the prosody. Non-linguistic uses include repair detection and correction (Schriberg & Lickley, 1992; Nakatani & Hirschberg, 1993), and modelling of the speaker-specific speech characteristics.

[1] Japanese, for example, makes use of a post-particle system; while English, on the other hand, employs greater use of stress in the signalling of information relations. At the deeper level of linguistic and paralinguistic feature representation, we assume that the two languages are equivalent.

Resources:
Since we require the prosodic interpretation to be done by machine, the main part of the first stage of the project involves the development of automatic procedures for the detection of prosodic features, and the mapping of these features to elements of the message.. Two small 'spontaneous-speech' corpora of 12 prompted dialogues are available to us; one produced by 107 speakers of English (Wood, 1992), and one by 19 speakers of Japanese. Both have been analysed for linguistic structure in the same format as used in the language-processing modules (Fais & Kikui, 1991; Nagata *et al.*, 1992). They include the following categories: Syntactic: declarative, interrogative, imperative, person, tense. polarity, etc.; semantic: probability, possibility, ability, potential, volition, permission, desire; and pragmatic: request, suggest, invite, reject, offer, response, acknowledge, inform, question, permit.
An example structure is shown below:

```
[[M [[PRAG [[HEARER !X10[[LABEL *HEARER*]]]
            [SPEAKER !X9[[LABEL *SPEAKER*]]]
            [TOPIC [[FOCUS !X7[[RESTR that [[RELN [(J-1]]]]]
                    [SCOPE [[RELN is [(J-IDENTICAL]
                            [OBJE !X7]
                            [ASPT !X8 STAT]
                            [IDEN [[RESTR [[RELN NAMED]
                                           [IDEN conference-office-1]]]]]]]
                    [TOPIC-MOD HA]]]]]
     [SEM [[RELN QUESTIONIF]
           [AGEN !X9]
           [RECP !X10]
           [OBJE !X11[[RELN BE-VI-5]
                      [OBJE !X6[[RESTR [[RELN THIS-PRON-1]]]]]
                      [ASPECT [[PERF -]
                               [PROG -]]]
                      [ASPT !X8]
                      [IDEN !X2[[RESTR [[RELN NAMED]
                                        [IDEN CONFERENCE_OFFICE-IDIOM-1]]]]]
                      [TENSE PRESENT]]]]]
     [SYN [[CAT S-TOP]
           [INV -]]]]] ... ( 57 lines of syntactic tagging omitted ) ... ]]]]]
```

The representation for 'conference office' in "Is this the conference office?".

Analysis-by-synthesis, or differencing from a generated default, provides a measure of the prosodic correlates of these features. To detect marked areas of the utterance, we first normalise pitch for height, range, and attack; duration for global and local variance; and energy for phone-type and context[2]. The correlations with fundamental frequency variation have been well studied in this respect. I shall next explain the contribution of segmental duration information.

CONTRIBUTIONS FROM DURATIONAL STRUCTURING

Previous work (Campbell, 1992, 1993) has shown that speakers signal the intended interpretation of an utterance not only through pitch-related intonational variation, but also through the durational structuring of their speech. Segmental durations are available to us from the initial speech-recognition stage, via post-processing. By normalising them to reduce phone-specific effects, the underlying prosodic structure becomes very clear. This can be used to assist in the automatic detection of pitch prominences.

[2] The energy component of the speech waveform can vary for reasons unrelated to the message (e.g., changing microphone-to-mouth distance), so may be less useful in an automatic analysis.

Underlying prosodic structure

Segments in prominent or focussed words and phrases in the utterance are typically lengthened, as are those preceding a prosodic phrase boundary. In English, reduced segments are significantly shortened. Differential lengthening of segments in onset and coda parts of the syllable enables us to distinguish between the two different lengthening contexts and therefore to distinguish prominence effects from pre-boundary effects. This information, which is easily obtained from the normalised durations, given the syllable structure, enables us to detect the intonational groupings in the utterance. Figure 1 diagrams the main steps of this process: raw durations are first z-score normalised to remove phone-type-specific durational effects, then the syllable-internal differences (slope of lengthening) are compared to distinguish prosodic boundary locations from stress-related lengthening. This is explained in more detail in the oral version of this paper and in Campbell '93. Boundary locations detected by use of duration differentials have revealed interesting inter-speaker differences concerning rhythmic vs. syntactic phrase structuring that go unnoticed on simple listening.

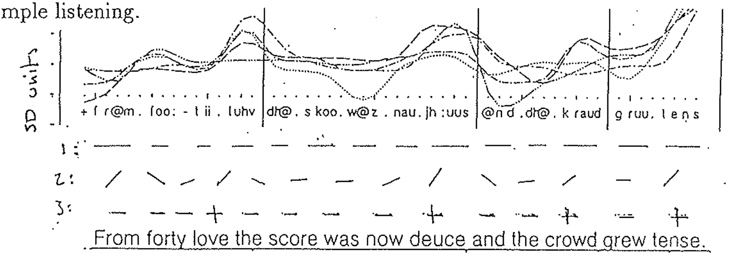


Figure 1: From normalised waveform through phone labels to boundary indication.

Pitch-prominence detection

We are testing several methods to extract meaningful phonological information from the raw fundamental frequency contour (Fujisaki & Sudo, 1971; Bagshaw, 1992; Taylor, 1993a). For Japanese, the Fujisaki-Sudo model specifies the location, strength and duration of each accent and phrase command in a small number of parameters that can be obtained by stochastic estimation procedures. These parameters include a speaker-specific/utterance-type-specific measure of the angle of onset/offset (attack) for each accent. This is normally fixed to a pre-defined value, but by estimating it along with the other parameters, we gain a useful indicator of speaking style and speaker characteristics. This, however, depends crucially on the accurate location of the accent and phrase commands; the numbers and locations of which must be specified in advance. Knowing the position and number of accents and phrases from the duration contour greatly increases the accuracy of this F0 decoding[3]. The residual from this estimation, along with the parameters specifying strength and type of accent form the raw data for our future analyses.

Unit selection for speech synthesis

In addition to facilitating pitch labelling, the prominences and boundaries detected in this way also improve the quality of our synthetic speech. The current method uses acoustic measures of goodness-of-fit when selecting units for concatenation from a labelled natural-speech source database (not enough is yet known about the fine details of segmental coarticulation to accurately predict the acoustic transitions by rule under parametric techniques). However, because no consideration is made of

[3] Unlike English, there is no durational correlate of accents in Japanese, but their detection with this method is relatively robust when given reliable phrase boundary location information.

the prosodic environment from which these units are selected, there is occasionally considerable distortion resulting from mismatch between the unit and the desired prosodic parameters. By labelling the source data according to prosodic features, and including these in the selection criteria, we can eliminate this cause of distortion. Looking to the future, with a corpus labelled for prosodic as well as segmental features, the task of speech synthesis may be even simpler. Since a small number of factors explain much of the variance in speech prosody and articulation, instead of predicting individual parameters directly, selecting appropriate units from a sufficiently large and well-labelled source database should yield natural durations and transitions that are optimal by default. Adequate labelling of the data should eliminate the need for prediction of the parameters.

CONCLUSION

This paper has introduced some recent work in the area of prosodic interpretation and highlighted the role of normalised segmental duration information in automatic labelling techniques for the acquisition and mapping of data. It showed how both pitch labelling and unit selection for synthesis can be helped by prominence and boundary detection from segmental lengthening contours. It is not yet clear what features of the utterance need to be mapped onto what higher-level aspects of the message, but we have laid the groundwork for an essential analysis.

ACKNOWLEDGEMENTS

I offer my sincere thanks to ATR for supporting this work, and to colleagues for helpful discussion.

REFERENCES

- P. C. Bagshaw (1992), An investigation of acoustic events related to sentential stress and pitch accents in English, pp 808-813, *Proc SST-92,* Brisbane, Australia.
- W. N. Campbell (1992), Multi-level Timing in Speech, *Unpublished PhD Thesis,* Sussex University, Department of Experimental Psychology..
- W. N. Campbell (1993), Automatic detection of prosodic boundaries in speech, *Speech Communication* (In Press).
- L. Fais & G. Kikui (1991), Determining surface forms for indirect speech acts in English, *ATR Technical Report,* TR-I-0235.
- H. Fujisaki & H. Sudo (1971) A model for the generation of fundamental frequency contours of Japanese word accent, Japan Acoustic Society, 27.9, 445-453.
- J. Hirschberg & J. Pierrehumbert (1986) The intonational structuring of discourse, *Proc ACL-86,* 136-144, Denver, USA.
- J. Hirschberg & G. Ward (1992) The influence of pitch-range, duration, amplitude, and spectral features on the interpretation of the rise-fall-rise intonation pattern in English, *Journal of Phonetics 20,* 241-252.
- I. R. Murray & J. L. Arnott (1993), Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *JASA 93,* 1097-1108.
- M, Nagata, M. Suzuki, & S Tachikawa 19920, First steps towards annotating illocutionary force types to a bilingual dialogue corpus, *ATR Technical Report,* TR-I-0298.
- T. Ogino & M. Hong (1992) 日本語音声の丁寧さに関する研究 (A study on politeness in Japanese speech) 215-258 in 日本語イントネーションの実態と分析 (The state-of-the-art and analysis of Japanese intonation), Mombusho, Tokyo, Japan.
- E. E. Schriberg & R. Lickley (1992), Intonation of filled pauses in spontaneous speech. *Proc ICSLP-92,* 991-994, Banff, Canada.
- M. Swertz, R. Geluykens, & J. Terken (1992) Prosodic correlates of discourse units in spontaneous speech, *Proc ICSLP-92,* 421-424, Banff, Canada.
- P. A. Taylor (1993), Automatic Recognition of Intonation from F0 Contours using the Rise - Fall - Connection Model, *Proc Eurospeech-93,* Berlin, Germany.
- N. M. Veilleux & M. Ostendorf (1993), Probabilistic parse scoring with prosodic information. *Proc ICASSP-93,* II-51, Minneapolis, USA..
- C. W. Wightman & M. Ostendorf (1992) Automatic recognition of prosodic features, *Proc ICASSP-92,* 321-324, San Francisco, USA..
- C. A. Wood (1992), The ATR-ITL/CMU Conference-registration task (Spontaneous speech), *ATR Technical Report,* TR-I-0328.

Towards an integrated view of stress correlates

Gunnar Fant and Anita Kruckenberg
Department of Speech Communication and Music Acoustics
KTH, Box 70014, Stockholm 10044.

ABSTRACT

Previous analysis of duration and of F_0 as stress correlates in Swedish prose reading, Fant and Kruckenberg (1989), Fant, Kruckenberg and Nord (1992), have been extended and work has been initiated on the study of intensity, voice source properties and segmental contrasts as prosodic parameters. The hierarchy of duration over F0 and intensity is established. Although a small average intensity difference is found between stressed and unstressed syllables, the major role of intensity, apart from determining the loudness level, appears to be that of supplementing overall F0 contours within breathgroups. A special phenomenon in Swedish is the inverse relation between emphasis and intensity of close vowels due to articulatory narrowing.

STRESS AND DURATION

In Swedish, the alternation between stressed and unstressed syllables constitutes quasi-rhythmical patterns that are mainly determined by language structure but also by the particular type of text and by stilistic and individual variations.

The most prominent stress correlate is duration. A stressed syllable is about 100 ms longer than an unstressed syllable of the same number of phonemes and the duration increases with the number of phonemes. Differences in inherent phoneme durations also enter but tend to be reduced with increasing syllable complexity. In addition we have to take into account a number of factors that contribute to syllable duration, in the first place prepause and phrase final lengthening, but also grammatical word class, accent type, syllabic word structures etc. Most of these have been included in a model of syllable duration now under development, Fant and Kruckenberg (1992).

The backbone of the system is the relation of syllable duration to the number of phonemes determined separately for stressed and unstressed syllables. In practice, almost all content words but also some of the function words carries a stress. An example of individual variations in stressed/unstressed contrast is shown in Figure 1. The two speakers differ little in the duration of unstressed syllables but more in the duration of stressed syllables. This situation is quite similar to that when our reference subject shifts from a normal to a more distinct speaking mode, Fant, Kruckenberg and Nord (1991b).

In order to eliminate differences due to variations in syllable complexity we have introduced a normalized duration, the syllable duration index S_i, which is scaled so as to provide a value of $S_i=1$ for average unstressed and $S_i=2$ for average stressed conditions. The particular value for a syllable of duration T is found by an interpolation or extrapolation

$$S_i = 1 + (T-T_{nu})/(T_{ns}-T_{nu}) \qquad (1)$$

where T_{nu} and T_{ns} are expected average unstressed and stressed values for the particular number of phonemes, n.

PERCEPTUAL SCALING

A continuous rating of perceived stress was established from experiments in which 14 subjects in two different sessions were asked to grade the relative prominence of syllables and words by making a pencil mark on a vertical line scaled from 0 to 30. They were told that 10 corresponded to average unstressed conditions. The experiment gave quite consistent results with standard deviations of single ratings of the order of 3 units only. The same technique was also used in experiments on continuous grading of perceived degree of prominence of syntactic boundaries, Fant and Kruckenberg (1989).

The corpus thus comprised all syllables in a 24 word sentence and all words in a nine sentence paragraph of the standard text. We found a high degree of correlation *(r=0.9)* between perceived syllable prominence, R_s, and the syllable duration index S_i amounting to $R_s = 6.4 + 5.5 S_i$, or in terms of a power function, $R_s = 12S_i^{0.5}$, which indicates a compression compared to the duration data. We also found a very high correlation between word prominence, R_w, and the prominence of the main stressed syllable in the word.

F_0 CORRELATES OF PERCEIVED STRESS

The F_0 contour contributes to the relative emphasis of syllables and words both by the depth of local word accent modulations and by overlaid sentence or focal stress.

The tonal patterns ascribed to Swedish word accents are HL* for accent I and H*L for accent II where the * indicates an allocation of the tone to the main syllable, Bruce (1977). A greater degree of prominence, i.e. focal accent, adds an H tone, thus HL*H for accent I and H*LH for accent II. The H of the HL* fall of accent I occurs in the preceding syllable and thus earlier than the H* of accent II. The H of focal accent II occurs in a following syllable while the L*H rise of the focal accent I already starts in the vowel of the main syllable.

We have quantified all F_0 measures in semitones, which gave us comparable values for female and male subjects. Our findings support the established importance of the accent II H*L drop, the magnitude of which showed a relative high degree of correlation *(r=0.7)* with perceived stress level, R_w. The speed of the drop showed a weaker correlation *(r=0.35)* with R_w. As expected, we found a relative stability of the L while the major part of the drop was due to a higher starting point H*. Because of difficulties in separating out the domains of successive accents we did not measure the relative height of the secondary peak H of accent II which is a wellknown prominence correlate. Our observations on interrelations between accent II F_0 features compare well with those of Engstrand (1989), but for greater modulation depth in our data.

Accent I F_0 modulations are harder to model. Our main parameter, the L*H relative increase, showed an $r=0.4$ correlation with R_w. Both the H and the L* of the HL* increased somewhat with stress and L*more than H. In the reading of iambic verse we have even found a reversal, i.e. L* higher than the preceding H.

INTENSITY

A study of relative intensities showed an average trend of 2 dB higher intensity in stressed than in unstressed syllables. A general hierarchy of duration versus F_0 and intensity as stress correlates was established.

One specific question that has been raised is to what extent intensity is related to properties of the vocal sound source and if source measures would be of special interest, e.g. for eliminating inherent differences in vowel intensity. A candidate for source strength is the negative peak of the volume velocity derivative at glottal closure, i.e. the rate of flow decrease at closure. This is labelled E_e, Fant, Liljencrants and Lin (1985). Although E_e and also F_0 are basic proportionality factors for intensity, there also enter source slope and formant bandwidths as important determinators adding to the effects of formant frequency and zero frequency patterns.

Intensity also shows an inverse relation to the relative emphasis of Swedish close vowels [u:], [ʉ:], [i:] and [y:] which are articulated with a gesture towards closure.

INTEGRATION OF PROSODIC PARAMETERS

A composite view of prosodic data of one sentence from our standard novel text is shown in Figure 2. The functions A, B and C below the spectrogram pertain to various methods of deriving the source amplitude $E_e(t)$. It was found that function C, the envelope of the negative part of the oscillogram, provides a rather close match to the proper inverse filtering. The conclusion is that inverse filtering is not needed for deriving approximate source amplitude functions, at least not for male speakers. A Hi-Fi recording of the speechwave is sufficient.

The $E_e(t)$ contour along the utterance conforms with the low pass, LP1000 Hz, intensity profile. The general trend in Figure 2 is that of a gradual decline of about 8 dB from the first stressed syllable to the last syllable which is also stressed but of greater subjective prominence as implied by the R_w function on the top of the figure.

There are two prominent examples of articulatory narrowing causing E_e reduction, in the middle of the long and stressed vowel [i:] of the word *skrivit*, and in the [u:] of the word *stor*, which as previously discussed increases with emphasis. This fact and the general trend of declination of intensity in the sentence complicate the use of intensity as a stress correlate. The $E_e(t)$ does not seem to have an advantage over intensity.

The F_0 accentual modulations in the accent II word *skrivit* and the accent I words *stor* and *sal* are apparent stress correlates adding to the basic duration correlates. In this sentence we found a correlation R_w*=4.8+5.6*S_i *(r=0.9)*.

In a study of a whole paragraph containing 9 complete sentences and 23 major groups separated by pause breaks we observed an average intensity downdrift of 9 dB from the highest value to a value sampled in the middle of the vowel of the final syllable of the group. Unstressed prepause syllables were found to be about 2 dB below average values when situated at a continuation juncture and up to 15 dB lower at sentence endings with a semantical break, in which case the prepause lengthening also was reduced.

Prepause lengthening was of the order of 50 to 150 ms for stressed syllables and 50-100 ms for unstressed syllables. Final lengthening as a possible contribution to stress is usually compensated by a final intensity decline and appears anyhow to be anticipated by the listener.

On the whole, the intensity contours largely follow the F_0 contours. A typical exception is the accent I induced L*H rise in F_0 during a focally stressed close vowel.

INTERACTION AS A PROSODIC CUE

Lack of articulatory closure and thus of vowel-consonant contrast is a sign of deemphasis but is also a personal speaker characteristic, Fant, Kruckenberg and Nord (1991a). Temporal contrasts as a prosodic feature is exemplified in Figure 3, which pertains to the word *behålla* uttered in focal position and in prefocal position. There is an apparent loss of intensity contrast in the prefocal position, which is the natural consequence of incomplete articulatory closure for [*l*] and an incomplete glottal abduction for the [*h*]. A related aspect of articulatory dynamics is reduction of spectrum pattern contrasts, e.g. vowel reduction. Advanced articulatory-acoustic modeling is needed for systematic studies of emphasis and deemphasis. Potentially, a complex of pattern details and interactions can be related to the modification of a single articulatory gesture, thus fascilitating the integration of segmental and suprasegmental aspects of speech prosody.

Acknowledgements: These studies have been supported by grants from the Swedish Council for Research in the Humanities and Social Sciences and the Bank of Sweden Tercentenary Foundation.

REFERENCES

G. Fant and A. Kruckenberg (1989), "Preliminaries to the study of Swedish prose reading and reading style," *STL-QPSR* 2/1989, pp. 1-83.

G. Fant, A. Kruckenberg and L. Nord (1991a), "Prosodic and segmental speaker variations", *Speech Communication* 10, 1991, pp. 521-531.

G. Fant, A. Kruckenberg and L.Nord (1991b), "Temporal organization and rhythm in Swedish", *Proc.12th Intern.Cong.Phon.Sc., Aix-en-Provence, 19-24 August 1991,* Vol.I, pp. 251-256.

G. Fant, A. Kruckenberg and L. Nord (1992), "Prediction of syllable duration, speech rate and tempo", *Proc. ICSLP 92, Banff,* Vol 1. pp. 667-670.

G. Fant, J. Liljencrants and Q. Lin (1985), "A four-parameter model of glottal flow", *STL-PSR 4/1985*, pp. 1-13.

G. Bruce (1977), *Swedish Word Accents in a Sentence Perspective,* CWK Gleerup, Lund 1977

O. Engstrand (1989), "F_0 correlates of tonal word accents in spontaneous speech: Range and systematicity of variation," *PERILUS*, Univ. of Stockholm, No X 1989, pp. 1-12.

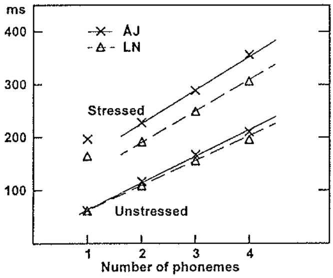


Figure 1. *Average syllable duration.*

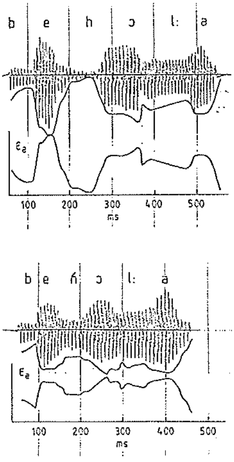


Figure 3. *Temporal contrasts comparing a word in focal (above) and in prefocal positions.*

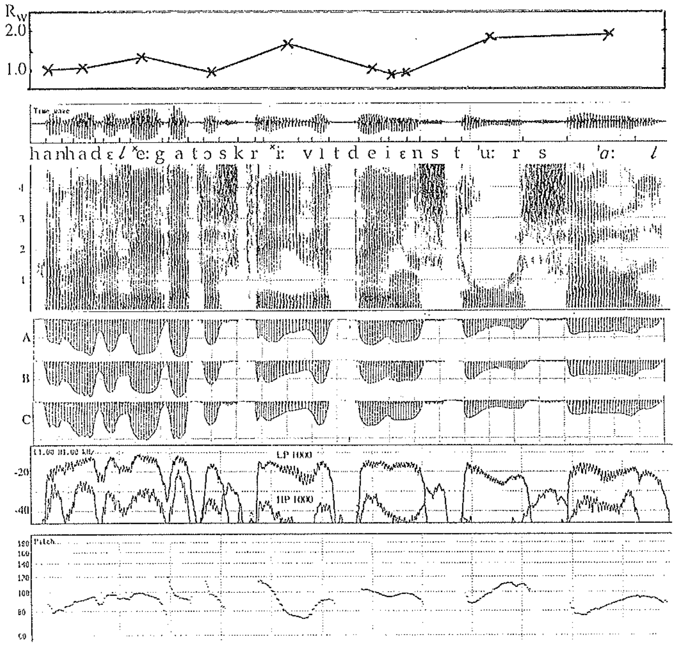


Figure 2. *Spectrogram with synchronous perceptual word prominence rating R_W, oscillogram, three different estimates of source amplitude E_e (A from inverse filtering, B from inverse filtering with constant settings, C from the negative side of the oscillogram), lowpass 1000 Hz and highpass 1000 Hz intensities, and F_0.*

The Grouping Function of F0 and Duration in two Prosodically Diverse Languages — Eskimo and Yoruba

Yasuko Nagano-Madsen
Department of Linguistics and Phonetics
Lund University
Sweden

ABSTRACT

This paper discusses how F0 and duration are utilized as local prosodic cues in two prosodically diverse languages — Eskimo and Yoruba. The analysis of text-reading and radio news-broadcasts has indicated that (1) F0 and duration are used in a mutually exclusive way in the two languages, (2) different relationships exists between F0 and duration for the two languages, but (3) there is a similar principle in the use of local prosodic cue in dividing a paragraph into smaller constituents.

INTRODUCTION

Languages are known to utilize a number of acoustic phonetic cues such as F0, duration, intensity, silence, and voice quality in order to signal prominence, boundary, and coherence in the stream of speech. In many European languages the acoustic variables of F0 and duration are often combined for such purposes, i.e. manifestation of lexical stress in Germanic languages, and boundary tone in French etc.. In this paper, I will discuss how F0 and duration are exploited as local prosodic cues in Eskimo and Yoruba, focusing the relation between the two acoustic dimensions and how they are utilized in dividing a paragraph into smaller constituents.

The present paper is a progress report of an on-going project called "MULTILINGUAL PROSODIC RULES, with specific reference to Eskimo, Japanese, and Yoruba". The immediate goal of the project is to extract rules and parameters which are necessary in order to describe the basic differences in the acoustic patterning of the three languages which are chosen on a typological basis. Eskimo represents a lexical quantity language where duration is the phonetic correlate (Mase and Rischel 1971, Nagano-Madsen 1992) whereas Yoruba represents lexical tone language for which F0 is the phonetic correlate. Eskimo and Yoruba can thus be seen as two extremes in the prosodic dimension, while Japanese comes somewhere in between. An underlying assumption to such an approach is that there is a certain correlation between properties of language and its spoken medium. A specific hypothesis for the current topic is that, when F0 or duration is used for signalling lexical property such as quantity and tone, there must be a limit in using the same acoustic cue for other purposes. Therefore, features other than those used for lexical information will play the major role for such purposes.

DATA AND ANALYSIS

The material used for the analysis consist of a reading of (1) a short text which was used earlier for duration analysis in Nagano-Madsen (1992), and (2) longer texts and news readings which were collected from naturally occurring radio broadcast. In the present paper, the results from three speakers for each (one text-reading and two news-readings for Yoruba and two text-readings and one news-reading for Eskimo), were analysed (total recording time of 25 minutes for each language). Durational analysis was carried out both using mingographs and CSL, while F0 analysis was done on the LUPP program installed on a Macintosh.

ACOUSTIC ANALYSIS

Duration

Durational characteristics of Eskimo and Yoruba were compared earlier by using short texts (Nagano-Madsen, 1992). The most notable difference between the two languages was the presence vs absence of extremely lengthened vowels. While in both languages, the duration of a large majority of vowels is concentrate between 30-120ms, Yoruba had many extremely long vowels.

In the present study, the news-broadcast of two Yoruba radio announcers were examined for this feature.The lengthened vowels were found to be present in both speakers. These vowels are extremely long, some even up to 500-600 ms. Table 1 shows the mean duration of lengthened vs non-lengthened vowels in Yoruba analysed from a female announcer. There was also a notable difference in the magnitude and frequency of the lengthened vowels in Yoruba between the two announcers (Figure 1).

Table 1. *Lengthened vs non-lengthened vowels in Yoruba. Analysis of first 5 minutes news-broadcast by female speaker (S2).*

	mean (ms)	SD	min - max	N
lengthened	300	99	150 - 643	95
not-lengthened	67	23	19 - 146	466

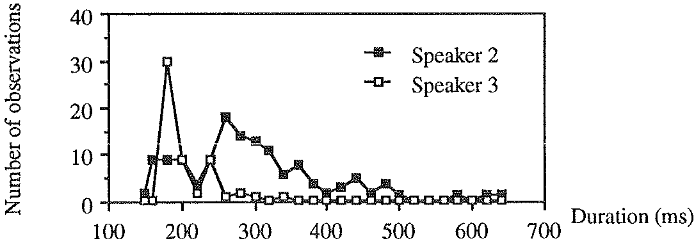


Figure 1. *Discrepancy in the frequency and magnitude of lengthened vowels between the two announcers of Yoruba. Analysis of 10 minutes news-broadcast for each speaker.*

The question arises as to where these lengthened vowels occur. Some words with lengthened vowels have been noted and discussed in Yoruba literature for some time (Ward 1956, Bamgbose 1966, Roland 1969). All these categories were present in the current material but the examples of lengthened vowels extend much further. It is beyond the scope of the present paper to discuss each categories and this is clearly the area which deserves much attention from Yoruba linguists. Though the exact function of these lengthened vowels may differ, i.e. in some cases it may be more appropreate to speak of emphasis and in other cases of boundary and coherence, one possible generalization which covers all the instances is to say that they all occur at initial position of major constituents such as clause VP etc.. Interestingly, Yoruba seems to use shortened vowels as well, though their occurrence is fewer and not as sizable as lengthened ones. The distribution of lengthened and shortened vowels are taken up under DISTRIBUTION.

Fundamental frequency (F0)

The durational prominence which is frequent in Yoruba was totally absent in Eskimo. Eskimo, instead, appears to use the F0 dimension maximally both as local and global intonation. Locally, each phonological word is characterized by a terminal tonal contour H-L which appears on the last two vowel morae. At phrase-final position, an additional H is added finally, shifting the word property H-L to penultimate and antipenultimate morae respectively. The phrasal H is typically lowered at sentence final position in text-reading material. The present analysis differs significantly from the previous one (cf. Rischel 1974). These features were common for all three speakers examined.

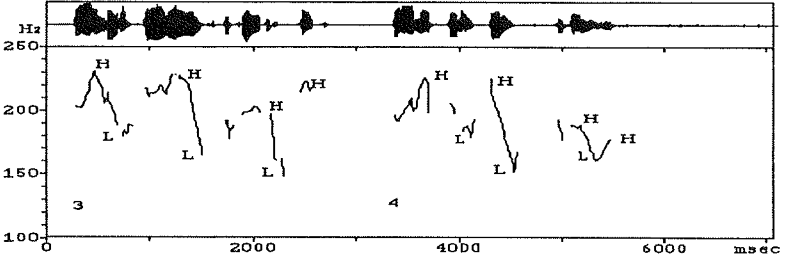


Figure 2. Manifestation of Eskimo phrase-internal and phrase-final contours: (3) *kaálì aqqanilínnìk ukióqàrpóq*, (4) *Kaalállù sisámàník.*

RELATION BETWEEN F0 AND DURATION

Perceptually, the local pitch prominence in Eskimo is signalled without marked durational increase. In order to examine this point, measurement was done for those vowels with or without pitch prominence (Table 2). The acoustic measurements show clearly that local pitch prominence in Eskimo is free from durational increment. The lengthening of vowels in Yoruba, however, seems not to be free from the F0 dimension. Table 3 summarizes the surface tone type of the lengthened vowels. There is a strong indication that lengthening is bound to high or mid tones.

Table 2. *Relation between terminal pitch prominence and duration. Analysis of total of 10 minutes speech.*

	mean (ms)	SD	min - max	N
H	71	19	39 - 126	123
not H	75	23	38 - 128	126

Table 3. *Relation between the lengthened vowels and (surface) tone type in Yoruba. Based on the analysis of two announcers reading 10 minutes each.*

tone	N of occurrences
H	165
M	29
L	0

DISTRIBUTION

In the previous sections, it was shown that Eskimo and Yoruba utilize different acoustic property as local prosodic cue. This section examines how these local prosodic prominences, be they F0 or duration, are distributed in relation to their position in a sentence (Tables 4 and 5). The Eskimo analysis shows that H-tone prominence is preferred in sentence-internal position and L-tone ending is preferred dominantly at sentence-final position. Furthermore, there was a strong preference to have L tone at paragraph-final position. Lengthened vowels in Yoruba were found exclusively at sentence-internal position. Shortened vowels occur both sentence internally and finally, but they occur consistently at constituent final position.

Table 4. *Distribution of High vs not-High local feature relative to sentence position in Eskimo. Three speakers pooled.*

prosodic feature	sentence internal	sentence final
High	103	27
Low	0	65

Table 5. *Distribution of lengthened and shortened vowels in Yoruba in relation to position in sentence. Three speakers pooled.*

prosodic feature	sentence internal	sentence final
lengthend	193	1
shortened	22	13

DISCUSSION

The results of the present study support the initial hypothesis which predicts the co-ordinative nature among prosodic properties to arrive at an efficient communication system. It has been shown that the acoustic property of F0 and duration are utilized in a mutually exclusive way to mark local prominence in the two languages. In Eskimo, where the durational contrast between single and geminated segments dominates the utterances, the local prominence is signalled exclusively by F0 without durational variation. Yoruba, on the other hand, shows wide range of freedom in the lengthening of vowels. Yoruba data also suggest that the initial hypothesis to be elaborated further since lengthening of vowels were bounded to either high or mid tones, indicating strong correlation between tone and duration.

The lengthening of vowels in Yoruba appears to have a more complicated function than pitch prominence in Eskimo, which regularly correlates with the end of phonological word form. Despite differences in the exact function of these local prominences, and despite differences in the exact acoustic devices the two languages employ, the signalling principles utilized in dividing a paragraph into smaller constituents were found to be similar. In both languages, it is the more prominent form such as high pitch or lengthened vowel which appear in sentence internal position, while in sentence final position the other end (low, or shortened) is preferred. However, it is quite common that sentences end with H tone (in Eskimo) and lengthened vowels (in Yoruba) when read in isolation. Therefore these features are better seen not as acoustic phonetic correlates of specific linguistic entities such as sentence or phrase, but rather they are signals which are dependent on a specific speaking style.

ACKNOWLEDGEMENTS

The present research is supported by the Swedish Research Council in Humanities and Social Sciences. I am grateful to Gösta Bruce for the inspiration to undertake the decomposition of Eskimo boundary tones. I am deeply indebted to Akin Akinlabi for various discussions and in particular for supplying me with a syntactic analysis of the text material on Yoruba. July Ikomi-Kumm helped me as consultant for transcription and translation of Yoruba news-broadcasts. Shiela Dooley-Colberg is kindly acknowledged for her generosity in discussing syntactic matters.

REFERENCES

A. Bamgbose (1966), *A grammar of Yoruba.* West African Language Monographs, no.5. (Cambridge University Press).

Mase, H. and J. Rischel (1971), "A study of consonant quantity in West Greenlandic", *ARIPUC.* University of Copenhagen, Vol. 5, pp.175-247.

Y. Nagano-Madsen (1992), *Mora and Prosodic Coordination. Phonetic study of Eskimo, Japanese, and Yoruba.* Doctoral dissertation. (Lund University Press, Lund).

J. Rischel (1974). *Topics in West Greenlandic Phonology* (Akademisk Forlag, Copenhagen).

E.C. Rowlands (1969), *Teach Yourself Yoruba.* (English Universities Press, London).

I. C. Ward (1956), *An introduction to the Yoruba language.* Reprinted. (W. Heffer & Sons Ltd., Cambridge).

F0 Troughs and Prosodic Phrasing

Anne Wichmann
Department of Languages
University of Central Lancashire, Preston PR1, UK

ABSTRACT

This paper describes prosodic boundary phenomena in naturally occurring speech. In two different speaking styles, read and spontaneous, a systematic low trough in the f0 contour is found to correlate with both syntax and topic structure. There is also evidence for systematic topline declination between such boundaries, but only in the read speech.

INTRODUCTION

There is considerable interest in the nature and function of prosodic boundaries. Knowles (1991) identifies a set of features (temporal, intonational and segmental) which co-occur with perceived tone group boundaries in the Lancaster/IBM Spoken English Corpus (SEC). This contains 50,000 words of prosodically transcribed speech. Bruce et al (1991), working on Swedish, investigate the temporal and intonational features of prosodic phrasing in association with syntactic boundaries.

The identification of prosodic boundaries forms the basis of earlier descriptions of larger prosodic domains, for example the "pitch sequence" (Brazil et al 1980) and "paratone" (Brown 1977). The phonetic cues to the boundaries of these domains are sought in pitch maxima on prominent syllables and in pitch movement. Brown (op cit) also claims an internal systematicity consisting of regularly descending pitch height on stressed syllables within each paratone.

There are therefore two issues to be addressed:
(i) What are the phonetic features which cause phrase boundaries to be perceived?
(ii) Do these boundaries define prosodic domains which also display systematic internal features?

The present paper describes the distribution of marked f0 minima (troughs) in relation to other prosodic features and in relation to the text itself. In its weakest form this feature is an f0 trough which is low in relation to preceding and following f0 minima; in its strongest form it is an f0 trough which is close to the speaker's base line and displays approximately the frequency in each case. The following describes the distribution of the strong form of this boundary feature in two different speaking styles, read speech (a news broadcast), and spontaneous speech. The aim of this research is to identify significant prosodic features which can be labelled in a speech data-base, providing information for a large-scale study of prosodic segmentation.

ANALYSIS I - NEWS BROADCAST

The section of text analysed here is the first news item (six sentences) in a broadcast news summary, taken from the Spoken English Corpus. For this speaker, the lowest in each series of f0 troughs was approximately 75Hz.

Syntactic correlates

In the section of text analysed there were 14 occurrences of the low f0 trough. Of these, all coincided with a syntactic boundary; 6 of these were sentence boundaries (one for each sentence); 5 were finite clause boundaries within a sentence, and 3 were sentence-internal phrase boundaries. Of the phrase boundaries, one marked the end of the NP subject of the first sentence, and two marked the beginning of a sentence-final adverbial (prepositional) phrase. Of the 15 finite clauses in the text, only four are **not** marked prosodically with a low f0 trough.

Textual correlates

The placement of a strong prosodic boundary at minor syntactic boundaries does not appear to be random. The f0 trough which co-occurs with the end of the noun phrase subject of a sentence has a textual function. The end of the first noun phrase at the beginning of a news item is often treated as prosodically separate; the final prominent syllable is frequently given a falling contour instead of the non-final contour (rise or fall-rise) normally expected at a non-final syntactic boundary. This alone gives the impression of a degree of finality, which is reinforced if the pitch falls to the speaker's base-line at this point. This is the prosodic equivalent of printing the first phrase of a newspaper paragraph in block capitals or bold print. The prominence given in this way, either visually or aurally, serves to mark a new topic in the text.

The f0 trough at the end of sentences is a predictable indication of syntactic completeness. Its occurrence before a sentence-final adverbial is typical for professional news broadcasts. Readers tend to mark a degree of finality at the earliest point at which the sentence is **potentially** complete. This may be related to mental processing of the text ahead; it may also be that there is a maximum or minimum length for a prosodic phrase (if the boundaries in question can be assumed to demarcate units of some kind).

Co-occurrence with other prosodic features

The prosodic transcription of this section of text contains 6 major tone group boundaries. All of these co-occur with a low f0 trough. There are also 23 perceived minor boundaries, only 6 of which are signalled in this way. Thus 12 of the 14 f0 troughs occur at perceived boundaries. The two occasions where the feature does **not** co-occur with a tone group boundary are before adverbial phrases. Pausing in read speech tends to be at syntactic boundaries; it is therefore not surprising that there is a high degree of co-occurrence in this text between low f0 troughs and pauses. Of the 14 f0 troughs, 11 co-occur with pauses, 6 at the end of sentences, four at clause boundaries and one at a phrase boundary. The remaining f0 trough at a clause boundary co-occurs with final syllable lengthening.

Internal regularities

Across each section of text between two f0 troughs there was a systematic declination of f0 maxima. This suggests that, in this speaking style at least, there may be a prosodic unit - a phrase or declination domain - which is larger than a tone group, and can be defined both in terms of its boundary features and in terms of internal regularity.

ANALYSIS II - SPONTANEOUS SPEECH

The material analysed here was part of a spontaneous, unrehearsed monologue by a female speaker. As with the first text, the base line of the f0 contour was analysed. The same marked f0 troughs were observable and displayed a consistent f0 value. (159Hz)

Syntactic correlates

Each of the f0 troughs co-occurred with a syntactic (phrase or clause) boundary. However, they occurred overall less frequently than in the read speech and there were many syntactic boundaries which were not marked in this way. In the sequence analysed there were 35 finite clauses. Only 13 of a total of 22 f0 troughs occurred at such clause boundaries; a further 9 occurred at phrase boundaries (6 before final adverbials; 1 before a subject complement in final position; two after an NP subject).

Textual correlates

The topic structure in this text was identified independently, partly semantically and partly by means of other prosodic features: for example there was a marked increase in speech rate immediately preceding a new topic and the new topic itself co-occurred with higher than average f0 maxima. While the read text dealt only with one topic, i.e. one news item, the spontaneous text contained a range of topics. Although the relation between syntactic boundaries and f0 troughs was weaker than in the read text, the association with topic structure remained the same. The f0 troughs marked, not surprisingly, the end of a topic. They also marked the end of the noun phrase or clause announcing a **new** topic. (see Figure 1)

Co-occurrence with other prosodic features

In the same stretch of speech there were 46 pauses. Of these only 8 co-occurred with an f0 trough. 14 pauses were hesitation phenomena and associated with repetition or pause fillers ("er") in the running text. 9 Occurred immediately **after** clause boundaries, i.e. after either a conjunction or discourse marker *(and, anyway, for example).* Low f0 troughs on the other hand, if they occurred at all, occurred **before** a conjunction or discourse marker, i.e. at the syntactic boundary. (see Figure 1)

Internal regularities

Unlike the read text, the systematic declination of f0 peaks did **not** occur in this text. The stretches between low f0 troughs were on average much longer than in the read speech, varying between one word and 68 words but averaging around 12 words. Unlike the read text, these stretches did not display a systematically declining topline, and therefore show no evidence of prosodic units which could be defined in this way.

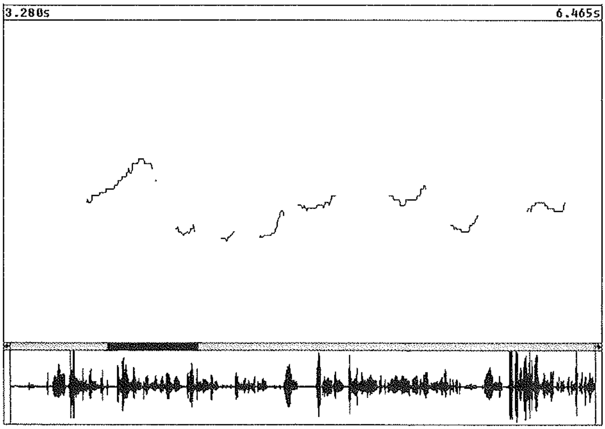


Figure 1. *Waveform and f0 contour of the text: "(She asks me things that) amaze me sometimes I mean for example ..". This is the start of a new topic. There is a low f0 trough on the first syllable of "sometimes" and a pause after "for example".*

DISCUSSION

These two speaking styles are markedly different both prosodically and syntactically. The spontaneous speech, for example, displays far fewer f0 boundary troughs than the read speech, and these do not co-occur to the same extent with other boundary features, particularly pauses. There are nonetheless similarities. The occurrence of low f0 troughs is in each case related in some way to syntax. Both texts also have in common that the low f0 troughs serve to indicate topic structure. This is particularly marked in the spontaneous speech. This suggests that the relationship to syntax is subordinate to, or a function of, the relationship to meaning. Further research is in progress to examine the distribution of f0 troughs in other speaking styles. The topic structure of both texts analysed so far is relatively simple. It remains to be seen whether the more complex topic structure of, for example, fiction is reflected in the nature of the prosodic boundaries.

Evidence for larger prosodic units is less convincing. Only·the read speech displays any internal regularity between boundaries, suggesting that such regularity may only be a feature of impersonal read speech.

IMPLICATIONS

The results of this investigation support the view that the base line of the f0 contour holds important information. As Bruce (op cit) suggests, a lowered base line is a signal of phrasing. My research indicates that in naturally occurring speech these low points may signal phrase boundaries not only because they are low in relation to the context but also relative to the speaker's own range. Further research is necessary to establish whether this is common to all speakers and all speaking styles. If this is so, the occurrences of f0 minima could be of use in speech recognition, by giving not only syntactic information but also information about structure at the level of the text. Such insights might also serve to produce a more realistic sounding intonation contour in synthesised speech.

REFERENCES

Brazil, D., Coulthard, M., Johns, C. (1980) *Discourse Intonation and Language Teaching* (Longman, Essex).

Brown, G. (1977) *Listening to Spoken English* (Longman, London).

Bruce, G., Granström, B., Gustafson, K., House, D. (1991) *Prosodic Phrasing in Swedish* in Working Papers 38, 5-17 Lund University, Dept of Linguistics

Knowles, G. (1991) "Prosodic labelling: the problem of tone-group boundaries" in S. Johannson, A-B. Stenström (eds) *English Computer Corpora* (de Gruyter, Berlin)

Phonetic Evidence of Narrow and Wide Temporal Scope for Prosodic Constituents in French

Louise Levac, Henrietta J. Cedergren, Hélène Perreault
Département de linguistique
Université du Québec à Montréal
C.P. 8888, Succ. A., Montréal, Québec, Canada H3C 3P8

ABSTRACT

This research addresses the issue of motivating surface prosodic structure in spontaneous French corpora through the use of three complementary methods: auditory, acoustic-phonetic and distributional analyses. This hybrid phonetic-phonological approach has allowed the identification of four distinct levels of prosodic grouping: the intonational phrase, the rhythmic group, morphophonological binding and the syllable.

INTRODUCTION

Recent developments in phonological theory have argued for the existence of prosodic structure as an interface component between surface syntactic representation and phonological representation. Prosodic structure defines the domains of application of phonological processes which cannot be described solely on the bases of their phonotactic environments; it delimits the domains within which patterns of prominence and patterns of timing are defined. While proponents of phrase phonology agree on the hierarchical nature of prosodic structure and assume that the syllable and the foot constitute lower levels of the hierarchy (Nespor & Vogel, 1986; Selkirk, 1986), the degree of hierarchical complexity mediating between syllables and intonational phrases is still under discussion.

This research investigates the motivation of hierarchical complexity in surface prosodic structure in spontaneous French. Auditory, acoustic-phonetic and distributional analyses provide evidence of four distinct levels of prosodic grouping: the intonational phrase, the rhythmic group, morphophonological binding and the syllable.

METHODS

A data-base has been constructed which consists of excerpts of running speech of approximately three minutes in length extracted from hour-long recordings of spontaneous conversations from a subsample of a socially balanced sociolinguistic corpus of Montreal French. The sample consists of eight speakers differentiated according to sex, age and social class. Speaker ages correspond to two generational categories: twenty to twenty five years of age or fifty five and over; an equal number of working class and middle class speakers were selected.

AUDITORY ANALYSIS

The orthographic transcription of the selected excerpts was parsed and coded for perceived prosodic grouping and prominence. Thus, as illustrated in Table 1, three levels of prosodic organization were identified: phonetic syllables, rhythmic groups and intonational phrases. Perceived prominence was distinguished as either demarcative, associated with the right boundary syllable of rhythmic groups or intonational phrases, or non-demarcative, secondary prominence, associated with non-final syllables.

Table 1. *Exemplification of prosodic transcription.*

(1) Ah! La boxe] Disons qu'elle comprend] *cinq entraînements principaux*]

(2) s ~e k) / ~a / t R ɛ n / m ~a / } p R ~ɛ / s p o /]
258 102 253.5 134 182 230 (ms)

(3)] = intonational phrase boundary, } = rhythmic group boundary,
/ = syllable boundary,) = secondary prominence.

ACOUSTIC ANALYSIS

The speech excerpts were digitized on a microcomputer at a sampling frequency of 16 kHz. Segment boundaries were identified by manually placed cursors on digital spectrograms time-aligned with a waveform; segment durations and labels were stored in an automatically generated file which was also coded for prominence and prosodic grouping.

Linear regression analysis was used to model the relationship between syllable duration and features of surface prosodic structure and syllable composition. Estimate values revealed characteristic differences in temporal effects that are sensitive to prosodic constituent type. Intonational phrase boundary is systematically implemented as the lengthening of the last two syllables in the prosodic constituent, while there are interspeaker differences for rhythmic groups. In contrast, non-demarcative prominence or secondary stress targets only a single syllable. Moreover, degrees of temporal marking contributes to the distinction of hierarchical levels.

DISTRIBUTIONAL ANALYSIS

We have further undertaken a limited distributional analysis to clarify the relationship between surface prosodic structure and underlying linguistic structure. This line of research was thus initiated by taking a long and hard look at lexical sequences that were produced with secondary prominence. All of the expressions with secondary prominence were classified according to morphosyntactic categories; they consisted of sequences of two nouns (N N), noun followed by preposition and noun (N de N or N à N), noun followed or preceded by an adjective (N A or A N), verb plus noun (V N), numeral adjective plus noun, and a few colloquial expressions.

A corpus of all similar morphosyntactic categorial sequences in the data set was then tested, first in a judgement acceptability experiment, and second in a relational analysis of normalized durations. The corpus consisted of 252 lexical sequences, such as, *Ville Laval, école anglaise, petit cousin, faire une carrière, trois fois, disons que, etc.*

Acceptability experiment

Two native speakers participated in the acceptability test. The first speaker's task was that of producing each lexical sequence with three distinct prosodic configurations: 1) a neutral prosodic configuration without secondary stress, 2) a prosodic configuration with secondary stress on the first lexical item in the sequence, and 3) a minor prosodic phrase boundary after the first lexical item. The second speaker's task was that of judging the acceptability of each production either as good (+), unacceptable (-) or uncertain (?).

As illustrated in Table 2, the results revealed that although secondary prominence is optional, all sequences where judged acceptable when produced with non-final prominence; this was also true for the neutral prosodic configuration. The test also revealed that a minor subset of the data was judged unacceptable when the lexical sequences were produced as separated by a minor prosodic boundary; thus providing evidence that prosodic phrasing assignment is constrained by the categorial status of lexical sequences. Compound lexical units were judged unacceptable when produced with an internal prosodic boundary in contrast to syntactic phrases.

Table 2. *Examples of acceptability judgements for lexical sequences. Data taken from young middle class male speaker.*

Lexical sequence	Neutral prosodic configuration	Internal minor prosodic boundary	Non-demarcative prominence
corde à danser	+	+	+
goût à (la) musique	+	?	+
Ville Laval	+	-	+

Relational analysis

Z-score normalization of the durations of all segments in a speaker's dataset was used to determine syllable lengthening or shortening in standard units. A relational analysis of word final syllables in each lexical sequence used for the acceptability test from a young middle class male speaker in the sample showed two patterns in the dataset. With few exceptions, word final syllables in sequences defined as phrases in the acceptability test revealed a pattern whereby the final syllable in the first lexical unit appeared to be subject to shortening as evidenced by a negative z-score value. In contrast, compound lexical units revealed a pattern whereby the final syllable in the first lexical unit appeared to be subject to lengthening as evidenced by a positive z-score value. These patterns which we interpret in metrical terms as W(eak) when the value is negative and as S(trong) when the value is positive are represented in Figure 1.

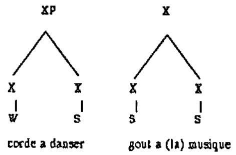


Figure 1. *Relational contrasts of lexical unit sequences and lexical phrases sequences as revealed by z-score values.*

Although, in general, the results of the acceptability test and the relational analysis of z-score transforms coincide, there are some exceptions. Further analyses which take into account the whole context of the utterances in the dataset should shed light on these exceptions.

CONCLUSIONS

We have presented evidence that surface temporal patterns in our dataset of spontaneous speech differentiate between properties of phrasing and secondary prominence. As revealed by statistical analysis, rhythmic group and intonational phrase boundaries are associated with lengthening of the last two syllables of the prosodic constituent, we call this modality of implementation wide scope prominence marking, while secondary prominence targets only one syllable, and thus is interpreted as narrow scope prominence marking.

A distributional analysis of a subset of the corpus has allowed us to examine some of the relationship between auditorily derived prosodic units and morphosyntactic categories. Thus lexical compounds in Montreal French cannot be split by a minor prosodic boundary. Moreover, lexical compounds appear to have a characteristic rhythmic pattern. The first lexical unit in the sequence is lengthened as measured by z-score normalization. This contrasts with sequences that may be interpreted as syntactic phrases. These sequences may be split by a minor prosodic boundary, and the first unit in the sequence is subject to shortening as measured by z-score values.

The structural constraints on prosodic phrasing are not paralleled by the distributional properties of secondary prominence; both lexical compounds and syntactic phrases may be produced with secondary prominence on the first unit. However secondary prominence, although optional, is positionally constrained; it targets the last syllable of the first lexical constituent, and thus provides evidence of an intermediate morphophonological prosodic domain in the dialect of French under analysis. Further research will extend and clarify the relationship between abstract phonological structure and phonetic patterns.

ACKNOWLEDGEMENT

This research was supported in part by the Social Sciences and Humanities Research Council of Canada under Grant No. 410-92-1840.

REFERENCES

F. Grosjean and A. Deschamps (1972), "Analyse des variables temporelles du français spontané, *Phonetica*, Vol. 26, pp. 129-156.

M. Nespor and I. Vogel (1986), *Prosodic Phonology* (Foris Publications, Dordrecht).

L. Santerre (1991), "Incidences du trait phonologique de durée vocalique sur la prosodie du français québécois", *Proc. 12th Internat. Cong. Phon. Sc., Aix-en-Provence, 19-24 August 1991*, Vol. 4, pp. 254-257.

E. Selkirk (1986), "On derived domains in sentence phonology", *Phonology Yearbook*, Vol. 3, pp. 371-405.

J. Vaissière (1991), "Perceiving rhythm in French?", *Proc. 12th Internat. Cong. Phon. Sc., Aix-en-Provence, 19-24 August 1991*, Vol. 4, pp. 258-261.

How to tell H% from L% in right-detached expressions in Norwegian

Thorstein Fretheim & Randi Alice Nilsen
Department of Linguistics
University of Trondheim
N-7055 Dragvoll, Norway

ABSTRACT

Native listeners' perception and identification of boundary tone - either low (L%) or high (H%) - in East Norwegian intonation contours is determined by more than just the presence of a falling or a rising tune at the end of the contour. Empirical evidence supports the claim that the H%-L% distinction in syntactically right-detached particles which are lexically specified for boundary tone is accessed most easily if H% is expressed inside and L% outside a focal Foot. This paper accounts for the observed constraints.

INTRODUCTION

We assume, with e.g. Pierrehumbert & Hirschberg (1990), that intonation contours are composed from *pitch accents* (word accents), *phrase accents*, and *boundary tones*. The paradigmatic opposition between a high (*H%*) and a low (*L%*) boundary tone in East Norwegian intonation has a more important function with some sentence elements than with others. It is especially important when it is used to differentiate right-detached particles that select *H%* from those that select *L%*. Syntactically right-detached ('tag') particles in spoken Norwegian are all attitudinal markers; some of them are realized on an *L%* slope, others on an *H%*-slope, while some permit either boundary tone, with a more or less determinate difference in pragmatic meaning between them.

THE EXPRESSION OF BOUNDARY TONES IN EAST NORWEGIAN

East Norwegian intonation is characterized by a syntagmatic contrast between *focal* and *nonfocal* phrase accents, the relevant prosodic phrase being the F(oot) (e.g. Fretheim 1992, Nilsen 1992). Each F starts with an obligatory *prosodic word* manifesting a left-edge word accent, which is *L** for Accent 1 and *H** for Accent 2, and is optionally followed by one or more unaccented word forms. The F also contains a phrase accent manifested as a right-edge *H*.

F constituents are grouped together under the IP category (Intonational Phrases) in the prosodic hierarchy. The IP is phonologically more akin to the *intermediate phrase* of Pierrehumbert and Hirschberg, than to their *intonational phrase*. The phrase-accentual *H* is raised to a higher F0 level at the end of an IP-final F, due to a *focus tone* which is a right-edge IP-level tonal phenomenon and a defining property of the IP.

A rising tune triggered by focus tone is a prerequisite of the expression of a minimal *L%-H%* distinction at the end of East Norwegian utterances. A postfocal F generated outside the IP category will lack focus tone and will therefore not be a possible vehicle for the generation of *H%*.

We contend that the East Norwegian system of intonational phrasing is not optimal for the expression of the *L%-H%* contrast. When the boundary tone of the IU (= Intonational Utterance) is *H%*, that tone will coincide temporally with the phrase-accentual *H* if realized inside IP. Our ability to perceive an abrupt decrease in F0 as a falling tune triggered by *L%* must depend at least partly on the extent of the F0 interval between the earlier maximum and the later minimum, and on the duration of the temporal interval between those maximum and minimum points in the intonation contour. An utterance-final syllable placed outside the IP domain enables a speaker to let the pitch drop from beginning to end in the syllable. On the other hand, an utterance-final syllable which is F- and IP-internal is realized on a F0 rise toward the target tone *H* before the F0 can start falling to *L%* in the same syllable. Under no circumstances is it possible to shift the F0 maximum from the IU-final syllable to the preceding syllable in order to leave more room for the fall to *L%*. One may ask whether the fall component of the local rise-fall contour of the final syllable can be extended in the temporal dimension to a point where the fall is perceived to be just as prominent as the falling tune on an IP-external particle. How much can the fall at the end of the utterance be stretched out temporally without loss of naturalness? Provided that the fall component of the rise-fall pattern aligned with the IU-final syllable is longer than the rise component, do we perceive the boundary tone as *L%* rather than *H%*?

We suspected that the answers to these questions might differ depending on whether the IU-final syllable is or is not a right-detached pragmatic particle that selects *L%*. Using utterances whose final syllable *da* could conceivably be interpreted either as a pragmatic particle or as a temporal adverb, we set out to test that assumption.

DA - A MULTIFARIOUS NORWEGIAN WORD

Most if not all uses of *da* in present-day Norwegian derive from the pro-adverb meaning 'then', which is used with reference to temporal as well as conditional clauses. As a right-detached particle, *da* has two discernable functions which, although they are related, should be kept lexically apart due to distinct sets of formal lexical properties. *L%-da* is an inference particle, and *H%-da* is a 'polarity reversal' particle (Fretheim 1989): the speaker attributes belief in the expressed proposition to the hearer, and challenges the hearer to reconsider his/her belief. While *L%-da* may be attached either to an interrogative or to a declarative, *H%-da* requires an interrogative 'host'. Also, in South-East Norwegian casual speech, the initial stop segment [d] of the polarity reversal particle may be elided but that of the inference particle may not.

The test sentence chosen was the interrogative *Bodde de der da?* (lit.: lived they there then?). As noted above, interrogatives license both *L%-da* and *H%-da*. Ten speakers of South-East Norwegian dialects were asked to listen to seven different auditory stimuli, each played three times. All seven utterances represented broad-focus intonation patterns with a focally accented time adverbial *der. Da* was inside the focal F in two utterances, and F- and IP-external in the remaining five.

Figure 1 is a F0 tracing of a male East Norwegian speaker's utterance of *Bodde de der da?*. The final F0 movement is to the right of the IP constituent and is distinctly falling. In Figure 2, *da* appears inside the focal F, which contains a succession of phonologically significant tones *H* and *L%*, a rise-fall contour produced in the course of the unaccented syllable *da*. (The parenthesis notation indicates the hierarchical IU structure. Vertical bars between tone specifications show F boundaries. The dotted vertical line in Figure 2 marks the beginning of the deliberately prolonged but not unnaturally long syllable *da* in that utterance. Observe also the [L] in the prefocal Accent 2 Foot, whose function is to preserve the so-called Obligatory Principle (OCP), which applies to Norwegian Feet but not across F boundaries.)

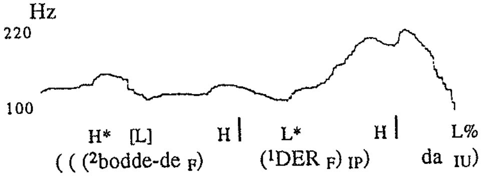


Figure 1

IP-external L%-da

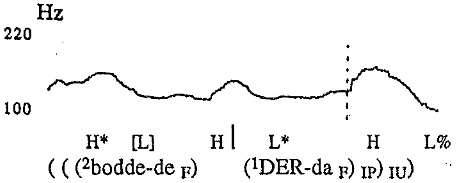


Figure 2

IP-internal L%-da

Figure 3 displays the F0 contour of the other utterance whose tag particle was placed inside the IP domain. This is the situation where the focus tone and the boundary tone coincide at the end of the IU.

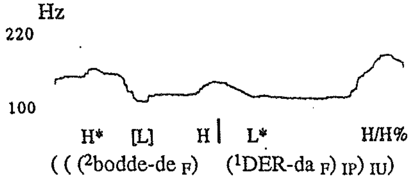


Figure 3

IP-internal H%-da

IDENTIFYING L% AND H%

The task of our ten South-East Norwegian informants was to select one of the three suggested interpretations *A-C*. They were told to select *C* only if they felt that neither *A* nor *B* was applicable. The stimuli were presented in a random order.

A - *Er du virkelig sikker på at det var der de bodde?*
('Are you really sure that that was where they were living?')
B - *Bodde de der, altså?*
(a request for confirmation of the assumption that they lived there)
C - *Bodde de dér på den tida?*
('Did they live there at that time?')

The particle in the intonation contour of Figure 1 was identified as the inference particle *L%-da* by nine out of ten informants; only one person associated that utterance with interpretation *A*, and no one chose the time adverb interpretation. The IU-final F0 fall in Figure 2 on the other hand caused only two informants to identify the particle as *L%-da*; four chose paraphrase *A*, which we interpret to mean that they identified *da* with *H%-da* in spite of the final fall, and another four found *C* to be the only possible paraphrase of what they heard when they were faced with the contour of Figure 2. Eight informants identified *da* in Figure 3 as the polarity reversal marker *H%-da*, while the remaining two chose the inference interpretation *B*.

Our informants' reactions to the intonation contour of Figure 2 is remarkable. Acoustically there is just as much of a final falling contour in Figure 2 as in Figure 1. How can we explain that a majority of informants paid attention to the falling tune at the end of Figure 1 but apparently ignored the fall at the end of Figure 2? The intonation structures of Figure 2 and Figure 3 share one feature that distinguishes both from Figure 1, namely the intonational phrasing. The rise to the focal maximum in the IP-final syllable is entirely within the unaccented syllable *da*. In other words, there is a rise-fall contour for *da* in Figure 2, as opposed to the straight fall through the right-detached particle in Figure 1. While a rise-fall tune in the IU-final syllable may be good enough for the realization of *L%* in most cases, it seems that the rise component must be eliminated altogether when the fall triggered by *L%* is on a right-detached pragmatic particle which selects *L%* realization. Even if the fall component of the rise-fall at the end of Figure 2 has a longer duration than the preceding rise and also a longer duration than the fall in Figure 1, the rise to the focal maximum appears to outweigh the ensuing fall.

Observe that there is no rule saying that a right-detached particle must be IP-external. *Da* was consistently identified as particle in Figure 3 where it is IP-internal; the favored interpretation of that utterance was *A*. There was more uncertainty in the informant group when what was intended to be *H%-da* was placed outside IP. The utterance with an IP-external rise showed five votes for *A*, two for *B*, and three for *C*.

To conclude, the right-detached inference particle *L%-da* must be F- and IP-external, because otherwise it would be impossible to produce a particle whose F0 movement is falling from the start. A tag particle that is lexically specified as an *L%* item does not tolerate a rise-fall contour on the particle. This finding supports the Boundary Tone Agreement Condition (BTAC) postulated by Fretheim (forthcoming). He found that when an East Norwegian utterance contains a sequence of right-detached particles, they must be either all *L%* or all *H%*, forming either a falling melody, or a rising melody, through all right-detached items. The result of our perception and comprehension test -- which is going to be followed up by a similar test in which synthetic stimuli will be used -- indicates that even when there is a single monosyllabic right-detached particle at the end of an utterance, the syllable must be perceived as producing a falling tune from beginning to end in order for the particle to be identified as an *L%* item. While the final fall in Figure 2 above may count as an *L%* fall if the word form is understood to represent the clause-internal time adverb, it does not count as *L%* if *da* is supposed to be the right-detached inference particle. Phonologically this is a weird situation. It seems that we are not able to identify the boundary tone in Figure 2 correctly unless we know whether the final lexical item is a pragmatic particle or the time adverb. If it is the latter, then the boundary tone in Figure 2 is *L%*; if it is the former, then the identity of the boundary tone may be indeterminable, due to the fact that the BTAC applies to right-detached particles like the inference marker *L%-da*. Our investigation has shown that the phonological analysis of the intonation structure of Figure 2 may depend on which *da* is being used.

Why did as many as four out of ten informants associate the utterance whose F0 contour is represented in Figure 2 with paraphrase *A*? One might expect it to be just as difficult to identify the IU-final rise-fall in Figure 2 with *H%* as with *L%*. The intutions of those four people combined with the large number of votes for an *H%-da* interpretation of Figure 3 suggest that the preferred realization of *H%-da* is an IP-internal realization. Utterances with an *L%* particle are seen to become acoustically maximally different from utterances ending in an *H%* particle when the latter kind of particle is integrated in the prosodic F domain and the former is placed outside F and IP, unaffected by the strong tonal constraints on the form of East Norwegian IPs.

REFERENCES

T. Fretheim (1989), "The two faces of the Norwegian inference particle *da*", in *Sprechen mit Partikeln*, ed. by H. Weydt (de Gruyter, Berlin), pp. 403-415.

T. Fretheim (1992), "Themehood, rhemehood, and Norwegian focus structure", *Folia Linguistica*, Vol. XXVI/1-2, pp. 111-150.

T. Fretheim (forthcoming), "The Norwegian 'Boundary Tone Agreement' condition", *Chicago Linguistic Society* 28, 1992.

R.A. Nilsen (1992), Intonasjon i interaksjon, unpublished doctoral dissertation, Department of Linguistics, University of Trondheim.

J. Pierrehumbert and J. Hirschberg (1990), "The meaning of intonational contours in the interpretation of discourse", in *Intentions in Communication*, ed. by P. R. Cohen, J. Morgan, and M.E. Pollack (The MIT Press, Cambridge), pp. 271-312.

Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis

Véronique Aubergé
Institut de la Communication Parlée, INPG/ENSERG - Université Stendhal, BP 25X, 38040 Grenoble, France
Email : auberge@icp.grenet.fr

ABSTRACT

We propose here a methodology and tools for the semi-automatic constitution of a intonative generation module. .

The first stage of the work was the analysis of a corpus recorded by a reference speaker and based on a set of linguistic presuppositions. These presuppositions are based on the concept of some structural ***rendez-vous*** *between the different levels of text on one part and the prosody. on the other part The processing of the data corpus was organized in a top-down hierarchy: sentences, clauses, groups and lexical units. The minimal symbolic unit is the syllable. For every level in the hierarchy, several initial classes of Fo contours are defined, each initially described by a maximal set of linguistic parameters. The validity of each class is first verified. Then the unification of the classes is systematically tested, using minimal pairs oppositions on the linguistic parameters. For every final class an average-contour is computed, which is a global form for this class. The result is a hierarchically structured* ***dynamic contour lexicon*** *of* ***global intonative forms*** *for which every representative is associated with a minimal set of distinctive attributes.*

Generation of prosody then consists in the calculation of prosodic patterns by the top-down cumulative superpositions of contours taken from the lexicon. An application is the automatic generation of prosody in a text-to-speech synthesis system, which must be adapted to a given application.

INTRODUCTION

The main difficulty that text-to-speech (TtS) synthesis has to deal with, is the generation of a well formed intonation on the sole basis of textual information and a few eventual extra-linguistic parameters concerning the speaker(s) and communication situation. Intonation is, nevertheless, a verbal language phenomenon (Bolinger, 1989). The transition from text –i.e., from written language– to speech is rather artificial or, at least, somehow special. Certain theories have suggested congruency relationships (that is, partial identity projections) between syntactic and intonative structures. We have adopted, in this study, a less restrictive perspective: intonation per se carries specific information and refers in fact to a particular performance structure of spoken language (Monnin & Grosjean, in press). The great coherency of the language structures implies that other linguistic structures are covered by intonation as well, in a somehow redundant fashion, especially in read-out text. This suggests the existence of *rendez-vous* nodes, from which sub-structures could be isolated. Specifically, this refers to the isolation of the sub-structures' global form, without expecting, however, to interpret its logic.

The aim was to propose a methodology with which a module could reproduce a specific utterance from a corpus (of a single speaker in a defined communication situation). This phonetic modeling corresponds to a lexicon of global intonative forms. It is structured according to a hierarchy of *rendez-vous* node levels. This lexicon, which is specific to the speaker and situation, links a symbolic input with a physical output, without explicitly involving the phonological level, although it is implicit in the structure of the lexicon.

CORPUS DESIGN

Prosodic coding

The prosodic parameters selected for this study are the fundamental frequency F0 and duration, and the minimal suprasegmental unit for both is the syllable (Campbell, 1992) even if this choice is not necessarily definitive (Barbosa & Bailly, 1993). The relative importance of F0 and duration in a given utterance context is subject to the choice of the speaker (Caelen, 1992).

For the segmental level, F0 was coded with three values (beginning, maximum excursion, end) for each syllabic nucleus. The duration was coded in two successive stages. For the segmental level, start and end points were stored for each vowel. For the supra-syllabic level, syllabic durations are measured on the stored syllabic F0 contours. Then, a measure termed *prosodic syllable length* is calculated for every syllable. Its intent is to quantify that part of the syllable duration which is not ascribable to inherent segmental phenomena, and it is calculated by subtracting the duration of a reference syllable from that of the syllable in the natural corpus.

Constraints: minimal pairs

The first step of the study presented here is the constitution of a corpus based on strict constraints of *minimal pair oppositions* which are supposed to give prominence to what we refer to as *rendez-vous* between the intonation and the linguistic structure nodes. The linguistic units are hierarchically organized: the highest level treated is the *sentence*, while the lower levels consist of the *clause*, the *group*, and *subgroup* (experimentally, it was confirmed that the subgroups exist inside the groups of 5 or more syllables long).

The factorial design of the corpus is based on the combination of minimal oppositions within attributes at each syntactic level. The basic opposition at the sentence level is one of sentence modality: declarative interrogative, or imperative. These modes are subdivided in fairly classic ways, declarative into positive and negative, and interrogative into direct, introduced, inverted, and elliptic. Sentence length in syllables was also treated, via sentences from 3 to 21 syllables long (see examples in Table 1).

At the clause level the oppositions concern syntactic dependency or embedding, linear position, and clause length. If either a clause is embedded or it is the matrix clause in which another clause is embedded, it is classed as *linked*; if not, it is *unlinked.* Unlinked clauses may be isolated, coordinated, or juxtaposed. Dependent clauses may be verb dependent or noun dependent (see examples in Table 1). The clause may be in initial, medial, or final position in the sentence, and length in syllables runs from 3 to 9.

The length of the groups varies from 2 to 15 syllables. Different values are defined for the nature of the group: nominal, verbal, adjectival, adverbial, and "grammatical word" group. The nominal group (NG) has been more most closely studied. When the NG is over 4 syllables long, it is arbitrarily decomposed into sub-groups. For small NGs (i.e., 4 syllables or fewer), the group is the terminal level. The functions of NGs are subject, object and complements. The absolute position inside the clause is an attribute of the group. A relative position attribute is used, in particular for the NG in front of the verbal group. During the analysis phase, this attribute was revealed to be redundant with the subject attribute.

A sub-group is a constituent of NG more than 4 syllables long. It can vary between 2 and 12 syllables. If the higher group is a syntactically simple one, the next level of constituents is the word. This type of sub-group is characterized by the categorial values of the word (grammatical word, adverb, adjective, noun, auxiliary, conjugated or composed verb) and by the relative position of the adjective in front of the noun. If the higher group is complex, the decomposition depends on the nature of the group structure: for an enumeration structure, the sub-groups each correspond to an enumerated constituent, while in case of a dependency structure the dependent constituent is treated like a group. Also, this last case, where the sub-group is "promoted" to group status, is the only case where the sub-group is not also a terminal level.

The morpho-syntactic values represented in the corpus are equivalent to the output of the automatic text analysis performed Stefanini et al.(1992) at CRISTAL (Grenoble).

Table 1. *Selected examples of the successive levels*

attributes in the sentence level		**examples**
interrogative	direct	Tu viens ?
	inverted	Viens-tu ?
attributes in the clause level		**examples**
linked	subord. verb->clause	Je verrai *si les enfants jouent*
attributes in the group level		**examples**
length	3 syllables	*Les enfants* jouaient sur la chaussée
pos. in the clause	initial	*Les enfants* jouaient sur la chaussée
function (NG)	object	Je vois *les enfants.*
	verb dependent	Les enfants jouaient *sur la chaussée*
attrib. in the sub-group level		**examples**
pos. in the group	before noun	Les *petits* enfants jouaient sur la chaussée.

METHODOLOGY OF ANALYSIS

The database, once recorded, was constituted according to the design described above, associating the phonetic labels, the syllable boundaries, the segmental and syllabic F0, the segmental and syllabic duration codes, and the linguistic attribute values.

The main hypothesis underlying this methodology is that each linguistic level (i.e., sentence, clause, group, and sub-group) corresponds to a *global intonation unit*, which can be performed by varying intonation forms (extracted from contours). At the lowest (terminal) linguistic level (sub-group, or group, as appropriate), the contours of the unit , which is also a terminal intonation unit, are defined directly for each syllable by the *syllabic F0* and the *prosodic syllable lengths* . At the higher linguistic levels, the contours of the unit, which is a non-terminal intonation unit, are defined by length attribute (number of syllables) in all cases, but by the *syllabic F0* exclusively for the first and last syllable of the unit considered (i.e., sentence, clause, or group). The intonation contours will thus be modeled by a simple declination line for the non-terminal units and by a (short) string of values for the terminal units.

These different intonation contours in the corpus were automatically segmented for each level, and were exhaustively and hierarchically grouped at all levels according to all the attribute values applicable to that contour at that level.

Further analysis consisted of verifying the homogeneity of intonation contours according to the different combinations of linguistic attributes, and this was done at every level. For the non-terminal levels, which are declination lines, statistical criteria were used, but in the absence of objective criteria for judging similarity of the terminal contours, where values are defined for each syllable, we used visual comparison.

Thus, at each linguistic level, classes of intonation contours are defined, each indexed by a set of attribute values. A representative contour, termed the mean contour (henceforth MC), is calculated for every class.

A lexicon, hierarchically following the linguistic levels, is then constituted with all the MCs, each indexed by the appropriate attribute values.

GENERATION OF INTONATION IN TtS

The text input to the synthesizer is processed sentence by sentence. The automatic text-to-phonetics and morpho-syntactic analysis yields the linguistic attributes. At the suprasegmental level, generating the intonation for an input sentence consists of the hierarchical calculation of a syllabic F0 pattern from the first, non-terminal level (the sentence) down to the terminal level (the group or the sub-group, as appropriate) and finally, of a prosodic syllable length pattern as well, though only at the terminal level. The declination line of the sentence is fixed first, and then this pattern becomes the input pattern of the clause level calculation module. The average-contours of the sentence's component clauses are juxtaposed to define a "local" declination line for the sentence, and this local pattern is warped to the input pattern. The resulting pattern becomes the input pattern for the group level calculation module. When groups are divided into subgroups, the group's MC is warped the same as those of the higher levels, and processing continues to the terminal subgroups. Since the average-contours of the subgroups and terminal groups are not simply declination lines, one is calculated in order to guide the warping process for the terminal group or subgroup. Finally, the prosodic syllable lengths are retrieved and associated to the F0 pattern. The output of the suprasegmental calculation of the intonation is a pair of functions specified in discrete syllabic steps, one for F0 and one for prosodic syllable lengths.

The last step is the "segmental" calculation of the intonation pattern. We have seen that F0 is coded with three values for each vowel. These values control the synthesis at the beginning, the middle and the end of the vowel. The acoustic curve of F0 is calculated by a Spline function interpolation between these points, and the curves for consonants or clusters are interpolated between vowels with the same Spline function.

The durations of the syllables are modified by the intrinsic and co-intrinsic values already used during the analysis of the corpus.

CONCLUSION

The hypothesized *rendez-vous* between prosodic strategy and the corpus structure design might have been invalid, either generally or for our speaker. In fact, the results show that the *rendez-vous* was quite clearly made and consequently the resulting synthesized intonation is high quality. That can be explained partly because of the specific situation chosen, a set of sentences read in isolation without any coherence outside the sentence level.

One unsatisfying point in this study is the duration processing. Presently, it is debatable whether it is the syllable or the group inter P-Center (Barbosa & Bailly, 1993) that is the appropriate unit to be chosen.

An application of this method of corpus constitution and analysis should be considered for other applications. Concerning a given language, an extension of this work could include building a database of MCs representative of different prosodic strategies (man-machine dialogue, multi-media bureautics, or remote control...).

The explicitation of the phonological level underlying the lexicon is one way to generalize such model (Hirst & Di Cristo, 1992). Another way is the systematic unification of the classes of MC, or the learning thrue a stochastic model such a neural network (Traber, 1992).

REFERENCES

V. Aubergé (1992), "Developing a Structured Lexicon for Synthesis of Prosody", In *Talking machines: theories, models and designs* , ed. by G. Bailly & C. Benoît (North Holland Pubisher), 307-322.

P. Barbosa, G. Bailly (1993), "Generation and evolution of rhythmic patterns of text-to-speech synthesis", *this volume,*

D.L. Bolinger (1989), *Intonation and its uses: Melody in gramar and discourse* Stanford, CA: Stanford University Press.

G. Caelen-Haumont (1993), " Dialogue homme-machine et intelligibilité : analyse des caractéristiques linguistiques et prosodiques des discours de lecture,*Séminaire Prosodie, GDR-PRC,* 113-128.

W. N. Campbell (1992), "Syllable-based segmental duration.", In *Talking machines: theories, models and designs* , ed. by G. Bailly & C. Benoît (North Holland Publisher), 211-224 .

F. Grosjean & J.Y. Dommergues (1983). "Les structures de performance en psycholinguistique," *L'Année Psychologique,* 83, 513-536.

D. Hirst & A. Di Cristo (1992), "Niveau de représentation et étiquetages prosodiques", *Séminaire Prosodie, GDR-PRC Communication Homme-Machine,* Aix-en-Provence.

P. Monnin & F. Grosjean (in Press), " Les structures de performances en français : caractérisation et prédiction," *L'Année Psychologique.*

M.H. Stéfanini, A. Berrendonner, G. Lallich & F. Oquendo (1992), "TALISMAN: A multi-agent system governed by linguistic laws for natural language processing", *Proceedings of COLING,* 490-497.

C. Traber (1992), "Fo generation with a database of natural Fo patterns and with a neural network." In *Talking machines* , ed. by G. Bailly & C. Benoît (North Holland Publisher), 287-304.

Generation and evaluation of rhythmic patterns for text-to-speech synthesis

P. Barbosa & G. Bailly
Institut de la Communication Parlée, U.R.A. CNRS n° 368
ENSERG/INPG -Université Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex 1, France

ABSTRACT

This paper presents a characterization of durational contours based on phasing relations between noticeable acoustic events and an internal clock. We generate segmental durations in two stages : the duration of a rhythmic programming unit is computed according to a reference clock and then is distributed among its segmental constituents. A perception experiment evaluates the necessity of the rhythmic patterns found in analysis. A method for mastering speaker's speech rate is described and is analysed to propose some guidelines for the integration of the pause phenomenon into automatic generation.

INTRODUCTION

"(...)intonation manages *to do what it does* by continuing *to be what it is* (...)" (Bolinger, 89). That is our point of view about prosody: it performs a linguistic task under biological constraints. Perception is guided by motor schemes: listeners use information from kinematic patterns (Viviani & Stucchi, 92).

We assume the existence of an underlying internal clock, a timekeeping function, used for synchronization of impulses transmitted to the muscles (Turvey *et al.*, 90). We show that the regularity of this clock is maintained through pauses (cf. section 3).

We generate segmental duration by a two-stage model (Campbell, 91) but our approach is different from Campbell's because duration is obtained by a control signal emitted at each Perceptual-center (PC).

1. THE INTER-PERCEPTUAL-CENTER GROUP

A two-rate, 88-sentence corpus was explored in order to study the rhythmic patterns of read sentences. It was designed for answering at: (1) are continuously increasing patterns (cf. fig. 1) needed to the perception of accentuation or are they just an artefact of production constraints ? (2) Are this typical configuration needed to the perception of any kind of isochrony in French connected speech ?

Pompino-Marschall's experiments have tried to estimate an absolute localization for PC using *syllable/beat* and *beat/syllable* sequences. Despite the diversity of the consonants the PC seems to be at the neighbourhood of the vocalic onset. The perception of momentary tempo is better characterized by inter-PC intervals.

Thus the PC location in our work is fixed at the vocalic onset. The importance of this event is largely developed in the literature (Dogil & Braun, 88; Stevens & Blumstein, 78; Fant & Kruckenberg, 89). The term PC will be maintained because of the allusion on perception and our hypothesis of an internal clock guiding the production of the programming rhythmic units. The lenghtening of IPCGs is characterized by a single factor *k*, by computing $\sum \exp(\mu_i + k.\sigma_i)$ = IPCG duration, where μ_i and σ_i are the mean and standard deviation of the log-transformed durations (in milliseconds) of the realizations of the phoneme *i* from a corpus of logatoms at comfortable rate. Rhythmic patterns of the corpora are characterized by *k* averaged for each IPCG and segmental durations are in turn computed for synthesis by the exponential expression above.

The analysis of the corpus evidences a rhythmic pattern by concatenation of elementary movements (cf fig. 1). These movements are monotonously increasing, mark clearly prosodic boundaries, start with a reset of *k* at 0 and exhibit a more or less exponential increase.

2. THE PERCEPTION EXPERIMENT

Method

Ten pairs of sentences were used for this experiment. They were listened in binoral presentation by eleven subjects working in the laboratory but not in synthesis domain. The duration of the test was between 10 and 15 minutes. Each pair contains a reference durational pattern (A) and a pattern to be tested (B). The two sequences (AB and BA) were listened in random order within a session. Listeners were asked to answer what sentence was the more natural by pressing on the keyboard "1" to the first one, "2" to the second one or "?" if a doubt persists. Listening may be repeated twice. An example of a sentence pair is showed below.

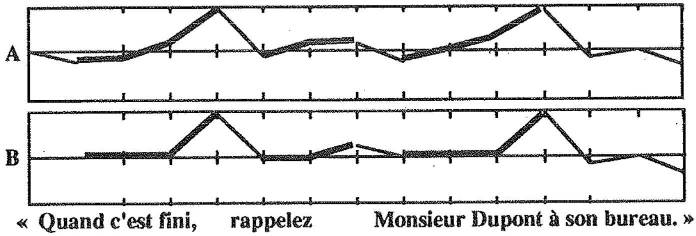


Figure 1. *Sentence 56 : Example of A and B configurations. Discrete values of k are connected by lines for sake of visibility. Prosodic groups are represented by thicker lines. The last IPCG of the utterance is not presented here.*

The A configuration was obtained by calculating the successive k-factors of all IPCGs in a sentence. In the B configuration k-factors preceding the accent were set to 0 and the ones associated with the accent were not modified. Segmental durations in the two configurations were computed by applying the formula mentioned above. Segmental durations in A patterns are different from the ones in the natural sentences (k is averaged in each IPCG) but since the IPCG durations are the same, the VO timing is identical in both ones. Silent pause durations and all other parameters were unchanged. A high quality speech analysis/resynthesis system was used to obtain the above parameters (Moulines, 92). We are testing the perceptual prominence of a gradual versus an abrupt pattern of accent realization.

Results

Considering all subjects 77% of A answers are obtained. Taking also the question-mark answers the result is 65% of A, 20% of B and 15% of question-mark answers.

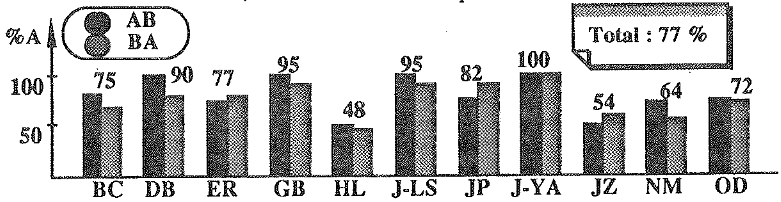


Figure 2. *Mean scores of A answers by subject and mean over all subjects*

There were no significant effect of presentation order. All listeners agree on the difficulty of the task : "During the experiment I thought I have changed the criterion of utterance choice. At the beginning, I have chosen the ones were more *constant* concerning the rhythm" (JLS), "I have chosen utterances that were the most *constant* at rhythm level." (OD), "Are you sure they are not identical ?" (JYA) !!

Results show subjects' preference for the gradual accent realization, despite the finesse of the task. The term *constant* was used to oppose the two stimuli.

Discussion

The clear general preference for the gradual pattern lead us to think that this configuration is necessary to the accent perception. Too small differences between A and B configurations explain the poor score of one of the utterances.

Lack of lengthening of previous IPCGs sounds abrupt. The internal clock hypothesis may explain the subjects' perceptual behaviour : shorter IPCIs are cues of an unexpected local acceleration. Gradual lenghtening contributes to the perception of isosyllabicity (Duez, 87 ; Lehiste, 77). But there is no implicit conclusion that human beings use k coefficients to produce the accent pattern.

3. TOWARD A MODEL INCLUDING PAUSE EMERGENCE

About the clock beats

Grosjean's performance structures are built using pause duration as a cue of the strength of corresponding prosodic juncture (markers here). Grosjean's approach do not take into account an underlying rhythmic activity which constrains pause durations to be realized by an integer number of clock units (Fant & Kruckenberg, 89): this is illustrated by the relatively low 86% correlation between cues obtained at two different speaking rates in their experiment (Monnin & Grosjean, in press).

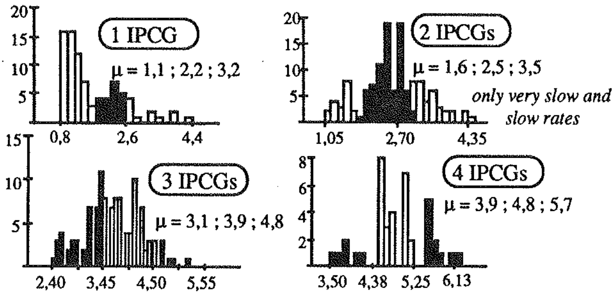


Figure 3. *Histograms and clusters by number of IPCGs in the PG*

A five-rate, 20-sentence corpus was recorded, pronounced by our speaker in order to evidence the influence of pause insertion on the rhythmic structure. This experiment aims at developing an automatic generation of duration including the pause phenomenon. The speaker was asked to answer to interrogative synthetic utterances with predefined sentences. These utterances were obtained using our text-to-speech system by multiplying μ_i and.σ_i by a phonation factor (Wightman *et al.*, 92) .

The analysis of this corpus confirms for all rates the general trend of rhythmic patterns: 80% of them are monotonously increasing. To study the pause emergence a characterization of total duration of PGs have been developed: the ratio between the total duration of each PG and the internal clock duration is computed.

The choice of an internal clock duration is particularly delicate, but necessary. How to choose an unaccented unit if lengthening is gradual over the PG ? Taking into account that: (1) the last IPCG in the prosodic group is clearly the main lengthened unit; (2) if there is a pause, we cannot separate silent and sound intervals : they are elements of the same phenomenon (Duez, 87); (3) the first IPCG of a PG is often shortened, the internal clock durations are computed for each utterance as the mean among the non-accentuated IPCGs. We assume that there is a reseting of the internal clock after the accent realization.

These ratios were submited to a cluster analysis for each PG length (cf. fig.3). Clusters are similar between the rates. The analysis was made over the five rates (in order to have statistically significant data).

Results

Mean values of the clusters are closely associated with integer number of clocks. The standard deviations represent between 10 to 20% of the respective means.

We then observed each cluster for each individual rate : (1) the additional clock units are associated with the presence of silent pauses in slow rates and only lengthening in the fastest rate (except for the strongest markers); (2) when there is a silent pause, lengthening of the preceding group is represented by an integer number of clock units; (3) the strongest markers are associated with the greatest values of the ratios.

4. COMMENTS AND PERSPECTIVES

Some deviations in the cluster sets may be due to wrong segmentations (several unvoiced plosives after silent pauses) and the choice of the clock unit.

Automatic generation of duration will depend on the strength of the prosodic marker associated with the juncture : strongest markers will receive more clock units. The main differences between the speech rates are : (1) the frequency of the internal clock clearly differenciates the rates (except very slow and slow rates) ; (2) in slow rates the subject seems to prefer lengthening plus silent pause to realize the accent whereas in fast rates, only lengthening ; (3) markers can be removed in fast rates to form a PG that will contain more IPCGs (marker deletion).

In this perspective, rhythmic patterns are monotonous decelerations which modulate in frequency a carrier clock.

REFERENCES

Barbosa, P. & Bailly, G. (1992) "Generating segmental duration by P-centers", *4th Workshop on Rhythm Perception and Production,* Bourges, France, June, 163-168.

Bolinger, D.(1989) *Intonation and its uses,* (Edward Arnold).

Campbell, W. N. & Isard, S. D. (1991) "Segment durations in a syllable frame", *Journal of Phonetics*, **19**, 37-47.

Dogil, G. & Braun, G. (1988) *The PIVOT model of speech parsing* , (Verlag, Wien).

Duez, D. (1987) "Contribution à l'étude de la structuration temporelle de la parole en français", *Thèse d'état.*

Fant, G. & Kruckenberg, A. (1989) "Preliminaries to the study of Swedish prose reading and reading style", *STL-QPSR*, **2**, 1-80.

Lehiste, I. (1977) "Isochrony reconsidered", *Journal of Phonetics*, **5**, 253-263.

Monnin, P. & Grosjean, F. (in press) "Les structures de performance en français : caractérisation et prédiction", *Année Psychologique.*

Moulines, E. (1992) "Synthesis models : a discussion". In : *Talking machines : theories, models and designs* (Bailly, G. & Benoît, C., Eds), 7-12.

Pompino-Marschall, B. (1992) "The P-center and the perception of rhythm in connected speech", *4th Workshop on Rhythm Perception and Production*, Bourges, France, June, 157-162.

Stevens, K. & Blumstein, S. (1978) "Invariant cues for place of articulation in stop consonants", *J. Acoust. Soc. Am.,* **64(5)**, 1358-1368.

Turvey, M.T., Schmidt, R.C. & Rosenblum, L. (1990) "Clock and motor components in absolute coordination of rhythmic movements", *Haskins Laboratories Status Report on Speech Research*, 231-242.

Viviani, P. & Stucchi, N. (1992) "Biological movements look uniform : evidence of motor-perceptual interactions", *Journal of Experimental Psychology : Human Perception and Performance*, **18(3)**, 603-623.

Wightman, C. W., Shattuck-Hufnagel, S. Ostendorf, M. & Price, P. J. (1992) "Segmental durations in the vicinity of prosodic boundaries", *J. Acoust. Soc. Am.,* **91(3)**, 1707-1717.

High Quality Intonation Algorithm for the Greek TTS - System

George Epitropakis, Nikos Yiourgalis, George Kokkinakis
Wire Communications Laboratory,
University of Patras, Greece

ABSTRACT

This article presents a complete algorithm for the generation of intonation (F_0 contours) for the Greek Text-To-Speech system, based on a multi-layer label structure that is constructed for the phonemes representing the input text. This structure consists of the phoneme's distinctive features, the position of the syllable that the phoneme belongs to, the prosodic label of the word that the phoneme belongs to, and the phoneme's prosodic context in the sentence. According to the contents of that structure, the algorithm assigns to each phoneme of the input sentence a target pitch level to be reached either at the beginning, or the middle, or the end of the phoneme. When all the phonemes have been assigned the appropriate F_0 level, the overall pitch contour is constructed by linear interpolation between the successive F_0 levels. Although the method proposed seems to be a rather abstract approach, it takes into consideration linguistic, phonotactics and metrical constraints of the input and not linguistic constraints alone. In addition, the method is especially suited for languages, such as Greek, which are inflectionally rich and have great freedom of word-order.

PURPOSE AND SCOPE

To improve the naturalness of the synthetic speech, numerous methods have been proposed, all of which differ in both the phonetic and the phonological representation of intonation (Hirst, 1992). However, from the most radically concrete positions of Pierrehumbert (1980) to the abstract representations of 't Hart, Collier, and Cohen (1991), and to the even more abstract approaches of Carlson and Granstrom (1973), it is evident that the choice of a particular prosodic model strictly depends on its purpose. The aim of our model is to produce high quality intonation contours for the Greek TTS-system, developed in Wire Communications Laboratory for the practical application of one-way information broadcasting i.e. reading newspapers and books.

To this end, the model is based on the results obtained from an extensive study of Greek intonation that was carried out on text material recorded from just one speaker, who was strictly instructed to speak in a "neutral" reading style. The pitch contours extracted from natural speech were "stylized" by following the guidelines of the "perceptual approach", developed at the Institute for Perception Research of Eindhoven (IPO), but no standardized pitch movements and pitch configurations were extracted. On the contrary, our methodology is based on the description of all the turning points produced in the "stylized" pitch contours in terms of discrete textual phenomena.

To extract and formulate the rules that establish the correspondence between the linguistic constraints and prosodic properties of any input sentence, we first drew up a list of 200 test sentences that covered the greatest possible syntactic structures of the language. The material to be recorded was created by combining simple types of Noun Phrases and Verb Phrases, and by increasing successively their complexity, always producing meaningful sentences. In addition, sentences with identical syntactic structures, but with different numbers of syllables were created. These sentences were recorded by 1 male speaker who was asked to speak in a neutral manner.

Through the steps of "perceptual equivalence" methodology ('t Hart, Collier and Cohen, 1991), a resynthesized version of each recorded sentence has been created where the original F_0 contour is approximated by the smallest possible number of straight-line segments, so that only the fluctuations that do essentially contribute to the perception of prosody are to be accounted for. The final step, was to determine all the turning points appearing in the resynthesized F_0 versions, and to describe them in terms of discrete textual events.

ACCENTS, PHRASING, AND PHONOLOGICAL REPRESENTATION

Accentuation and phrasing are two of the weak points in TTS-systems (Traber 1992), partly because syntactic information is not enough to derive a reasonable stress pattern and phrasing, and partly because the accentuation and phrasing rules are not elaborate enough. As far as the accentuation of the Greek language is concerned, there is no problem, because word accents are included in the orthographic representation of the input. However, the phrasing is a more complicated problem for the following reasons:
a. Greek is an inflectionally rich language and the problems start from the correct and unambiguous text labelling achieved in real-time.
b. It is a language where different syntactic structures can be used to express the same meaning. In addition, the position of constituents in a sentence is almost free and the constituents can be moved to reflect pragmatic factors.

As a consequence, it was our deliberate choice to lay greater emphasis on phonotactics and metrical data in the development of the model rather than upon syntactic phenomena. However, a syntactic-prosodic parsing of the input is carried out (Epitropakis, 1993), in order to determine the phrasal units of the sentence that are necessary to determine some of the pitch movements that are not signalled by phonotactics and metrical events.

We finally came out to a set of linguistic information (table 1) according to which each phoneme is labelled with a target pitch level, so that all the pitch configurations presented in the results obtained from the resynthesized F_0 contours data-base analysis to be synthesized. Generally, each phoneme of the input is labelled according to:

- certain distinctive features (vowel/consonant, stressed/unstressed, etc.),
- the lexical position of the syllable that the phoneme belongs to (antepenultima, penultima, ultima), and
- the phoneme's relative position to prosodically significant events, where the boundaries between sentences, sentence and relative phrase, verb phrase (VP) and noun phrase (NP), VP and adverbial phrase (AP), VP and prepositional phrase (PP) are taken into account.

1	First phoneme of the phrase	16	Boundary (VP-AP) vowel
2	Last phoneme before the first accent of the phrase	17	Voiced consonant before the end of the sentence
3	Accented phoneme	18	First vowel after the boundary (VP-AP)
4	Phoneme between two accents	19	Boundary (VP-PP) vowel
5	First pre-boundary (NP-VP) phoneme	20	Pre-boundary (VP-PP) vowel
6	Second pre-boundary (NP-VP) phoneme	21	First vowel after the boundary VP-PP
7	Penultima	22	Voiced consonant before the boundary VP-PP
8	Boundary phoneme (NP-VP)	23	Final phoneme of the phrase
9	Ultima vowel after an accented penultima	24	Pre-boundary (VP-AP) vowel
10	First phoneme after the boundary NP-VP	25	Boundary (,) vowel
11	Voiced consonant before an accent	26	First pre-boundary (,) vowel
12	First phoneme just after the accent	27	Second pre-boundary (,) vowel
13	Pre-final phoneme of the phrase	28	First vowel just after the boundary (,)
14	Final vowel of the phrase	29	Voiced consonant before the boundary (,)
15	Voiced consonant before the NP-VP boundary	30	Voiced consonant before the boundary VP-AP

Table 1. *Linguistic properties that are used for labelling the input phonemes.*

A set of rules that could constitute the "grammar of the intonation" of the Greek language for neutral speaking style has been extracted. These rules are of type:

$$a,\ b,\ c,\ ..\ \text{-->}\ F_0\ level,$$

where *a,b,c,..* is the above presented linguistic information labelling the phonemes of the input. F_0 *level,* is a label used to assign at each phoneme one of the following tone levels: *Base, mid and top.* It is not an absolute F_0 value (in Hz), but a dynamic estimation of the corresponding phoneme's pitch value in one of the three declined lines used for the Greek language (baseline, midline and topline). Both the starting point (Hz) and the declination slope in semitones/sec of these lines have been experimentally extracted by statistical analysis of the original pitch contours of the speech data-base. Equations 1 and 2 give the starting point B_{start} and the slope B_{slope} of the baseline respectively according to the sentence's duration *t*. Additional information for the determination and the implementation of the necessary reset points has also been extracted.

$$B_{start} = 120 * e^{(-0.061/t)} \quad (1) \qquad B_{slope} = 3.05 \ / \ (t{+}0.505) \quad (2)$$

INTONATION ALGORITHM

The intonation algorithm assigns to each phoneme of the input sentence a target pitch *level* to be reached at the middle or the end or the start of the phoneme, according to the rules of the intonational grammar.

For example, given the input *"O aEt'Os tu f'ilu mu tu t'aKi pu 'idamE xTEs, pEt'ai psil'a"* (=the eagle of my friend Takis, which we saw yesterday, flies high) after the linguistic analysis (Epitropakis, 1993), which obtains data such as those shown in table 1, the algorithm generates the overall pitch contour in the following steps:

1. Firstly the declined lines and the appropriate reset points are determined.
2. Input in the algorithm is the phonetic representation of the input sentence plus the corresponding labels describing the phonemes properties. The appropriate F_0-levels are assigned to the corresponding phonemes according to the rules of the grammar (table 2 gives the rules used in this example). A total of 67 such rules constitute the grammar for Greek intonation. The phonological representation for the example given, is as follows: *O[1] aE[2,7]t'O[3]s tu[4,12] f'i[3,7]lu[9,12] mu tu[26] t'a[3,7,27]Ki[8,9,12,25] pu[28] 'ida[5,7,12,26]mE[6,27] xTE[8,25]s, pE[10,28]t'a[3,7]i[9,12] psil[11,17]'a[3,14,23,30]*

Linguistic attributes	F_0-level and timing
1	Base at the start of the phoneme
2,7	Base at the end of the phoneme
3	Top at the end of the phoneme
3,7	Mid at the start of the phoneme
3,7,27	Base at the end of the phoneme
3,14,23,30	Top at the middle and base at the end of the phoneme
4,12	Base at the end of the phoneme
5,7,12,26	Top at the start of the phoneme
6,27	Base at the end of the phoneme
8,25	Top at the end of the phoneme
8,9,12,25	Top at the end of the phoneme
9,12	Top at the end of the phoneme
10,28	Base at the end of the phoneme
11,17	Mid at the middle of the phoneme
26	Base at the end of the phoneme
28	Base at the end of the phoneme

Table 2. *The rules of the intonational grammar used for the given example.*

3. When all the phonemes have been assigned a F_0-level, the overall pitch contour is constructed by linear interpolation between the successive levels. Figure 1 gives the overall pitch contour (solid line) constructed for the example sentence according to the F_0-levels (solid dots) that have been determined. The original pitch contour is also given (dotted line).

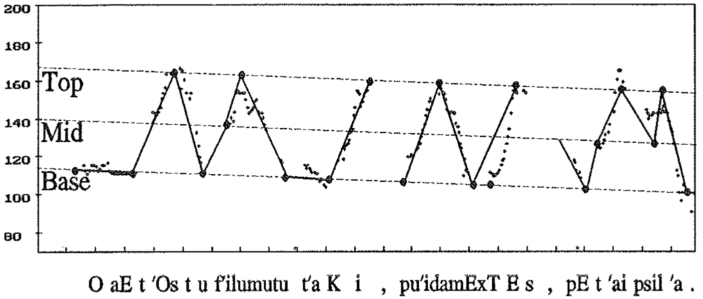


Figure 1. *Synthesized pitch contour for the example mentioned above.*

EVALUATION RESULTS - CONCLUSIONS

A complete methodology for constructing intonation contours in the context of TTS-systems has been presented. It has been shown that a successful intonation model can be created for a language such as Greek, which has peculiar difficulties, by combining syntactic, phonotactics and metrical data. The particular combination of data exploited in this model permits to achieve great accuracy and to be computationally tractable in real-time. To evaluate this methodology, two different tests have been carried out:

The first test estimates the performance of the algorithm in correctly determining the prosodically important turning points in the pitch contours. For this purpose, a set of sentences have been recorded and the appropriate turning points have been manually determined. Then the proposed algorithm has been used for the same set. Analysis showed that 71% of these points were correctly determined. A percentage of 18% were not been determined, but the resulting overall pitch contour was almost equivalent to the expected contour. Finally, 11% of the turning points have been determined incorrectly or have been totally missed. This omission led to unacceptable pitch contours. Generally, the method proved to be robust, but for the extraction of the grammar rules only the rules with high statistical correlation (above 80%) in the speech data-base, are included. The inclusion of heuristic rules with lower occurrence must be taken into account in the future.

The second test concerns the evaluation of the speech output quality. For this reason, preliminary tests with 4 listeners were carried out. Two different text subjects were synthesized by the Greek TTS-system consisting of 39 sentences (650 words). The results show that the naturalness of the resulting speech is very high. Refinement of the rules and Spline function interpolation are in progress in order to further improve the speech quality. In addition, a novel phrasing model (Michos et al.) is under consideration.

REFERENCES

R. Carlson and B. Granstrom (1973), "Word accent, emphatic stress and syntax in a synthesis by rule scheme for Swedish", *QSPR-STL*, Vol. 2-3, pp.31-36, KTH.

G. Epitropakis, N. Yiourgalis, and G. Kokkinakis, "Prosody control of TTS-Systems based on linguistic analysis", *Eurospeech '93*, (to be presented).

D. Hirst (1992), "Prediction of prosody: An overview", *Talking Machines: Theories, Models, and Designs*, Baily, Benoit, and Sawallis (eds), pp.199-204, Elsevier Science Publishers B.V.

J. t'Hart, R. Collier, and A. Cohen, "A perceptual study of intonation", Cambridge University Press.

S. Michos, G. Epitropakis, N. Fakotakis, and G. Kokkinakis, "A novel phrasing method for high quality prosody in TTS-systems", to be submitted to ICASSP '94.

J. Pierrehumbert (1980), "The Phonology and Phonetics of English Intonation", *MIT Ph.D. dissertation.*

C. Traber (1992), "F_0 generation with a database of natural F_0 patterns and with a neural network", *Talking Machines: Theories, Models, and Designs*, Baily, Benoit, and Sawallis (eds), pp.199-204.

Prosodic Modelling for a Text-to-Speech System in Spanish

Eduardo López-Gonzalo and Luis A. Hernández-Gómez.
Dpt. SSR, E.T.S.I.Telecomunicación. Univ. Politécnica de Madrid.
Ciudad Universitaria s/n, 28040 MADRID.
e-mail: eduardo@gaps.ssr.upm.es luis@gaps.ssr.upm.es

ABSTRACT

In this contribution, we present the results in modeling Spanish prosody for a text-to-speech system based on units concatenation using a TD-PSOLA synthesizer. For this purpose, we have developed a new methodology to transfer to a text-to-speech system the prosody of one speaker considering both fundamental frequency and duration jointly, trying to take into account their interactions. The results on a corpus test over the system shows very good intelligibility and naturalness.

INTRODUCTION

Our aim was to produce a prosodic model of the speaker who recorded the acoustic database for a TD-PSOLA (F. Emerard et. al. 1992) synthesizer trying to capture all his characteristic features. For this task we designed a data-driven methodology that is shown in Figure 1.

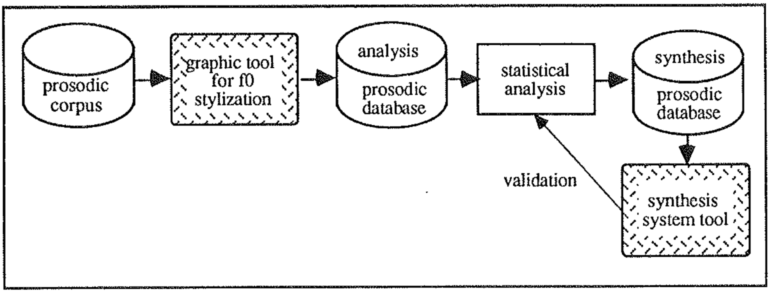


Figure 1. Methodology for prosodic modeling.

This methodology is guided by the main objective of producing a model of natural prosody allowing its artificial simulation. This is a more delimited task that prosodic modeling of the Spanish language but we think it is sufficient in the context of high quality text-to-speech synthesis. The rest of the paper is organized to describe this methodology.

CORPUS DESIGN

First of all, we designed a prosodic structure for Spanish that is an adaptation to the one proposed by V. Aubergé (1990). Based on a first analysis of our data and on the general linguistic knowledge of Spanish we considered two prosodic units: the "prosodic word" that is defined as the sequence of syllables associated with one accent and the "prosodic proposition" that is defined by the sequence of syllables between two pauses (regardless of its duration).

In order to constitute an organized corpus of minimal pairs of attributes, we defined a set of attributes for the two prosodic units. The "prosodic proposition" was characterized by its number of syllables, its number of prosodic words and its prosodic type depending on the syntactic relation between consecutive propositions (9 defined types, two of them complementary to put a contrast prosody in consecutive propositions of the same type). The "prosodic words" were defined by three attributes: the number of syllables, the position of the accent (3 classes) and the ordinal position inside a proposition.

This special corpus should take into account all the different types of propositions formed by all different prosodic words (V. Aubergé 1992). In order to limit the number of sentences to analyze, we impose some restrictions to emphasize the more important characteristics to be modeled. We mainly take into account the position of first and last accent of propositions, and study relations in sentences with only two propositions. Relations between sentences in a paragraph were not considered. The other parameters were kept as broad as possible. The resulting corpus was 144 sentences long, ranging from very short sentences to very large sentences taken from newspapers.

CORPUS STYLIZATION

This corpus was recorded at the CNET by a selected Spanish native speaker to produce a linguistic procesor for ELAN INFORMATIQUE. Then we analyzed it in order to extract for each sentence the pitch contour with marks representing the location of the vowels. For this task, we programmed a variation of a super resolution pitch detector (Y. Medan et.al. 1991).

This representation is known to contain many irrelevant information that is filtered by our auditory system, so we "stylized" these contours (R. Collier 1990) in such a way we keep only pitch movements that are perceived by a common listener. These movements are described following their direction (rise or fall), their slope in semitones per ms, extension in time (one or more syllables) and timing respect the tonic vowel. A frequency-based acoustic module elaborated for the synthesizer (E. Rodríguez, E. López and C. García 1993) was used in an analysis-synthesis mode with a graphical tool developed for the stylization of contours. The tool shows the temporal evolution of the original pitch in a semilogaritmic scale, and let you approximate it by straight lines (Figure 2). Then the acoustic algorithm is used to produce the original speech with the new pitch contour. Whenever a difference is perceived in prosody, we make a better approximation of the pitch contour. The output file of this module codes the fundamental frequency contour and duration aligned with the phonetic transcription of the sentence. The representation of the pitch contour in a vowel is given with a resolution of three points per vowel by means of 5 parameters: the initial, internal and final fundamental frequency value in the vowel and two time durations.

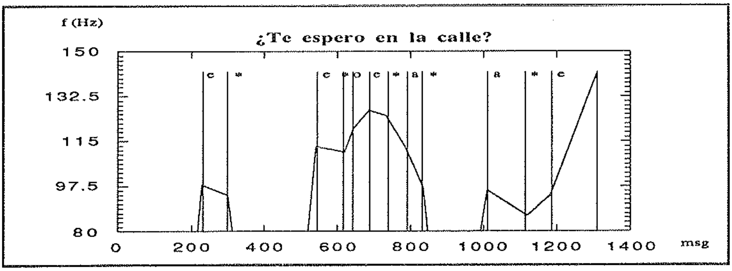


Figure 2. Presentation format of the graphical tool for a stylized contour.

CORPUS ANALYSIS

The result of the stylization process is put in a prosodic database specially designed to measure statistically the real consistency of the suprasegmental data produced by the speaker. A total of 23 fields were needed, the resultant structure is similar to the one proposed by Emerard et. al. (1992) but adapted to our prosodic structure. The database has 3500 vowels, 1700 consonants (including semivowels) and 166 pauses. The duration is stored for the consonants and the pauses. The stylized movement of pitch and the normalized duration is stored for each vowel. Besides that we store all attributes for accessing the prosodic structure (defined above) and phonetic context. The normalization is calculated dividing the duration of each vowel by the mean value of this type of vowel (5 classes in Spanish). Following this approach the data stored have been produced by the speaker. We make a statistical analysis between consistent prosodic words in a certain kind of proposition in order to minimize the interaction between segmental and suprasegmental parameters. The results of this analysis showed us a very good consistency to model absolute frequency contours and normalized suprasegmental duration in each class. As an example in Figure 3, it is shown the last four syllables (a vowel per syllable) of the last prosodic word in a proposition at the end of a declarative sentence for the three positions of the accent considered (accent in the last vowel, accent in the last but one vowel or accent in the last but two vowel).

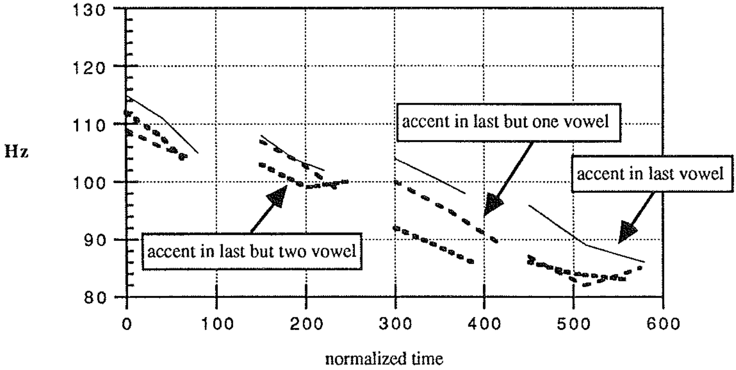


Figure 3. Pitch and duration modeling for the last prosodic word in a declarative sentence, considering the three positions of the accent.

It should be noted that in our model, pitch and duration are modeled in a joint way considering the effect of the position of the accent and position of the pause because of the defined prosodic structure. This model takes into account the possible interaction between duration and pitch in prosody realization. It is observed in Figure 3, that there is a residual effect of suprasegmental duration as predicted by other duration models with the position of the accent. It was also observed from the resultant model that there was a lengthening in the vowel located in a syllable previous to a pause.

For this first model, the duration of each consonant (and semivowel) was considered to be the mean value calculated from the database.

The mayor correlation in the duration of the pauses was found to be with the type of orthographic sign and the number of syllables of the proposition but there was a large dispersion from the mean value. A closer look to the data ordered by the duration value of the pause show us that the duration of the pauses had more to do with the syntactic structure of the sentence.

SYNTHESIS TOOL

In order to validate and refine the model, it seemed sensible to us to use a text-to-speech system as a tool for leading our analysis in the database.

For this purpose, we developed a very flexible text-to-speech system (E. López 1993). The input text is parsed to mark the "prosodic words" and the major “prosodic propositions” based in some independence coefficients between two consecutive words that take into account both rhythmical and syntactical constrains. The same procedure has been integrated in TTS systems for Spanish (E. Rodríguez et.al., E. López et. al. 1993).

The prosody calculation is realized by a look-up in the synthesis database containing the results of the statistical analysis explained before. We choose this data based approach for being more flexible that a rule based system so we could test quickly our hypothesis. This original contour is then post processed with an interpolation of pitch contours between consecutive vowels, and a micromelody for voiced consonants and semivowels is added. The length of the pauses are calculated in function of the independence coefficient.

The model was tested in an implementation of a TTS system based in diphone concatenation with a TD-PSOLA algorithm. We designed a small corpus of isolated sentences considering two types of propositions with all different prosodic words. The naturalness was good recognizing perfectly the peculiar "style" of the speaker, although it was noted some deficiencies in segmental duration of consonant clusters.

For this reason we calculated from the analysis database a better segmental model of consonant duration based in some multiplicative coefficients as a function of phonetic context. This new model was considered in the synthesis tool improving the naturalness of the model.

As a future research, we are going to follow the same methodology to adapt the independence coefficients and consequently produce a better modeling of pauses in large sentences or even paragraphs. We are presently studying a syntactic theory developed by J. Vergne (1992) that can be adapted to our TTS system in order to compute better coefficients.

CONCLUSIONS

We have discussed a data-driven methodology that has been shown very successful to transplant the prosody of a speaker into a text-to-speech system. The complementation between statistical analysis and a synthesis tool was found to be very useful to improve the model.

REFERENCES

F. Emerard et. al. (1992) "Prosodic processing in a text-to-speech synthesis system using a database and learning procedures" in Talking Machines: Theories, Models and Applications" Elsevier 1992

V. Aubergé (1990) "Semi-automatic constitution of a prosodic contour lexicons for the text-to-speech synthesis" Proceedings in ESCA Workshop on Speech Synthesis. Autrans 1990.

E. López Gonzalo (1993). "Técnicas de procesado linguístico-prosódico y acústico para conversión texto-voz mediante concatenación de unidades" Doctoral Thesis. Universidad Politécnica de Madrid. (In preparation).

E. López-Gonzalo, G. Olaszy and G. Nemeth. "Improvements of The Spanish Version of the Multivox Text-To-Speech System". Accepted for publication in Eurospeech 1993. Berlin.

Y. Medan, E. Yair and D. Chazan (1991). "Super resolution pitch determination of speech signals" IEEE Transactions on Signal Processing, vol. 39, nº1. January 1991.

R. Collier "Multi-Lingual Intonation Synthesis: principles and applications" Proceedings in ESCA Workshop on Speech Synthesis. Autrans 1990

G. Bailly "Integration of rhythmic and syntactic constraints in a model of generation of French prosody" Speech Communication 8 1989 pp. 137-146.

E. Rodríguez-Barga, E. López-Gonzalo and C. García-Mateo (1993)."A Text-to-Speech System for Spanish with a Frequency Domain Based Prosodic Modification" International Conference on Acoustics Speech and Signal Processing" (ICASSP) Minneapolis (USA). 1993.

J. Vergne. "Syntax as clipping blocks: structures, algorithms and rules". Jornadas de la Sociedad Española para el Procesado del Lenguaje Natural (SEPLN), Granada 1992.

Prosodic Control in a Text-to-Speech System for Italian

Silvia Quazza, Pier Luigi Salza , Stefano Sandri, Alberto Spini
CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G.Reiss Romoli 274, 10148 Torino, Italy

ABSTRACT

The paper illustrates a TTS-oriented model of prosody generation for Italian, which tries to account for the relations linking the prosodic parameters to relevant linguistic features. A proper rule development system allows to express contextual rules and to maintain the alignment between different linguistic representation levels.

INTRODUCTION

The prosody generation model of the TTS system for Italian developed at CSELT (for more details see Balestri et al., 1993) tries to account for the relations linking phoneme duration and fundamental frequency to phonetic structure, stress positions, syntax, and application-dependent pragmatic requirements of the message. High-level linguistic information is provided to the prosodic processor by the TTS module devoted to text processing, or by the application software which may select the appropriate prosodic style or insert syntactico-prosodic markers in the text of the message. By analyzing the surface structure of the text and by making use of a lexicon, the text processor is able to perform the following steps: text segmentation according to punctuation marks; translation of numbers and abbreviations in letters; recognition of the grammatical category for verbal forms and functional words; lexical stress assignment; syntactic boundary detection. Acronym pronunciation rules and phonetic transcription rules are then performed. The TTS prosodic processor marks any syllable carrying lexical stress, functional words excluded, and groups words into prosodic phrases relying on punctuation marks and NP/VP boundaries. So far, only one stress level is considered, whereas two different strength categories are assigned to phrase boundaries, depending on the presence/absence of pauses. Weak phrase boundary is recognized between short NP and VP, before numerical expressions and between the two words of a double surname, and involves no pause. Strong phrase boundary requires the insertion of a short pause (50-300 msec.) and is located in correspondence with comma or colon and anyway between long NP and VP. Sentence boundary corresponds to semicolon, full stop, exclamative and question marks, and is always accompanied by a long pause (500-1000 msec.). Finally, the prosodic processor distinguishes six sentence melodic modalities.

The rules managing such information are written in an abstract formalism and translated into C procedures by SCYLA (Lazzaretto and Nebbia, 1987), a specially designed rule development system which allows: to express contextual relations between sequentially organized elements by means of variables affecting different descriptive linguistic levels, hierarchically organized; to maintain data structures alignment between the different descriptive representations of the sentence; to associate duration and f0 values to each phoneme through a sequence of steps. As long as the position of a given segment is crucial for effectively determining prosodic parameter variations, suitable counters are activated which detect step-by-step the units' location within the higher level unit they belong to, i. e. phonemes into syllable, syllables into phrase and sentence (counting stressed syllables is equivalent to counting phonological words), phrases into sentence. The variables pertaining to the several descriptive levels can be indifferently used for specifying the context in any procedure. A proper audio-visual debugger is available for rule testing, allowing to scan up to ten parameters for each phoneme, to interactively change parameter's values in every point of the phoneme string and to listen to the speech output.

DURATION RULES

The linguistic model underlying the duration rules (Salza and Sandri 1986, Salza 1988), is based on the principle of superposition of effects and on the possibility of modifying speaking rate, as previously illustrated (Nebbia, 1990). Standard rate corresponds to that used by speakers in aloud reading (about 5 syllables per sec.), but it can be continuously varied from -50% to +50%. If the sentence length is <= 5 syllables an automatic control slows the rate by 20%. Further specifications have been recently included into the general model coming from the analysis of dedicated natural speech corpora for the application of TTS to the reading of surnames, acronyms and addresses in the automation of telephone reverse directory service. New rules proved to contribute dramatically to the improvement of TTS intelligibility (Balestri et al., 1992).

Three separate procedures are devoted to assign, respectively, Intrinsic Duration of phonemes (Di), Phonetic Context Coefficients (Cph) and Syntactic Context Coefficients (Csy) to a target which is usually the single phoneme, rarely a whole class of speech segments. The context is expressed by means of the variables listed below, affecting five descriptive levels. In parentheses are signed the procedures involved.

Level Phoneme: individual phoneme label; phoneme articulatory (manner and locus) classes; voicing; intrinsic length, i. e. full vowel, semiconsonant, reduced vowel, single consonant, geminated consonant (Di, Cph).

Level Syllable: stressed/unstressed; open/closed (Cph, Csy).

Level Word: grammatical category (Cph, Csy).

Level Phrase: unit length, i. e. short phrase, full phrase; boundary strenght, i. e. weak phrase boundary, strong phrase boundary, sentence boundary (Cph, Csy).

Level Sentence: melodic modality, i. e. affirmative, interrogative, etc.; further specification, i. e. isolated word, long sentence, reverse directory service (Cph, Csy).

By making efficient use of SCYLA facilities, the maximum context extension now reached by the rules is 5 units, target phoneme included.

The möst important phonetic context (microprosodic) phenomena covered by rules are the following. Both stressed and unstressed vocalic segments appear longer in vowel clusters than in diphthongs and in interconsonantal position. Increasing of consonant duration is apparent when they are in clusters. Stress realization involves: vowel lengthening from 50% to 150%, according to the surrounding phonetic context, semiconsonants longer in stressed diphthongs than in unstressed ones, slight duration increasing of intervocalic consonants. Like in many european languages, vowel duration is fairly influenced by voicing and articulatory class of the following consonant.

As for so called syntactic context effects, adjacency to sentence and phrase boundaries determines a generalized duration lengthening of the involved syllables, on the average more noticeable in final than in initial position, larger in final position for vowels, in initial position for consonants. Sentence/phrase final segments lengthens from 20% to 180%, stressed segments less than unstressed ones. In presence of weak phrase boundary rules assign to segments in pre- and post-boundary positions slight lengthenings (from 20% to 40%). Heavier increasings are assigned in presence of strong phrase boundary. Moreover, the final vowel of the first word in long NP lengthens by 60%. Sentence boundary determines even more consistent duration lengthenings on adjacent segments. In isolated words every variation is stronger than in long sentences.

INTONATION RULES

Intonation rules have been developed relying on a multi-layered description of intonation (Avesani, 1990). An abstract phonological description of the linguistically relevant f0 modulations, which can be traced back to Pierrehumbert intonation model (Pierrehumbert, 1980), is well suited to represent the correspondence between tonal elements and structural events such as stressed syllables or syntactico-prosodic boundaries. A less abstract approach ('t Hart et al., 1990) has been adopted both as a methodological guideline for the experimental analysis of natural f0 curves and as a framework for rule implementation. This detailed phonetic and acoustic interpretation of tones is perspicuous in drawing the exact shape of pitch variations, which seems to be perceptually relevant and should be reproduced by the TTS system. Finally, the actual computation of f0 values takes into account those (physiological) features, such as pitch range and declination, which can be mathematically modeled as global trends of the f0 curve.

The intonation rules interpret the abstract tonal elements as pre-defined configurations of stylized 'pitch movements', classified according to their direction, dimension and timing with respect to stressed syllables and boundaries. Separate procedures, to be sequentially applied to the whole sentence, respectively:

- represent sentence intonation as a sequence of pitch configurations determined by the syntactico-prosodic structure;
- convert each configuration into a sequence of pitch movements by associating to the proper phonemes the targets of pitch variations, expressed as differences (in semitones) with respect to a reference baseline;
- combine targets with the current baseline, which depends on pitch range, declination and reset points;
- linearly interpolate target f0 values in order to obtain the complete f0 curve.

Declarative sentences are realized as a sequence of 'pointed hat' configurations aligned with stressed syllables, with a 'continuation rise' on the last syllable preceding a phrase boundary and with a 'flat hat' final configuration on the last two words. Phonologically, declarative contours can be represented as follows:

((H*) L H%) (H*) H* H+L* L L%

Phonetically, H* is interpreted as a rise starting on the pretonic syllable and reaching a peak at the end of the stressed vowel, followed by a fall on the postonic syllable. The continuation rise L H% is realized as a low rise on the last vowel before the boundary. The flat hat H* H+L* L L% is a rise on the pretonic and stressed syllables of the penultimate word, followed by a gradual slight fall till the last stressed syllable, a steep fall on the last stressed vowel and a further slight fall till the end of the sentence. Special cases of declarative sentences are isolated words, which are realized with a high plateau followed by a steep fall on the stressed vowel and a further fall on the postonic syllables.

Interrogative sentences are realized with peculiar initial and final configurations and with 'reduced pointed hat' configurations on intermediate stressed syllables. As a final configuration a fall is realized on the syllables preceding the last stressed vowel, followed by a steep rise on the stressed vowel and a high plateau till the end of the sentence (higher in yn-questions). Different initial configurations are assigned to yn- and wh-questions. A rise-fall is realized on the first two words of yn-questions, reaching its peak at the end of the first stressed vowel and coming back to the baseline after the second stressed syllable. Two kinds of wh-questions are distinguished. If the wh-word is unstressed a rise is realized reaching its peak at the beginning of the first stressed vowel of the sentence, followed by a steep fall on the stressed vowel and the postonic syllables. If the wh-word is stressed a gradual fall is realized on it, starting well over the baseline at sentence beginning and reaching the baseline on the pretonic syllable of the second word.

CONCLUSION

Current research aims at realizing a more sophisticated text-analyzer which would give a richer and more reliable representation of the prosodic structure of sentences. Experimental analyses are also in progress in order to enhance the modelling of duration and intonation in long sentences, accounting for: destressing and graduation of stress levels, contextual speaking rate variations, identification of semantic focus, closer correlations among variations of different parameters for realizing prominence. As an alternative to rule systems, adaptive systems are also under investigation. Some encouraging results have already been obtained in the application of Neural Nets and CART to the prediction of phoneme duration.

REFERENCES

C. Avesani (1990), "A contribution to the synthesis of Italian Intonation", *Proc. ICSLP '90, Kobe, November 1990,* Vol. 2, pp. 833-836.

M. Balestri, E. Foti, L. Nebbia, M. Oreglia, P.L. Salza and S. Sandri (1992), "Comparison of natural and synthetic speech intelligibility for a reverse telephone directory service", *Proc. ICSLP '92, Banff, October 1992*, Vol. 1, pp. 559-562.

M. Balestri, S. Lazzaretto, P.L. Salza and S. Sandri (1993), "The CSELT system for Italian text-to-speech synthesis", to be presented at *EUROSPEECH '93, Berlin, September 1993.*

S. Lazzaretto and L. Nebbia (1987), "SCYLA: Speech Compiler for Your LAnguage", *Proc. EUROSPEECH '87, Edinburgh, September 1987*, Vol. 2, pp. 381-384.

L. Nebbia, "Text-to-Speech Synthesis System for Italian: an Overview", *Proc. VERBA '90, Rome, January 1990*, pp. 326-333.

J. Pierrehumbert (1980), *The Phonology and Phonetics of English Intonation*, Ph.D. dissertation (MIT, Cambridge).

P.L. Salza (1988), "Durations of Italian diphthongs and vowel clusters", *Language and Speech,* Vol. 31, Part 2, pp. 97-113.

P.L. Salza and S. Sandri (1986), "Microprosodic timing rules for consonant clusters in Italian", *Proc. ICASSP '86, Tokyo, April 1986,* Vol. 3, pp. 2035-2038.

J. 't Hart, R. Collier, A. Cohen (1990), *A Perceptual Study of Intonation* (Cambridge University Press, Cambridge).

Lexicon and prosodic structure in a text-to-speech system

Sverre Stensby, Berit Horvei and Georg E. Ottesen
SINTEF DELAB, N 7034 Trondheim, Norway

ABSTRACT

This article is a presentation of a pronunciation lexicon and its use in improving prosody in a text-to-speech system. The inclusion of a lexicon is a natural way of giving synthesis systems information of word class and word pronunciation. The grammatical information is used to disambiguate homographs and to form a basis for prosodic structuring. The realized prosody is based on patterns for rhythmical units named feet and utilizes the information of word tone and pronunciation found in the lexicon.

INTRODUCTION

The scope of this paper is to present a pronunciation lexicon and its use in improving prosody in a text-to-speech (TTS) system. The paper addresses lexical entry, syntactic analysis, and prosodic structuring and realization. Parts of the work have been published previously, Stensby (1991, 1992, 1993).

LEXICON

When we read most of the words are known to us. We thus know their meaning and how to pronounce them. The inclusion of a lexicon in TTS-systems is a natural way of giving such systems similar information about the words in the text.

By lexicon we mean a dictionary in a machine readable form. Our lexicon is based on full word forms inflected by rule. Each entry includes the inflected word form, the pronunciation, the word class and grammatical function, and reference to the basic part of the word, i.e. the lexeme. The pronunciation is included because letter-to-sound rules will never be foolproof. For words not found in the lexicon, so-called unknown words, the pronunciation is found by rules.

Many complicated changes in pronunciation occur for the individual inflected forms of the same lexeme. These irregularities are the main reason for using a full word lexicon and not utilizing morphology in this first version. The method of inflection by rule allows for easy control of the produced words.

Basic word list

The basis for the lexicon is a small Norwegian dictionary and a list of the most common full word forms in newspaper text. The pronunciation and grammatical information given in the original dictionary are improved and extended. The grammatical information includes the word class and inflectional pattern for the word. The pronunciation includes indication of the stressed syllable and word tone. The inclusion of the word tone is important since there are no infallible rules for allocating this information.

We distinguish between open and closed word classes. The set of words belonging to the closed classes is mainly stable with time and is on the whole covered by a finite number of words. Thus it is possible to include practically all words belonging to the closed classes in a dictionary. In this lexicon the set of closed classes are the infinitive marker, auxiliary verb, determiner, conjunction, preposition, interjection, and pronoun. These comprise in sum 525 words. The open word classes are noun, adjective, numeral, name, verb, and adverb. This gives jointly a basic word list of 12000 words, mainly consisting of lexemes.

Generation of inflected full word forms

Norwegian is partly an agglutinating language with many inflected forms created by adding suffixes. The inflected full word forms are generated by rule from the words in the base word list. Typically 4, 7 and 6 full word forms for nouns, adjectives and verbs respectively are generated, though there may be additional forms due to alternatives. To cover all these patterns 70 inflection models are included in the system. The resulting lexicon of inflected full word forms has 48000 entries. An example is shown in Table 1.

Table 1. *Example of generated full word forms for the verb grave (= dig). The apostrophes indicate the stressed syllable and the type of tone.*

Word	Pronunciation	Lexeme	Gram. category
grave	[grˋɑːvə]	grave	infinitive
graver	[grˈɑːvəɾ]	grave	present tense
gravde	[grˋɑvdə]	grave	past tense
grov	[grˈuːv]	grave	past tense
gravd	[grˈɑvd]	grave	perfect
gravende	[grˋɑːvənə]	grave	present participle
grav	[grˈɑːv]	grave	imperative

Table 1 illustrates that the change in pronunciation may be complicated even in words with regular inflection. Especially the word tone may change. The symbols ˈ and ˋ are used for tone 1 and tone 2 respectively.

Tones

Norwegian, Swedish, and some other languages have a limited use of tone, and most Norwegian dialects distinguish between two tones named tone 1 and tone 2. The difference is manifested in the stressed syllable where a fall in the fundamental frequency (F_0) is typical for tone 2. In Norwegian there are more than two thousand pairs of words with identical sound segments which are distinguished by tone alone. Many of these are homographs and need to be disambiguated.

The distribution of the tone 1 and 2 on the word initial syllable and a word internal syllable is shown in Table 2. Tone 2 predominantly occurs on the first syllable, while tone 1 is uniformly distributed between the initial and word-internal syllables. This information may be utilized in the task of allotting word tone to unknown words.

GRAMMATICAL ANALYSIS

The main task of the grammatical analysis is to disambiguate homographs and thus find the correct pronunciation and grammatical function. The system disambiguates

Table 2. *Relative occurrences of tone 1 and 2 in the lexicon depending on the accented syllable.*

Word tone	1	2
Word initial syllable	0.22	0.51
Word internal syllable	0.26	0.01

homographs of different word classes due to differences of their function in the sentence. The word class is used in assigning the prosodic structure.

In a text there may be words not found in the lexicon. Such unknown words are assumed to belong to one of the open word classes since all words belonging to the closed classes are presumed to be included in the lexicon. This restriction allows for treating the unknown words as homographs of the open word classes. The allocated word class may also be used by the pronunciation rules.

Parsing

The text is analysed in a multi-pass associative parser. In each pass the text is compared with phrase patterns, and a phrase is created when a match is found. The least ambiguous patterns are sought first, the prepositional phrase is an example of such a pattern.

PROSODIC STRUCTURING

The prosodic structuring utilizes the word class of the words, the established phrases, and rules for allotting sentence accent and focal accent. The prosodic structuring works locally within sentences. A unique division into sentences requires semantic analysis beyond the scope of this work. We therefore define every punctuation mark as a sentence divider, even though this gives a large number of sentences.

Focal accent (the highest prominence) is allotted to new information only, and is primarily located late in the sentence. This gives broad focus which we regard as the best when the assignment is based on simple assumptions only.

Focal accent is given to the last new noun or adjective in a sentence, if any, while ordinary accent is given to noun, adjective, verb or adverb. The word classes are given preference in the mentioned order. There is at most one accent per phrase and one focal accent per sentence.

The distinction between new and known is determined by a table of the lexemes occurring earlier in the passage. By use of the corresponding lexeme and not the word form itself, the words may be recognized in different grammatical forms. There is no forgetfulness in the present system, but this might be implemented simply by remembering the last few words of a passage, Horne (1992).

PROSODIC REALIZATION

The prosodic realization is based on patterns for F_0 and rhythm over accented and focally accented feet. A foot is a rhythmical unit starting with a pitch-accented syllable. The tone of the foot-initial word is assigned to the foot. The distinction between accent and focal accent is signalled by the F_0 at the end of the foot. The end of a foot coincides with the end of the text, or with the beginning of the next foot, or with a so-called foot-external section. The latter is a section which joins a foot and a pause or two feet.

A comparison of F_0 and timing in an authentic sentence (solid line named Arne) and a synthetic sentence (dotted line named Modell2) is shown in Figure 1. The figure is generated by a system for studying prosody in texts read aloud, Ottesen (1992), Horvei et.al. (1993). In translation the sentence is "The painter worked for a month".

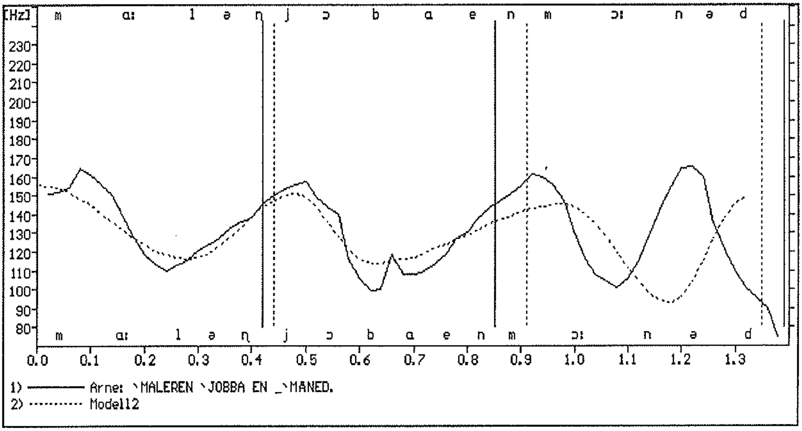


Figure 1. *Comparison of authetic and modelled fundamental frequency contour. Foot boundaries are indicated by vertical lines. The time scale is in seconds.*

The grammatical analysis, the prosodic structuring, and the prosodic realization are done by rule. The structure is indicated by vertical lines and apostrophs in Figure 1 and the _ indicates the focal accent. The established prosodic structure is identical to the reading, and there is good agreement between the authentic and synthesised contours apart from a slight difference in timing and in the final F_0.

ACKNOWLEDGEMENTS

This research has been funded by the Norwegian Telecom and was carried out in cooperation with the Department of Linguistics at the University of Trondheim.

REFERENCES

M. Horne (1992), "Semantic and pragmatic conditioning of accenting in Swedish - implications for speech synthesis", *Proc. Nordic Prosody VI, KTH, Stockholm.*

B. Horvei, G. Ottesen and S. Stensby (1993), "Analysing prosody by means of a double tree structure", *Proc. EUROSPEECH 93, Berlin, Germany.*

G. Ottesen (1992), "A method for studying prosody in texts read aloud", *Proc. ICSLP 92, Banff, Canada*, pp. 1271-1274.

S. Stensby (1991), "Prosody in a rule-based Norwegian text-to-speech system", *Proc. EUROSPEECH 91, Genova, Italia*, Vol 3, pp.1149-1152.

S. Stensby (1992), "Prosody in a text-to-speech system for Norwegian", *Proc. Nordic Prosody VI, KTH, Stockholm.*

S. Stensby (1993), "Struktur av prosodi i talesyntese", *Avhandling for graden doktor ingeniør, Institutt for teleteknikk, Akustikk. Norges Tekniske Høgskole, Universitet i Trondheim.* (In Norwegian. Abstract in English).

Synthesizing Intonation using the RFC Model

Paul Taylor
ATR Interpreting Telecommunication Laboratories, 2-2 Hikaridai,
Seika-cho, Soraku-gun, Kyoto, JAPAN.
(email: paul@itl.atr.co.jp)

ABSTRACT

This paper describes an implementation of the rise/fall/connection (RFC) model of intonation for use in the speech synthesis component of a machine interpretation system. The phonological input is described, as is the algorithm which converts this to a rise/fall/connection description, and eventually to an F_0 contour. As we wish to model different speakers' intonational characteristics, some details are described on the method for adapting the model's parameters to match the behaviour of a given speaker.

1 INTRODUCTION

The English speech synthesis system currently under development at ATR is part of a machine interpretation system. In the design of the phonetic intonation module of this system (which converts phonological intonation descriptions to F_0 contours) we had two basic goals: comprehensive coverage of the intonational phenomena of English and natural sounding intonation that reflects an individual's speaking style.

These design criteria dictate that our system has a different structure and operation to intonation systems found in most text-to-speech (TTS) systems (e.g. [1]). The system has to synthesize a wide range of intonational effects as we want to be able to ask different types of questions and statements. It is not simply enough to produce a "neutral declarative" style of intonation. However, as we were dealing with machine generated utterances we have a richer, more accurate linguistic description of that utterance than would be typically be available from a text-input system. Thus our system has to perform more complex tasks than most TTS intonation systems, but to balance that we have access to more detailed information about the utterance.

In addition to synthesizing all the intonational effects of English we want to capture particular speakers' phonetic realisation characteristics. This requires that we must have some method of deriving a speaker's intonational characteristics from a database of that speaker's voice. Thus our system has to be capable of analysing F_0 contours, extracting intonational characteristics and using these characteristics for synthesis.

2 BASIC SYNTHESIS SCHEME

The synthesis algorithms are based on the *rise/fall/connection* ("RFC") model of intonation [8]. This model was designed to be formal in its workings so that all its operations are explicitly defined and therefore easy to implement on computer. The model was designed for both analysis and synthesis: the synthesis mapping takes a phonological input and produces an F_0 contour, the analysis mapping takes a F_0 contour and produces a phonological description. Algorithms have been developed which can automatically label F_0 contours, and it has been shown that contours generated using the model are very similar to naturally occurring ones [8], [9]. This paper focuses on the practical requirements needed to use this model in a speech synthesis system.

There are several theoretical advantages [8] this model has over others but the principle reason for its use here concerns the analysis capability of the system. As precise, detailed F_0 contour labelling is straightforward, we can easily use large quantities of data to determine the settings and parameters of the model for a particular speaker, and thus imitate that speaker's phonetic realisation behaviour.

2.1 INPUT

The input to the system is a specification of the intonational phonological description of the utterance. This is made up from four things: *tune association, tune type, pitch range* and *phrasing*.

Tune association simply indicates where the phonological elements occur. Pitch accents are associated with single syllables, and boundary elements (see below) are associated with phrase boundaries.

Tune type refers to which *elements* and*features* in the HLCB system are being used. The HLCB tune description system uses four basic elements. Pitch accents are of either type **H** (high) or **L** (low), phonologically relevant connection elements (see below) are marked **C**, and rapid rises at phrase boundaries are marked **B**. Within these classes, features are used for sub-classification. **H** accents are the most common type found in utterances and have three features , "downstepped" (**H**$_d$, "late" (**H**$_l$) (for accents with peaks late in the syllable, corresponding to the British school rise-fall [7]) and "elevated" (**H**$_e$) for extra high accents (high fall in the British school) [3]. The **L** class has a single feature "antecedent" (**L**$_a$) which indicates whether the **L** has a steep fall leading into it. **C** elements are normally of no phonological significance, and simply "fill in" the contour between pitch accents and boundary elements, but can be marked with the feature "rising" to distinguish them from the usual (level or falling) behaviour. **C**$_r$ elements are often found after the nuclear accent, e.g. **H**$_d$ **C**$_r$ is similar to a fall-rise accent in the British school. **B** elements can occur at the starts of phrases (often termed "declination resets"), is which case they are marked as being "initial", or at the ends of phrases where they are unmarked.

Phrasing structure is in principle recursive in that there ,are an arbitrary number of phrase types and each phrase may include any other phrase [4]. In practice, we define a small number of phrase types (currently four) and only include daughter phrases of the same or a lower type in a parent phrase. The phrase information is mainly for use with the duration module, which determines phrase-final lengthening and pause durations [2]. The intonation module itself only makes use of the phrase boundary locations and is not concerned with the actual type of phrase.

Each phrase can take a number of parameters which are constant for the duration of that phrase. The most important parameter for the intonation module is *pitch range*. A (variable) number of pitch range levels are defined, each with a name, a maximum and a minimum F_0 value. The pitch range of a phrase determines the starting F_0 value of that phrase and the maximum and minimum values that the F_0 contour can vary between. A phrase may be marked as lying in a particular pitch range or be left marked. If marked, that pitch range type specifies the starting pitch of the phrase, and the minimum and maximum values that the contour can very between. If left unmarked, the starting F_0 of the phrase is the same as the last F_0 of the previous phrase. In the unmarked case the maximum and minimum values are determined by moving up through the phrase tree until a parent phrase is found with a marked pitch range, and that pitch range is then used. The top level phrase always has a marked pitch range. This pitch range system is similar to the register system of Ladd [5], but here we only specify one register per phrase rather than one for each accent. We don't have to specify a register setting for each accent as it is thought that our tune description system is powerful enough to account for the prominence variation of accents within the phrase.

2.2 Example Input

```
(Phrase ((Type S) (PitchRange two))
    (Phrase ((Type C) ())
        (okay              (H (d 1)) (B ()))
    )
    (Phrase ((Type C) ())
        (Phrase ((Type P) ())
            (i'll                      )
            (send              (H (d)))
            (you                       )
        )
        (Phrase ((Type P) (PitchRange three))
            (a                 (B (i)))
            (form              (H (d)))
        )
    )
)
```

A typical input to the system is given above. (Only the information relevant to the intonation module is shown, and thus words are given in their text form rather than their phonological form.) The scope of each phrase is shown by its bracketing. The first entry in the phrase is a list of features for that phrase, including its type (mainly used in the duration module) and its pitch range. The names and values of the pitch ranges are pre-defined. Here we simply use the names "one", "two", "three" etc. For example, pitch range "two" is defined as having a minimum F_0 values of 90Hz and a maximum of 220Hz. After each word there is an optional list of HLCB elements, with the features for each specified in brackets.

3 SYSTEM OPERATION

The system creates F_0 contours by taking the HLCB description and producing a RFC description, which is a linear sequence of *rise*, *fall* and *connection* elements, each with a duration and amplitude. From this RFC description equations are used to generate F_0 contours.

The equation for the fall element is given below and the equation for the rise element is the same as this but reflected in the y (F_0) axis. Connection elements are realised as straight lines.

$$\begin{array}{ll} f_0 = A(1-2(t/D)^2) & 0 < t < D/2 \\ f_0 = 2A(1-t/D)^2 & D/2 < t < D \end{array} \tag{1}$$

Where t is time, A is the amplitude and D is the duration of the element.

3.1 HLCB and Pitch Range to RFC

HLCB elements re-write to RFC elements. **H** becomes a rise element followed by a fall element, **L** becomes a fall element followed by a rise element, **B** becomes a rise element, and **C** becomes a connection element.

A pitch accent requires five parameters; the rise amplitude and duration, the fall amplitude and duration, and some indication of how these elements are aligned with the segmental part of the utterance. At present, accent alignment is defined in terms of the distance between the start of the vowel in the syllable and the position of the peak (juncture between rise and fall elements in a **H** accent) or trough (juncture between fall and rise elements in a **L** accent). A *definitions table* (which is read from file at run time) determines basic default parameter values for each element. This table also specifies the modification that the features make on the elements (see section 4).

Most of the phonological tune specification is in terms of **H**, **L** and **B** elements. **C** elements need only be marked if they carry the feature "rising" as a **C** that is unmarked is assumed as the default. After the **H**, **L** and **B** accents are realised, rising **C** elements are added, and then any remaining part of the utterance which is not covered by an element is designated as being a connection element. After this, the entire utterance is specified in terms of RFC elements.

Pitch range information is used to determine the absolute (i.e. with respect to 0 Hz) amplitudes of the elements. The maximum and minimum values for the phrase pitch range are determined by the method described in section 2.1. If a phrase is explicitly marked as being in a particular pitch range, the defined minimum F_0 of that pitch range is used as the absolute amplitude of the start of the first element, otherwise the end amplitude of the previous element is used. In most cases the first element in a phrase is a $\mathbf{B}_i$ which serves to raise the F_0 level from the pitch ranges's minimum to a more medial level.

Once the starting amplitude of the phrase has been decided, the absolute amplitude of each element is calculated by adding its amplitude (which is negative in the case of fall elements) to the end amplitude of the previous element. If this operation results in the F_0 level exceeding the specified maximum or minimum of that phrase, the amplitudes are constrained so as the F_0 contour is kept within the designated pitch range. An additional phrase-final lowering rule (similar to that described in Liberman and Pierrehumbert [6]) rule states that if the last accent in a phrase is a $\mathbf{H}_d$ this always falls to the bottom of the pitch range.

3.2 RFC to F_0

The conversion of an RFC description to an F_0 contour is straightforward. The list of RFC elements is processed left to right and the equations are used to produce a continuous contour. Depending on the F_0 input for the particular synthesizer, this F_0 contour can be further processed. For example, some synthesizers require there to be a zero F_0 during unvoiced segments, and this can easily be achieved simply by masking the contour and setting it to zero in these regions. As yet, no segmental perturbations or micro-prosody effects are incorporated into the system.

4 DEFINING ELEMENTS AND FEATURES

At present, nearly all the operation of the system can be determined at run time by specifying how HLCB descriptions are mapped onto RFC descriptions. These parameters are kept in the definitions file which may be altered at will. The basic re-write rules cannot be varied, but the default sizes of the HLCB elements, and the effect the features have on those elements are all variable. It is also possible to specify what units these parameters are given in, e.g. both Hertz and semitones can be used to described amplitudes. Some typical parameters definitions are given below. Each statement consists of a variable, an operator, and a value. The "=" operator implies a straight assignment, an operator such as "+=" means "the existing value plus the new value" (in a similar syntax to the C programming language).

```
(define Element H                    (define Feature late
      (rise_amp    = 50 Hz)                (rise_amp       += 0 Hz)
      (fall_amp    = 50 Hz)                (fall_amp       += 0 Hz)
      (rise_dur    = 120 ms)               (rise_dur       += 100 ms)
      (fall_dur    = 160 ms)               (fall_dur       += 0 ms)
      (peak_pos    = 0.5 rel)              (peak_pos       *= 1.8 rel)
)                                    )
```

The first list defines the default **H** accent properties. In this example the accent amplitudes are defined in Hz, and the peak position is defined as occurring half way through the syllable, although it is possible to specify peak position in absolute ms terms also. The second list defines the behaviour of the late feature. In this definition, the values defined for the **H** accent are modified, such that everything is left unchanged except for the `rise_dur` which is increased by 100 ms and the `peak_pos` which is shifted back by a factor of 1.8. The feature definitions can also have non-modifying assignments, for example one could define downstepping accents as having no rise component, which would be achieved by setting stating `(rise_amp = 0 Hz) (rise_dur = 0 ms)`.

Current work is concentrated on deriving the definitions automatically by using the RFC analysis system on large sets of F_0 contours [9].

References

[1] J. Allen, S. Hunnicut, and D. Klatt. *From Text to Speech: the MITalk System*. Cambridge University Press, 1987.

[2] W. N. Campbell and S. D. Isard. Segmental durations in a syllable frame. *Journal of Phonetics*, 19:37–47, 1991.

[3] D. Robert Ladd. Phonological features of intonation peaks. *Language*, 59:721–759, 1983.

[4] D. Robert Ladd. Intonational phrasing: the case for recursive prosodic structure. *Phonlogy Yearbook 3*, pages 311–340, 1986.

[5] D. Robert Ladd. A model of intonational phonology for use with speech synthesis by rule. In *European Conference on Speech Technology*. ESCA, 1987.

[6] Mark Liberman and Janet Pierrehumbert. Intonational invariance under changes in pitch range and length. In Mark Aronoff and Richard T. Oehrle, editors, *Language Sound Structure*. MIT Press, 1984.

[7] J. D. O'Connor and G. F. Arnold. *Intonation of Colloquial English*. Longman, 2 edition, 1973.

[8] Paul A. Taylor. *A Phonetic Model of English Intonation*. PhD thesis, University of Edinburgh, 1992.

[9] Paul A. Taylor. Automatic recognition of intonation from F_0 contours using the rise/fall/connection model. In *Proc. Eurospeech '93, Berlin*, 1993.

Studies of Intonation and Discourse

Julia Hirschberg
AT&T Bell Laboratories
2D-450 600 Mountain Avenue, Murray Hill NJ 07974, USA

ABSTRACT

Research on intonation and discourse falls into two major categories: work on the intonational correlates of discourse structure and work on accent and information status. In both categories, problems of specifying an adequate and independently motivated discourse model hinder evaluation of results and generalization across experiments. Also, much work remains to be done on combining these results into general models capturing the mapping between intonational features and discourse features.

INTRODUCTION

Most research on intonation and discourse to date has fallen into one of two categories: investigations of the intonational correlates of topic structure or studies of the relationship between information status and intonational prominence. Much of this work has involved empirical experimentation or corpus-based research. For the latter, the need for large, shareable, prosodically labeled corpora is viewed as increasingly important, given the labor-intensive natural of corpora labeling. To promote the development of such corpora, some efforts have been made to agree upon common labeling standards, such as the current TOBI standard for Standard American English which has recently been proposed (Silverman et al.,1992). While the term 'discourse' might seem to suggest spontaneous conversation, in fact, most work in this area has defined discourse more generally as 'utterances in context'; so, monologues, elicited speech, read speech, and radio speech, have been more frequently examined than natural dialogue.

PROSODIC CUES TO DISCOURSE STRUCTURE

Most researchers who work on discourse accept that it is structured into segments; disagreement arises primarily over the nature of the larger units into which segments are grouped, and the relationship among individual segments. To date, most studies of prosody and discourse structure have focussed on how intonational and acoustic variation signals segment boundaries and conveys larger 'topic' structures.

The notion that discourse structure is signalled by variation in intonational features such as pitch range, timing, and amplitude — and probably variation in some combination of these features — has been widely believed for years. However, there has been surprisingly little empirical testing of this belief. A major problem is noted by Brown et al. (Brown et al.,1980, p. 27) in discussing some production studies designed to elicit acoustic and intonational cues to discourse structure : "... until an independent theory of topic-structure is formulated, much of our argument in this area is in danger of circularity." In fact, most speech-based empirical studies have assumed a particular structure for the discourses they examine or use as stimuli and have then looked for acoustic-intonational indicators of these assumed structures, usually using the experimenter's intuitive notions about changes in topic and topic-subtopic relations. Alternatively, they have used lexcal phenomena believed to be constrained by discourse structure (such as pronominal forms) or supposed to explicitly indicate structure (such as cue phrases) as indicators of structure, even when these hypotheses themselves remain to be tested.

One of the features most frequently mentioned as important to conveying some kind of 'topic structure' in discourse is PITCH RANGE (the distance between the maximum of the FUNDAMENTAL FREQUENCY (f0) for the vowel portions of accented syllables in the phrase and the speaker's *baseline*, defined for each each speaker as the lowest point reached in normal speech over all . In a study of speakers reading a story, Brown et al. (Brown et al.,1980) found that subjects typically started new topics relatively high in their pitch range and finished topics by compressing their range; they hypothesized that internal structure within a topic was similarly marked. Lehiste (Lehiste,1975) had reported similar results earlier for single paragraphs. Silverman (Silverman,1987) found that manipulation of pitch range alone, or range in conjunction with pausal duration between utterances, could enable subjects to disambiguate utterances that were intuitively potentially structurally ambiguous reliably; for example, he used a small pitch range to signal either continuation or ending of a topic or quotation, and expanded range to indicate topic shift or quotation continuation. Avesani and Vayra (Avesani and Vayra,1988) also found variation in range in productions by a professional speaker which appear to correlate with topic structure, and Ayers (Ayers,1992) found that pitch range appears to correlate more closely with hierarchical topic structure in read speech than in spontaneous speech. Swerts et al (Swerts et al.,1992) also found that f0 scaling was a reliable indicator of discourse structure in spoken instructions, although the structures tested were quite simple. Duration of pause between utterances or phrases has also been identified as an indicator of topic structure in (Lehiste,1979; Chafe,1980; Brown et al.,1980; Silverman,1987; Avesani and Vayra,1988; Swerts et al.,1992; Passoneau and Litman,1993). Brown et al. found that longer, 'topic pauses' (.6-.8 sec.) marked major topic shifts (Brown et al.,1980, 57). Passoneau & Litman (Passoneau and Litman,1993) also found that presence of pause was a good indicator of segment boundaries in Chafe's pear stories, when tested against their own subjects' segmentations of these stories. Another aspect of timing, speaking rate, was found by Lehiste (Lehiste,1980) and by Butterworth (Butterworth,1975) to be associated with perception of text structure: both found that utterances beginning segments exhibited slower rates those completing segments were uttered more rapidly. Amplitude was also noted by Brown et al. (Brown et al.,1980) as a signal of topic shift; they found that amplitude appeared to rise at the start of a new topic and fall at the end. Finally, contour type has been noted (Brown et al.,1980; Swerts et al.,1992) as a potential correlate of topic structure. In particular, (Hirschberg and Pierrehumbert,1986) suggested that so-called 'downstepped' contours (In which one or more pitch accents which follow a complex accent are uttered in a compressed range, producing a 'stairstep' effect.) commonly appear either at the beginning or the ending of topics. Empirical studies reported in (Swerts et al.,1992) showed that 'low' vs. 'not-low' boundary tones were good predictors of topic endings vs. continuations.

Recently, Hirschberg & Grosz (Hirschberg and Grosz,1992; Grosz and Hirschberg,1992) have addressed the problem of acoustic-prosodic correlates of discourse structure, inspired by the need to test potential correlates against an independent notion of discourse structure, as noted by (Brown et al.,1980). We looked at pitch range, aspects of timing and contour, and amplitude to see how well they predicted discourse segmentation decisions made by subjects using instructions based on the Grosz and Sidner 1986 (Grosz and Sidner,1986) model of discourse structure. Our corpus consisted of three AP news stories previously recorded by a professional speaker. Subjects labeled either from text alone (Group T) or from text (with all orthographic markings except sentence-final punctuation removed) and speech (Group S); average inter-labeler agreement for structural elements varied from 74.3% to 95.1% for subject decisions such as where segments began and ended. Decisions subjects all agreed upon were then correlated with variation in the acoustic-prosodic features mentioned above, as well as features such as change in f0 from preceding phrase, subsequent as well as preceding pause, absolute and relative amplitude, and type of nuclear pitch accent. We found statistically significant associations between aspects of pitch range, amplitude, and timing with segment beginnings and segment endings both for labelings from text alone and for labelings from speech.

For phrases labelled as beginning segments (We collapsed this category with phrases identified as SEGMENT MEDIAL POPS, those phrases which immediately followed a segment final phrase.) identified by Group T, we found significant effects for pitch range and subsequent pause; for Group S significant effects were found for pitch range, subsequent pause and preceding pause. So, segment beginnings do appear to be signalled by expanded range and timing, as previous studies had suggested. For phrases ending a segment, for both Group T and Group S, we found a single intonational correlate, subsequent pause; longer subsequent pauses are significantly associated with segment-final phrases. These findings confirm those noted above that pitch range and timing variation are important in signaling topic structure, and demonstrate that these relationships hold when topic structure has been independently determined from consensus subject labeling, which is based upon an independently-motivated theory of discourse.

We further found that segment beginnings and endings could be reliably identified from the same acoustic and prosodic features with considerable success. For example, automatically generated prediction trees distinguished segment beginnings from other phrases in 91.5% of cases, using only a simple combination of constraints on duration of preceding pause (> 647 msec.) and pitch range (< 276 Hz.). They distinguished segment-final phrases from other phrases in 92.5% of cases, using subsequent pause (> 913 msec.), amount of f0 change from prior phrase (< 93%), and overall rate for the story (> 4.76 sps).

While these initial studies were encouraging, they also revealed some problems with our experimental design: First, due to the speech corpus we employed, we had no access to the speaker's own intentions with respect to structure at the time of recording. Inferring these intentions from labelers' performance on text was much too indirect to be satisfying. The subject matter of the recordings, news stories, proved unexpectedly difficult to segment for our subjects. We also felt we had inadequate means to compare inter-labeler segmentation; clearly segment beginnings and endings only capture part of what is going on in a discourse. We will be addressing these problems in the next phase of the study.

INTONATIONAL PROMINENCE AND INFORMATION STATUS

How speakers decide which words to accent and which to deaccent is an open research question. While syntactic structure was once believed to determine accent placement, it is now generally held that syntactic, semantic, and discourse/pragmatic factors are all involved in accent decisions (Bolinger,1972; Bardovi-Harlig,1983). Word class, grammatical function, syntactic constituency, and surface position are still believed to influence accent location (Ladd,1979b; Erteschik-Shir and Lappin,1983), and there are some recent empirical results supporting this (Altenberg,1987; Terken and Hirschberg,1992). But it has also been found that less easily defined phenomena falling into the broad category of INFORMATION STATUS, including CONTRASTIVENESS (Bolinger,1961; Bing,1983; Bardovi-Harlig,1983; Couper-Kuhlen,1984), FOCUS (Jackendoff,1972; Rooth,1985; Baart,1987; Dirksen,1992; Wilson and Sperber,1979; Enkvist,1979; Gussenhoven,1983; Rochemont and Culicover,1990; Horne,1987; Zacharski,1992; Eady and Cooper,1986), and the GIVEN/NEW distinction (Brown,1983; Fuchs,1984; Kruyt,1985; Fowler and Housum,1987; Terken and Nooteboom,1987; Nooteboom and Kruyt,1987; Koopmans-van Beinum and van Bergem,1989; Horne,1991) influence accent decisions, with most of the empirical studies currently focussing on the last category. All of these types of information status are defined in terms of the structure of the discourse context. At least implicit in the notion of what is 'in focus' or what is 'given' in these accounts is some assumption about how discourses are structured, and what identifies an item as focussed or as given or as contrastive in its context. (Clearly, the cue of accentedness itself cannot be used as such an indicator for the study of accent itself, so some independent notion of discourse structure must be appealed to in order to establish the discourse variables to be tested.) Often, these models are not made explicit, or are greatly simplified for the purposes of the experiment; it is not always clear, thus, how results will generalize. Also, it is not easy to compare results when researchers have made different assumptions, as, about what defines 'givenness'.

The role of accent in reference resolution has been more speculated upon than studied, although observations such as Lakoff's (Lakoff,1971) classic *'John called Bill a Republican and then HE insulted HIM'* have long been noted. Some empirical work has also been done (Gleitman,1961; Hirschberg and Ward,1991; Horne,1985). However, given the heavy emphasis on this topic in text-based studies of discourse, there would appear to be richer fields to plow here than have yet been touched.

As yet there have been few attempts to combine potential or attested determinants of accent location into unified models of accent assignment, perhaps because the role of individual factors is still an open question. But it is important to start viewing our knowledge of the contribution of individual phenomena, such as 'givenness', within the larger framework of contributions from other discourse features and from syntactic and semantic features. There is of course considerable practical incentive to find solutions to these problems for text-to-speech synthesis. Many current research systems implement algorithms which attempt to make use of discourse-level information for accent assignment (Carlson and Granstrom,1973; Horne,1987; Hirschberg,1990; Monaghan,1991; Quené and Kager,1992); message-to-speech systems have also employed their richer sources of discourse information to improve prominence location (Davis and Hirschberg,1988; House and Youd,1990).

CONCLUSION

So, we have some evidence of some intonational and acoustic features that appear to signal certain aspects of discourse structure, such as topic beginnings and endings. And we have some notion about which discourse-level factors influence the decision to accent an item. In neither case do we know which factors are more important or more reliable than others. Nor do we know what sort of interaction there is among different cues. Nor do we know much about speaker or listener variability. Future work on intonation and discourse must thus address the following questions: What discourse factors influence intonational decisions, and how do these discourse factors interact with other components of the grammar? What sort of individual variation exists in these models? What assumptions are we making about our underlying models of discourse phenomena when we study the mapping between intonation and discourse? Are they justified?

REFERENCES

Bengt Altenberg (1987), *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*, Vol. 76 of *Lund Studies in English*. Lund University Press, Lund.

Cinzia Avesani and Mario Vayra (1988), Discorso, segmenti di discorso e un' ipotesi sull' intonazione, In *Att del Convegno Internazionale "Sull'Interpunzione"*, Florence.

Gayle M. Ayers (1992), Discourse functions of pitch range in spontaneous and read speech, Presented at the Linguistic Society of America Annual Meeting.

J. L. G. Baart (1987), *Focus, Syntax and Accent Placement*, Ph.D. thesis, University of Leyden, Leyden.

K. Bardovi-Harlig (1983), Pronouns: When 'given' and 'new' coincide, In *Papers from the 18th Regional Meeting*. Chicago Linguistic Society.

J. M. Bing (1983), Contrastive stress, contrastive intonation and contrastive meaning, *Journal of Semantics*, 2:141–156.

Dwight Bolinger (1961), Contrastive accent and contrastive stress, *Language*, 37:83–96.

Dwight Bolinger (1972), Accent is predictable (if you're a mindreader), *Language*, 48:633–644.

G. Brown, K. Currie, and J. Kenworthy (1980), *Questions of Intonation*. University Park Press, Baltimore.

G. Brown (1983), Prosodic structure and the given/new distinction, In D. R. Ladd and A. Cutler, eds., *Prosody: Models and Measurements*, pp. 67–78. Springer Verlag, Berlin.

B. Butterworth (1975), Hesitation and semantic planning in speech, *Journal of Psycholinguistic Research*, 4:75–87.

R. Carlson and B. Granstrom (1973), Word accent, emphatic stress & syntax in a synthesis by rule schenme for Swedish, *STL-QPSR*, 2(3):31–35.

W. L. Chafe (1980), The deployment of consciousness in the production of a narrative, In W. L. Chafe, ed., *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Navvative Production*, pp. 9–50. Ablex Publishing Corp, Vol. 3, Advances in Discourse Processes.

Elizabeth Couper-Kuhlen (1984), A new look at contrastive intonation, In Richard J. Watts and Urs Weidmann, eds., *Modes of Interpretation: Essays Presented to Ernst Leisi*, pp. 137–158. Gunter Narr Verlag, Tubingen.

J. R. Davis and J. Hirschberg (1988), Assigning intonational features in synthesized spoken directions, In *Proceedings of the 26th Annual Meeting*, pp. 187–193, Buffalo. Association for Computational Linguistics.

A. Dirksen (1992), Accenting and deaccenting: A declarative approach, In *Proceedings of COLING-92*, pp. 865–869.

S. J. Eady and W. E. Cooper (1986), Speech intonation and focus location in matched statements & questions, *Journal of the Acoustical Society of America*, 80:402–415.

N. Enkvist (1979), Marked focus: Functions and constraints, In S. Greenbaum, G. Leech, and J. Svartvik, eds., *Studies in English Linguistics for Randolph Quirk*, pp. 134–152. Longmans, London.

Nomi Erteschik-Shir and Shalom Lappin (1983), Under stress: A functional explanation of English sentence stress, *Journal of Linguistics*, 19:419–453.

C. A. Fowler and J. Housum (1987), Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction, *Journal of Memory and Language*, 26:489–504.

A. Fuchs (1984), Deaccenting and default accent, In D. Gibbon and H. Richter, eds., *Intonation, Accent and Rhythm*, pp. 134–164. Walter de Gruyter, Berlin.

L. Gleitman (1961), Pronominals and stress in English, *Language Learning*, 11:157–169.

B. Grosz and J. Hirschberg (1992), Some intonational characteristics of discourse structure, In *Proceedings of the International Conference on Spoken Language Processing*, Banff, October. ICSLP.

Barbara J. Grosz and Candace L. Sidner (1986), Attention, intentions, and the structure of discourse, *Computational Linguistics*, 12(3):175–204.

Carlos Gussenhoven (1983), *On the Grammar and Semantics of Sentence Accents*. Foris Publications, Dordrecht.

J. Hirschberg and B. Grosz (1992), Intonational features of local and global discourse structure, In *Proceedings of the Speech and Natural Language Workshop*, pp. 441–446, Harriman NY, February. DARPA, Morgan Kaufmann.

J. Hirschberg and J. Pierrehumbert (1986), The intonational structuring of discourse, In *Proceedings of the 24th Annual Meeting*, pp. 136–144, New York. Association for Computational Linguistics.

J. Hirschberg and G. Ward (1991), Accent and bound anaphora, *Cognitive Linguistics*, 2(2):101–121.

J. Hirschberg (1990), Using discourse context to guide pitch accent decisions in synthetic speech, In *Proceedings of the European Speech Communication Association Workshop on Speech Synthesis*, pp. 181–184, Autrans, France.

M. Horne (1985), English sentence stress, grammatical functions and contextual coreference, *Studia Linguistica*, 39:51–66.

Merle Horne (1987), Towards a discourse-based model of English sentence intonation, Working Papers 32, Lund University Department of Linguistics.

M. Horne (1991), Accentual patterning in 'new' vs 'given' subjects in English, Working Papers 36, Department of Linguistics, Lund University, Lund.

Jill House and Nick Youd (1990), Contextually appropriate intonation in speech synthesis, In *Proceedings of the ESCA Workshop on Speech Synthesis*, pp. 185–188, Autrans. ESCA.

Ray S. Jackendoff (1972), *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge MA.

F. J. Koopmans-van Beinum and D. R. van Bergem (1989), The role of 'given' and 'new' in the production and perception of vowel contrasts in read text and in spontaneous speech, In J. P. Tubach and J. J. Mariani, eds., *Proceedings of the European Conference on Speech Communication and Technology*, pp. 113–116, Edinburgh. Eurospeech, CEP, Vol. 2.

J. G. Kruyt (1985), *Accents from Speakers to Listeners: An Experimental Study of the Production and Perception of Accent Patterns in Dutch*, Ph.D. thesis, University of Leyden.

D. R. Ladd (1979b), Light and shadow: A study of the syntax and semantics of sentence accents in English, In L. Waugh and F. van Coetsem, eds., *Contributions to Grammatical Studies: Semantics and Syntax*, pp. 93–131. University Park Press, Baltimore.

George Lakoff (1971), Presupposition and relative well-formedness, In *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, and Psychology*, pp. 329–340. Cambridge University Press, Cambridge UK.

I. Lehiste (1975), The phonetic structure of paragraphs, In A. Cohen and S. G. Nooteboom, eds., *Structure and Process in Speech Perception*, pp. 195–203. Springer, Heidelberg.

I. Lehiste (1979), Perception of sentence and paragraph boundaries, In B. Lindblom and S. Oehman, eds., *Frontiers of Speech Research*, pp. 191–201. Academic Press, London.

I. Lehiste (1980), Phonetic characteristics of discourse, Paper presented at the Meeting of the Committee on Speech Research, Acoustical Society of Japan.

A. Monaghan (1991), *Intonation in a Text-to-Speech Conversion System*, Ph.D. thesis, University of Edinburgh, Edinburgh.

S. G. Nooteboom and J. G. Kruyt (1987), Accent, focus distribution and the perceived distribution of given and new information: An experiment, *Journal of the Acoustical Society of America*, 82(5):1512–1524.

R. Passoneau and D. Litman (1993), Feasibility of automated discourse segmentation, In *Proceedings of ACL-93*, Ohio State University. Association for Computational Linguistics.

Hugo Quené and René Kager (1992), The derivation of prosody for text-to-speech from prosodic sentence structure, *Computer Speech and Language*, 6:77–98.

Michael S. Rochemont and Peter W. Culicover (1990), *English Focus Constructions and the Theory of Grammar*. Cambridge University Press, Cambridge UK.

Mats Rooth (1985), *Association with Focus*, Ph.D. thesis, University of Massachusetts, Amherst MA.

K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg (1992), TOBI: A standard scheme for labeling prosody, In *Proceedings of the Second International Conference on Spoken Language Processing*, Banff, October. ICSLP.

K. Silverman (1987), *The Structure and Processing of Fundamental Frequency Contours*, Ph.D. thesis, Cambridge University, Cambridge UK.

M. Swerts, R. Gelyukens, and J. Terken (1992), Prosodic correlates of discourse units in spontaneous speech, In *Proceedings*, pp. 421–428, Banff, October. International Conference on Spoken Language Processing.

J. Terken and J. Hirschberg (1992). Deaccentuation and persistence of grammatical function and surface position. Ms.

J. Terken and S. G. Nooteboom (1987), Opposite effects of accentuation and deaccentuation on verification latencies for given and new information, *Language and Cognitive Processes*, 2(3/4):145–163.

Dierdre Wilson and Dan Sperber (1979), Ordered entailments: An alternative to presuppositional theories, In C.-K. Oh and D. A. Dinneen, eds., *Syntax and Semantics*, Vol. 11, pp. 229–324. Academic Press, New York.

Ron Zacharski (1992), Generation of accent in nominally premodified noun phrases, In *Papers Presented to the 15th International Conference on Computational Linguistics*, pp. 253–259, Nantes. International Conference on Computational Linguistics.

On the Prosodic Prediction of Discourse Finality

Marc Swerts
Institute for Perception Research (IPO)
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

ABSTRACT

It is investigated whether the approaching end of a spontaneously produced description is presignalled by prosodic means. Experiment I tries to determine to what extent listeners are able to estimate (on the basis of prosodic cues) how far a given utterance is situated from the end of a description. Experiment II is set up to systematically test - by means of manipulated synthetic prosody - the cue strength of speech melody and duration as predictors of discourse finality.

INTRODUCTION

One of the important functions of prosody is the demarcation of units of discourse. For instance, the end of discourse segments can be signalled by means of low boundary tones (e.g. Brown et al. 1980, Swerts et al. 1992). However, prosodic cues such as these tones are rather local in the sense of being positioned right before or at the actual boundary. It is a relevant question whether important breaks in the flow of information can also be presignalled, i.e. announced some time before they actually occur, so that listeners can anticipate them. Anecdotical evidence for the possibility of anticipation can e.g. be found in the small delay between the end of discourse segments and the onset of applause in political speeches.

Grosjean (1983) has already shown that subjects, basing themselves solely on prosodic cues, are surprisingly accurate at estimating the upcoming ending of a sentence. As Grosjean's study was limited to prosodic prediction at the sentence level in read-aloud speech, the present investigation is set up to test whether his findings can be generalized to (i) larger-scale discourse units (ii) in spontaneous speech. The analysis is centred on a particular type of descriptive language use, i.e. route descriptions.

EXPERIMENT I

In this experiment it is explored whether listeners can exploit prosodic cues of an utterance to estimate how far it is situated from the end of a (spontaneously produced) route description.

Speech materials, elicitation method

A Dutch speaker was asked to describe routes from given starting points to given end points on the basis of (schematic) city maps. Four of these descriptions were selected for further experimentation, the selection being based on the absence of clear non-prosodic (lexical) cues to discourse position. The seven last clauses of each description were isolated from their contexts, giving 28 utterances to be used in the listening experiment.

Perception test

Two lists of randomly ordered utterances were created, containing four repetitions of each utterance. Both lists were played to ten listeners each (chosen from students and staff of IPO) who were asked to estimate for each stimulus how many clauses (0 to 6), similar in length and nature to the one presented, followed in the original context. The interval between two utterances was 4 s, in which time period listeners had to respond. The listening test was preceded by the presentation of a short fragment of a route description (which was not used in the actual experiment) to get the subjects accustomed to the speaker's voice and speech register.

The listening results, averaged over the different types of clauses (-6 to -0) of the four route descriptions, are given in figure 1, which is a three-dimensional plot of a confusion matrix.

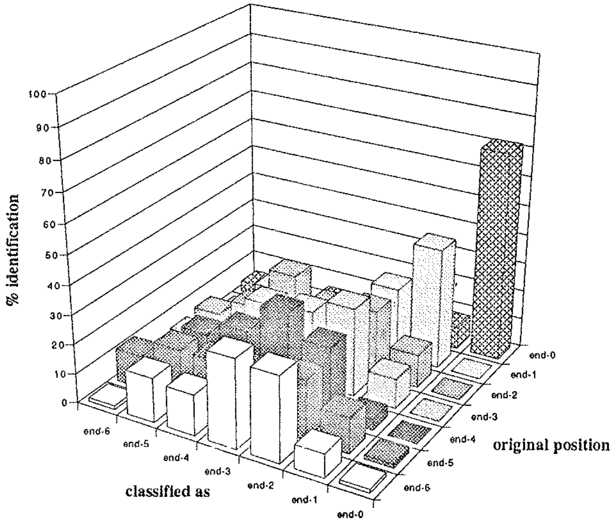


Figure 1. *3-D plot of perceptual results (averaged over the 4 route-descriptions)*

Figure 1 brings to light that in this speaker's route descriptions listeners only begin to hear that the end is approaching at one clause from the final one. As can be seen in the different bar chart diagrams, the answers to the clauses -6 to -2 are similarly distributed around the middle of the scale. The picture for clause -1 is different in that there is a clear preference of the listeners to label it - correctly - as the prefinal clause. Finally, the last clause of the total route description can most easily be classified correctly, which is reflected in the large amount of responses for the -0 class.

Acoustic measurements

Casual listening by two prosody researchers suggested that the most prominent cues were the pitch and duration of the last word of a clause. Therefore, fundamental frequency (Fo) was determined at the end of each clause and the relative duration of the last word of each clause was measured and compared to the average length of the same word read aloud four times in isolation by the same speaker. Average values were computed for each clause position over four descriptions (see figure 2).

It can be noticed that on the whole there is a close correspondence between the results of the perceptual test (see figure 1) and the acoustic variables measured. An ANOVA reveals that clauses -6 to -2 do not differ with respect to the two acoustic features (except for the difference in end frequency between -5 and -4), whereas the two other clauses, -1 and -0, differ from any of the others as to these two features, again with one exception: the difference in end frequency between -4 and -1 is not significant. From the prefinal clause onwards, a clear, though non-significant, tendency can be observed: end frequency gets lowered stepwise and the last word becomes increasingly shorter.

EXPERIMENT II

This experiment was set up to independently test the cue validity of both end frequency and relative duration of the last word in an utterance, keeping other (prosodic) variables constant, in order to explore whether these parameters are sufficient to influence subjects' finality judgments.

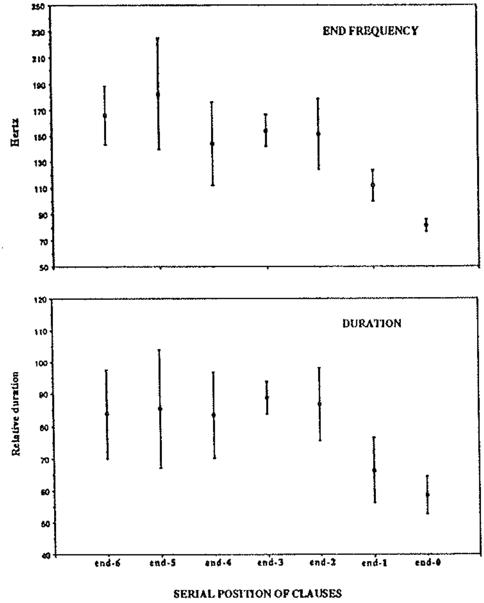


Figure 2. *Results of acoustic measurements (explanations in text)*

Speech materials, prosodic variables

The Dutch sentence "en dan gaan we rechtsaf" (and then we turn right), with accent on the syllable "rechts-", uttered by a male speaker at a normal speaking rate, was recorded with a 10 kHz sampling frequency at 12 bits. Melody and duration were varied by means of a wave-form manipulation technique (Charpentier and Moulines 1989), resulting in nine different prosodic patterns, as shown in figure 3.

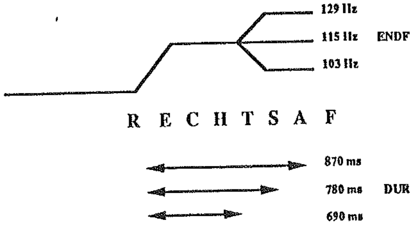


Figure 3. *Graphical representation of the stimuli used*

The different melodic contours that the utterance was provided with were identical from the beginning of the utterance up to the pitch accent in the syllable "rechts-". From that point, pitch could either rise on the syllable "-af" (and end in 129 Hz), fall (and end in 103 Hz) or remain at the higher declination line (and end in 115 Hz). The duration of the word "rechtsaf" was either original (i.e. 870 ms), or reduced to 90% (780 ms) or 80% (690 ms), respectively.

Procedure

Two lists of randomized utterances were created, containing four repetitions of each utterance. Both lists were played to seven listeners each (all 14 subjects selected from students and staff of IPO). The interval between two successive utterances was 4 s, in which time period subjects had to respond. After having heard a particular sentence, listeners had to say how near to the end of a route description the sentence was uttered. They had to express their judgment on a ten-point scale with 10 meaning 'very close to the end' and 1 'far from the end'. The actual experiment was preceded by a random presentation of five of the utterances to give the listeners an impression of the stimuli.

The mean ratings, averaged over the responses of 14 listeners, are given in table 1, together with their standard deviations.

Table 1. *Results of perception test (explanations in text)*

Dur ╲ Endf	103 Hz	115 Hz	129 Hz
690 ms	7.32 (2.11)	6.67 (2.13)	5.39 (2.47)
780 ms	7.02 (2.49)	6.13 (2.18)	4.89 (2.51)
870 ms	6.30 (2.70)	5.48 (2.41)	4.32 (2.37)

Table 1 indicates that both prosodic variables had an effect on the listeners' perception of finality: a given utterance sounds more final the lower its end frequency (though all the boundary tones are not-low in the terminology of e.g. Brown et al. (1980)), and the faster the last word of the utterance is spoken.

DISCUSSION AND CONCLUSION

Summarizing the results of these experiments: speech melody, i.e. end frequency, and length, i.e. relative duration of the last word of an utterance, were shown to give information about the serial position of an utterance within a discourse unit. More specifically, in the monologues analyzed, intonational and durational properties distinguish between three classes of utterances: final, prefinal and non-final ones. The picture may be different for other speakers or for other speech genres, for instance dialogues. In the latter, prosodic predictors of finality may also play a role to indicate the ending of a turn: this phenomenon may then partly explain the often fluent transitions between speakers' turns, i.e. without much overlap nor delay (Levinson 1983).

ACKNOWLEDGEMENTS

The author is also affiliated with the University of Antwerp (UIA). The research was funded in part by the Belgian National Science Foundation (NFWO). René Collier, Ronald Geluykens, Angelien Sanderman, Margreet Sanders and Jacques Terken are thanked for stimulating discussions.

REFERENCES

G. Brown, K. Currie and J. Kenworthy (1980), *Questions of intonation* (Croom Helm, London).

F. Charpentier and E. Moulines (1989), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Proceedings EUROSPEECH '89, 2*, pp. 13-19.

F. Grosjean (1983), "How long is the sentence? Prediction and prosody in the on-line processing of language", *Linguistics*, Vol. 21, pp. 501-529.

S.C. Levinson (1983), *Pragmatics*, (CUP, Cambridge).

M. Swerts, R. Geluykens and J.M.B. Terken (1992), "Prosodic correlates of discourse units in spontaneous speech", *Proceedings ICSLP '92, Banff, Canada*, pp. 421-424.

Intonation and Prosodic Coherence in Greek Discourse

Antonis Botinis
Phonetics Laboratory, Dept. of Linguistics
Philosophical Faculty, University of Athens
GR-15784 Athens, Greece

ABSTRACT

The present contribution is part of a larger investigation on tonal structuring in spontaneous Greek discourse. In this paper, we present our observations on prosodic coherence and the role of intonation in signalling different types of prosodic boundaries. Our analysis is concentrated on tonal sequences that may be used at intonation unit boundaries. The results support the view that various combinations of prosodic parameters are used for discourse segmentation, but they also bring into light evidence that different tonal manifestations at prosodic boundaries may imply different prosodic coherence and discourse interpretation.

INTRODUCTION

This paper reports our latest results on the use of intonation for prosodic coherence in spontaneous Greek discourse. Our current research on Greek prosody is in the framework of an ongoing project on speech technology (LOGOS, 1991-93) in which the University of Athens (Phonetics Laboratory) participates with three additional partners: the Institute for Language and Speech Processing (Athens), Athens Polytechnic (Dept. of Computer Science), and Knowledge S.A. (Patras). Research on Greek (e.g. Botinis 1989, 1992) as well as investigations on different languages (e.g. Bruce et al. 1991 for Swedish) refer to alternative strategies a speaker may use for prosodic segmentation. In Greek, a regular tonal pattern at prosodic boundaries is an abrupt pitch-change, from a high pitch-level associated with continuative accent of the preceding prosodic unit to a low pitch-level for the following prosodic unit. However, in this paper we report our latest observations, with reference to another prosodic pattern, i.e. a complex rise-fall pitch pattern realised as one pitch gesture; the pitch-rise is associated with the preceding prosodic unit, whereas the pitch-fall is associated with the following prosodic unit. We assume that the different tonal sequences at prosodic boundaries represent different types of prosodic coherence and discourse interpretation.

SPEECH ANALYSIS

The speech material presented in the present article consists of small speech units from a spontaneous dialogue between a programme leader and a programme participant (both males) that exemplify different patterns of intonation sequences associated with different types of prosodic coherence. The dialogue was recorded on a consumer-quality cassette recorder and analysed on the CSL of Kay Elemetrics Corp. at a sampling rate of 10 KHz at the Phonetics Laboratory, University of Athens. Our analysis is organised into four complementary stages: (1) text-discourse analysis in terms of syntactic-units, turn-units, and discourse structure; (2) auditory-phonetic analysis in terms of auditory-prosodic units; (3) acoustic-phonetic analysis in terms of acoustic-prosodic units; and (4) perceptual analysis and analysis-by-synthesis. In the present paper we shall confine ourselves to the acoustic-phonetic analysis, concentrating to specific prosodic units and their discourse interpretation.

TONAL SEQUENCES IN PROSODIC BOUNDARIES

Discourse prosody has taught us that a speaker may segment his turn-unit in unpredictable ways and, quite often, at great variance with syntax. At the present, we do not have sufficient knowledge "... as to what makes a speaker segment his speech in one way rather than another and, furthermore, what are the consequences of segmentation for communicative meaning?" (Botinis 1992, p. 49). Our current research on prosodic units on spontaneous speech has revealed a new pattern at intonation unit boundaries and, at the same time, has thrown light on recent observations about their interpretation. We shall thus present our material in four successive steps that represent our accumulated work on discourse prosody.

First, Figure 1 represents the programme participant's pitch-contour of the speech-unit /'ezisan ke meγa'lurγisan/ '(they) lived and grew bigger' which is within a larger turn-unit. This speech-unit shows two pitch-gestures each beginning at the stressed syllable of the corresponding stress group. Our present interpretation of this pitch-pattern is that the speaker wants the listener to consider this speech-unit as two relatively independent communicative entities and, for that purpose, the speaker assigns one pitch-gesture to each stress group. On the other hand, these communicative entities, in this example two lexical items, are closely related since they have the same pitch pattern and not any higher pitch/prosodic disjuncture but the two pitch-gestures.

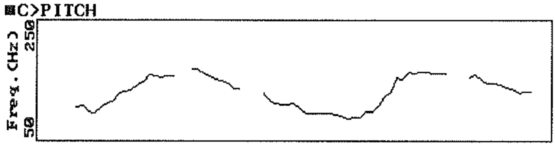


Fig. 1.'e z i s a n k e m e γ a 'l u r γ i s a n
'(They) lived and grew bigger'

Second, Figure 2 shows the programme participant's pitch-contour of the speech-unit /kata δi'afores epo'xes egata'staθikan sti makeδo'nia/ 'in different times (they) settled down in Macedonia' which is also within a larger turn-unit. This speech-unit appears with hardly any pitch-change (but microprosodic interferences) except for the boundary of the speech-unit which is associated with a major pitch-up. In our earlier analysis (Botinis 1989, 1991) we have referred to this pitch-gesture as "continuative accent" that is realised on the final syllable(s) of the boundary and may have its turning point either at a stressed or an unstressed syllable; it has been attributed to a turn-keeping function with forward directionality. Our present position on this pitch pattern is that the speaker intends to convey the information of this speech-unit as a single unit and not as several; hence, the flattening of the tonal structure up to the major pitch-up at the boundary. This flattening denotes a high degree of prosodic coherence whereas the major pitch-up contributes to a relative strong independence of this speech-unit as well as its structuring with what it follows.

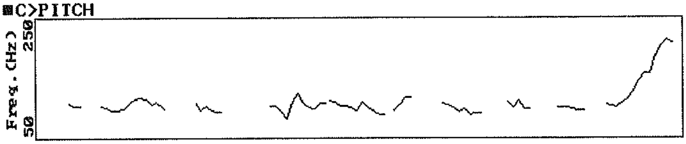


Fig. 2. kata δi'afores epo'xes egata'staθikan sti makeδo'nia
'in different times (they) settled down in Macedonia'

Third, Figure 3 represents the programme participant's pitch-contour of the speech-unit /'opos 'θaleγe ka'nis kriti'kos/ 'as one might say a Creten' which is also within a larger turn-unit. In this speech-unit the word /ka'nis/ appears with a major pitch-up which reaches its maximum at the word boundary. This pitch-gesture is not a focal accent, since the focal accent is realised as a major pitch-down at its domain of application (see Botinis 1992), but neither is it a pitch-gesture associated with a stress group (cf. Fig. 1). Furthermore, we may not consider it as a turn-keeping cue since the syntactic structure where that would be most probable is not complete at this point for this particular context. On the other hand, there is an abrupt pitch-change to a low pitch-level associated with the onset of the following word /kriti'kos/, which then takes on the same tonal pattern as /ka'nis/. We interpret this abrupt pitch-change as a prosodic disjuncture, the function of which is to assign the word /kriti'kos/ an independent information constituency.

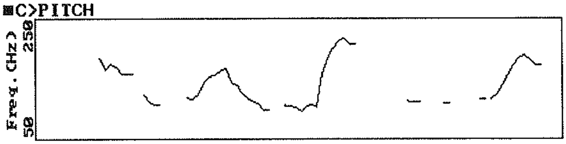


Fig. 3. 'opos 'θaleγe ka'nis kriti'kos
'as one might say a Creten'

Last, Figure 4 represents the programme participant's pitch-contour of the speech-unit /me sxo'lia me ekli'sies me 'xrimata/ 'with schools, with churches, with money' which is also within a larger turn-unit. We observe two major pitch-ups associated with the boundaries of the prepositional phrases /me sxo'lia/ and /me ekli'sies/, which have similar tonal patterns similar to those of Figure 3, rather than the stress groups of the coresponding words (cf. Fig. 1). On the other hand, there is no abrubt pitch-change to a low level associated with the onset of the following material, either with /me ekli'sies/ or /me 'xrimata/, the kind of which is represented at Figure 3. Instead, we observe a pitch-down which forms a single pitch-gesture in combination with the pitch-up of the preceding prosodic unit. We will call this "phrasal cross-over". Thus the series of pitch patterns is not repeated at the final boundary of the last prepositional phrase /me 'xrimata/, since this phrase forms the last paratactic information item. We interpret phrasal cross-over as an additive prosodic disjuncture whose function is to give the relevant material an independent information constituency while at the same time adding it to a larger information unit.

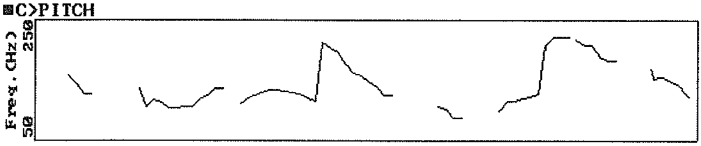


Fig. 4. m e sxo'lia m e ekli'sies me 'xrimata
'with schools, with churches, with money'

CONCLUSIONS

Our latest research partly presented in the present paper has opened new ways to look into discourse prosodic units. A major question which has puzzled us over the last years has been the range of freedon a speaker has to segment his speech, ranging from a single word to a whole turn-unit. We think that the speaker segments his speech into prosodic units in order to apply to them a (relatively) independent information structure, and this determines the options he may use.

On the other hand, the different prosodic strategies a speaker may use to segment his speech appear to us not as a means to an end in themselves but primarely to produce different cohesion patterns for communicative purposes in accordance with the information structure he wants to convey. Thus, a basic question put by researchers on phrasing (e.g. Bruce et al. 1991) with regard to a probable hierarchy of prosodic parameters represents a first aspect of prosodic segmentation; a second, the one we have presented, is the different patterns a prosodic parameter, and particularly intonation, may take for distinctive cohesion patterns in discourse communication.

ACKNOWLEDGEMENTS

The present research is carried out in the framework of the STRIDE HELLAS project LOGOS: Development of language and Speech Industry in Greece, with financial support from the EEC and the Greek Ministry of Education. Gabriel Drachman revised my English and made content comments.

REFERENCES

A. Botinis (1989), "Discourse Intonation in Greek", *Working Papers* 35, pp. 5-25, Dept. of Linguistics and Phonetics, Lund University.

A. Botinis (1991), "Intonation Patterns in Greek Discourse", *Proc. 12th Internat. Cong. Phon. Sc., Aix-en-Provence, 19-24 August 1991,* Vol. 4, pp. 286-289.

A. Botinis (1992), "Accentual Distribution in Greek Discourse", *Travaux de l'Institut de Phonetique d'Aix,* Vol. 14, pp. 13-52.

G. Bruce, B. Granstrom, K. Gustafson and D. House (1991), "Prosodic Phrasing in Swedish" *Working Papers* 38, pp. 5-17, Dept. of Linguistics and Phonetics, Lund University.

Analysis of Global Pitch Contour Domains at Paragraph Level in Spanish Reading Text

Juan María Garrido
Departament de Filologia Espanyola
Facultat de Lletres, Edifici B
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain
e-mail: ilfh7@cc.uab.es

ABSTRACT

In this paper a study of F0 patterns at paragraph level is presented. The goals of this work are the search of a domain for the 'basic intonation unit' within the paragraph and the validation of the existence of an F0 pattern for the whole paragraph. The results suggest that it is necessary to define a special linguistic unit that covers the scope of these patterns. They also seem to indicate that there is some general tendency to a global declination in the paragraphs, specially in the bottomline, and that there exists some kind of partial reset working at the internal boundaries of the paragraph.

INTRODUCTION

Intonation[1] at paragraph level has been studied for different languages such as Danish (Thorsen, 1985) or English (Ladd, 1988) among others, but not for Spanish. This work presents the preliminary results of a study of paragraph intonation of Spanish text readings. It has been focused on two topics related to the intonation in paragraphs: a) how can paragraph F0 contours be split into smaller units, and b) whether it is possible to define a global paragraph F0 pattern.

Intonation units within the paragraph

The F0 contour of a whole paragraph can be seen as a series of one or several global F0 patterns, (henceforth **pitch contours**). Assuming that F0 contours have a general tendency to decline, **F0 resets** can be a cue to indicate the beginning of a new pitch contour. An F0 reset is found when the general F0 level of a contour is reset (placed at a higher level).

Each pitch contour has an associated domain in the paragraph, that we will call here **contour domain**. It is not clear yet, what is the nature of this domain. Some studies assume that pitch contours match syntactic units, mainly clauses or sentences. Recent studies in phonology (i.e. Selkirk 1984, Nespor & Vogel, 1986, among others) have noted the lack of an exact matching between syntactic and intonative units, and suggest a phonological unit, the 'intonational phrase' as the basic domain for pitch contours. A phonetic unit, the breath group, has also been proposed as a domain for intonational units in Spanish (Navarro, 1939). However, the results of recent studies suggest (Schuetze-Coburn *et al.*, 1991, for example) that there is not an exact matching between the position of pauses and the position of F0 resets in discourse. Breath groups would not be then always the domain for the pitch contours.

A F0 pattern for the whole paragraph

Several studies on intonation in paragraphs have reported the presence of some kind of global pattern through the different pitch contours of a paragraph. This pattern could be defined as a general 'supra-declination' line linking the onset of the initial pitch contour

[1] Intonation is considered here as the evolution of the F0 contours at a non-local level (phrases, sentences and beyond).

and the offset of the final pitch contour in the paragraph (Thorsen, 1985). Thorsen's study shows that initial and final values of contours in paragraphs are kept quite constant independently of the length of the text, and that within the paragraphs, the starting values of each contour depend on the number of sentences and the length of each one.

Finally, it has also been hypothesized that the starting value of the pitch contours is also dependent on the degree of coherence between the syntactic components before and after the reset. According to this hypothesis, the reset will be higher if it is placed between two sentences (**total reset**) than if it is placed between two coordinated clauses, for example (**partial reset**).

GOALS AND METHOD

Goals

The goals of this work are following: 1) To find out the different pitch contours in a series of paragraphs, by looking for the different F0 resets; 2) to analyze the domain of these contours, in order to define which is the type of unit (syntactic, phonological, phonetic) underlying them; 3) to study the structure of these pitch contours, and to analyse its relation with the rest of contours in the paragraph, including the hypothesis of the partial reset.

Material

A corpus of 'real' speech material has been collected, including news and comments read by professional speakers (extracted from recordings of radio and TV news bulletins) and by non professional ones (readings of journal news). The chosen paragraphs have been labeled according to its duration in number of syllables (10 categories established), the number of sentences in the paragraph (6 categories), the type of speaker (professional/non professional) and the type of read text (news/comments of news). The combination of these labeling served to classify each paragraph.

The results presented in this study have been obtained only from a small part of this corpus. The analyzed paragraphs have been extracted from news read by professional speakers. The range of categories was also limited to a maximum of 4 sentences per paragraph (category 3), and a maximum of 326 syllables (category 6). An example of each category was selected for the analysis, giving a total of 11 paragraphs analyzed, ranging from 1 to 4 sentences, and from 51 to 287 syllables per paragraph. This is a rather small amount of analyzed material, so the results should have to be considered as preliminary.

Measurements and analyses

The presence of a F0 reset was detected if these three conditions were met: a) there was a boundary that could imply a reset (pause, syntactic boundary); b) the F0 value of the last F0 maximum before the potential boundary was lower than the F0 value in the first maximum after the boundary; c) the last F0 minimum before the boundary was lower than the corresponding value at the first minimum after the boundary. Minima and maxima were defined as points in syllabic nuclei that showed lower or higher F0 values, respectively, that their immediately preceding or following syllables.

Once a pitch contour was defined in this way, F0 and time values corresponding to its first maximum (IMax), the first minimum (IMin), the final maximum (FMax) and the final minimum (FMin) were taken. Duration of the contour was also calculated. Finally, the number and position of pauses in the contour, and the syntactic type of boundary were noted as well.

Initial and final maxima and minima were used to trace top and bottom lines for each contour and for each paragraph, in the way that can be observed in Figure 1. The difference between the first and the last maximum, on the one hand, and between the first and the last minimum, on the other, were computed, as well as the slope of these lines, in Hz/sec., both for each contour and for each paragraph as a whole.

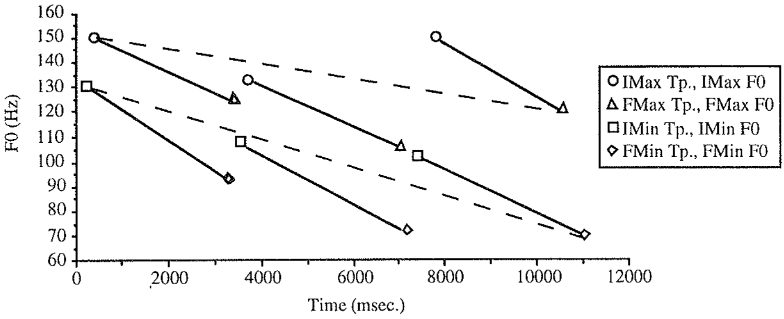


Figure 1. *Top and bottom lines calculated for the pitch contours of one of the analysed paragraphs. Solid lines correspond to the different pitch contours, and dashed lines to the whole paragraph.*

Finally, the level of reset in toplines was determined by subtracting the F0 value at FMax (last maximum before the reset) from the value at IMax (first maximum after the reset). The same procedure was used with FMin and IMin to calculate the level of reset in the bottomline.

RESULTS

The analysis of the 11 paragraphs led to its segmentation into 44 different pitch contours. The results of the analyses performed on these contours are summarized below.

Contour domains

The analysis of the relation between breath group and pitch contours led to the following results:

1) There is a clear tendency to find pauses at the end of the pitch contours (36 times, 81.81 % of the cases).

2) This does not mean that there exists a high degree of matching between breath groups and contour domains, because it is not rare to find domains containing more that one breath group (29.54 % of the cases presented 2 or more pauses in a single contour), or only a part of it (18.18 % of the contours showed no pause at all, even at the end of the contour).

The analysis of the relation between syntactic structure and contour domains showed that:

1) There is a tendency to find resets at the end of sentences (20 cases, 45.45% of the total), but more that 50% was found in other boundaries: before a verbal complement (6 times, 13.63% of the cases), at the end of a subordinate clause (5 times, 11.36%), between the Subject and the Predicate (11.36%), after parenthetical phrases (6.81%) after phrases moved to the front of the sentence (2 times, 4.54%), between two coordinate elements (4.54%) or between the verb and its object (1 time, 2.27%). These results show that there is a tendency in complex paragraphs to find the resets at the end of major syntactic boundaries (sentences or clauses), but it is also possible to find them at lower-level boundaries.

2) The analysis of the syntactic categories underlying the contours led to the conclusion that in a high number of cases (19 times, 43.18 % of the cases) it was not possible to find a syntactic structure that matched the domain of the pitch contour. When it was possible, this category was a Sentence or a Clause (17 times, 38.63%

together). There appeared other categories as NP Subject or VP Predicate (3 times each) with a lower percentage of occurrences.

Definition of a global paragraph unit

The analysis of the top and bottom lines for the whole paragraphs gave the following results:

1) There is a tendency to the declination in both the top and the bottom lines of the paragraphs (mean declining slope in both: -1.79 Hz/sec. for the topline and -3.4 Hz/sec. for the bottomline). This tendency is stronger in the bottom than in the topline, mainly due to the fact that some of the paragraphs showed a positive topline (maximum value: 1.53 Hz/sec. of rising slope). There is also less variation in the slope of the bottomline than in the slope of the topline, according to the standard deviation for each line (1.971 for the top and 1.803 for the bottom).
2) There is a high correlation (-0.805) between number of syllables in the paragraph and the slope of its bottomline: the longer the paragraph is, the steeper is the slope of its bottomline. This could indicate that initial and final values of the bottomline are more or less constant, while the slope is adapted to reach the same value at the end of the paragraph.

Finally, as far as the analysis of partial reset is concerned, there seems to be a certain degree of correlation between syntactic structure and level of reset in both the top and bottom lines. The differences in the mean values for the level of reset at sentence (40 Hz for the topline and 40.22Hz for the bottomline), clause (30.25 Hz for the top and 34.4 Hz for the bottomline) and Subject/Predicate (19.4 Hz for top and 26.8 Hz for bottom) boundaries seem to indicate that the weaker is the syntactic relation between the constituents at both sides of the reset, the higher is the level of reset at that boundary. The analysis of these means also suggests that the variation is higher in the top that in the bottom line.

DISCUSSION AND CONCLUSION

The results presented here, although preliminary, seem to indicate, on the one hand, the existence of a special prosodic unit that covers the domain of pitch contours. Position of pauses and syntactic structure seem to have an influence in the intonational parsing of paragraphs, but there is not a one-to-one relation. On the other hand, these results suggest too the existence of a global pattern for the whole paragraph, according to the findings of earlier studies (Thorsen, 1985), and a relation between syntactic structure and level of reset in the pitch contours within the paragraphs. Further research is being carried out to confirm and formalize these findings.

ACKNOWLEDGEMENT

This research is being supported by a grant from the *Universitat Autònoma de Barcelona.*

REFERENCES

D. R. Ladd (1988).- "'Declination reset' and the hierarchical organization of utterances", *Journal of the Acoustical Society of America*, Vol. 84, pp. 530-544.

T. Navarro (1939).- "El grupo fónico como unidad melódica", *Nueva Revista de Filología Hispánica*, Vol I, pp. 77-107.

M. Nespor & I. Vogel (1986).- *Prosodic Phonology* (Foris, Dordrecht).

S. Schuetze-Coburn, *et al.* (1991).- "Units of intonation in discourse: a comparison of acoustic and auditory analyses", *Language and Speech*, 34, 3: 207-34.

E.O. Selkirk (1984).- *Phonology and Syntax: The Relation between Sound and Structure* (The MIT Press, Cambridge).

N. Thorsen (1985).- "Intonation and text in Standard Danish", *Journal of the Acoustical Society of America*, Vol. 77, pp. 1205-1216.

Local and global prosodic cues to discourse organization in dialogues

Ronald Geluykens and Marc Swerts*
Institute for Perception Research (IPO)
P.O. Box 513, NL-5600 MB Eindhoven

ABSTRACT

It is experimentally investigated to what extent speakers use prosody to signal topic- and turn-boundaries non-ambiguously to a listener. Results show that local and global melodic features are employed to structure both tnformation and interaction.

INTRODUCTION

This paper reports on an investigation into prosodic cues to dialogue structure. In monologue, prosody often signals how the discourse is structured in terms of topical organization (see Swerts & Geluykens, in press). In dialogue, prosody also plays a role in the turn-taking mechanism (Sacks et al 1984). In this paper, we attempt to find out experimentally to what extent these two dimensions, information and interaction, interfere with one another, and how speakers employ their prosodic resources to regulate both dimensions in such a way that they are signalled non-ambiguously to the listener.

Research to date has generally been limited to uncontrolled, spontaneous speech (Brown et al. 1980; Schaffer 1983); to our knowledge, no work exists which tries to investigate the two above-mentioned dimensions independently in a controlled manner. Previous research has also tended to concentrate on local pitch cues (especially falling vs. rising pitch); non-local cues remain virtually uninvestigated. Since more global cues appear to play a role in monologue (Swerts & Geluykens, in press), it seems reasonable to assume that they are also relevant in dialogue discourse.

ACOUSTIC STUDY

Experimental set-up

A series of experiments was set up employing strings of differently coloured geometrical figures (see also Swerts & Collier 1992). Each time two subjects (from a total of ten) were seated in a sound-proof studio, without visual contact, and had to perform three experiments (see also Table 1).

In the first condition (C1; monologue), the speaker had to describe from left to right strings of geometrical figures (as in figure 1), in such a way that the 'topical breaks' between the individual strings became apparent for the hearer; the latter's task was to indicate the perceived breaks on an answer sheet. In condition C2 (dialogue), both subjects acted as speakers and had to produce strings without any internal breaks, and signal to the other participant when their turn was finished; the other speaker then had to take over the floor as soon as s/he felt it was possible. By means of experiments C1 and C2, we wanted to elicit both information ('topic') and interaction ('turn') signals in their purest form.

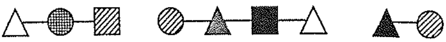


Figure 1. *Example of a series with strings of geometrical figures (different shadings actually correspond to different colours)*

In condition C3, the two tasks were combined. Speakers had to produce strings of figures (as in C1) and make the topical breaks apparent to the listener; at the same time, they had to indicate (as in C2) when their description was finished, so that the dialogue partner could take over. Listeners were instructed (i) to transcribe the topic boundaries within each speaker-turn, and (ii) to take over the floor at the appropriate turn boundary. In this way, topic-finality and turn-finality were varied to some extent independently (since topic-finality did not necessarily imply turn-finality).

Table 1. *Overview of experimental set-up for the 3 conditions (C1, C2, C3)*

	speech mode	speaker instruction	hearer instruction
C 1	monologue	signal series breaks	transcribe breaks
C 2	dialogue (simple)	signal end of turn	take over floor
C 3	dialogue (complex)	{signal series breaks {signal end of turn	trancribe breaks take over floor

Auditory analysis of pitch movements

To begin with, we analyzed auditorily the pitch contours in the different discourse locations of the elicited speech, i.e. string-internally, string-finally, and series-finally. It appeared that the contours could most easily be distinguished into those ending in High (H), Mid (M) or Low (L) level of a speaker's register. Results for the distribution of these contours are in Table 2 (representing which of the contours each of the ten speakers used in the majority of cases in the different discourse locations; see also Geluykens and Swerts 1992).

Table 2. *Pitch contours (H, M, L) in various discourse positions*

end of string		-	+	+			-	+			-	+	+
end of series		-	-	+			-	+			-	-	+
C1	M	10	0	0	C2	M	10	0	C3	M	10	0	0
	H	0	5	0		H	0	2		H	0	9	0
	L	0	5	10		L	0	8		L	0	1	10

Table 2 reveals that, in the three conditions, the end of a series is always marked by means of a Low-contour (in C2 and C3 end of series coinciding with a shift in speaking turn) and descriptions of string-internal figures are always provided with an Mid-contour. However, if we look at the string-final (but not series-final) contours, it appears that there is an even distribution of High or Low in C1, but a clear preference for High in C3. Apparently, in the latter condition, a speaker knows that he risks to be interrupted by his partner if he uses a Low.

Table 3. *Mean end-frequencies of different pitch contours [+Sd]*

	monologue	dialogue (simple)	dialogue (complex)
M [total]	+0.08	+0.11	+0.05
M (string ends in H]	-1.71 [2.91]	-2.85 [0.88]	-1.18 [2.41]
M (string ends in L]	+1.56 [2.38]	+1.38 [1.84]	+1.49 [2.04]
H	+7.40 [2.30]	+10.20 [1.17]	+7.92 [2.86]
L	-5.17 [1.56]	-5.56 [1.62]	-4.57 [1.99]

Acoustic measurements

To give acoustic support to our auditory transcriptions, we determined instrumentally the end frequency of each pitch contour. To make comparisons across speakers more easy, the average distance (in semitones) between this end frequency and the speaker's average pitch was calculated. Results are presented in Table 3.

Results show that, in all three conditions, average end-frequencies of string-internal contours (row 1) are very close to average frequencies, and very far removed both from end-frequencies of L's (row 5) and H's (row 4). This makes end-frequency a reliable indicator of discourse position. Secondly, if one compares internal end-frequencies which occur in strings ending in H (row 2) with those occurring in strings ending in L (row 3), an interesting picture emerges, in that the former have a lower average end-frequency than the latter; this is true in all three conditions. In other words, internal tones are maximally different from the end-tone while staying near the middle of the speaker's pitch range. This has important repercussions, since it would appear that the final position of a figure is, as it were, pre-signalled in the end-frequencies of the prefinal figures. Since end-pitch (fall or rise) is an important cue to discourse location, especially in condition 3, what appears to be the case is a non-local way of signalling whether the ongoing series is going to be turn-final or not. In the following section, we will investigate, among other things, whether this pre-signalling has perceptual cue value for the listener .

PERCEPTUAL EVALUATION

Experimental set-up

In order to evaluate the perceptual cue value of the acoustic characteristics (local and global) discussed in the previous section, we conducted a perception experiment using the speech produced in condition 3 (which was central to our investigation) as input.

Stimuli were prepared in the following manner. From four speakers, we selected strings of 2 up to 5 figures occurring in different discourse positions, viz. turn-initial, turn-medial, and turn-final. We then employed these as stimuli in various formats. First of all, strings were presented in their entirety, including the final pitch contour. Secondly, we chopped off parts of the utterance, starting with the description of the last figure, and continued to do this until some stimuli had only one geometrical figure (the first one) left. Ten test subjects (students and staff at IPO) were then asked to listen to all stimuli, presented in random order, and indicate on a score sheet whether they thought a particular stimulus occurred initially, medially, or finally in a series of at least three of such strings. Subjects could listen more than once to each stimulus.

Results and Discussion

Looking at the results for complete strings, presented in Table 4, it appears that listeners' scores are significantly higher than chance (X^2=217.391, p<.001). Moreover, if one examines scores in more detail, this significance appears to be mainly due to their almost perfect perception of finality versus non-finality. We have therefore reinterpreted the results, conflating the two non-final judgments into one 'non-final' category, as represented in Table 5.

Table 4. *Perceptual evaluation of discourse position within series (complete strings)*

perceived as:	initial	medial	final
initial	**47**	31	2
medial	29	**49**	2
final	4	0	**76**

The differences visualized in table 5 turn out to be highly significant (X^2=205.009, p<.001). In other words, listeners are able to use prosody as a perceptual cue for finality in dialogues. It remains to be seen, however, what it is precisely that causes this result. Do listeners respond solely to the nature of the final pitch contour, or is there something more global, such as relative end-frequencies of internal pitch movements, which they take into account? Results for the incomplete strings, which were constructed to test precisely this potential non-local cue, should give us an answer to this question. Since the relevant distinction appears to be 'final' versus 'non-final' (see above), we have once again conflated the results into these two categories (Table 6).

Table 5. *Perceptual evaluation of finality vs. non-finality (full strings)*

perceived as:	non-final	final	total
non-final	**156**	4	160
final	4	**76**	80
total	160	80	240

Table 6 shows that, even in the absence of final pitch cues, listeners are able to perceive finality of strings well above chance level (X^2=64.496, p<.001). Given the fact that most of these strings were reduced more extensively than by just deleting the final figure, with for some only the first figure remaining, this shows rather convincingly that listeners are able to make perceptual use of non-local prosodic characteristics of the speech signal (though performance does drop compared to full strings, as one would expect, the latter also containing local finality cues).

Table 6. *Perceptual evaluation of finality vs. non-finality (incomplete strings)*

perceived as:	non-final	final	total
non-final	**425**	135	560
final	135	**145**	280
total	560	280	840

CONCLUSION

This limited experimental study has yielded a number of results concerning prosody in dialogue. First of all, it was shown that speakers are able to use prosody in such a way that both the informational and interactional dimension is signalled in a non-ambiguous manner. Speakers make use of local pitch contours to do so, but also appear to pre-signal the type of final pitch movement by means of the end-frequencies of the non-final pitch movements.

In a perception test, the perceptual relevance of both local and non-local cues was investigated. It appears that both local and global prosodic features play a role in the perception of finality: even in the absence of final pitch contours, listeners are still able to predict finality to some extent. This may be due to non-final end-frequencies, but other factors, such as declination and accent distibution, may also play a role (see Geluykens & Swerts 1992).

*both authors are also affiliated with the University of Antwerp (UFSIA and UIA, respectively) and with the Belgian National Science Foundation (NFWO). René Collier is thanked for useful comments on an earlier version of this paper.

REFERENCES

Brown, G., K. Currie & J. Kenworthy (1980)*Questions of intonation.* (London: Croom Helm).

Geluykens, R. & M. Swerts (1992) Prosodic topic- and turn-finality cues. *Proceedings of the Workshop on Prosody in Natural Speech Data,* University of Pennsylvania, August 1992.

Sacks, H., E.A. Schegloff and G. Jefferson (1984) A simplest systematics for the organization of turn taking in conversation. *Language,* Vol. 50, pp. 696-735.

Schaffer, D. (1983) The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics,* Vol. 11, pp. 243-344.

Swerts, M. & R. Collier (1992) On the controlled elicitation of spontaneous speech. *Speech Communication,* Vol. 11, pp. 463-468.

Swerts, M. & R. Geluykens (in press) The prosody of information units in spontaneous monologue. To appear in *Phonetica.*

Why Sentence Modality in Spontaneous Speech is More Difficult to Classify and why this Fact is not too bad for Prosody

A. Batliner[1], C. Weiand[1], A. Kießling[2], E. Nöth[2]
[1] L.M.-Universität München, Institut für Deutsche Philologie, Schellingstr. 3, 80799 München, F.R. of Germany
[2] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), Martensstr. 3, 91058 Erlangen, F.R. of Germany

"You crazy," said Max. It was either a statement or a question.
(John le Carré: Tinker Tailor Soldier Spy)

"So you're our man, then," he said. It was half statement, half question.
(Josef Skvorecky: The Engineer of Human Souls)

ABSTRACT
We show in this paper that the labeling of sentence modality in German, esp. of questions vs. non-questions, is more difficult for spontaneous than for read speech and easier for non-elliptic than for elliptic utterances. However, the prosodic marking *of sentence modality is more important in elliptic utterances that occur more often in spontaneous speech.*

INTRODUCTION
Until now, most research has been done on controlled, read speech (i.e., non-spontaneous speech, henceforth NSP), and so far, little work has been reported on spontaneous speech (SP) in German. In an experimental design for the recording of NSP, sentence modality, e.g. question/non-question (Q and NQ respectively), can be controlled beforehand via the careful construction of the linguistic context, explicit instructions or simply via punctuation marks. In SP, however, sentence modality has to be determined afterwards, using different criteria - syntactic, semantic, contextual, or prosodic; the corresponding cues are not always present, especially because SP often contains elliptic utterances. In this paper, we will concentrate on the marking of the Q/NQ dichotomy in SP and NSP as well as in elliptic and non-elliptic utterances (ELs and NELs respectively). Related work and comparable results for English are reported e.g. in [4].

MATERIAL AND EXPERIMENTAL DESIGN
Two pairs of speakers (3 female, 1 male) who didn't know that they were recorded for prosodic research had to solve different problems in a "blocks world". The experiment was designed in a way that resulted in absolutely SP (short clarification dialogs with many turn takings). The utterances were transliterated and classified along the lines of a formal syntactic model, cf. [1]. The four cross-classified main groups were Qs vs. NQs and ELs vs. NELs. From the whole material those utterances were chosen for further investigation that met the following criteria: a sufficient signal quality and no specific non-syntactic phenomena like hesitations which are normally only found in SP. We chose all Qs, all ELs, and out of the NQs all non-statements that met the criteria, and roughly the same number of NEL statements. After 9 months, the same 4 speakers read the chosen utterances - their own utterances and those of the partner, given in written form and embedded in a sufficiently large context. Recording conditions were comparable to a quiet office environment. The 1329 utterances (approx. 30 minutes of speech, 1/3 SP, 2/3 NSP) were digitized with 12 Bit and 10 kHz. The number of the four main sentence types is the following (in parenthesis, NELs/ELs): Qs: 566 (332/234), **statements**: 623 (266/357), **commands**: 128 (108/20), **exclamations**: 12 (9/3); i.e. NQs in total: 763 (383/380). Using three different F0 algorithms, a F0 contour was computed and corrected manually to obtain a reference contour. From the corrected F0 contour the following features were

extracted: Onset, offset, maximum, minimum, range, mean, standard deviation, and regression coefficient. These features were normalized with respect to the average F0 value of the utterance. A perception experiment was performed where 10 naive listeners had to classify each utterance as Q or NQ. For more details, cf. [2] and [3].

CLASSIFICATION OF NQs VS. Qs

The classification problem was already mentioned in the introduction. We assume that for ELs, the prosodic marking is more important than for NELs, because other features such as e.g. word order are missing. This assumption is reasonable but as far as we can see it has up to now not been verified for German. It would, however, almost be a sort of "self-fulfilling prophecy" if the object of investigation (prosodic marking) is used as crucial criterion for the classification. There is no simple way out of this "classification paradox". We decided therefore to use three different classification procedures:

1. **Linguistic classification**, where the sentences were classified according to a formal syntactic model by an expert who listened to the utterances as well (formal classification without contextual knowledge).
2. **Perceptual classification**, where a group of naive listeners had to determine the sentence modality of the utterances presented in isolation ("out of the blue"-sentences).
3. **Context classification**, where the sentences were classified by another expert with the help of contextual features (content criteria and dialog structure, e.g., what does the speaker know, what is the reaction of the listener, etc.) and with the help of syntactic features, but without listening to the utterances, i.e. without prosodic knowledge (functional classification).

The context classification was conducted for the SP part of the material; their NSP counterparts could be grouped automatically into the same class because they were embedded into the same context. We established four classes, NQs and three Q classes:

1. **NQs:** All utterances that are not followed by an answer, a confirmation, etc.; it is obvious that the speaker is in possession of the information at stake but not the partner.
2. **possible Qs (Q*poss*):** Utterances followed by an answer; the context shows that both speaker and partner are in possession of the information at stake. The context and/or lexical information (e.g. modal particles) give no clues whether the speaker is confident about that what he/she says or not. Quite often the speaker is simply paraphrasing something the partner has said just shortly before.
3. **probable Qs (Q*prob*):** Utterances followed by an answer, but not clear-cut Qs; the context shows that, in contrast to Q*poss*, the speaker obviously does not know whether he is right or wrong, but the partner does. Often, the speaker uses a modifying particle, e.g. *vielleicht* (*perhaps*).
4. **Qs:** clear-cut questions, i.e. utterances followed by an answer, etc., mostly with an agreement of contextual and grammatical criteria (e.g. WH-questions). It is obvious from the context that the information needed by the speaker is in possession of the partner but not of the speaker.

The following example can illustrate both Q*poss* and Q*prob*: speaker: *"The green block is on the red one."* – partner: *"Yes, that is right."*. Depending on the different contextual information, cf. above, the first sentence is assigned either to Q*poss* or to Q*prob*. With only syntactic information, the first sentence had to be classified as a clear-cut statement. The reaction of the partner makes it possible that the first utterance could be a declarative Q. Without prosodic and/or contextual information, the conflict cannot be solved, because almost any statement can be followed by a confirmation or by a negation.

RESULTS AND DISCUSSION

As for the context classification, a systematic difference between ELs and NELs can be seen in figure 1 for SP. Note that the classification for the NSP counterparts is identical, cf. above: in the clear-cut categories NQs and esp. in Qs, there are more NELs than ELs. It is the other way round in the two other categories (approx. 25% of the cases); i.e. ELs are really less clear-cut than NELs.

In figure 2 and 3, the height of the F0 offset in semitones (st) subtracted by the F0 mean of the utterance as the most stable prosodic feature indicating the Q/NQ-dichotomy, is plotted for the four context categories. For NSP (figure 3), there is almost a linear relationship between offset and Q-proneness: the more Q-prone, the higher the offset. There is, however, no difference in SP between ELs and NELs for NQs; for Qs in SP (figure 2), the offset is markedly higher in ELs than in NELs.

In figure 4, the perception results are compared with the context classification; as almost no difference could be noticed between SP and NSP, they are plotted together. The ordinate shows the frequency of the cases, the abscissa perceived NQs and Qs for the four context classes. A perceived NQ is defined if less than five out of the ten listeners classified an utterance as Q; the other cases are classified as Q. In approx. 5% of the cases, cf. the small bars for NQ and Q, there is disagreement between context and perceptual classification due to an inherent difficulty in the context classification and/or an equivocal prosodic marking of the utterances; for details, cf. [3].

Figure 5 and 6 put the F0 offset in relation to the perception experiment. The abscissa shows the number of listeners that categorized an utterance as Q, the ordinate shows – analogously to figure 2 and 3 – the average of the height of the F0 offset in semitones (st) in relation to the F0 mean of the utterance. There were not many scores in the region between 2 and 8 and extreme values would have a distorting influence on the mean of the offset. This region is therefore combined and projected onto the value 5. For ELs, there is a linear relationship between F0 offset and Q-score: the higher the offset, the more listeners classified the utterances as Qs. The linearity is more pronounced for NSPs (figure 6) than for SPs (figure 5), and for SPs, the offset is markedly higher in the rightmost region, i.e. for Qs. For NELs, this relationship is much less clear. Obviously, Q-proneness is marked

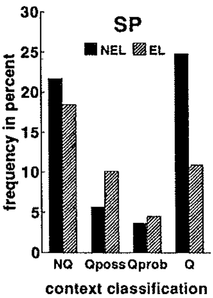


Figure 1

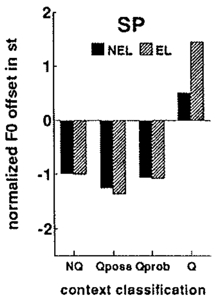


Figure 2

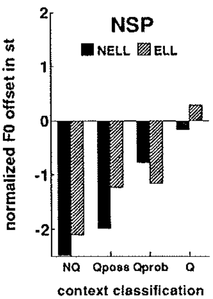


Figure 3

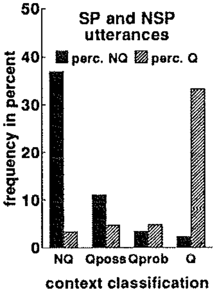


Figure 4

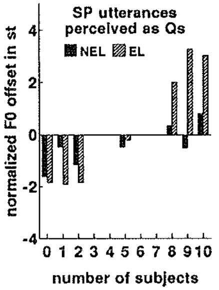


Figure 5

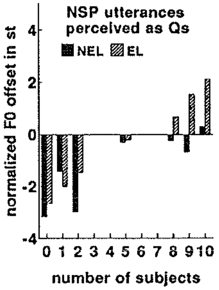


Figure 6

much more with prosodic means in ELs than in NELs.

FINAL REMARKS

Coming back to the first part of the title of this paper, it is now clear why sentence modality in SP is more difficult to classify than in NSP: even if the Q/NQ-dichotomy holds for most of the utterances, one should say goodbye to a straightforward and clearcut dichotomy. In quite a number of cases (approx. 20%, cf. Q*poss* and Q*prob* in figure 1 and figure 4), contextual and prosodic features point towards a category in between Qs and NQs that is illustrated in the two quotations above: sometimes, the category can not be decided upon (le Carré, Q*poss*), sometimes, it is really just something in between (Skvorecky, Q*prob*). That holds especially for ELs. Note that ELs do occur much more often in SP than in NSP; in our material, however, both are strictly parallelized. In real life, this difference will thus show up even more clearly. There was no pronounced difference between NSP and SP, although NSP behaved more regularly. There is, however, throughout a difference between ELs and NELs: sentence modality in ELs is more often marked by prosodic means. This fact corroborates the second part of our title: as ELs do occur quite often in SP, prosody will be needed much more in automatic speech recognition – if one really wants to deal with SP.

Acknowledgements

This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research project ASL/VERBMOBIL and by the *Deutsche Forschungsgemeinschaft* (*DFG*). Only the authors are responsible for the contents of this paper.

References

[1] H. Altmann, A. Batliner, and W. Oppenrieder, editors. *Zur Intonation von Modus und Fokus im Deutschen.* Max Niemeyer Verlag, Tübingen, 1989.

[2] A. Batliner, B. Johne, A. Kießling, and E. Nöth. *Zur prosodischen Kennzeichnung von spontaner und gelesener Sprache.* In G. Görz, editor, *KONVENS 92*, Informatik aktuell, pages 29–38. Springer-Verlag, Berlin, 1992.

[3] A. Batliner, A. Kießling, and E. Nöth. *Die prosodische Markierung des Satzmodus in der Spontansprache – Methodologie und erste Ergebnisse.* Technical Report: ASL-Süd–TR–14–93/LMU, Februar 1993.

[4] N. Daly and V. Zue. *Statistical and Linguistic Analyses of F0 in Read and Spontaneous Speech.* In *Int. Conf. on Spoken Language Processing*, volume 1, pages 763–766, Banff, Canada, 1992.

Prosodic Diversity in Mandarin Natural Discourse

Li-chiung Yang
Department of Linguistics, Georgetown University
Washington, D.C. 20057 U.S.A.

ABSTRACT

Expressive rhythm and melodic structures permeate natural discourse. This study examines the prosodic system of Mandarin Chinese by acoustically and perceptually analyzing discourse data. Pitch shapes often reflect discourse intentions, underlying cognitive states, and accommodation to intonational conflict and progression. It is found that rhythm, iconicity and inversion act fundamentally to connect sound with meaning.

INTRODUCTION

Intonation ultimately depends on the sympathetic unity of emotions which are called up by specific sound characteristics. Intonation mirrors the complexity of cognitive and emotional states throughout a dialogue, and small waves of finely focused micro-emotions simultaneously exist within larger waves of more broadly focused emotions. The complex cognitive and emotional layers are reflected in a rapidly changing process of intonational progression. For example, the arched rise fall shape of dawning realization iconically represents the transformation from surprise and uncertainty to certainty and acceptance. The progression of intonation parallels the progression of emotion.

SHAPES OF INTONATION

Pitch shapes of wiggles, twists, turns and waviness are characteristic of animated and involved speech, and often reflect discourse intentions, underlying cognitive states, and accommodation to intonational conflict and progression. Wiggles, upturns and downturns also contribute to a feeling of pleasure and agreeableness. Twists in pitch shape can arise from the strain to accommodate divergent emotions. Smooth wiggles and waves may result from a more relaxed state of the vocal cords and twists from tension of the vocal cords.

Micro pitch turns are also used to resolve the conflict between intonation and lexical tone. Counterbalancing micro-turns at the end of syllables can also enhance rhythm. A tiny pitch reversal at the end of a syllable often indicates either an emotional state whose intonational shape is opposite to the predominant pitch shape of the lexical tone, or a tiny indication of lexical tone attached to an opposite intonational shape of the main part of the syllable. The pitch shape of a rising second tone syllable in an environment of rejection or negativity often takes on a falling shape with a slight pitch upturn at the syllable end. The extreme sensitivity of the ear and brain in obtaining critical information from these micro-turns emphasizes the importance of even slight pitch movements in intonation.

DISCOURSE INTONATIONAL RHYTHM

Rhythm is established by the repetition of specific aspects of sound or sound patterning at regular time intervals. Rhythmic repetition can be inexact, as in a natural progression of vowel sounds from open to closed, or vice versa. Rhythm can also be evoked by the timewise balance of natural sound inverses, such as high vs low, strong vs weak, sharp vs gradual, and quick vs slow.

Poetic elements are pervasive in natural discourse and are manifested through repetition, sound harmony and melodic pattern variations. Reduplication can be seen in the example "wo3 kan4 zhao3 zhao3 kan4" meaning "I'll look to see". There is a great contrast in the pitch rise of the first *kan* and the pitch fall of the second *kan*. The first *kan* is self-questioning and tentative, and so rises. By the second *kan*, the speaker has already decided on a course of action, is more definite and certain, and so the pitch falls. The second *zhao3* of the reduplicated form *zhao3 zhao3* psychologically completes the reduplication and so has a short fall. Such intonational couplets of rising and falling form often accompany proposition and completion pragmatic structures. This association of fall and low with completion and concluding may also explain why unit final third tones are often realized as a simple fall and may partially explain why tone sandhi (33-->23) occurs in speech.

Reduplication makes things more casual and lighter. Reduplication, repetition and rhythmic speech often serve to increase familiarity and intimacy. This may be because by using rhythm, the speaker adds a pleasant sound sensation to the ear beyond the purely semantic meaning. Another possible reason for reduplication is that it allows one of the syllables to take on greater intonational variation. *Intonational* repetition occurs frequently in speech. In the example of Figure 1 "London hen3 re4, jiu3 shi2 du4" [London's very hot, ninety degrees], the striking and exaggerated intonational pitch pattern of *London hen re* is repeated in the phrase *jiu shi du*. The reasons for such patterns of repetition are that an expression of emotion can be emphasized or prolonged by repetition of the physiological speech sensation associated with an initial emotion. Time spent in sensory perception may also increase the cognitive impression.

Rhythm connects. Through assignment of equal or hierarchical acoustic forms, syllables, words, or phrases achieve equal or hierarchical status. In this way, rhythm can be used for organizational effect, and a hierarchy of grammar and meaning can be achieved. For example, a frequent phrase or sentence pattern found in my data is to first establish a specific rhythm, and then to conclude by breaking the rhythm, sometimes by using an inverted rhythmic form. In the example [you4 zhun3] [wai4 guo2] [xue2 sheng1] [shu3 jia4] [han2 jia4] [ke3yi3 da3 gong1], "allow foreign students to work during summer and winter vacations", syllable couplets were formed by rhythmic timing through *han jia*, then the rhythm was broken to conclude by the timewise merging of *ke* and *yi* and the lengthening of *gong*. In another example the phrase "yi1 bai3 duo1 ge0 zhong1 guo2 xue2 sheng1" [over one hundred Chinese students] was said with a H-L H-L H-L L-L pitch level sequence, concluding with two lows. Establishing a rhythm and then breaking it performs a cognitive or discourse function, and the break serves to signal the conclusion of the rhythmic unit. The notion of rhythm is fundamentally related to cognitive signal-processing efficiency. Our cognition and senses may be more attuned to signals of a periodic nature.

Rhythm is enhanced by symmetry. In "kou3 shi4 ye3 hai2 mei2 kao3, lun4 wen2 hai2 mei2 xie3" [Oral exam I still haven't taken, dissertation I still haven't written], the phrases *kou shi ye hai mei kao* and *lun wen hai mei xie* are rhythmically paired by the equal amount of time spent on each phrase and their parallel syntactic construction. The 2nd half of the phrase is slightly lower in pitch and amplitude for completion, but there is no downdrift. That preserves the poetic equality.

In discourse, there is often a rhythmic acceleration as one works away from the point of most emphasis. In the example "bu2 guo4 ta1 li3 mian4 you3 san1 shi2 ji3 ge0 college" [but inside they have thirty some colleges], there is strong emphasis on *san shi*. Within *san shi ji ge*, *san* carries the main focus and therefore is the longest, and the subsequent syllables are progressively shorter. The speech tempo is often iconic in that we dwell on things which are more important or pleasant, and spend less time on deemphasised points.

Transitions between rhythm are expressive. In the example of Figure 2 "qi1 nian2 nian4 bu2 nian4 de0 wan2 bu4 zhi1 dao4" [Whether or not I can finish studying in seven years, I don't know], the phrase *qi nian nian bu nian de wan* is said as a smooth descending pitch unit and then changes to a short and choppy *bu zhi dao*, with downward pitch force in each syllable. Negativity is signalled by the repetitive downward pitch falls of *bu zhi dao*. Choppy or clipped speech in general seems to iconically represent a reluctance to speak and was often associated with embarrassment, unhappiness, or complaining.

In the example "bu2 guo4 zhi4 shao3 ni3 gan3 jue2 zai4 mei3 guo2 zhi4 shao3" [But at least you feel that in America, at least], an initial dramatic disagreement is signalled by the very high pitch and strong amplitude on *zhi shao*. By repeating and normalizing the same phrase *zhi shao* at the end, with lower amplitude and pitch, the speaker is making a more convincing case. The phrase *zai mei guo* sounds like intonational repetition of *ni gan jue*, and the particular repeated pitch pattern of the fall rise is very expressive of "trying to convince" in this example.

ICONICITY AND INVERSION

The notion of iconicity is one which can unify many aspects of intonation. There are different levels of iconicity, and sounds reflect iconicity to different degrees. The data reveal that regular associations between sounds and other non-sound entities exist. Iconicity at one level is connected with the body, providing speech examples such as throat constrictions and high pitch levels. Beyond that, speech sounds are associated to non-sound entities in a metaphoric way or by parallelism, as in sound symbolism. Iconic elements often seem to reflect the cognitive nature of the mind and the perceptual demands of physiology, e.g. in repetition or in lengthening of duration for importance, and iconic elements are related to emotion: we dwell on things we like, we shorten and disregard things we don't like.

Iconicity leads directly to the *principle of inversion*. If speech is iconic, if sounds iconically represent mental and physical states, then when one wants to express an opposite emotion or state, one uses the opposite sound. Sound opposites include high vs low pitch, strong vs weak amplitude, quick vs slow tempo, long vs short duration, smooth vs non-smooth rhythm, rising vs falling pitch, etc. In discourse, the flow of changing conditions and states may often call for a need to overcome and reverse the preceding intonational effect and may result in intonational flip-flop, that is, an inverse intonation may be used just to counteract the just preceding unit.

Why should inversion be so important in speech? In natural speech, discourse structures of climax and resolution or intensification and normalization appear constantly, both on a small scale of a few syllables and on a large scale across phrases or sentences, as in stories or episodic narrative. Patterns of disequilibrium and equilibrium are therefore inherent in communication. Inversion then acts as the metaphoric correlate of these forces. It is iconicity which binds together the physiological and cognitive structures which govern prosody.

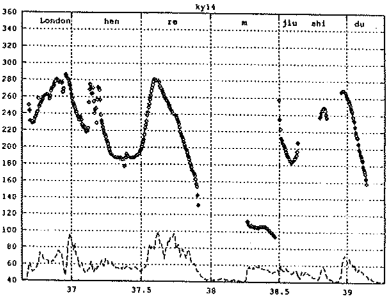


Figure 1.

London hen3 re4 jiu3 shi2 du4
[London's very hot, ninety degrees]

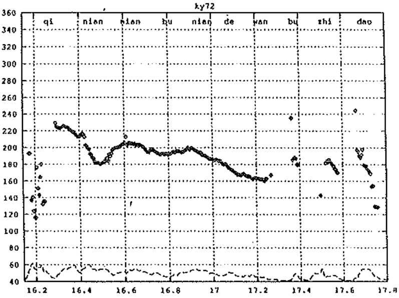


Figure 2.

qi1 nian2 nian4 bu2 nian4 de0 wan2

bu4 zhi1 dao4

ACKNOWLEDGEMENTS

This research was conducted at the Phonetics Lab of Stanford University. I am grateful to Bill Poser for his generous support, to Tom Veatch for help with the WAVES program, and to Stanford Linguistics Department for supporting this research.

REFERENCES

D. Bolinger (1989), *Intonation and It Uses* (Stanford University Press).

Y.-R. Chao (1968), *A Grammar of Spoken Chinese*. (University of California Press, Berkeley).

I. Fonagy (1971), "Double coding in speech", *Semiotica* 3:189-222.

L.-C. Yang (1992), "A semantic and pragmatic analysis of tone and intonation in Mandarin Chinese", in *ICSLP 92 Proceedings* (University of Alberta Press), pp. 655-658.

Rhythmic patterns and lexical parsing in French

Nicole Bacri and Marie-Hélène Banel
Laboratoire de Psychologie Expérimentale, URA 316 CNRS-
Université René Descartes, 28 rue Serpente, 75006 Paris, France

ABSTRACT

Two experiments examined the effects of usual and reverse rhythmic patterns on two-syllable sequence segmentation. The usual short-long pattern impeded parsing of monosyllabic words embedded in disyllabic words. In Experiment 2, focusing subjects'attention on the timing structure strengthened this effect. Implications for speech recognition models are discussed.

INTRODUCTION

In a language with fixed stress placement, as French, in which stress usually falls on the last syllable in polysyllabic words and is therefore predictable, stress pattern may be used for lexical parsing more readily than in a language with variable stress placement as English: Indeed, prosodic patterning of an utterance could facilitate speech segmentation into lexical units. A non-emphatic rhythmic group in French is best characterized by a sizeable lengthening of its final syllable as compared with non-final syllables. Although there exists a tendency to stress word-initial syllable in certain speaking styles (e.g., radio broadcasts), short-long (iambic) structure appears reliably as the basic rhythmic structure in French (Fletcher 1991, Vaissière 1991). The role of this pattern in speech perception could be tested by examining whether inversion of usual disyllabic word stress information affects or not lexical parsing.

Previous research has shown that, in French, the syllable is the basic unit of segmentation and is used in early lexical processing (Cutler, Mehler, Norris and Segui 1987). In this line of research, structural parameters, as the number of syllables, appear to affect lexical access more than durational parameters, which permit contrasting stressed and unstressed syllable (Dupoux and Mehler 1990). In this view, stress pattern, either normal or reverse, should have no effect on lexical parsing: Rhythm should be processed post-lexically.

Nevertheless, in English, strong syllables trigger segmentation and initiate lexical access, even when in non-initial position (Cutler and Norris 1988). Rhythm may be used to improve perceptual processing (Pitt and Samuel 1990). These results suggest that syllabic-timing differences that characterize a rhythmic group at the word level could play a role in French. Since French is a trailer-timed language (Wenk and Wioland 1982), final lengthening in a two-syllable sequence should induce listeners to merge the first short syllable with the second long syllable and derive a coherent lexical unit. The aim of the present study is to test whether rhythmic expectancies play a role in segmenting two-syllable sequences into one or two words depending on their stress pattern.

If such is the case, the prosodic pattern is an important cue to word identity: A short-long (iambic) pattern would facilitate one-word perception, a long-short (trochaic) pattern would facilitate two-word perception. It may be the case, however, that syllable lengthening *per se* facilitates speech processing: According to this view, longer syllables might help access in monosyllabic words, whatever their within-word position. On the other hand, a "structural" model, according to which prosodic information is processed post-lexically, would predict that monosyllabic words are accessed faster than disyllabic words and all the faster as they are more frequent, whatever their rhythmic or durational patterning.

EXPERIMENT 1

The role of rhythmic pattern in speech segmentation has been tested through a comparison between the effects of the usual iambic pattern and of the reverse trochaic pattern on parsing judgements and response times. Within each rhythmic condition, embedded monosyllabic word frequency was varied in order to appreciate the relations between this lexical parameter and rhythm processing.

Method

Test stimuli were 48 two-syllable sequences, each disyllabic word embedding two monosyllabic words ("marmotte": "mare", "motte"). Sequences were built by concatenating monosyllabic items recorded in isolation by a male speaker. Hence, the two-syllable stimuli contained no inter-syllabic coarticulation cue. Frequency of disyllables was medium-low, frequency of monosyllables was either very high or low. For each combination of monosyllable frequency (HH, HL, LH, LL), two rhythmic patterns were realized, short-long and long-short. Duration differencé between both versions of a syllable was about 35%. Intensity was equalized and F0 was held as flat as possible. Each of the four combinations of syllabic structure, CV or CVC, was equally represented for each frequency range. 96 fillers were added. Half of the fillers were disyllabic words without embedded words, and the remaining half were sequences of two monosyllabic words that did not constitute a disyllable.

Eight subjects were presented stimuli and fillers (two counterbalanced orders), one item every 3 seconds. They were instructed to determine as quickly as possible whether they heard just one disyllabic word or more words. Response times (RT) were measured from the sequence acoustic offset.

Results and discussion

Iambic rhythm induced more disyllable identifications than the reverse rhythm did (73% vs. 54%; $F(1, 7) = 13.5$, $p<.01$). The effect of monosyllable frequency was also significant ($F(3, 21) = 4.6$, $p<.01$) as was the Rhythm x Frequency interaction (Fig. 1, Exp. 1). In fact, frequency did not yield any difference for iambic pattern, whereas for trochaic pattern a high frequency word in final sequence position induced far more multi-word responses than a low frequency final word did (55% vs. 37%). Syllabic structure yielded no significant effect. "One-word" responses were in the average 62 ms faster than "multi-word" responses, especially for iambic patterns (92 ms faster), but these differences in RT were not statistically significant. RTs were not correlated with the duration of whichever syllable.

These results suggest that rhythm is used as a cue to segmentation: Lexical decomposition is facilitated by a long-short pattern, impeded by a short-long pattern. Processing of this last pattern is insensitive to frequency effects: In spite of the suppression of intersyllabic coarticulation cues, a high frequency monosyllabic word in initial position does not trigger lexical access more often than does a low frequency word. Likewise lengthening does not facilitate monosyllabic word identification. However, the RT data suggest that the stimulus set failed to induce strong rhythmic expectancies: RTs for iambic patterns are not faster globally than for trochaic patterns. The next experiment was designed to bias subjects'attention towards the timing structure of the test stimuli, in order to investigate the time course of rhythm processing.

EXPERIMENT 2

Method

Experiment 2 reduplicated Experiment 1, except that all the fillers presented the same structure: The two-syllable sequences were constructed by concatenating two monosyllabic words of the same length and 80% did not constitute a disyllabic word (10 subjects).

Results and discussion

An ANOVA revealed a significant effect of rhythm both on parsing judgements (Figure 2, Exp. 2) and on RTs. Iambic rhythm induced more numerous and faster one-word responses than trochaic rhythm did (73% vs. 62%, F(1, 9) = 24.3, p<.001; 786 ms vs. 911 ms, F(1, 9) = 36, p<.0005). One-word responses were reliably faster than multi-word responses (808 ms vs. 948 ms). Syllabic structure complexity facilitated multi-word identification, but had no effect on RTs. Finally, as in Experiment 1, RTs were not correlated to syllable length.

Assuming that rhythm is a cue to lexical parsing, focusing listeners'attention on timing structure had the predicted effect: Rhythm facilitates or impedes segmentation. The lack of interaction between rhythmic pattern and frequency suggests that rhythm processing is not post-lexical, but is processed in parallel with lexical access.

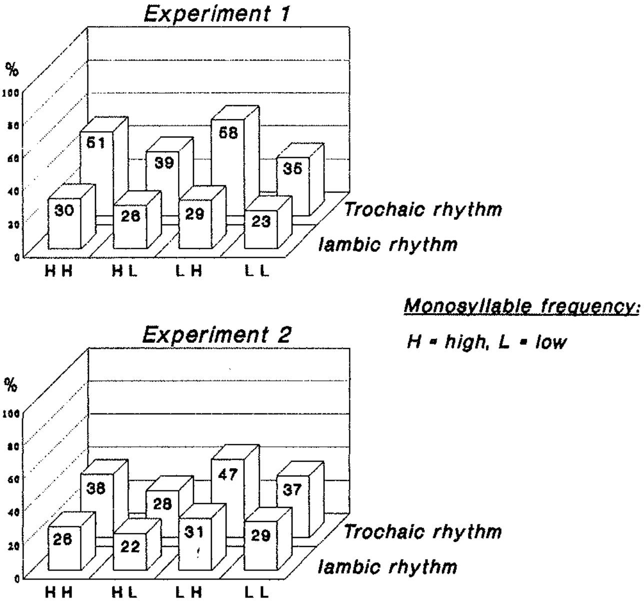


Figure 1. *Percent of multi-word responses as a function of rhythm and frequency.*

GENERAL DISCUSSION

Cutler and Clifton (1984) claimed that stress plays no role in lexical access whereas Pitt and Samuel (1990) concluded that the advantages for "expected-stress" syllables are small. In French, rhythm regularity may provide a stable structure that the processing system can use to anticipate word boundaries. However rhythmic expectancies take time to develop, and the disyllabic representation is not activated faster than the monosyllabic one for iambic patterns except when attention is attracted on timing structure. A reverse pattern deceives listeners'expectancies and improves parsing of disyllables into their embedded components. Thus rhythm appears as selectively tuning speech processing towards lexical parsing rather than affecting directly lexical access.

ACKNOWLEDGEMENTS

Financial support for this research was provided by the GDR 957 "Sciences Cognitives de Paris" (CNRS).

REFERENCES

A. Cutler and C. Clifton (1984), "The use of prosodic information in word recognition", in *Attention and Performance X*, ed. by H. Bouma and D. Bouwhuis (Erlbaum, Hillsdale), pp. 183-196.

A. Cutler, J. Mehler, D. Norris and J. Segui (1987), "Phoneme identification and the lexicon", *Cognitive Psychology*, 19, pp. 141-177.

A. Cutler and D. Norris (1988), "The role of strong syllables in segmentation for lexical access", *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 14 (1), pp.113-121.

E. Dupoux and J. Mehler (1990), "Monitoring the lexicon with normal and compressed speech: Frequency effects and the prelexical code", *Journal of Memory and Language*, 29, pp. 316-335.

J. Fletcher (1991), "Rhythm and final lengthening in French", *Journal of Phonetics*, Vol. 19 (2), pp. 193-212.

M. Pitt and A. Samuel (1990), "The use of rhythm in attending to speech", *Journal of Experimental Psychology: H.P.P.*, Vol. 16 (3), pp. 564-573.

J. Vaissière (1991), "Rhythm, accentuation and final lengthening in French", in *Music, Language, Brain and Speech*, ed. by R. Carlson, L. Nord and J. Sundberg (Macmillan Press, New-York).

B. Wenk and F. Wioland (1982), "Is French really syllable-timed?", *Journal of Phonetics*,Vol. 10, pp. 193-216.

Duration and Energy indices : a Convergent Deep Organization of Grouping and Phrasing

Geneviève Caelen-Haumont
Institut de la Communication Parlée, CNRS URA 368,
INPG et Université Stendhal,
46 avenue F. Viallet, 38031 Grenoble Cédex 1 France

ABSTRACT

The aim of this paper is to present the results of a study on duration and energy indices in a context of text readings with tasks. This study is not traditional because the data are not analyzed in the respect of the prosodic organization of all the linguistic units which compose the phrases. In fact the study is based on the deep relative organization of duration and energy values of the lexical items (internal and external relationships) at different phrase levels. On the whole the study takes into account 10 indices.

INTRODUCTION

Since a few decades, many papers have been devoted to duration organization. Whatever the perspective in phonology and in acoustic studies, most of these works took in consideration the internal organization of phrases. For instance in the perspective of phonetical studies, Bruce (1985) identifies two functions of prosody, i.e. weighting (stressed / unstressed syllables) and grouping. This paper concerns with the function of grouping in french. The new perspective which is reported here considers 1° the deep acoustical organizations of prosodic indices 2° the internal and external lexical organization of these indices in different types of phrases 3° the grouping and the phrasing functions of 7 duration indices and 3 energy indices *(Energy* and not *intensity* because this parameter is extracted from a model of ear).

EXPERIMENTATION, AIMS AND METHOD OF ANALYSIS

The study of duration and energy organization corresponds to the second part of a more important analysis concerning an assessment of the relations between prosodic indices (Fo, energy and duration), and linguistic structuration (syntax, semantic) and pragmatic conditions (phases of discourse, reading constraints...). For this purpose, a 50-word text was elaborated : *"D'éminents biologistes et d'éminents zoologistes américains ont créé pour des vers géants un nouveau phylum dans l'actuelle classification des nombreuses espèces vivantes. Ces longs vers prospèrent sur le plancher marin des zones sous-marines profondes. Des sources thermales chaudes y maintiennent une température moyenne élevée."* The experiment bore on 3 readings (3 tasks: 1° natural and intelligible reading 2° very intelligible reading 3° extremely intelligible reading, relevant for man-machine interaction) of the text by 12 speakers. A data base was elaborated from these 36 utterances, and was tagged with about 40 000 labels related to the various linguistic and prosodic analysis levels.

ENERGY AND DURATION INDICES

Definition

Duration and energy indices were converted in a four-tiered space, in order to discart surface phonetic information in the search of underlying structures. The minimal and maximal numerical values were automatically detected in the context of each discourse. Duration indices are based on all phonetic segments of the lexical words, and energy ones on their vowels. Note that the analysis being only based on lexical words, the study does not concern the group duration and the group energy, but the internal and external organizations (duration and energy) of the lexical items at phrase, sentence and text levels.

Duration indices are seven. Calculated in the entire word, they are: whole duration (WWD), mean syllabic duration (mSD), maximum syllabic duration (MSD), absolute value of syllabic duration range (|ΔD|). In the last syllable of the word (and monosyllables), the whole duration (WSD), mean phonetic duration (mPD) and the whole syllabe duration + the following pause (SDP) are found. As for energy indices, there are few of them. Only calculated in the whole word, they are the absolute value of vowel energy range (|ΔE|), the mean energy (mE), and the maximum of vowel energy (MWE).

Forms

On a general point of view, energy and duration indices are constructed on the same pattern: the values (1 to 4) are arranged in an order which is ascending for duration indices (lenghtening), and generally descending for energy indices (less loudness). Though the results were calculated in previous studies, on the whole structure of phrases, these present findings concerning the lexical duration structure of groups support these previous results (Caelen-Haumont, 1978; Pasdeloup, 1992).

This dynamic process concerns duration and energy parameters. Note that the step between two successive values may not be proportional, but the order is respected. The resetting is more or less important between the final value of the precedent group and the first one of the following group. So this organization enables us to give each group an internal structuration based on this progressive order, and an external one based on the breaks of this process. Table 1 below shows an example of this pattern issuing from the 2 first phrases of the text and 12 speakers.

For the analysis it is useful to distinguih between 3 types of phrase units : the micro-phrase (mP), the phrase of the most superficial level (P) and the macro-phrase (MP). P concerns the syntactic one which just dominates the level of words, and also pseudo-syntactic one. The pseudo-syntactic group is constitued for prosodic reasons when the syllable number of one of the two phrases is less than 5 (ex: *"ont créé pour des vers géants" ...).* MP is the combination of n phrases and mP is a part of the phrase (P) made up by two sub-units which are strongly linked in the domains of syntax and semantics: for instance the units of pseudo-syntactic phrase that we described just above, or as a second example, a noun with an adjective. This process always occurs in the case of a compound noun, but it is not the only case. It is also used when an adjective is syntactically and semantically closer from the noun than a second one. See for example table 1 below, speaker IN for DL index, and speaker BR for Em index.

Table 1. *An example of DL index and EM index coefficients coded in a four-tiered scale. Data across 6 speakers, the first 2 phrases of sentence 1, task 1:"d'éminents biologistes et d'éminents zoologistes américains ..."*

Idx	DL	DL	DL	DL	DL	DL	Em	Em	Em	Em	Em	Em
spea-kers	BR	CA	CO	FA	FO	IN	BR	CA	CO	FA	FO	IN
G1 é	2	2	2	2	2	2	3	3	4	3	2	3
b	3	3	4	3	3	3	2	2	3	2	1	2
G2 é	2	2	2	2	1	2	3	2	4	3	2	2
z	3	3	3	3	2	4	2	2	3	3	1	2
a	3	3	3	3	2	3	3	2	3	2	1	2

In order to explain the following results, we make a distinction between the notions of *lexical phrase segmentation* and *lexical phrase demarcation.* The function of the segmentation is grouping, while the function of the demarcation is phrasing.

Lexical phrase segmentation

The notion of phrase segmentation is used each time the boundaries of a series of values in a progressive order (admitting plateaux) correspond to the limits of a phrase (P) and macro-phrase (MP), whatever syntactic or not the combination of phrases is. For instance, whatever the grouping of the verb with the precedent or following phrase is, those cases were considered relevant for phrase segmentation tasks.

In those conditions the mean results, evaluated on the number of syntactic and pseudo-syntactic phrases (i.e, 396 phrases) and concerning the number of lexical phrases corresponding to these segmentation rules, is 86%. That mean is calculated over the 12 speakers, the 3 tasks, and the 3 energy indices. Most of energy indices generally overrun 90% in sentences 1 and 2, but not in sentence 3 because of the great number of micro-phrases which occur in speaker utterances, and make the results drop drastically. On the whole and even in sentence 3, |ΔE| produces the best scores (89%). As it concerns duration indices (all the indices except SDP), the mean number of syntactic and pseudo-syntactic phrases which were correctly segmented, is 83%. Among all of them, the whole duration of the word (WWD) has the best score (93%). In sentence 3, just like energy indices, the results concerning the ratio of phrases and macro-phrases drop steeply.

On the whole, these results indicate that the internal and external organizations of prosodic indices in the space of phrase lexicon is thoroughly constructed. While the targets of minimum and the maximum loudness are well positionned according to the internal space of the lexical item, to the relations of lexical items in the phrase, and to the relations of phrases at the sentence level, the speakers simultaneously organize at the same time the structuration of duration. This structuration is organized on the basis of relative lengthening of each successive lexical word in the phrase.

Another interest is that of determining what the phrase and macro-phrase distribution is. For energy indices, the amount of phrases is 60%, the amount of macro-phrases is 40%. For duration indices, the amount of phrases is 81%, the amount of macro-phrases is 19%. Among the list of these indices, some are more analytic than others. For energy parameter, the best index in this respect, is |ΔE| (82%); for duration parameter, the best ones are two indices which are very close, MSD (85%) and |ΔD| (84%). For those last 3 indices, the ratio phrase/macro-phrase is greater than for the other indices (about 3 times). These results show once more that the targets of the maximum loudness (and minimum loudness) and the maximum duration in the lexical word are fundamental : the prosodic organization of the lexical dimension is centered on them and their relative places in this hierachical structure within the phrase.

On lexical phrase segmentation, the sentence effect is very strong. For energy and over the 3 tasks, the mean proportion of phrases increase in a great expense for sentence 1 to 3 (56->58->91%), and the same phenomenon occurs for duration (62->62->92%). On the other hand, the task effect is insignificant. At the most, the analytic perspective decreases of 6%, for the benefit of macro-phrases constitued of 2 minimal phrases. The macro-phrases of more than 2 phrases are very rare (about 4%) for both parameters. So concerning lexical organization of duration and energy in phrases of various length, the general characteristics between indices are very close whatever the parameter may be.

Lexical phrase demarcation

Lexical phrase demarcation imposes a new condition, namely a syntactic one. Thus a macro-phrase is considered syntactic if 1° the ascending or descending pattern of the value index (including plateaux) corresponds exactly to its syntactic boundaries 2° the successive phrases which compose the macro-phrase have the same hierarchical level in the syntactic structure and if they are issued from the same father. In addition a new index is introduced, the last syllable duration plus the following pause (SDP).

In the context of phrase demarcation, it seems necessary to take into account the third linguistic unit, i.e. the micro-phrase. Let us consider the different cases where a micro-phrase was performed by speakers. Over 7 events in the text, 2 concern the case of pseudo-syntactic phrase, made up of two syntactic groups, one of them having too few syllables. So this micro-structuration takes into account syntactic organization. Over the next 5 cases, 4 are made up of compound-nouns or expressions the noun and adjective meanings of which are very dependent *(sources thermales, température moyenne ...)*, and the last one is made up of two adjectives, one of them qualifying two coordinate noun phrases. Thus one can say that all those phenomena are syntactic ones. They operate on less extended units, the relations of word dependency are stronger, but nevertheless syntax is always concerned. The other cases of micro-phrases which do not correspond to those cases are not counted as syntactic ones.

For the 3 energy indices, over the 12 speakers, the 3 tasks and the 3 sentences, among all the productions corresponding to syntactic phrase level, it appears that the phrases with the least extension are the more syntactic: 99% of decreasing or increasing values performed in the context of phrases correspond to a syntactic organization, and so are 83% of them corresponding to a micro-phrase, and 68% corresponding to macro-phrases. As for the best duration indices (WWD, |ΔD|, mPD, SDP), the results are similar to energy ones: phrase level, 99%, micro-phrase level, 92%, macro-phrase level, 53%.

Over all sentences, tasks, speakers, and different phrase units, |ΔE| appears to be the best index of phrase demarcation (91%), while for duration, the best indices are |ΔD| (92%) and SDP (91%). From task 1 to 3, the ratio of phrases and micro-phrases with a syntactic organization of energetic values, is regularly increasing for the three indices (as the case may be, 3 to 18% more), while that of macro-phrases decreases (21 to 25% less). For duration, from task 1 to 3, the 4 best indices show a great steadiness for phrase level, a slight tendency to decrease for micro-phrase, and for all indices except SDP, a tendency for macro-phrases to increase (10 to 37% more). Then it seems that the function of energy indices is more analytic than that of duration indices when the speech rate becomes slowler in tasks 2 and 3, with many pauses in addition for task 3.

It is interesting to note that DSP is the best index for macro-phrase demarcation when task 1 is running (76%) with relevant pause durations, but that its score becomes worse and worse (from 76 to 53%) as pauses get more numerous. On the whole, |ΔE| seems to be specialized in the demarcation of the phrases of less extension (phrases and micro-phrases), and MWE is the more resistent index for the hardest tasks. On the other hand, the 4 best duration indices are all specialized in the demarcation of phrase level (99 or 100%), SDP both for micro-phrase (100%) and macro-phrase demarcations, though the score of the latter is not very high (66%). For energy and duration indices, the task has no significant effect: from tasks 1 to 3, the range spreads from 87 to 90%.

CONCLUSION

For both duration and energy parameters the two functions of grouping and phrasing exist. The grouping function is based on the search of cohesion in discourse. The phrasing function is built on the linguistic objectives of parsing.

The 7 duration indices and the 3 energy indices show a deep lexical organization very similar in their forms and functions. Within this complex (and at the same time, simple) prosodic organization, all phrases (and probably most of the micro-phrases) are demarcated. For the grouping function, the global duration of lexical items appears to be a sufficient cue, while the phrasing function requires within this global duration, more precise targets, such as those performed by SDP and |ΔD|. |ΔE| exerts both grouping and phrasing functions. For the phrasing fonction, the targets of minimum and maximum duration and energy in the lexical items are crucial, and most of the speakers control those targets with a great precision. Then the demarcative function is built, for both parameters, on the precise control by speakers of the minimum and maximum values. This control at both levels, i.e. energy and duration, is operated for each parameter, in the context of at least 2 intricated scales : the scale within the word, and the scale within the phrase.

REFERENCES

G. Bruce (1985), Structure and Functions of Prosody, *Proc. of French-Swedish Seminar on Speech, Guérin et Carré eds.*, Grenoble, pp. 549-559.

G. Caelen-Haumont (1978, 1981), *Structures prosodiques de la phrase énonciative simple et étendue,* Thèse de doctorat de 3ème cycle, Université du Mirail, Toulouse, Hamburger Phonetische Beïtrage, band 34, Hamburg Busk.

V. Pasdeloup (1992), "Durée intersyllabique dans le groupe accentuel en français", Proc. of 19èmes JEP, Bruxelles, pp. 531-536.

Functional, acoustical and perceptual analysis of vocal hesitations in spontaneous speech

Isabelle Guaïtella
Institut de Phonétique d'Aix-en-Provence
Laboratoire Parole et Langage URA CNRS 261

ABSTRACT

The aim of this study is to reconsider and to show the interest of the analysis of a phenomenon which is specific to spontaneous speech: tvocal hesitations. These latter are often assimilated to silent pauses. Acoustical investigations led us to shed some light on the perceptual and communicative functions of vocal hesitations. This study is based on a general description of the rhythmic organisation of spontaneous speech. We demonstrate that both functions and acoustic manifestations of vocal hesitation are specific.

INTRODUCTION

Among all processes of expression which can be used by speakers in order to plan their discourse (repetitions, silent pauses, etc...; see Butterworth & Goldman-Eisler, 1979; Butterworth, 1980; Siegman 1979; Guaïtella 1991), only vocal hesitation cannot signify something else than the perceptible trace of the speaker's planning activity. We think that the vocal manifestation of hesitation plays a specific role in the rhythmic organisation of communication

HYPOTHESIS

Our hypothesis is that when a segmental duration cue is used simultaneously with a f0 cue, the syllable is alway accented. On the contrary, when a duration cue alone is used, it can be a vocal hesitation. As a consequence, the whole rhythmic organisation has to be analysed in order to bring to light the specificities of a functional element (for the rhythmic model, see Guaïtella 1991).

APPLICATION: STUDY OF INTERVIEW MATERIAL

This study deals with an extract of an interview, recorded in a sound-proof chamber. The speaker is a female, speaking a standard French. The topic is about a journey. Data are obtained from this material and completed by some examples taken from various other materials (other situations and other speakers) which are here to confirm that the results cannot be explained by individual specificities.

F0 configurations and contexts of apparition of hesitation

We observed the totality of the hesitations according to the context of appearance and disappearance (i.e. preceded or followed by text or silence), their duration and their f0 pattern.

We observed that hesitations could appear in all contexts, i.e.:

- preceded and followed by silence (22.5%)
- preceded by speech and followed by silence (50%)
- preceded by silence and followed by speech(5%)
- preceded and followed by speech (22.5%)

Four fundamental frequency patterns were observed:

a) - Drop in pitch only (or eventualy flat contour), see figure 1.
b)- Drop in pitch including or ended by creaky voice.
c)- "Step-drop" of pitch, i.e. minor modulations of pitch which cannot be confused with a rising contour, followed by greater falls.

d)- Decreasing following or ended by a "start", i.e. major rising contour.

For the last configurations (d) we consider that a single vowel is divided into two parts: the hesitation following or preceding an accented syllable. This point of view can be justified by perception and by the "linguistic logic" of the phenomenon: we can easily imagine that the speaker could hesitate on a vocalic segment, then accent the same segment to initialise what he is going to say (the opposite strategy is also possible).

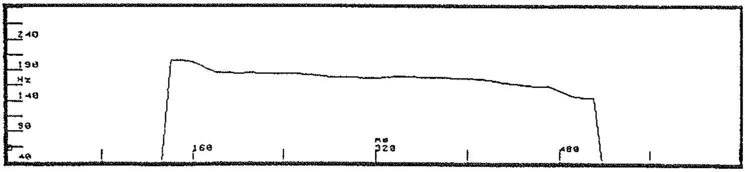


Figure 1. *Example of drop in (text: "et", and).*

Hesitation and duration variation

The durations of the hesitations are much much greater than those of the other syllables (Tables 1 et 2). The shortest duration is 197 ms and the longest is 1157 ms.

legend:

T : presence of text
: presence of pause
- # : hesitation preceded and followed by pauses
\ : progressive drop in pitch
k : presence of creaky voice
m : presence of modulations when decreasing
Am : presence of a setting of a rising contour

Table 1. *Durations of vocal hesitations according to the context (in ms). moy = mean; min = minimum, max = maximum.*

	moy	min	max
T - #	465.4	197	712
# - T	574	420	728
T - T	413.9	237	579
# - #	657.7	393	1157

Table 2. *Durations of vocal hesitations according to the prosodic configurations (in ms). moy = mean; min = minimum, max = maximum.*

	moy	min	max
\	476.4	197	930
K	567.2	237	1157
m	527	420	712
Am	582.7	468	756

We can observe that duration is much longer for the hesitations preceded and followed by silent pauses.

The duration of the hesitations (the overall mean duration is 533 ms) is greater than the mean of syllable durations (184.5 ms) and less than the mean duration of silent pauses (817.1 ms). This result can be considered the consequence of the presence of three classes of segmental duration: syllable duration, hesitation duration, pause duration.

Hesitation and pitch

A falling contour is always observed on hesitations (a few cases of flat contours were also found, but never rising ones). However, the amplitude of the fall can vary (Tables 3 and 4).

Table 3. *Pitch fall according to the context (in Hz). moy = mean; min = minimum, max = maximum.*

	moy	min	max
T - #	- 28.6	- 10	- 67
# - T	- 46.5	- 42	- 55
T - T	- 22.75	0	- 50
# - #	- 45	- 25	- 65

Table 4. *Pitch fall according to the prosodic configurations (in Hz). (For the "settings" we have taken into account the last value before the rise or the first value before the fall; for the "modulations" we have considered the global decreasing of pitch). moy = mean; min = minimum, max = maximum.*

	moy	min	max
\	- 30.75	0	- 65
K	- 32.65	- 10	- 50
m	- 34.8	- 20	- 67
Am	- 39	- 27	- 60

When hesitation is preceded and followed by text, it shows a smaller fall. In the same way this lowering is often shorter, which seems logical. Hesitations preceded by a pause (and followed by text or pause) show a deeper fall and longer duration. The context of appearance of hesitations seems to influence the quantity of the fall. On the other hand, this phenomenon does not seem to depend on the prosodic configurations.

PERSPECTIVES AND DISCUSSION: CONSEQUENCES FOR THE PERCEPTION OF RHYTHM.

Hesitation corresponds to the desire of the speaker to “keep the floor” while preparing what he is going to say (Maclay, Osgood, 1959). What is specific to vocal hesitation is its vocal but non-verbal nature. They are certainly necessary for maintaining contact by keeping talking during the period of time which the speaker saves for conceptualisation. During vocal hesitations voice sounds like a sustained note. However, while this sustained note shows, at acoustic level, a dcreasing of pitch, it corresponds to a diminution of sub-glottal pressure, i.e. the physiological dimension of declination. The acoustical realisations of vocal hesitations are of interest because they seem to prove their physiological origin.

This could explain why hesitations are hardly "detectable" by listeners in a standard situation of communication. It is well known that listeners only notice hesitations if they are extremely frequent. But, if we consider that the vocal reality of hesitations is only physiological, the non-verbal nature of hesitations has also to be admitted. However, it does not mean that hesitations have no communicative functions.

Moreover, the specific duration scale of hesitations could facilitate a specific perceptual treatment. This interpretation is logical if we consider that hesitations are not dependant (in most cases) on the speaker’s will.

We can admit that there exists a subconscious agreement among speakers in order to not consciously detect hesitations. If we maintain that hesitations are inevitable and essential in communication, speakers often consider them as failures of reasoning.

Their role in communication (and especially in perception) is probably to create a "time of rest" during which both speaker and listener can review what has been said and extrapolate what is going to be said.

CONCLUSIONS

Lengthening of syllable duration, without break of pitch, cannot be considered as a cue for accentuation but - at least for the cases of extreme increasing - an hesitation cue considered as another segmental category. Determining a duration threshold could permit the discrimination of non-accented syllables (with increase in duration) from vocal hesitation. The parameter of duration appears to be an accentuation cue but remain a secondary cue associated with a pitch break. The existence of hesitation phenomena confirm the prime role of the localisation of direction of slope on the syllable in the perception of perception.

REFERENCES:

Butterworth B., 1980, "Evidence from Pauses in Speech", in Butterworth *Language Production*, vol.1 *"Speech and Talk"*, Academic Press, pp. 155-76.

Butterworth B., Goldman-Eisler F., 1979, "Recent Studies in Cognitive Rhythm", in Siegman, Feldstein, *Of speech and time Temporal Speech Patterns in Interpersonal Contexts*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 211-24.

Guaïtella I., 1991, *Rythme et parole: comparaison critique du rythme de la lecture oralisée et de la parole spontanée*, Doctoral Thesis, Aix-en-Provence.

Maclay H., Osgood C.E., 1959, "Hesitation phenomena in spontaneous English speech", *Word*, 15, pp. 19-44.

Siegman A.W., 1979, "Cognition and Hesitation in Speech", in: Siegman, Feldstein, *Of Speech and Time*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 151-78.

On the temporal domain of focal accent

Vincent J. van Heuven
Holland Institute of Generative Linguistics/
Phonetics Laboratory, Leiden University
Cleveringaplaats 1, PO Box 9515, 2300 RA Leiden, The Netherlands

ABSTRACT

Accented words are longer than their unaccented counterparts. Our contribution studies the domain of this lengthening effect: is it the (stem) morpheme, the (complex) word, or the phrase? We present preliminary evidence, from the literature and from a recent experiment by ourselves, that the lengthening domain extends beyond the monomorphemic word, but not beyond the level of the compound.

INTRODUCTION

The function of pitch accents in languages such as English and Dutch is to focus the listener's attention on the contents of a specific linguistic domain (syllable, word, word group, or even a whole sentence). The principal phonetic correlate of focus is a prominence-lending change in pitch on the syllable that constitutes the prosodic head of the focus domain. Thus, when a whole word is in focus, as in *Did the police arREST or merely susPECT the butler?* (capitals indicate accent, underlining indicates focus), the speaker executes a conspicuous pitch change on the lexically stressed syllable of the words in focus. As a secondary effect, an accented word is linearly stretched in time by about 15% (i.e. across stressed and unstressed syllables alike) relative to a version of the same word spoken outside focus (Eefting, 1993; Sluijter, van Heuven and Neijt, 1992). It is not the case, however, that every content word in a larger focus domain should bear an accent. Speakers typically accentuate only the prosodic head of a word group that constitutes a larger focus domain. This option is called integrative accent. For example, the entire NP *the old man* is in focus in the utterance *Did you see the old MAN or the BAby?*, even though only the prosodic head of the NP (*man*) is accented. The realisation of the NP under integrative focus is exactly the same (temporally and melodically) as with narrow focus on the head only, as in *Did you see the old MAN or the old WOman?* It is therefore not the case that materials in focus are pronounced more slowly and deliberately than materials outside focus. Moreover, Sluijter et al. (1992) showed that pronouncing a single word with an accent on another syllable than the lexical stress yields linear time expansion of the entire word, albeit with a large shift in relative duration between the stressed and non-stressed syllables. These results can be summarised as follows: all the syllables (stressed and unstressed alike) in a word containing a pitch accent (whether on the stressed syllable or not) are stretched in time. An unaccented word, even when in focus, is not stretched (relative to its realisation outside focus).

In the present research we seek to establish the maximal domain for the time-expansion effect of accent in more detail. Specifically, we address the question whether the lengthening domain concerns the word level or the morpheme level. We shall pursue the answer by studying the effects of (i) narrow focus on the prosodic head of adjectival compounds versus (ii) integrative focus on the same compounds.

As in English, in Dutch compound adjectives the prosodic head is the rightmost stem morpheme. The crucial question is whether there is a timing difference between compounds with narrow focus on the final morpheme and with integrative focus on the entire compound, as in the following (English) question/answer pairs:

Q1. *Is that tea orange-flavoured or orange-scented?*
A1. *The tea is orange-flavoured.* (narrow focus on final stem morpheme)

Q2. *Is that tea orange-flavoured or plain?*
A2. *The tea is orange-flavoured.* (integrative focus on entire compound)

If the lengthening effect is word-based, the entire compound *orange-flavoured* should stretch in both A1 and A2, i.e. whether under narrow or under integrative focus. If, on the other hand, the domain of the lengthening effect is the stem morpheme, only the final morpheme that constitutes the prosodic head (*flavoured*) should lengthen. In order to determine the extent of the lengthening effect we shall adopt as a base-line condition a question/answer pair such as Q3/A3, where the compound adjective is spoken outside focus:

Q3. *Is the tea or the coffee orange-flavoured?*
A3. *The tea is orange-flavoured.* (no focus on any part of compound)

PRODUCTION EXPERIMENT

Among other words (which do not concern the question addressed in this paper), two Dutch compound adjectives (with monosyllabic and disyllabic stem morphemes, respectively) were included in non-final position in question/answer pairs, with (i) integrative focus on the entire compound, (ii) narrow focus on the prosodic head (final morpheme), and (iii) with no focus on the target (for examples cf. perception experiment). The six relevant words types were recorded twice by five speakers of standard Dutch, in different random orders. Each target was spoken as a lexical word and repeated once more in reiterant speech, replacing the lexical word by an otherwise identical sequence of syllables /na:/. *Knalgeel* /knAl#'Ge:l/ 'bright-yellow' was repeated as /na:'na:/; *donkerécru* /dONk@r#e:'kry:/ 'dark-écru' as /na:na:na:'na:/.

Table 1 presents the durations of the first and second parts of the compound adjectives (20 tokens per condition), as measured with a high-resolution digital waveform editor, for each of the relevant focus conditions.

Table 1. *Duration (in ms) of morphemes and words in lexical and reiterant versions broken down by focus condition. Durational differences from base-line condition (no focus) are indicated in percent.*

	lexical			reiterant		
focus	1st part	2nd part	word	1st part	2nd part	word
no	295	266	561	217	279	496
narrow	+2%	+13%	+7%	+0%	+4%	+1%
integrative	+5%	+15%	+9%	+8%	+4%	+5%

Accented words (in focus) are longer than unaccented words, and the effects are stronger for lexical words than for the reiterant versions. Crucially, both the final morpheme (prosodic head) within the compound, and the initial non-head morpheme are elongated, which seems to point to the entire word as the basis of the lengthening effect under accent. However, the lengthening of the initial non-head morpheme is stronger (by a factor 2) when the non-head morpheme is in (integrative) focus than when it is outside the narrow focus on the head.

We conclude from these results that the scope of the lengthening effect of accent extends to the complex word, and is not limited to the stem morpheme. In contradistinction to what was found earlier for monomorphemic words (Sluijter at al., 1992) compounds show a difference in temporal organisation between integrative and narrow focus: the non-head portion of the compound is stretched more under integrative focus. In the perception experiment to be discussed next, we shall consider the perceptual consequences of this difference in temporal organisation.

PERCEPTION EXPERIMENT

The answer portions of the recordings made above (second tokens of lexical versions only) were presented to 20 native Dutch listeners over headphones. Listeners were given answer sheets containing, for each answer sentence, written versions of the three questions that originally preceded the answer sentences so as to elicit the three different focus distributions. For example, when the stimulus answer sentence was *Ik heb die muur knalgeel geverfd* 'I have that wall bright-yellow painted', whether produced with narrow focus, with integrative focus or without focus on *knalgeel*, the listeners read three written questions:

(1) *Heb je die muur knalgeel of grijs geverfd?* (integrative focus)
'have you that wall bright-yellow or grey painted?'

(2) *Heb je die muur knalgeel of knalrood geverfd?* (narrow focus)
'have you that wall bright-yellow or bright-red painted?'

(3) *Heb je die muur knalgeel geverfd of behangen?* (no focus)
'Have you that wall bright-yellow painted or papered?'

The listeners' task was to decide for each spoken sentence on the tape whether it was most likely to be the answer to question type (1) asking for integrative focus on the target compound, or to question type (2) asking for narrow focus on the target, or to question type (3) asking for no focus on the target. Each stimulus type was presented twice in counterbalanced order. In all, 1200 responses were collected (5 speakers * 2 targets * 3 focus distributions * 2 orders * 20 listeners). The results of the perception experiment are summarised in table 2.

Table 2. *Frequency of perceived focus responses as a function of focus distribution intended by speaker.*

	perceived focus		
intended focus	integrative	narrow	no
integrative	157	243	0
narrow	117	283	0
no	0	0	400

Sentences with non-focused targets were matched with question type (3) without a single error. This performance will be due rather to the absence of a pitch accent on target rather than to its shorter duration. we shall not discuss this condition any further. Still, even in the remaining part of the contingency table the distribution of the responses deviates significantly from chance, X^2=8.4 (p=.004), which indicates that our listeners were able, to some extent, to perceive the difference between integrative and narrow focus on the adjectival compounds as intended by the speakers. The effect is clearly speaker dependent (data not indicated in table 2). Speakers #3 and #4 hardly made any difference between narrow and integrative focus (no significant association between intended and perceived focus), speakers #1 and #5 have a moderate association, while speaker #2 is surprisingly successful (over 75% correct) in communicating his intended focus distribution.

We conclude from the results of this perception experiment that at least some speakers are able to make a communicatively relevant distinction in morphologically complex words between integrative focus on the entire compound and narrow focus on just the prosodic head within the compound. Note that'listeners proved unable to make this distinction in similar listening tests using the monomorphemic tokens collected by Sluijter et al. (1992).

CONCLUSION

We conclude from the production and perception experiments reported above, that the domain of the lengthening effect of accent extends beyond the morpheme or the monomorphemic word: lengthening applies to all the segments in a compound word. However, other than what we found earlier in monomorphemic words, the lengthening effect is weaker for the non-head morpheme than for the head morpheme. Moreover, if the head morpheme is in narrow focus, the lengthening of the non-head if weaker still (by a factor 2 or more) than when both head and non-head morpheme are in focus. This difference between integrative versus narrow focus in Dutch compounds is communicatively relevant, at least for some speakers.

At least two caveats are in order here. First, the conclusions that were drawn above are partly based on data collected in our own experiments, and partly on data published in the literature. There is a risk involved in so far as different speakers and lexical materials were used in the various experiments. What is needed is a single, large-scale experiment in which all the relevant factors are are examined using carefully controlled materials and the same set of speakers across experimental conditions. Second, there need not be a causal relationship between the observed differences in temporal structure between integrative and narrow focus on adjectival compounds and the perception of the contrast. Further experiments are necessary in which acoustic parameters other than timing are kept under strict control.

REFERENCES

W.Z.F. Eefting and S.G. Nooteboom (1993), "Accentuation, information value and word duration: effects on speech production, naturalness and sentence processing", in *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation*, ed. by V.J. van Heuven and L.C.W. Pols (Mouton de Gruyter, Berlin), pp. 225-240.

A.C.M. Sluijter, V.J. van Heuven and A.H. Neijt (1992). Lexical stress and focus distribution as determinants of temporal structure, in B.L. Derwing, J.J. Ohala (eds). *Proc. 2nd Internat. Conf. Spoken Lang. Processing, Banff, Alberta, 12-16 October 1992*, Vol. 1, pp. 349-352.

Establishing Prosodic Structure by Measuring Segment Duration

Elise Hofhuis
Department of Language and Speech
Nijmegen University
Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

ABSTRACT

In this paper we describe two experiments which show how knowledge about segment durations can de used to decide between prosodic theories. One experiment shows us that unstressed word-final heavy syllables do not form separate feet. A second experiment demonstrates that word initial monosyllabic feet are not feet postlexically.

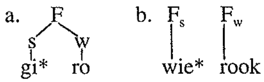
INTRODUCTION

Prosodic structure is generally motivated on the basis of its relevance to segmental phonological rules. However, it is also reflected in the durational structure of speech. For example, Gussenhoven and Rietveld (1992) found that English listeners expect the duration of preboundary syllables to increase with the rank of the prosodic boundary they precede. Many other experiments report preboundary lengthening in speech production, such as Lindblom & Rapp (1973), Klatt (1975) and Lehiste (1979). Although these experiments are concerned with morpho-syntactic boundaries, the observed lengthening can be interpreted as a reflection of prosodic structure, since higher prosodic boundaries by and large coincide with important morpho-syntactic boundaries. In this paper we will present two experiments that illustrate how knowledge about segment durations can be used to decide between theories of foot structure.

EXPERIMENT 1

Dutch treats open (VV) syllables as light and closed VC, VVC- and VCC-syllables as heavy. The Dutch stress system is left dominant, quantity sensitive and right-to-left. Heavy penultimate ,syllables attract the main stress without exception (e.g. Aláska, *álaska), while open penultimate syllables can be skipped (e.g. dóminee). As a result of this fact, final syllables without main stress are included in a binary foot with the penultimate main-stressed syllable when they are open, but closed syllables are assumed to form feet by themselves (see van der Hulst (1984), Kager (1989) and Trommelen & Zonneveld (1989), among others). This distinction is illustrated in (1a,b):

(1)



For the proposals in (1) there has been no independent evidence. In fact, there are no obvious prominence distinctions between the final syllables of (1a) and (1b). Moreover, Gussenhoven's (1993) investigation of the chanted call in Dutch and several segmental foot domain rules suggests that monosyllabic feet only occur word finally in Dutch. This means that both words in (1) should be analysed as having structure (1a). It was the aim of experiment 1 to find out wether we could find phonetic evidence for the foot structure of words like (1b) as compared to (1a).

Method

It is a well documented fact that the length of stressed syllables is inversely related to the number of unstressed syllables within the foot. This was described by Nooteboom (1972) for Dutch. Thus, we assume that the stressed syllable of a monosyllabic foot is generally longer than the same syllable in a polysyllabic foot. If we generalize this assumption, we can establish the foot structure of words like (1a) and (1b) by measuring the segment durations of their first syllables. If the first syllables of words like (1b) were to have longer durations than those of words like (1a), it would be reasonable to assume that the first syllable of (1b)-type words does constitute a separate foot. However, if their first syllables are shown to be equal in duration, we must accept that both types of words should be analysed as in (1a), as Gussenhoven suggests.

Three minimal pairs of bisyllabic s/w words were selected. One had an open first syllable (2a), one a closed first syllable (2b) and one had a final syllable closed by an ambisyllabic consonant (2c). The words in each pair differed only in the weight of their second syllables. For comparative reasons we also included versions of the pairs with second syllables containing schwa.

Table 1. *Material for experiment 1.*

	2nd Light	2nd Heavy	2nd schwa
a.	Syra (si:ra)	sieraad (si:ra:t)	sieren (si:r@)
b.	basta (bɑsta)	bastaard (bɑsta:rt)	basten (bɑst@)
c.	mamma (mɑma)	mammoet (mɑmu:t)	mammen (mɑm@)

The words were embedded in a carrier sentence in postfocal position. They were spoken ten times by two male speakers of Dutch, 22 and 23 years old, who were paid a small fee for their services. The words were recorded in two sessions in a sound proof studio. Segments were measured by hand using the SESAM segmenting program of the Department of Language and Speech of Nijmegen University.

Results

We performed separate ANOVA's on the first syllable durations of the words in table 1(a), (b) and (c). The syllable durations are listed in table 2 below. Factors were speaker (2) and nature of the second syllable (light, heavy or schwa).

The nature of the second syllable had a significant effect on the duration of the first syllable for the (a) words, but only before schwa, before which it was longer (see table 2), not in the other two conditions (F(2,53)=5.05, p=.010). The lengthening effect schwa can have on preceding stressed vowels was also found by Nooteboom (1972). There was no speaker effect.

For the words in table 1(b) we found no significant difference in duration of the first syllables in any condition (F(2,52)=1.43, p=.248). No speaker effect was found.

The first syllables of the (c) words in were not significantly different in any of the three conditions (F(2,54)=1.068, p=.351). There was a significant speaker effect: the syllables of one speaker were on average 10 ms longer (F(1,54)=7.31, p=.009). However, there was no interaction involving the speakers.

Table 2. *Results of experiment 1: first syllable durations in ms.*

	2nd light	2nd heavy	2nd schwa
1st syll. open	261	260	274
1st syll. closed	269	271	277
1st syll. ambi.	226	231	226

EXPERIMENT 2

Another implication of Gussenhoven's (1993) findings concerns words with w/s stress patterns. They are usually analysed as consisting of a stressless monosyllabic foot followed by a foot bearing main stress. If monosyllabic feet cannot occur word initially this analysis cannot be maintained. Gussenhoven suggests that these word initial feet should be analysed as appendices to the word, and do not receive foot structure. The phonological data seem to corroborate this analysis. One way of finding additional phonetic evidence for the analysis is to look at the neutralisation of vowel duration in the appendix. It is a distributional fact of Dutch that the opposition between long and short vowels is only maintained in foot-initial and word-final position, which suggests that no VV-V opposition exists outside these contexts. Therefore, a lack of durational opposition between tense and lax vowels in word-initial position may be interpreted as indicating the absence of foot structure. The experiment was designed to find out whether this neutralization is an acoustically relevant process.

Method and results

The only minimal pair we could find for this experiment was the pair *anale* (ana:l@) 'anal', with a long first vowel, versus *annalen* (ɑna:l@) 'annals', with a short vowel and ambisyllabic consonant. The two words were embedded in the same carrier sentence as the one in experiment 1, and spoken 10 times by the same speakers.

We carried out separate ANOVA's on the vowel and on the following /n/. Factors were speaker (2) and vowel type (long or short). We found no significant difference in duration between the long and short vowels (F(1,36)=.051, p=.823). There was no significant difference between the speakers either. In the second analysis the /n/ was found to have nearly equal durations in both words (2 ms difference) but one speaker's /n/'s were significantly longer (F(1,36)=55, p<.001). We also found some interaction between the two factors (F(1,36)=5.83, p=.021) due to the fact that for one speaker the /n/ was longer after the "long" vowel (7 ms) while it was longer after the short vowel (3 ms) for the other speaker. This was not a substantial difference, however. The segment durations are listed in Table 3 below.

Table 3. *Results of experiment 2: durations in ms.*

	duration of vowel	duration of /n/
'anale'	73 ms	68 ms
'annalen'	72 ms	66 ms

CONCLUSION

The results of experiment 1 clearly show that the weight of an unstressed syllable does not influence the duration of a preceding main-stressed syllable. Thus, we must conclude that the words in table 1 have the same binary foot structure. As was mentioned above, the foot-domain rules that are described in Gussenhoven (1993) independently motivate this analysis.

Experiment 2 demonstrates the total neutralisation of durational differences between long and short vowels in word-initial unstressed syllables. We think this can be explained by the absence of foot structure above these syllables, and that these syllables should be analysed as appendices to the foot. We assume that this can cause VV-syllables to drop a V-slot. This analysis was argued for independently by Gussenhoven (1993) on the basis of observations about the chanted call in Dutch, and segmental phonological rules.

Both experiments indicate that it is possible to corroborate phonological theories using phonetic evidence. It is our intention to explore this possibility further in future experiments.

ACKNOWLEDGEMENTS

I wish to thank my supervisors, Toni Rietveld and Carlos Gussenhoven, for their assistance and their comments on earlier versions of this paper.

REFERENCES

C. Gussenhoven and A. Rietveld (1992), "Intonation contours, prosodic structure and preboundary lengthening", *Journal of Phonetics,* Vol. 20, pp. 283-303.

C. Gussenhoven (1993), "The Dutch foot and the chanted call", to appear in *Journal of Phonetics.*

H. van der Hulst (1984), *Syllable structure and stress in Dutch* (Foris, Dordrecht).

R. Kager (1989), *A metrical theory of stress and destressing in English and Dutch* (Foris, Dordrecht).

D. Klatt (1975), "Vowel lengthening is syntactically determined in connected discourse", *Journal of Phonetics,* Vol. 3, pp. 161-172.

I. Lehiste (1972), "The timing of utterances and linguistic boundaries", *Journal of the Acoustic Society of America,* Vol. 51, pp. 2018-2024.

B. Lindblom and K. Rapp (1973), "Some temporal regularities of spoken Swedish", *Papers from the Institute of Linguistics,* Vol. 21 (University of Stockholm).

S. G. Nooteboom (1972), *Production and perception of vowel duration* (University of Utrecht, Utrecht).

M. Trommelen and W. Zonneveld (1989), *Klemtoon en metrische fonologie* (Coutinho, Muiden).

"Roger", "Sorry", "I'm still listening": Dialog Guiding Signals in Information Retrieval Dialogs

A. Kießling[1], R. Kompe[1], H. Niemann[1], E. Nöth[1], A. Batliner[2]

[1] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5),
Martensstr. 3, 91058 Erlangen, F.R. of Germany

[2] L.M.-Universität München, Institut für Deutsche Philologie,
Schellingstr. 3, 80799 München, F.R. of Germany

ABSTRACT

During any kind of information retrieval dialog, the repetition of parts of information just given by the dialog partner can often be observed. As these repetitions are usually elliptic, the intonation is very important for determining the speakers intention. In this paper prototypically the times of day repeated by the customer in train table inquiry dialogs are investigated. A scheme is developed for the officers reactions depending on the intonation of these repetitions; it has been integrated into our speech understanding and dialog system EVAR (cf. [6]). Gaussian classifiers were trained for distinguishing the dialog guiding signals confirmation, question and feedback; recognition rates of up to 87.5% were obtained.

INTRODUCTION

Dialog systems for information retrieval are potential applications for human-machine communication. In human-human dialogs, very often (parts of) the information just given by the speaker is repeated by the partner. It can be observed e.g. in train table inquiries that the customer (henceforth C) repeats the times of arrival or departure just given by the officer (O). Very often only the intonation of this repetition of the time of day (RTD) shows the intention of C and thus governs the continuation of the dialog. In the scenario of our speech understanding and dialog system EVAR (an experimental automatic information system on train tables) the transmission of these times is a pivot point. Of course, a user-friendly system should be able to react adequately (cf. [7]). Let's e.g. consider the following dialog: O: *"... leaves Ulm at 17 23."* C: *"17 23./?"*. In the case of a rising intonation ('?') O – or the system, respectively – has to repeat the time of day, because C wants to have the time acknowledged. In the case of a falling intonation ('.') no specific reaction is necessary and the system can e.g. give the next part of information. This paper describes a corpus of *"real-life"* train table inquiry dialogs, the frequency of occurrences of RTDs, and their different intonational marking and functions. We will show how these functions can be determined automatically, and briefly how this prosodic information has been introduced into EVAR.

MATERIAL

Our investigations were based on a corpus of 107 "real-life" train table inquiry dialogs, recorded at different places, most of them conducted over the phone; for more detail cf. [1]. In most of the cases (88) the callers did not know that they were recorded. 92 dialogs concerned train schedules, the rest had other topics like fares. These 92 dialogs contained 215 utterances of C with in total 227 RTDs of arrival or departure, i.e. more than two repetitions per dialog on the average. In all but 3 cases the repetition concerned the time of day O just gave. There are two forms of time of day expressions possible in German: with or without the word *Uhr* (e.g. *"17 Uhr 23"* or *"17 23"*).

DIALOG GUIDING SIGNALS

By repeating the time of day, C has different aims, i.e. he wants to signal O different kinds of information. Depending on the specific kind of information, mostly expressed by the intonation, the reaction of O and thus the continuation of the dialog is governed. We observed three different functional roles of the RTD: *confirmation*, *question* and *feedback* (cf. figure 2).

- Using a **confirmation**, C wants to signal O, that he got the last information, e.g. the time of arrival. Functionally, this corresponds to the word *"Roger"* in the radio traffic. Usually, the intonation ($F0$-contour) at the end of such an utterance is falling (F, cf. figure 2a). A confirmation can be frequently observed after the end of a turn of O, just at the beginning of the turn-taking by C.

  Ex: O: *You'll arrive in Munich at 5 32.*

  C: **5 32.**

- The function of a **question** is *"Sorry, please repeat"*. C signals O that he didn't understand, i.e. that he didn't get the time of day completely or that he just wants to ask O to confirm the correctness (*"correct me if I'm wrong"*). The prototypical $F0$-contour is rising (R, cf. figure 2b). These questions often occur as short interruptions during the answer phase of O.

  Ex: O: *...you'll leave Hamburg at 10 15...* *...yes, 10 15, and you'll reach...*

  C: **10 15 ?**

- By using a **feedback**, C usually wants to signal O *"I'm still listening"*, *"I got the information"* and sometimes *"slow down, please!"* or *"just let me take down the information"*. It is normally characterized by a constant or slightly rising $F0$-contour (continuation rise, CR, cf. figure 2c) and like the question it is usually found during the answer phase of O.

  Ex: O: *...the next train leaves at 6 35...* *...and arrives in Berlin at 8 15.*

  C: **6 35 –**

Note that one has to distinguish **function** (confirmation, question, and feedback) and intonational **form** (F, R, and CR) although in prototypical cases there is an unequivocal mapping of form onto function. The dialog guiding function of a confirmation is similar to a feedback, but their intonational form is different. Normally, questions can be distinguished easily from confirmations. Feedbacks, however, are sometimes likely to be confused with questions or even with confirmations. In our material, in 100 of the 227 repetitions of C the reaction of O (confirmation of the correctness, repetition, correction or completion of the time of day) was governed by nothing but the intonation of C. In the remaining cases, there were other indicators like interrogative particles. In 64 of the 100 cases, the time of day occurred isolated; thus, the only possibility for O to interpret the intention of C correctly, is by using the intonation of these elliptic RTDs as a cue, because other grammatical indicators like word order or *Wh*-words are missing.

THE REACTION SCHEME FOR THE DIALOG SYSTEM EVAR

From the corpus we developed a scheme (cf. figure 1) showing the reactions of O depending on the intonation of the RTD of C. In the scheme it was also taken into account if the customer repeated the time of day completely, incompletely or incorrectly. The scheme was integrated into our speech understanding and dialog system EVAR [6], which in our application plays the role of O. In figure 1, for an example of an information given by the system the possible ways of repeating the last given time of day are shown in the first two columns: C can repeat the time of day not at all or completely or incompletely and correctly or incorrectly, using different intonations. In the case of the system, the word recognition module has to provide the dialog module with the analyzed word chain that has to be compared with the previous time of day given by the system. Depending on the different possible combinations, the third column shows the next dialog step, the reaction of the system. For the integration into our system a prosody module has been added to the linguistic knowledge base, a semantic network. All necessary prosodic knowledge has been specified as a set of concepts and attributes within the formalism. Since the same network contains all linguistic knowledge sources of EVAR as well, appropriate links between the prosody module on one hand and the syntactic, semantic, pragmatic and dialog module on the other hand can be easily established. For more details, cf. [6].

THE PROSODY MODULE

The task of the prosody module is to determine automatically the intonation type, i.e. F,

System answer: "... *In München sind Sie dann um 17 Uhr 32.*"
"... *You'll arrive in Munich at 5 32 p.m.*"

RTD		prosody–module		system reaction	
no utterance		——		——	
incorrect		——		correction	(*'Nein, um 17 Uhr 32.'*)
complete & correct		R	(*'17 Uhr 32?'*)	confirmation	(*'Ja, um 17 Uhr 32.'*)
		CR	(*'17 Uhr 32–'*)	——	
		F	(*'17 Uhr 32.'*)		
correct & incomplete	only minutes	R	(*'32?'*)	confirmation	(*'Ja, um 17 Uhr 32.'*)
		CR	(*'32–'*)	——	
		F	(*'32.'*)		
	only hours	R	(*'17 Uhr?'*)	completion	(*'17 Uhr 32.'*)
		CR	(*'17 Uhr–'*)		
		F	(*'17 Uhr.'*)	——	

Figure 1: *The reaction scheme for* RTD*s within the dialog system EVAR*

R, CR, that are mapped onto the functional roles of the RTD, i.e. confirmation, question, and feedback. From the automatically computed $F0$-contour [5] the following 4 features are extracted: the slope of the regression line of the whole (cf. the lines in figure 2a-c) and of the final part of the $F0$-contour, and the differences between the offset (the $F0$-value of the last voiced frame) and the values of the regression lines at this offset position (related work and comparable features are e.g. reported in [7] [3] [4]). Gaussian classifiers with full covariance matrix were trained to classify into the three classes F, R, and CR and thus – prototypically – into the functional roles confirmation, question, and feedback.

DATABASES FOR THE CLASSIFIER

Two databases were recorded and digitized with 16 kHz and 14 bit: In database A one female and three male speakers (not "naive", because they are working in prosody) read 90 complete time of day utterances each (all with the word *"Uhr"*; 30 questions, confirmations, and feedbacks each). As this database was used for training, misproductions (e.g. a question was intended, but a falling $F0$-contour was produced) and erroneous $F0$-contours were discarded. Thus a total of 322 utterances could be used for training. In database B two female and two male "naive" speakers read 50 time of day expressions each. Neither misproductions nor erroneous $F0$-contours were sorted out; this database gives therefore a good impression about how the system works in real life.

EXPERIMENTAL RESULTS AND DISCUSSION

Three experiments were performed. In the first experiment database A was used for testing in a leave-one-out mode (i.e. 3 speakers were used for training, the other for testing). In the second experiment the classifier trained on database A was tested on database B. Different feature combinations (e.g. computing the slope of the 2nd regression line over the last, the last two or the last three voiced regions) were tried. The results for the best feature combination where the 2nd regression line was computed over the last two voiced regions are shown in table 1 and 2. In the leave-one-out experiment (table 1) for all 3 cases (the rows marked by R, CR, and F; number of occurrences in parentheses) good recognition rates could be achieved (average recognition rate: 87.5%). For the speaker-independent test with the naive speakers (table 2) we obtained an average recognition rate of 71.3%. The decrease in performance is due to the fact that no utterances were discarded and that the naive

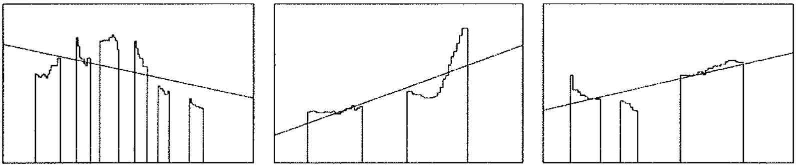
Table 1: *Results for leave-one-out*

	R	CR	F
R (97)	81.4	18.6	0.0
CR (107)	7.5	87.9	4.7
F (118)	1.7	5.1	93.2

Table 2: *Results for database* B

	R	CR	F
R (70)	87.1	7.1	5.7
CR (64)	21.9	37.5	40.6
F (66)	3.0	7.6	89.4

speakers obviously had enormous difficulties in the controlled production of a continuation rise: whereas questions and confirmations were recognized with approximately the same recognition rate (88%) as in the first experiment, it was much more difficult to classify the feedbacks correctly. As a final experiment the classifier trained on database A was tested on a subset of the above mentioned "real-life" material. Due to the sometimes very noisy telephone quality, only 32 isolated RTD's could be used for classification. Their reference type (R, F, CR) was determined by auditory tests. For classification, the same features as described above were extracted from the digitized signal. All the 10 confirmations, all the 5 questions and 7 of the 17 feedbacks were classified correctly.



2a: F, Confirmation **2b:** R, Question **2c:** CR, Feedback

Figure 2: *Prototyp. F0-contours (F, R, CR), their functional roles, and regression lines*

FINAL REMARKS

In [2] we show that the prosodic marking of sentence modality is more distinct in elliptic utterances than in non-elliptical utterances. Therefore we expect our modeling of question, confirmation, and feedback with R, F, and CR to work reasonably well not only with RTDs in train table dialogs, but also within other scenarios, where short elliptic utterances in clarification dialogs are used (like e.g. prizes in fare dialogs). However, our modeling is not exhaustive, because if e.g. in a confirmation a contrastive accent is positioned on the last syllable, or vice versa, in a question on the first syllable, our model will possibly not work adequately. Moreover, RTDs might not be purely isolated. They do often occur together with additional particles (like *"yes"*, *"no")* or with repetitions of city names. In future, we plan to take into account the other possibilities of accentuation as well as non-isolated RTDs.

Acknowledgements: This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research project ASL/VERBMOBIL. Only the authors are responsible for the contents of this paper.

References

[1] A. Batliner, A. Kießling, R. Kompe, E. Nöth, and B. Raithel. *Wann geht der Sonderzug nach Pankow? (Uhrzeitangaben und ihre prosodische Markierung in der Mensch-Mensch- und in der Mensch-Maschine-Kommunikation).* In *Proc. DAGA '92*, volume B, pages 541-544, Berlin, 1992.

[2] A. Batliner, C. Weiand, A. Kießling, and E. Nöth. *Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody.* In this volume.

[3] N. Daly and V. Zue. *Acoustic, Perceptual, and Linguistic Analyses of Intonation Contours in Human/Maschine Dialogues.* In *Int. Conf. on Spoken Language Processing*, pages 497-500, Kobe, 1990.

[4] N. Daly and V. Zue. *Statistical and Linguistic Analyses of F0 in Read and Spontaneous Speech.* In *Int. Conf. on Spoken Language Processing*, volume 1, pages 763-766, Banff, Canada, 1992.

[5] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. *DP-Based Determination of F0* contours from speech signals. In *Proc. ICASSP*, volume 2, pages II-17-II-20, San Francisco, 1992.

[6] R. Kompe, A. Kießling, T. Kuhn, M. Mast, H. Niemann, E. Nöth, K. Ott, and A. Batliner. *Prosody takes over: A prosodically guided dialog system.* To appear in: *Proc. Eurospeech93*, Berlin, Sept. 1993.

[7] A. Waibel. *Prosody and Speech Recognition.* Morgan Kaufmann Publishers Inc., San Mateo, California, 1988.

Cross-Speaker Influences on Intonation in Dialogue

Jacqueline Kowtko
Human Communication Research Centre
2 Buccleuch Place
Edinburgh EH8 9LW, U.K.

ABSTRACT

Following up previous claims that a mapping exists between intonation contour and discourse function, this paper provides evidence that another factor is involved, that of cross-speaker influence: there are systematic relationships between adjacent speakers' contours.

INTRODUCTION

Recent work on intonation in dialogue tends to follow one of two opposite approaches: it either describes very general discourse functions or identifies very specific discourse contexts.

The former approach is taken by McLemore (1991), in a study of phrase-final tunes in monologue and conversation. McLemore finds that the tunes indicate certain general discourse functions: *rising* tune *connects*, *level* tune *continues*, and *falling* tune *segments*. Context determines how each of these tunes operates. For instance, phrase-final rise—indicating non-finality or connection—can manifest itself as turn-holding, phrase subordination, or intersentential cohesion.

The latter approach is adopted by Hockey (1992). She examines three types of contour in terms of two contexts in task oriented dialogue, distribution of pronominal anaphora and turn-taking behaviour.

In order to further the understanding of intonational function, the present work attempts to combine these two approaches. This in turn requires an independent description of dialogue context as the basis for a robust account of intonational function. Such an independent description is the conversational games analysis outlined in Kowtko, Isard and Doherty-Sneddon (1992).

DIALOGUE CONTEXT

Kowtko *et al.* (1992) propose a repertoire of interactional exchanges, called *conversational games* (deriving from a tradition of literature originating in Power, 1974), which can be identified in dialogue. Within each game, we can identify individual *moves*, which are defined in terms of speaker intention and dialogue function. This analysis makes it possible to describe an utterance or part of an utterance as a specific move at a specific point within a specific game.

The repertoire of games and moves is based upon a map task (See Anderson *et al.*, 1991, for a detailed description) in which one person who has a map with a path marked on it describes this route to another person with a similar map who then draws the path onto their map. A barrier separates the two participants. The nature of the task is such that the speaker's intentions in the conversation are fairly obvious. Kowtko *et al.* (1992) report that one expert and three naïve judges achieve 83% agreement when classifying conversational moves in two map task dialogues.

Six games appear in the dialogues: Instructing, Checking, Querying-YN, Querying-W, Explaining, and Aligning. They are initiated by the following moves:

INSTRUCT	Provides instruction
CHECK	Elicits confirmation of known information
QUERY-YN	Asks yes-no question for unknown information
QUERY-W	Asks content, *wh-*, question for unknown information
EXPLAIN	Gives unelicited description
ALIGN	Checks alignment of position in task

Six other moves provide response and additional feedback:

CLARIFY	Clarifies or rephrases given information
REPLY-Y	Responds affirmatively
REPLY-N	Responds negatively
REPLY-W	Responds with requested information
ACKNOWLEDGE	Acknowledges and requests continuation
READY	Indicates intention to begin a new game

Since the task involves one player telling the other how to draw a path, the conversation naturally consists of many Instructing games. Games occur in series and may nest within one another. Response and feedback moves may loop within a game.

The prototypical game consists of an initiating move, a response move, and an optional feedback move. The majority of games (84% from a sample of 3 dialogues, $n = 65$) match the simple prototype. Games that do not match this structure are still well-formed, containing extra moves, additional response-feedback loops, or nested games. Very few games (less than 2%) break down as the result of a misunderstanding or other problem.

Here is an example of a prototypical Instructing game. The vertical bar indicates the boundary of a move:

A: Right,|| just draw round it.
READY || INSTRUCT
B: Okay.
ACKNOWLEDGE

INTONATION

Once we have analysed the game structure of a dialogue, we can look for relationships between move type and intonation contour. Kowtko (1992) takes this approach with promising results. However, this procedure presupposes that discourse function, as defined by move type, is the principal factor in determining the choice of contour. This assumption is consistent with much recent work on functional factors influencing intonation (e.g. Hockey, 1991, 1992; Litman and Hirschberg, 1990; McLemore, 1991) and is supported by earlier work on game structure and intonation in task-oriented dialogue (Kowtko, 1992), but it ignores another factor that may be significant, namely the influence of the previous speaker's contour.

The present results use data from map task dialogues: single words which compose moves within themselves (*mmhmm, uh-huh, okay, yup, yes, no, almost, fine, right, okay, aye*[1]). These words typically surface as 6 of the 12 moves in the games analysis: ALIGN, REPLY-Y, REPLY-N, REPLY-W, ACKNOWLEDGE, and READY. The data set consists of 100 out of 151 single word moves spoken by four conversants in two entire dialogues. To avoid interference with pitch accents in larger intonational phrases, words which form partial utterances are excluded (the other 51 moves). The intonation of each word has been transcribed as high level, low level, rise, fall, rise-fall, or fall-rise.

When categorized according to *move* (specific function) and position in *game* (discourse context), trends emerge from the data. Results are summarised in Table 1.

Table 1: Intonation Associated with Move

Move	*Preceding Move*	*Tune*	*Data*
ALIGN		Rise	6 of 7
REPLY-Y, or REPLY-N, or REPLY-W	ALIGN ALIGN, embedded	Fall Level	6 of 7 4 of 6
REPLY-Y, or REPLY-N	QUERY-YN	Fall	13 of 16
REPLY-Y	CHECK	Level	1 of 1
ACKNOWLEDGE	INSTRUCT	Level	18 of 36
ACKNOWLEDGE	EXPLAIN	Level	2 of 3
ACKNOWLEDGE	CLARIFY	Fall ⋆	4 of 7 3 of 7
ACKNOWLEDGE	ACKNOWLEDGE	⋆	3 of 4
ACKNOWLEDGE	REPLY-Y, or REPLY-N, or REPLY-W	⋆	10 of 12
READY		Fall	1 of 1

It has been proposed (e.g. Brazil *et al.*, 1980, Brown Currie and Kenworthy, 1980) that the pitch range of one speaker can influence the pitch range of another. Results in the starred (⋆) categories of Table 1 suggest that the intonation contour of one speaker can influence the intonation contour of another speaker. When an ACKNOWLEDGE move follows a response move (CLARIFY, ACKNOWLEDGE, or a REPLY), the relative final height of its intonation contour matches the relative final height of the contour in the last utterance spoken by the other conversant (70% of the time). Final heights are judged within a speaker's own pitch range.

CONCLUSION

The data here supports the view that intonation contours may be influenced by those of the previous speaker's utterance. While this is a preliminary study it nevertheless provides sufficient evidence of interesting trends to support further work.

[1] Participants in the map task were undergraduates at Glasgow University, and therefore spoke Scottish English.

REFERENCES

A H Anderson, M Bader, E G Bard, E Boyle, G Doherty, S Garrod, S Isard, J Kowtko, J McAllister, J Miller, C Sotillo, H Thompson, and R Weinert, 'The HCRC Map Task Corpus,' *Language and Speech*, **34**:351-366 (1991).

D Brazil, M Coulthard and C Johns, *Discourse Intonation and Language Teaching*, Longman, London (1980).

G Brown, K Currie, and J Kenworthy, *Questions of Intonation*, Croom Helm, London (1980).

B A Hockey, 'Prosody and the interpretation of "okay",' Presented at the *AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation*, November (1991).

B A Hockey, 'Prosody and the interpretation of cue phrases,' Proceedings of the IRCS Workshop on Prosody in Natural Speech, University of Pennsylvania, IRCS Report No.: 92–37, 71-77 (1992).

J C Kowtko, 'Comparing intonational form with discourse function: A study of single word utterances,' Proceedings of the IRCS Workshop on Prosody in Natural Speech, University of Pennsylvania, IRCS Report No.: 92–37, 79-82 (1992).

J C Kowtko, S D Isard and G M Doherty-Sneddon, 'Conversational games within dialogue,' Research Paper HCRC/RP-31, Human Communication Research Centre, University of Edinburgh (1992).

D Litman and J Hirschberg, 'Disambiguating cue phrases in text and speech,' COLING-90 *Proceedings*, 251-256 (1990).

C A McLemore, *The Pragmatic Interpretation of English Intonation: Sorority Speech*, Ph.D. dissertation, University of Texas at Austin (1991).

R Power, *A Computer Model of Conversation*, Ph.D. dissertation, University of Edinburgh (1974).

Word-prosodic features in Estonian conversational speech: Some preliminary results

Diana Krull
Stockholm University, Institute of Linguistics
S-106 91 Stockholm, Sweden

ABSTRACT

Estonian has three distinctive degrees of quantity: short, long and overlong. This paper reports an investigation on the temporal and tonal correlates of quantity in the natural conversation of one Estonian speaker. The results show statistically significant differences between quantities only for the temporal correlates. The tonal correlates had a considerable overlapping between quantities.

INTRODUCTION

Over the years, a large amount of investigation has been addressed to the Estonian quantity system. A number of theories about the nature of the three distinctive quantities and their acoustic correlates have been put forward. (For an overview, see Lehiste 1970; a new theory is presented by Eek and Help 1987). The phonetic material on which the theories could be based, consisted of so called "laboratory speech", that is, words or sentences prepared by the investigator and read by the speaker. More recent work has shown that more naturally produced speech can differ considerably from such laboratory productions (see e.g. Lindgren, Krull and Engstrand (1987). In particular, language dependent differences in the stability of temporal cues to quantity have been found (Engstrand 1992). The temporal cues for Swedish seem to be less stable than in Finnish. The probable reason is that in Swedish, quantity is not signalled by temporal cues alone, there is also a clear difference in vowel quality.

It can be hypothesized that if an acoustic parameter is used as a primary or only cue to phonological distinctions in a language, then the freedom of the speakers of the language will be restricted. That is, differences in the acoustic parameter should remain robust in any speaking style. Therefore, an investigation of natural Estonian conversation is important for two reasons: It can help to test the hypothesis, and, at the same time, throw some more light on the question relative importance of different cues to quantity in Estonian.

Estonian has three phonologically distinct degrees of quantity, short (Q1), long (2) and overlong (Q3). They are signalled by the duration ratio between the first (main stressed) and the second syllable of a word. The typical ratio for Q1 is 2:3, for Q2 3:2 and for Q3 2:1 (Lehiste 1960). To distinguish Q3 from Q2, listeners use an additional, tonal cue (Lehiste 1970): falling F0 for Q3 and flat or slightly rising for Q2. Earlier results with words read in isolation and in a carrier phrase (Krull 1993) showed that the duration relation between syllables remained stable even when the syllables involved were shortened. In most cases, also the falling F0 contour remained stable.

Other tonal cues described in the literature but not studied in Krull op. cit., are F0 stepdown from the end of V1 to the beginning of V2 (Lehiste 1970b), and an earlier location in time of an F0 maximum within V1 for Q3 (Eek 1990).

The aim of the present study is to assess the stability of these cues in natural conversational Estonian speech, to begin with, of one speaker.

METHOD

The subject was a male phonetician, native speaker of standard Estonian, resident in Estonia. Seated in an anechoic chamber together with the author and prompted by a few short questions, he related episodes from his childhood, schooldays and travels. The talk - over an hour and a half of a lively near-monologue - was recorded digitally. Lexical non-compound Q2 and Q3 words of the form (C)V1CV2(CV3...) were located and sampled at a rate of 10 kHz/s. V1 was a long or overlong vowel/diphthong, C a short consonant. This form was chosen because the exact duration of single short consonants in syllable initial position is of no consequence for the quantity degree and therefore V1 and V2 can be used to represent syllables (Lehiste 1960). The three-way quantity contrast is connected to the two initial syllables of a word, of which the first carries the main stress.

For the analysis, the Kay CSL 3400 system was used. First, the duration of the vowels was measured. The beginning and end of the vowel were defined as the onset and and offset of a clear formant pattern. In the case of stops, the burst release was defined as the beginning of the vowel. The aspirative phase sometimes appearing at the end of vowels, especially in word final position, was not included. Thereafter, the duration ratio V1/V2 was calculated.

Next, F0 was measured at the beginning and end of V1 and V2. After stops, the beginning was now defined as the voice onset instead of the burst. If there was an F0 maximum within V1, its frequency and temporal location were marked. Utterance final words were excluded. The remaining material consisted of 157 words, 48 of Q1, 45 of Q2 and 64 of Q3. Words where V1 was an (over)long vowel were measured separately from words where it was a diphthong. Moreover, disyllabic words - which were in a clear majority - were measured separately from words of three and four syllables.

Mann Whitney U-tests showed no significant change in syllable duration due to the lengthening of the word as was reported by Krull (1993) for laboratory speech, nor any significant difference between long vowels and diphthongs. Therefore, these groups were analyzed together.

RESULTS

Duration ratios and F0 data are shown in Table 1. The data were, in general, not normally distributed, therefore median values are given instead of means, and the Mann-Whitney U-test was used. The test showed statistically significant differences between Q1-Q2, and between Q2-Q3, both for the durations of V1 and V2 and for the V1/V2 ratios. For all cases $p<.001$. The ratios are shown in the so called "box and whiskers" plot in Figure 1.

The corresponding differences in F0-change, on the other hand, were not statistically significant, although there was a tendency for F0 to fall more during V1 in Q3 than in Q2 words. There was also a tendency for F0 to stay unchanged between the end of V1 and the beginning of V2 in Q3 words, and to fall in connection with Q2, but even here the difference was small and the variation considerable.

An F0 maximum within V1 was found in 27% of the Q2 words and in 59% of the Q3 words. The median F0 rise was about 8.9% of the vowel-initial value for Q2 and 7.5% for Q3. The location of the maximum on the time axis was earlier in Q3 than in Q2 words: for Q2 median location of the maximum lay at a point 55% of the entire duration of V1, for Q3 the corresponding value was 44%. However, the difference in location was not statistically significant.

Finally, the material was checked for possible correlations between the different acoustic correlates to quantity. No such correlation - positive or negative - was found.

Table 1. *Median, mininimum and maximum of the duration of V1, V2, and of the V1/V2 ratio. Percent F0 change between the beginning and end of V1, and between the end of V1 and the beginning of V2.*

	Duration				**ΔF0 (%)**	
	N	V1 (ms)	V2 (ms)	V1/V2	V1	V1-V2
Q1 med	48	82	109	.7	0.3	2.4
min		56	69	.3	-7.0	-35.3
max		115	194	1.3	44.4	28.9
Q2 med	45	134	78	1.8	.8	-6.1
min		74	37	1.0	-40.1	-42.2
max		244	149	3.2	52.5	21.9
Q3 med	64	180	56	3.2	-4.5	-2.5
min		48	23	1.0	-33.8	-28.3
max		323	132	8.1	62.5	32.2

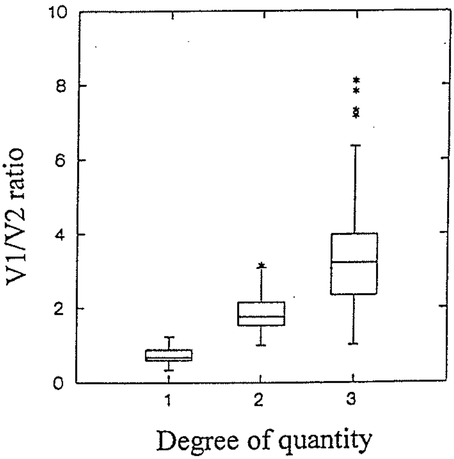


Figure 1. *Box-and-whiskers plot of the V1/V2 ratio in Q1, Q2 and Q3. The boxes represent the spread of 50% of the sample, 25% (quartile) above and 25% below the median. The whiskers (above and below the boxes) are drawn to the nearest value not beyond a standard span of the quartiles. The points beyond that are represented by asterisks and circles (For details see Velleman and Hoaglin 1981, p. 65f).*

DISCUSSION

The hypothesis presented in Introduction, if true, would predict that the temporal correlates to Q1 and Q2, and at least one - temporal or tonal - correlate to the distinction between Q2 and Q3 would remain stable. That is, there should be some acoustic differences between quantities found in "laboratory speech" that remain stable even in a more spontaneous speaking style.

The results of this investigation showed that - for this one speaker - the most stable acoustic difference between quantities was temporal. Not only the duration ratios, but even the absolute durations of V1 and V2 were significantly different between quantity degrees. Although there was always a certain overlap, it involved relatively few items. A comparison of these results with temporal data from CV(V(V))CV words read in a carrier phrase by the same speaker (Eek 1975) showed, moreover, that the temporal differences between quantities were enhanced rather than weakened in a more natural speech style.

The tonal cues, on the other hand, were present only as a statistically non- significant tendency. The stability of the temporal correlates to quantity in Estonian therefore suggests that they may be the most crucial cues for signalling quantity. However, further investigation with more speakers is necessary.

REFERENCES

Eek, A. (1975). "Observation on the duration of some word structures II". *Estonian Papers in Phonetics (EPP)* 1975, Academy of Sciences of the Estonisn S.S.R., Tallinn, 7-51.

Eek, A. (1990). "Units of temporal organization and word accents in Estonian". *Linguistica Uralica XXVI*, 251-264.

Eek, A. and Help, T. (1987). "The interrelationship between phonological and phonetic sound changes: A great rhythm shift of old Estonian." *Proceedings of the 11th ICPhS,* Tallinn 1987, Vol. 6 pp. 218-233.

Engstrand (1992). "Durational correlates of quantity in Swedish and Finnish: Data from natural speech". *Papers from the Sixth Swedish Phonetic Conference, Gothenburg, May 20-22, 1192*, 47-50.

Krull, D. (1993). "Temporal and tinal correlates to quantity in Estonian", *PERILUS XV*, Institute of Linguistics, Stockholm University.

Lehiste, I. (1960). "Segmental and syllabic quanity in Estonian", *American Studies in Uralic Linguistics.* Indiana University, Bloomington, Vol. 1, 21-82.

Lehiste, I. (1970a). *Suprasegmentals*. MIT Press, Cambridge, Mass.

Lehiste, I. (1970b). "Experiments with synthetic speech concerning quantity in Estonian", *Congressus Tertius Internationalis Fenno-Ugristarum*, Pars I, Acta Linguistica Tallinn.

Lindgren, R., Krull, D. och Engstrand, O. (1987). "Akustiska studier av fonetisk variation i svenskan", i P. Linell, V. Adelsvärd, T. Nilsson och P.A. Pettersson (eds.), *Svenskans beskrivniong **16**. Förhandlingar för att dryfta frågor rörande svenskans beskrivning.* Vol **2**, 326-338.

Velleman, P. and Hoaglin, D. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. D. Duxbury Press, Boston, Mass.

Prosodic aspects of reported speech

Sylvie LEANDRI
Université Paris7
UFR Linguistique, Laboratoire de Phonétique

ABSTRACT

Trying to characterize reported speech compare to direct speech, I discuss some prosodic properties thanks to a statistic study of the mean f0 values measured at different level in the utterances.

INTRODUCTION

My aim is to characterize phonetically so-called "heterogeneous" utterances (reported speech), which eventually will contibute to give them a linguistic status. My first hypothesis came from regular observations I made while studying French spontaneous speech. To support them on a "phonetics-lab" French basis, I built a corpus containing different types of speech-styles in two types of environment: a long one (tales); and a short one (isolated sentences). Here I am dealing with the second type.

In a preliminary study (on the "tales" corpus), I observed that contrary to common thoughts and compare to other types of utterances, a reported sentence (direct style) is not preceded by a significatively long pause; I noticed also that important cues can be observed on boundaries.

In this study, I compare first the global characteristics of utterances according to the context which is used to define their speech style. Then I compare what happens at the boundary parts between the introductory sentence and the "target" utterance.

CORPUS

The corpus is made of five different "target" utterances read by 5 male and 6 female speakers in 4 different contexts which are expected to induce different speech- styles:
N= Neutral Declarative (control context, isolated target) .
D= Direct style (X said: "target").
I = Indirect style (target introduced by "que").
Q= Quotation (embedded target with no special introductory marks).
The length of the "targets" vary between 8 and 17 syllables; the syntactic structures are quite different; the expressive and semantic weights of the utterances vary also.
The target sentences are:
" C'était sur que ça marcherait pas"; "Selon les pompiers, la bombe aurait été placée dans les cuisines"; " Il reviendra ce soir à 5 heures"; "J'étais la honte de la famille"; "Hier soir, tous les syndicats appelaient à la grève".
The speakers had to read for example:
a) J'étais la honte de la famille. (N)
b) Je le dis sans malice: " J'étais la honte de la famille". (D)
c) Il était évident que j'étais la honte de la famille. (I)
d) Son nouveau tube 'J'étais la honte de la famille' entre 9ème au Top 50. (Q)

The expected realization of each target sentence is supposed to be triggered,or induced, by its context "N", "D", "I", "Q".

RECORDING CONDITIONS AND ANALYZE FACILITIES

The corpus was recorded in the recording room of the phonetics laboratory on a Marantz tape-recorder, then sampled at 11khz with a MacRecorder.
The analysis was made with E. Keller's program "Signalyze 2" on a McI Quadra.
The statistics were computed with "Stat-View".

VARIABLES CONTROL

The first measurements were necessary to evaluate the two variables:
a) male / female speakers . b) different syntactic structures.
A two factors Anova computed on the "mean f0" of the target sentences shows that the male/female factor is relevant (P < 0,005) (obviously we are dealing with two different registers), while the syntax factor is not: These two factors are independant.
On the "N" sentences group, the anova made on mean f0 between the different speakers, shows the same partition between male and female readers. Thus it is not possible to use parametric statistics tests.

COMPARISON BETWEEN THE DIFFERENT "TARGET" SENTENCES

The parameters which I choosed to be measured are: mean f0 of the whole sentence, mean f0 of the 2 first syllables (300ms), and of the two last syllables (500ms).

In each case, there is a double question: Are the "targets" different on an absolute basis, or are they different according to the fact that they have been produced by the same speaker in different contexts.

To answer the first question, the H value of the KRUSKAL-WALLIS test is computed, while for the second question the FRIEDMAN test (χ^2) is used as the two compared values are paired.
As the final f0 values are compared on two groups only (N and Q), the U value of the MANN-WHITNEY test is used in the first case , and the WILCOXON test in the second.
The results are given in Table 1:

Table 1: *Comparison of f0 values according to the contexts.*

(P is the percentage of probability for the result to be at chance level.
* indicates that the test shows signicant differences between the different contexts).

	unpaired tests	paired tests

onset syllables	H=5,11
	P=16%
		Chi2=29,76
		P=0% *

global mean f0	H=4,49
	P=21%
		Chi2=29,92
		P=0% *

Final syllables "Q"vs"N"	U=294 et Z=-2,31
	P=2% *
		Z=-4,7
		P=0% *

According to the 1st column, the general and the onset mean f0 values are not significatively different, while the final f0 value is clearly different.
According to the 2nd column (paired tests), the differences are clearly significant.
The two following graphs illustrate Table 1:

figure 1: *target sentences "3"*

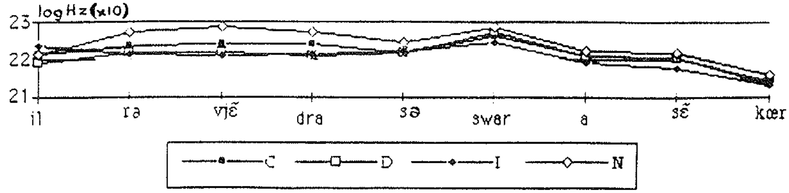


Fig.1 shows that the first part is distinctive: the onset slope is negative in the case of "I", while it is positive in the other cases. This is generally true even if it is not always so well marked (P= 0% between "N" and "I"; P< 4% between "N,I,D" and "Q,P").

figure 2: *target sentences "4"*

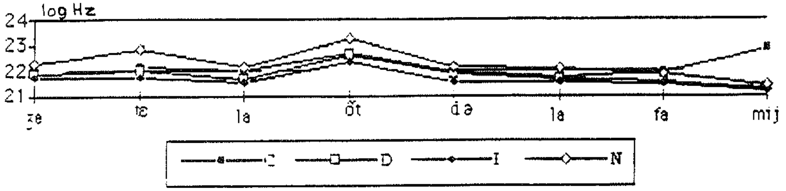


Fig. 2 shows the positive slope on the 2 last syllables in the case of embedded sentences ("Q").
As far as global mean f0 values are concerned, it appears that "N" target utterances have higher values (N > C > D > I).
These first results indicate that we must consider what happens at boundary.

COMPARISON BETWEEN THE "TARGET" AND ITS CONTEXT

Here, paired tests are not necessary as it is the same utterance which varies in different contexts.

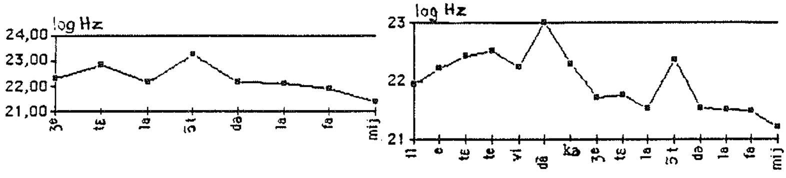
According to the results, the 2 parts (introductory and target) are clearly different, except in the case of the mean f0 values in the "D" group and the onset f0 values in the "I" group.

Of course, one must be cautious in interpreting these results as the utterances and their length being different, there is an interaction between the value of f0 heigths and the declination line. This is particularly the case of the "Q" embedded sentences; but if we compare the mean f0 values of the introductory part, of the target part and of the context final part, the differences are significant.

These observations lead me to think that there is an interaction with the place of the measured segments within the utterances, and that comparing the mean values reveals the height variations due to the place.

We can now have a look at what happens at boundaries, having in mind that there may be a pause between the two sentences.

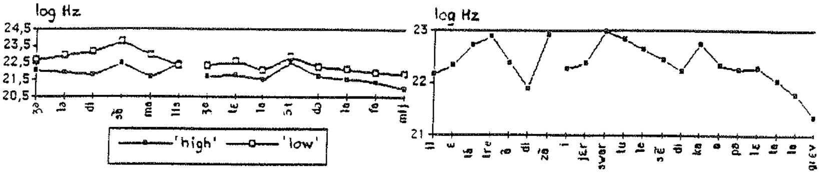
figure 3
("N"4 & "I"4)



If we compare "N" and "I" groups, we can see that in the case of "I" sentences, the onset of the target has a negative slope, going on with the general fall of the introductory part. If we think of what is generally said, from a linguistic point of view, about indirect style, we may conclude that "I" utterances are "Monolithic", while the 2 others, "D" and "Q", are "dual".
If we remember that the mean f0 of the "I" utterances has generally lower values, we may say that these targets are lowered along the expected declination line, because they are not the beginning of an utterance but always the ending part of some introductory sentence.

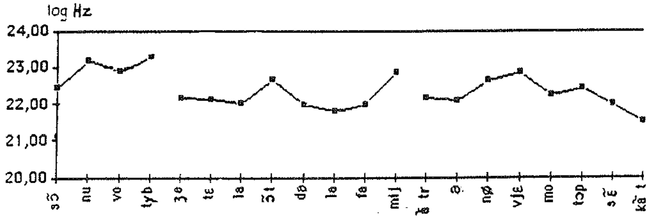
figure 4

("D"4 & "D"5



In "D" utterances there is a supplementary feature: a pause. In a previous study, I claimed that there is a relationship between the existence of a pause and the value of f0 on both sides of the boundary. It is difficult to precise the nature of this relationship because speakers may adopt 2 different strategies: In one case, which I call the "high" solution, there is a rise before the pause and the target has a lowered onset. In the second case, there is no preceding rise and the onset of the target is on the same level, or even on a slightly higher one.

figure 5 ("Q"4)



In the "Q" group the same double possibility is used, and there is always a pause, at least perceived if not silent. In this group it must be emphasised that the transition from the target to the end part of the context is always made according to the "high" strategy, with or without a pause.

CONCLUSION

I hope to have shown in this study that there are prosodic differences which I tried to systematize between different types of reported speech.

Indirect style can be contrasted with the others as it is constituted by one prosodic group only with a continuously falling f0 line on the whole utterance, the introductory morpheme "que" being never stressed.

In direct style, there is a correlation between an inter-sentence pause and the values of f0 heights before and after the pause; two strategies can be observed: rise/fall vs. continuous fall.

It will be worth to examine more precisely the case of "Quotations", compare to "free indirect style" in order to dicuss their status in linguistic theory.

Intonational grouping, boundaries, and syntactic structure in French

Piet Mertens
K.U.Leuven, Department of Linguistics
Blijde-Inkomststraat 21, 3000 Leuven (Belgium)

ABSTRACT

This paper describes the units and rules involved in (1) the lexicon-prosody interface, (2) the intonation grammar, (3) prosodic grouping, and (4) the syntax-prosody interface. All units are depicted in a multi-layered representation of prosodic structure.

INTRODUCTION

Given an arbitrary sentence several intonation patterns can be generated that are equally well-formed with respect to the grammar of intonation, although in a particular contextual setting some may be more appropriate than others. The acceptability of an intonation pattern depends on several factors: (1) the syntax of intonation units, (2) the alignment of these units with the segmental level, (3) the interaction between prosodic structure and syntactic structure, (4) the semantics of intonation units.

This paper will be restricted to the second and third aspects. The intonation model used and the semantics of basic intonation units have been described in Mertens [1987, 1990, 1992, 1993]. Although this paper will be about French, it states which parameters (in the sense of Hirst [1991:306]) of the model are language-specific and which are not.

At the level of prosodic form, the intonation group (IG, for short) is just a sequence of one or more tones that complies with the internal syntax of the unit. Our analysis for French is shown below. The tone paradigms (NA, AI, AF), their composition, and the internal structure are language-specific. The brackets indicate optional parts.

```
IG → ((NA) AI) (NA) AF (NA)
```

The structure of the IG was defined on the basis of the data in a corpus, in such a way that each utterance could be analysed as a sequence of IGs. But the definition of the IG does not explain how sequences of IGs can be generated for arbitrary sentences, i.e. how the unit should be mapped onto the segmental chain. It will be shown that, in order to do so, an intermediate layer is needed between the syllable and the IG, as well as a rule on the grouping of syllables. Moreover an additional layer on top of that of the IGs is needed to represent the prosodic grouping of IGs; there will also be rules stating the grouping mechanism and the constraints from (linear) syntactic structure.

GENERAL OVERVIEW OF DESCRIPTIVE UNITS AND RULES

The segmental and suprasegmental chains appear to be interwoven in a rather complex way. The suprasegmental structure results from the use of tones which in turn are constrained by both morphemic and syntactic properties of the segmental chain. These relationships call for a *multi-layered representation* of prosodic structure in which each layer introduces a new descriptive unit of a higher level, resulting from the grouping of units of the preceding level. The following table gives an overview of the factors at play for each of these layers, and indicates the corresponding prosodic units.

sub-units grouped	resulting unit	stress type	factors	rule nature
-	syllable/word	potential (word stress)	morphemic/ lexical	deterministic
syllables/words	stress group	potential	syntactic	deterministic
stress groups	intonation group	observable (accent)	syntactic semantic	non-deterministic
intonation groups	package	observable (accent)	syntactic semantic	non-deterministic

(1) A *word* (or a *syllable*, cf. infra) is either *clitic* or *non-clitic* depending on lexical or morphemic properties (cf. infra). Non-clitic words are commonly said to carry *word stress*, but this is indeed merely a potential stress since non-clitic words need not be stressed always. (2) Clitic words are reorganized around the non-clitics, on the basis of the syntactic relation between the clitics and the non-clitic. The resulting unit, the *stress group* (SG), still carries virtual stress for the reason given above. (3) An *intonation group* (IG) is obtained when the speaker selects a particular tone sequence (from the sequences allowed by the intonation grammar) and combines it with a part of the segmental chain corresponding to one or more contiguous SGs. (4) Whereas the use of a tone sequence gives rise to the formation of an IG, the actual choice of the stressed tone (among other tones from the AF position) will determine the possible *grouping* of this IG with adjacent IGs in the chain. We use the term *package* to refer to a group of one or more IGs linked by the prosodic grouping mechanism. The following sections describe the units and rules for the different layers, starting with the lower one.

THE [±clitic] FEATURE

When a word can by its own form an SG or can constitute the centre of an SG and hence become an IG, it is non-clitic. Several criteria have been put forward to define the classes of clitic and non-clitic words: (a) *lexical identity*: for each lexical morpheme, state its [± clitic] feature (e.g. "man" is [-clitic]), (b) *grammatical category* (e.g. nouns are [-clitic]), (c) *syntactic function* (e.g. the subject is [-clitic]), (d) *syntactic category* (e.g. the noun phrase is [-clitic]). The difficulties for each of these criteria are discussed to a great length in Mertens 1993. In the following we will assume criterion (b) can be used.

In French, a language with *bound stress*, stress position is *predictable* from the limits of the SG; so there is no need to specify which syllable in the word carries stress. (This would be necessary in *free stress* languages.) For this reason, one can speak of [±clitic] syllables, where a non-clitic syllable is one that carries stress in a non-clitic word.

STRESS GROUP FORMATION

In the following examples clitic syllables are indicated by dots on the layer "WS" (for word stress) and non-clitic syllables by o signs; on the layer "SG", boundaries of SGs are indicated by brackets, and stressable syllables by the minus sign.

```
(1)      tu le vois, cet admirable bateau?    vois-tu le problème?
    WS    .  .  o     .  .  . o    .  o         o   .  .   .  o
    SG  ( .  .  - )  (.  .  . - ) (.  - )     ( .   -)(.   .  -  )
```

Rule 1: A stress group is made up of a non-clitic word N_1 and all contiguous clitic words that are governed by N_1 (as in *tu le vois* or in *vois-tu*) or by another non-clitic word N_2 which in turn governs N_1 (as in *cet admirable*).

As can be seen from *vois-tu* in (1), the stressable position can shift away from a non-clitic to a clitic syllable, at least in bound stress languages. It is this phenomenon which explains the need for two levels (WS and SG) in those languages.

However, when a clitic is separated from its governing non-clitic by one or more constituents with which it has no syntactic relation, that clitic too forms an SG (as for *qui*).

```
(2)      René,  qui, tu le sais bien, ne nous aime pas tellement ...
     WS   . o    .    .  .   o    o    .   .   .    .   .    o
     SG ( . -) ( - ) (.  .   - )(.- ) (.   .   - ) (- )(.    -  )
```

INTONATION GROUP FORMATION

Example (3a-c) shows different forms of the same utterance, with a varying number of IGs. The IGs are surrounded by brackets on the layer labelled "IG". A plus sign indicates a syllable with a tone of the AF tone paradigm. The speaker can map an IG to every SG (as in (3a)), or he can combine many SGs into one IG (as in (3b)), provided the following syntactic constraint is met: an SG should be grouped with the SG it depends upon syntactically before it can be grouped with any other SG. So (3c) is wrong because *ainsi* is governed by *voir* and not by *attriste*.

```
(3)      de la voir ainsi  m'attriste beaucoup
     WS   .  .   o   .  o     .  o      .   o
     SG ( .  .   -) (.  -) (  .  -  ) ( .   -)
(a)  IG ( .  .   +) (.  +) (  .  +  ) ( .   +)
(b)  IG ( .  .   .   .  +) (  .  .      .   +)
(c)* IG ( .  .   +) (.  .     .  +  ) ( .   +)
```

Rule 2: An IG is made up of one or more contiguous SGs that are syntactically governed by an element that appears in the linear chain making up the IG.

Since a sequence of SGs can be arranged in several ways, i.e. with a variable number of IGs, the question arises about the semantic effect of the different arrangements. Multiple SGs merged into a single IG show a larger semantic cohesion; the IG forms a semantic unit. Moreover when tones with paradigmatic effects (such as *focus* or *paradigmatic contrast*) are used, their scope is limited to the elements within the IG, so the limits of the IG are essential for the semantic interpretation of the intonation pattern.

Mertens [1992] gives a further rule (Rule 3) that accounts for phonotactic constraints (the interference between syllable count and tones) in IG formation.

PROSODIC GROUPING AND PACKAGES

Any series of IGs shows an internal grouping which depends on the tones being used. In French this grouping is determined by the tones of the AF position, whereas in Dutch it is linked to tones of unstressed positions. Thus the grouping mechanism is once again language-specific. The rule for French follows (cf. Mertens[1990], *Règle de dominance*):

Rule on prosodic grouping: For any two successive IGs: if the tone in the AF position of the last IG dominates that of the first IG, then there is an embedding effect of the first IG within the second; otherwise, the two IGs are independent (juxtaposition).

Prosodic grouping is recursive: it can be applied to the units formed by a previous step. The term *package* is used to refer to the result of a grouping operation; a package contains one or more IGs.

PROSODIC GROUPING AND SYNTACTIC STRUCTURE

Many authors note that prosodic grouping has to conform to syntactic structure. It is commonly held that the prosodic boundaries (and therefore the prosodic grouping) should be *proportional* to the syntactic boundaries. This *correspondence view* is challenged by spontaneous speech data, as shown in Mertens [1993], where the following types of non-agreement are found.

1. Since the limits of packages are determined by those of the IGs, and eventually by those of the SGs, and since the latter can be either a constituent, part of a constituent, or more constituents, packages need not have the dimension of constituents.

2. The correspondence view implies the impossibility of having a major prosodic boundary at an internal boundary of a complex syntactic constituent. However, the speech data shows cases where a package merges a first constituent with only a part of the next constituent.

```
(4)     la lecture n'était pas eh un niveau auquel on s'intéressait
    WS   .  .  o     .  o   o  .   .  .  o   o  .   .   .  . .   o
    SG  (.  .  - ) ( .  -) (-) (.  .  .  -) (.  -) (.   .  . .   -)
    IG  (.  .  + ) ( .  .   +) (.  .  .  +) (.  +) (.   .  . .   +)
    IG   l...l HH   l....l HH  l. ....l  HH  ll/LL \l........l   HH
    P   (   1    ) (    2    ) (     3     ) ((    )    4         )

        quand on faisait une théorie de la littérature
    WS    o   .    . o    .    .. o   .  .  .  . . o
    SG  ( - )(.    . - ) (.    .. -) (.  .  .  . . - )
    IG  ( .   .    .  +) (.    .. +) (.  .  .  . . + )
    IG   l.........l/LL  \l.....l HH  l.........\l L-L-
    P   ((      5      )      5     )
        (((     6      )      6     )        6       ))
```

3. Since the prosodic grouping can only render a limited amount of hierarchical relations, the mechanism is unable to reproduce the entire syntactic structure, even for sentences of moderate syntactic complexity. So, at some point the agreement will fail.

4. Especially with contrasting tones (such as HL) one can find one or more constituents as the unstressed part of an IG, even with syntactically dominant elements.

These facts point to a new criterion for the syntax-prosody agreement.

Rule 4: IGs can be grouped in a package, and packages into larger packages, if the grouped elements are linked by a valency relation. There is no requirement for the inclusion of complete constituents.

The explicit representation of the levels of WS, SG; IG, and packages in the prosodic structure enable the relations between intonation, syntax, and morphology to be pinpointed and will hopefully bring about a better understanding of their interaction.

REFERENCES

Hirst, D. (1991) Intonation models: towards a third-generation, *Proc. ICPhS* 12, vol. 1, 305-310.

Mertens, P. (1987) *L'intonation du français. De la description linguistique à la reconnaissance automatique.* Unpublished Ph.D., 2 vol., 317+90 pp.

Mertens, P. (1990) Chap. IV. *L'intonation,* in .Blanche-Benveniste, C.; Bilger, M.; Rouget, C. & Eynde, K. van den (1990) *Le français parlé, Etudes grammaticales.* Paris: Ed. du CNRS, p.159-176.

Mertens, P. (1992) L'accentuation de syllabes contiguës. *ITL-Review of Applied Ling.* 95-96, 145-165.

Mertens, P. (1993, to appear) Accentuation, intonation et morphosyntaxe, *Travaux de Linguistique*

Syntax, Pauses, and Temporal Relations in the Final Part of the Sentence

Peter Molbæk Hansen, Niels Reinholt Petersen, and Ebbe Spang-Hanssen
Department of General and Applied Linguistics
University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen S., Denmark

ABSTRACT

The paper is concerned with sentence final phenomena, viz. the question of final lengthening, and the occurrence of pauses in Danish news reading. The main findings were (1) that a complementary relationship exists between preboundary lengthening and the occurrence of a silent interval, sentence internally as well as between sentences; and (2) that there is a tendency for pauses to be overrepresented toward the end of the sentence. It is speculated that the latter tendency is related to information structure.

INTRODUCTION

In a previous investigation of the pattern of occurrence of perceived pauses in 174 read-aloud sentences from a Danish news broadcast (Molbæk Hansen, Spang-Hanssen & Reinholt Petersen, forthcoming) we examined the relation between syntactic boundaries and the occurrence of sentence *internal* pauses and the acoustic realisation of such pauses.

The present paper reports on further analysis of the material. We focused on pausal phenomena in the final part of the sentence, more specifically, we were interested in the following two questions: (1) does Danish have sentence final lengthening, and (2) does the distribution of pauses in the final part of the sentence (henceforth late pauses) deviate from that of pauses in the sentence as a whole?

SENTENCE FINAL LENGTHENING?

In the investigation referred to above, we showed a perceived pause at a syntactic boundary to be realised *either* as a moderate lengthening (ca. 3.5 centiseconds) of the last syllable followed by a silent interval, *or* as a marked lengthening (almost 10 centiseconds) of the last syllable before the boundary without a following silent interval.

Here we shall compare this result with observations of syllable durations before the (stronger) boundaries *between* sentences. As in the previous investigation syllable durations were measured in the five syllables preceding the boundary. The median sentence final syllable durations are displayed in figure 1 together with the corresponding median durations before sentence internal boundaries.

The boundaries between sentences were always accompanied by silent intervals, and it is seen from figure 1 that the sentence final syllable durations correspond very closely to the durations obtained for sentence internal durations before silent intervals. Thus in Danish there seems to be a tendency towards a complementary relation between preboundary syllable lengthening and the insertion of a silent interval irrespective of the type of syntactic boundary associated with the pause.

PAUSES IN THE FINAL PART OF THE SENTENCE

In our previous investigation we located pauses in the material and described their relationship with the syntactic boundaries derived from a rather finegrained surface syntactic analysis. The great majority of the pauses (1136) did in fact occur at the syntactic boundaries as defined by us, while some pauses (88) did not occur at such boundaries. Further, there seemed to be a tendency for pauses of both types to be more frequent towards the end than at the beginning of the sentence.

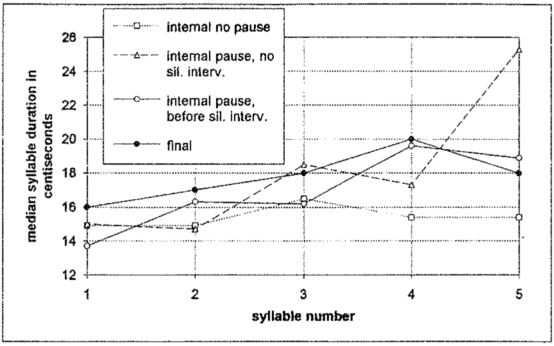


Figure 1. *Median duration (in centiseconds) of the last five syllables before sentence internal and external boundaries.*

In order to examine this matter further, we determined the position of all pauses. The position of a pause is defined as the number of syllables preceding it in the sentence. Since, of course, sentence lengths vary, the pause position was expressed in percent of the total number of syllables in the sentence, henceforth referred to as *relative position.*

Syntactically defined late pauses

Figure 2. shows, for all boundary types pooled, the distribution of syntactically defined pauses on 20 percent intervals over the sentence. The height of the bar within an interval indicates the difference between the percentage of pauses observed in the interval and the percentage to be expected if the pause probability associated with a boundary were evenly distributed over the sentence.

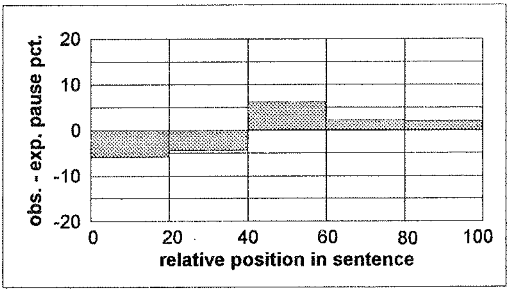


Figure 2. *Distribution of syntactically defined pauses on 20 percent intervals over the sentence. For further explanation, see text.*

As appears from figure 2 the distribution of pauses over the sentence is slightly skewed: the tendency for a syntactic boundary to be accompanied by a pause is greater in the final part of the sentence than in the beginning.

Syntactically undefined late pauses

Figure 3 shows the distribution over the sentence of the pauses which did not coincide with syntactic boundaries, as defined by us. It is evident that there is a pronounced overrepresentation of these pauses in the final part of the sentence: 40 percent of them occur within the last 20 percent of the sentence, and 60 per cent occur within the last 40 percent.

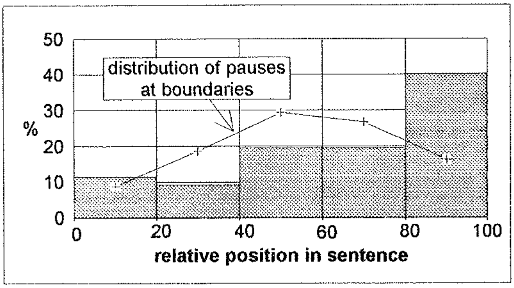


Figure 3. *Distribution of syntactically undefined pauses over the sentence. For comparison, +———+ indicates the distribution of syntactically defined pauses.*

One fourth of the 88 syntactically undefined pauses occurred between preposition and prepositional object and this position might, of course, have been included in our inventory of syntactically defined boundary types, but this would not have changed the general picture: 12 of these, i.e. more than 50 percent, occurred in the last 20 percent of the sentence. The remaining 66 undefined pauses occurred at various positions which were difficult to classify in a syntactically meaningful manner.

DISCUSSION

The results reported above suggest a general tendency for pauses to occur more frequently late in the sentence, whether at syntactically well defined boundaries or not. The explanation of this tendency is thus not to be sought in syntax. In our view, it should rather be considered in terms of the information structure of the sentence and of the text as a whole.

Besides having a grammatical structure related to the participants in the action described (*subject*, *object*, *indirect object*, etc.), a sentence also has a structure related to the information to be conveyed. This information structure is generally considered to be a binary structure, the first part consisting of *given* knowledge, and the second part conveying the *new* knowledge. The first part, the *topic*, is used to identify what the sentence is about, whereas the second part, the *comment* or *rheme*, gives the information intended to make a change in the hearer's mental model. For these concepts, see e.g. Brown and Miller (1991).

The new information is also the information which is most prominent. Therefore, new information/rheme has to do not only with end position, but also with prominence or focus, and consequently with various linguistic means of expressing relative prominence: stress, pitch (Quirk et al. (1972), and certainly also pauses.

The comment part of the sentence may just add new information, but clearly the new information has a tendency to be more clearly signalled when it *deletes* or *restricts* old knowledge. It has been claimed, for instance, that negations are often placed between topic and comment (Sgall et al. (1986)).

If the syntactically undefined late pauses are viewed in this light, a somewhat different pattern appears: For instance, in 9 of the 12 cases of pause after preposition, the prepositional object introduces a *new referent* or *theme*, and in the remaining 3 cases of this kind the prepositional object contains the only information on *time* and *place* in the sentence. A few of the remaining pauses occur just before the *main verb* (which introduces the comment/rheme), and in the remaining cases there is a clear tendency for syntactically unusual pauses to occur in the neigbourhood of *restrictive expressions* with contrastive stress (*not yet, not before, not only, exclusively, purely, for the moment*, etc.) or after adjectival or adverbial modifiers of nouns/adjectives *not representing new information*, i.e. it is the *modifiers* which represent the new (restrictive) information in these cases.

Needless to say, our last remarks are highly speculative, and as long as we have no unchallengeable way of identifying topic, comment, etc., such remarks may seem empirically vacuous. There can be no doubt, however, that more knowledge can be gained on the relation between information structure and prosodic structure (in a wide sense) by investigating systematically how readers treat texts specifically designed to have a well defined information structure.

REFERENCES

K. Brown and J. Miller (1991), *Syntax: A Linguistic Introduction to Sentence Structure* (Harper Collins Academic, London).

P. Molbæk Hansen, N. Reinholt Petersen, and E. Spang-Hanssen (forthcoming), "Syntactic Boundaries and Pauses in Read-aloud Danish Prose", to appear in *Nordic Prosody VI.*

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik (1972), *A Grammar of Contemporary English* (Longman, London).

P. Sgall, E. Hajicová, and J. Panevová (1986), *The Meaning of the Sentence in its - Semantic and Pragmatic Aspects* (Reidel, Dordrecht).

Accenting on Pronouns and Proper Names in Spontaneous Narrative*

Christine Nakatani

Aiken Computation Laboratory, Harvard University
33 Oxford St., Cambridge, MA 02138, USA
Email: chn@das.harvard.edu

ABSTRACT

This paper reports empirical findings on discourse factors affecting the accentuation of pronouns and proper names. We account for the distributions of accent on referring expressions in a narrative sample by proposing a discourse constraint on the interpretation of accent function. Empirical support for this view is presented based on analysis of the discourse functions of accented subject pronouns within the framework of CENTERING THEORY.

INTRODUCTION

In this paper, we account for the distributions of accentuation patterns on pronouns and proper names in a spontaneous narrative. Our account relates the separate but interacting contributions of linguistic form, grammatical position and discourse structure within the framework of CENTERING THEORY.

Previous prosody research has identified GIVEN/NEW information status as an important but not sole predictor of accentuation of referring expressions in elicited spontaneous speech (Brown 1983, Terken 1984). Other factors affecting accentuation include grammatical function and surface position (Terken and Hirschberg 1992) and TOPIC or DISCOURSE STRUCTURE (Brown 1983, Fuchs 1984, Terken and Nooteboom 1987, Hirschberg 1991, *inter alia*). Our account of the accenting patterns on pronouns and proper names integrates several of these factors in a novel interpretion of the discourse function of accent.

THE NARRATIVE STUDY

For the narrative study, 481 noun phrase referring expressions in a 20 minute long, American English spontaneous narrative[1] were analyzed for accentuation, grammatical function (e.g. subject, direct object, object of preposition), and form of referring expression (e.g. proper name, pronoun, definite/indefinite noun phrase). Previous research has shown there is a general tendency for GIVEN information to be accented and NEW information to be unaccented (Brown 1983, Terken 1984). It is also generally thought to be the case that the information status of pronoun referents is GIVEN, and the information status of proper

*This work was partially supported by a National Science Foundation (NSF) Graduate Research Fellowship and NSF Grant No. 447557525 at Harvard University. The author thanks Barbara Grosz, Julia Hirschberg and Mark Liberman for helpful comments.

[1] The narrative was collected by Virginia Merlini for the purpose of studying American gay male speech and was made available by Mark Liberman at the University of Pennsylvania Phonetics Laboratory.

names is NEW. (But see Discussion section). Finally, there is a general tendency for grammatical subjects to represent GIVEN (DISCOURSE-OLD) information and grammatical direct objects, NEW (DISCOURSE-NEW) information (Prince 1988). Our study considers the role of accentuation in relation to grammatical position and form of referring expression.

RESULTS

Overall results shown in Table 1 support the hypothesis that NEW information is generally accented and GIVEN information generally unaccented. Although this trend is significant (p<.001, chistat = 112, df = 1), the accentuation on 19% of the narrative's 294 pronouns and proper names is not consistent with the information status thought to be conveyed by the form of the referring expression itself. These overall results are broken down by grammatical function in Table 2 and Table 3. We conclude that there exist two major non-anomalous linguistic categories of referring expressions whose distributions in naturally occurring discourse need to be accounted for: accented subject pronouns and unaccented object proper names. Below, we present an initial account of these distributional data and then examine the discourse functions of the former class in greater detail.

Table 1: Accenting on all pronouns and proper names

		Accented	Unaccented
Proper names	(n=135)	81%	19%
Pronouns	(n=159)	19%	81%

Table 2: Accenting of pronouns and proper names in SUBJECT position

		Accented	Unaccented
SUBJECT proper names	(n=55)	48	7
SUBJECT pronouns	(n=111)	25	86

Table 3: Accenting of pronouns and proper names in DIRECT OBJECT position

		Accented	Unaccented
DIRECT OBJECT proper names	(n=22)	11	11
DIRECT OBJECT pronouns	(n=15)	1	14

AN ACCOUNT OF THE ROLE OF ACCENTUATION

The data reported above can be explained by the following constraint: If grammatical function and form of referring expression convey conflicting GIVEN/NEW statuses, then accentuation must "reinforce", or agree with the GIVEN/NEW status conveyed by the form of referring expression. A corollary of this hypothesis is that for the cases where a referring expression of a certain form is realized as its preferred grammatical function (e.g. pronouns as subjects, proper names as non-subjects), the speaker is free to use accenting to convey linguistic information apart from GIVEN/NEW status, such as topic shift, emphasis or contrast.

To test the above corollary, we analyzed the 25 cases of accented subject pronouns and determined the possible discourse functions served by accentuation. We found six cases of emphasis or contrast and three cases requiring limited inference to determine the pronoun referent.

The remaining cases can be viewed within the framework of CENTERING THEORY (Sidner 1979, Joshi and Weinstein 1981, Grosz et al. 1983, Grosz and Sidner 1986) as falling into two major classes. For one class (seven cases), accentuation accompanied a GLOBAL shift in discourse context and signaled reference to a previous discourse CENTER that was not salient in the immediate discourse context. Accented pronouns signaling global shifts occurred as the first grammatical subject following the completion of an embedded discourse segment.[2] Also, the pronoun referent crucially was not realized in the immediately preceding utterance for these cases. For the other class (nine cases), accentuation marked a LOCAL shift in attention away from the current discourse CENTER to a new discourse entity that was indeed salient in the immediate discourse context. In contrast to the first class, the pronoun referent always occurred in the immediately preceding utterance, but never as the BACKWARD-LOOKING CENTER.

Figure 1: Examples of accented pronouns (in capitals) signaling local and global shifts in attention

> *so Masson became the new curator –*
> *he flies to london and – you know –*
> *he's already met Anna Freud and therefore*
> *he has access to the secret cupboard of Freudian letters jin jin jin jin jin jin*
> *and naturally Anna assumed that uh –*
>
> > SHE was a brilliant woman too –
> > she did more a lot of work in child psy- psychiatry and psychoanalysis
>
> *assumed that HE would keep this information*
> *you know within the confines of the psychoanalytical group*
> *well as Masson was studying these letters he realized ...*

In Figure 1, the first accented pronoun, *SHE*, illustrates a local shift establishing *Anna Freud* as the CENTER of the embedded subsegment. The second accented pronoun, *HE*, marks a global shift in attention back to the outer segment. This accented pronoun realizes the previously established CENTER of the outer segment, namely *Masson*.

DISCUSSION

This study furthers our understanding of discourse factors affecting accentuation for the genre of spontaneous narrative (cf. Altenberg 1987). Related hypotheses have been put forth to explain why GIVEN information may be accented. For example, Hirschberg 1991 notes that the accenting of GIVEN proper names in a large speech corpus can be explained by the proposal by Sanford et al. 1988 that proper names may be used to "refocus" the speaker's attention on previously established discourse entities that lack prominence in the immediate discourse context. Our study builds upon these previous

[2] Segmental structure was determined by the author using speaker intention, cue phrases, and referring expressions as discourse structure cues (cf. Hirschberg and Grosz 1992). Cue words signaled segment pops in five out of the seven cases of global shift.

findings in two ways. First, we found that accentuation on *pronouns* can cue similar shifts in attention, suggesting that accentuation cannot simply be associated with form of referring expression but rather makes an independent contribution to the structuring of information in discourse. Second, the distinction between LOCAL and GLOBAL shifts in attention makes precise two different notions of lack of discourse prominence in terms of discourse structural properties: roughly, LOCAL shifts occur at the initiation of subsegments and GLOBAL shifts occur at the closing of embedded segments.

We conclude that principles describing the discourse role assigned to accentuation must be sensitive to interactions between grammatical function and form of referring expression, and must incorporate notions of discourse structure to account for two distinctive classes of accented pronouns at discourse segment boundaries. Analysis of other kinds of referring expressions is needed to further test the proposed discourse constraint on accentuation and to refine our understanding of the interaction of accentuation and structural properties of discourse.

REFERENCES

Altenberg, B. *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*, Lund University Press, Lund.

Brown, G. 1983. "Prosodic structure and the Given/New distinction", in (eds.) Cutler, A. and D. R. Ladd.

Fuchs, A. 1984. " 'Deaccenting' and 'default accent' ", in (eds.) Gibbon, D. and H. Richter, *Intonation, Accent and Rhythm*, Walter de Gruyter, Berlin.

Grosz, B. J., Joshi A. K. and S. Weinstein 1983. "Providing a unified account of definite noun phrases in discourse", *Proceedings of 21st Meeting of the Association for Computational Linguistics*, Cambridge, MA.

Grosz, B. J. and C. Sidner 1986. "Attention, intentions, and the structure of discourse", *Computational Linguistics*, 12(3).

Hirschberg, J. 1991. "Pitch accent in context: Predicting intonational prominence from text", to appear in *Artificial Intelligence*.

Hirschberg, J. and B. J. Grosz 1992. "Intonational features of local and global discourse structure", *Proceedings of DARPA Workshop on Speech and Natural Language*.

Joshi A. K. and S. Weinstein 1981. "Control of inference: Role of some aspects of discourse structure centering", *Proceedings of International Joint Conference on Artificial Intelligence*, pps. 385-387.

Prince, E. 1988 (in press). "The ZPG letter: subjects, definiteness, and information-status", in Thompson, S. and W. Mann (eds.), *Discourse Description: Diverse Analyses of a Fund Raising Text*, Elsevier Science Publishers, Amsterdam.

Sanford, A., Moar, K. and S. Garrod 1988. "Proper names as controllers of discourse focus", *Language and Speech*, Vol. 31, pp. 43-56.

Sidner, C. 1979. *Toward a Computationa Theory of Definite Anaphora Comprehension in English*, MIT Technical Report AI-TR-537.

Terken, J. 1984. "The distribution of pitch accents in instructions as a function of discourse structure", *Language and Speech*, Vol. 27, Part 3, pp. 53-73.

Terken, J. and S. G. Nooteboom 1987. "Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information", *Language and Cognitive Processes*, Vol. 2, Nos. 3/4, pp. 145-163.

Terken, J. and J. Hirschberg 1992 (manuscript). "Deaccentuation and persistence of grammatical function and surface position", AT&T Bell Laboratories, Murray Hill, NJ and the Institute for Perception Research, Netherlands.

Prosodic Aspects of Political Rhetoric

Paul Touati
Department of Linguistics and Phonetics
Helgonabacken 12, S-22362 Lund, Sweden

ABSTRACT

In this paper, I addressed the issue of rhetorical prosody in French politics. Overall pitch variation is analyzed as expression of a paradigmatic contrast between a pre-electoral persuasive speech and a pathos-oriented post-electoral press-conference, and a syntagmatic change within the specific setting of the pre-electoral speech. As a result of this analysis, I wish to propose a two-fold categorization of overall pitch variation in French: one in terms of range and the other in terms of register.

INTRODUCTION

In public discourse in general, and in politics in particular, the art of persuasion is often based on a dual rhetorical transaction. It consists on the one hand of a clear identification of the audience to be persuaded – the voters –, and on the other hand of the careful exercize of target-oriented rhetoric by the persuader – the politician himself. In Touati (1991) I explored how some rhetorical features were translated into prosodic categories and specified as acoustic-phonetic properties in the particular setting of a pre-electoral television debate. Here, I addressed again this issue of rhetorical prosody by analyzing contrasts in overall pitch as produced by a French politician (J. Chirac) in two different settings (a pre-electoral speech versus a post-electoral press-conference) and within the specific setting of the pre-electoral speech. As a result of this analysis, I wish to propose a two-fold categorization of overall pitch variation in French: one in terms of range and the other in terms of register.

RHETORICAL PROSODY IN POLITICAL SPEECH

The nature and span of rhetorical prosody may appear with greater clarity if prosodic features could be captured when produced in rather different and specific rhetorical transactions: pre-electoral television debate, pre-electoral speech, post-electoral press-conference; in short and well time-defined discourse events. During pre-electoral periods, persuasive monologues typify the speaking style adopted by politicians. Rhetorical features such as intensifiers, parallelisms and meta-discursive comments (i.e. incidental comments, parenthetical comments or direct quote-comment structures) are then used with dexterity. Intensifiers are often used in order to highlight certain aspects of the argumentation. One way for the speaker to intensify what he is saying is to underline individual words tonally and rhythmically. The function of parallelism is to facilitate monologue processing by reducing information density and increasing redundancy. It is achieved by repeating certain words or phrases. Meta-discursive comments are used to relate the speaker himself, his contender, or a new argument to the topic which is under discussion. Meta-discursive comments are often part of a topic-comment structure produced with specific rhythmic and tonal contrasts. Focal accents apparently work as intensifiers, contrasts in overall pitch being used to express meta-discursive comments and pauses helping to synchronise parallel prosodic configurations. Obviously these three strategies seem to account for typical prosodic means used in French political rhetoric. In the present paper I choose to analyse the opposition pre-electoral versus post-electoral speech because this is when persuasion (when a politician aims to gain votes) gives why to objective pathos (when a politician comments his political victory or defeat). It seems then that overall pitch range is used at a paradigmatic macro-level to establish this contextual opposition. On the other hand, overall pitch variation is also used at a syntagmatic micro-level as in the pre-electoral speech. To persuade and to gain votes seems to demand more refined rhetoric and richer prosody. Chirac's pre-electoral speech is structured in a nearly classical rhetoric manner, its 'dispositio' consisting of an ex-abrupto exordium (prooimion), a narration of the facts (diégèris), a confirmation of the facts (pistis), and a peroration (épilogos) (see Figure 1 in Touati 1993, and for the notion of 'dispositio' and an introduction to rhetoric see Barthes 1970). Interestingly, building blocks for the dispositio are provided by the recurrent use of a direct quote-comment structure (hereafter DQCS) where each DQCS is segmented into three units: the quoting (lead-in) segment, the quoted segment and the comment segment.

ACOUSTIC-PROSODIC ANALYSIS

The acoustic-prosodic analysis was done within a methodological framework developed in a multilingual research project called Contrastive Interactive Prosody ('KIPROS') conducted at the Department of Linguistics and Phonetics in Lund, Sweden (see Bruce and Touati 1992 for a presentation of the research methodology and exemplifications from Swedish and French). The recorded material were digitalized and analyzed using the LUPP program (see Eriksson 1990). The acoustic segmentation of the recordings was made by visual inspection of the tonal configuration and interactive listening to the segmented speech. Every 10 ms, a Fo value was computed for each segment of speech. The data files were exported to a statistical package. For this analysis I examined average Fo (means and standard deviation), the absolute Fo minimum and maximum values and the range.

Overall pitch contrasts in different settings

Results for the analysis of pitch contrasts in different settings are presented in Table 1. As shown in Table 1, the pre-electoral speech was uttered with a higher mean Fo while the post-electoral press-conference was uttered with a lower mean Fo. Absolute Fo minima are similar. On the other hand, maxima values are lower and range is more reduced in the post-electoral press-conference. By using this long term reduced overall pitch range, J. Chirac is acting in the role of a generous and tranquil winner or, more to the point, seeking to impose on the audience such an image of himself.

Table 1. *Pitch range and register in two different settings. Average Fo (means and standard deviation), absolute Fo minimum and maximum values and Fo range in two different settings (A: pre-electoral speech, B: post-electoral press-conference; values are in Hz).*

	MEAN	ST DEV	MIN	MAX	RANGE
A	163	38	79	250	171
B	117	22	71	209	138

Overall pitch changes in direct quote-comment structure

Results for the analysis of pitch changes in the DQCS are presented in Table 2. As shown in Table 2 and as an example in Figure 1 (1) , quoting segments were uttered with a high Fo mean, with a very high Fo minimum and maximum, and a reduced Fo range. On the other hand, quoted segments were signalled, in contrast to quoting segments, by a lower Fo mean, lower absolute Fo minimum and maximum and larger Fo range. Comment segments generally show a strong tendency to contrast with quoted ones in the same way that quoted segments contrasted with quoting ones (comment segments are realized by a lower Fo mean, lower absolute Fo minimum and maximum, and larger Fo range). There is an exception: the last comment segment has a reduced Fo range (see Figure 1 (2)). This is probably due to its position as comment in the épilogos (final part) of the speech.

Table 2. *Pitch range and register in DQCS. Average Fo (means and standard deviation), absolute Fo minimum and maximum values and Fo range in DQCS (A: quoting segment, B: quoted segment and, C: comment segment; values are in Hz).*

1) Quoting the socialists

	MEAN	ST DEV	MIN	MAX	RANGE
A	183	29	147	221	74
B	172	28	95	204	109
C	155	38	86	228	142

A (ils nous disent) B(on continue) C (c'est tout voilà l(e) programme)
(they are telling us) (we continue) (That's all, that's the programme)

2) Quoting the President

	MEAN	ST DEV	MIN	MAX	RANGE
A	195	23	156	241	85
B	160	27	93	215	122
C	160	45	83	210	127

A((et) le Président de la République nous dit) B (il faut voter pour nous car on n(e) change pas une équipe qui gagne) C (sans complexe c´est tout)
(And the President of the Republic is telling us)(you vote for us because you don't change a winning team)(it's that simple, that's all)

3) Quoting the Prime Minister

	MEAN	ST DEV	MIN	MAX	RANGE
A	176	34	133	245	112
B	157	34	94	226	132
C	151	28	102	189	87

A ((quant) au premier ministre il nous dit) B(au secours la droite revient) C (point final)
(As for the Prime Minister, he is telling us) (help! the right wing is back) (Period)

CONCLUSION

During our work in KIPROS, we assumed that changes in overall pitch were essentially changes in range and were mainly achieved by raising/lowering the Fo peaks. This assumption is confirmed here in the analysis of different fragments of rhetorical prosody. It seems that pitch range variation is used when establishing a paradigmatic contrast between pre-electoral speech and post-electoral speeech or a syntagmatic change between direct quote segments and comment segments. But, as it was possible to observe in the pre-electoral speech, changes from a higher register to a lower register were used in order to express the opposition quoting segments-quoted segments (variation in register in spontaneous French was observed by Mertens 1987). Our analysis of overall pitch in a pre-electoral speech as opposed to a post-electoral press-conference and within the specific setting of the pre-electoral speech provide evidence that we need a two-fold categorization of overall pitch variation in French: one in terms of range and the other in terms of register. I also propose an adjustment of the KIPROS transcription system concerning overall pitch in the following way:

Pitch range:	[→] = same range	[↑] = expanded range	[↓] = reduced range
Pitch register:	[⇒] = same register	[⇑]= higher register	[⇓]= lower register

REFERENCES

R. Barthes (1970), "L'ancienne rhétorique. Aide-mémoire", in *Communication,* Vol. 16

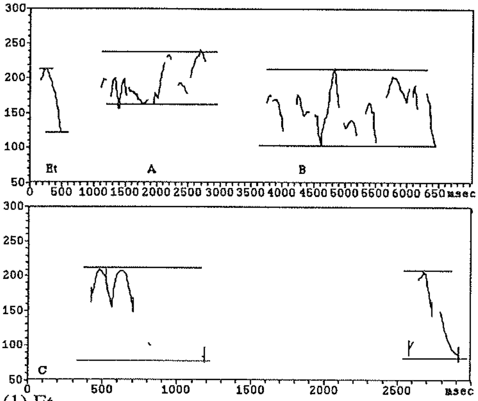
G. Bruce and P. Touati (1992), "On the analysis of prosody in spontaneous speech with exemplification from Swedish and French", in *Speech Communication,* Vol. 11:4-5, pp. 453–458.

L. Eriksson (1990), "New phonetic programmes for Macintosh", *Working Papers,* Vol. 36, pp. 73–80.

P. Mertens (1987), *L'intonation du français. De la description linguistique à la reconnaissance automatique* (Katholieke Universiteit Leuven).

P. Touati (1991), "Temporal profiles and tonal configurations in French political speech", *Working Papers,* Vol. 38, pp. 205–219.

P. Touati (1993), "Overall pitch and direct quote-comment structure in French political rhetoric", in *RUUL,* Vol. 23, pp. 98–101.

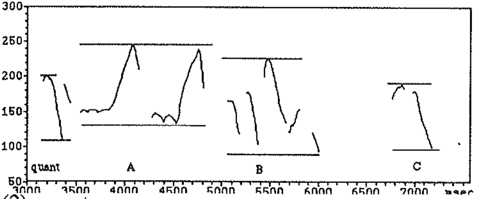


(1) Et

A. ⬇ ⇑ le Président de la République nous dit:

B. ⬆ ⇓ "il faut voter pour nous car on n(e) change pas une équipe qui gagne",

C. ⬆ ⇒ sans complexe , c'est tout.



(2) quant

A.⬇ ⇑ au premier ministre il nous dit:

B. ⬆ ⇓"au secours la droite revient ",

C. ⬇ ⇒ point final.

Figure 1. *Fo range and register in DQCS (A: quoting segment, B: quoted segment and, C: comment segment; values are in Hz), (1) Quoting the President (pistis) and (2) Quoting the Prime Minister (épilogos).*

Intonational Consequences of Varying the Accent Location

Aart De Zitter & Bert Van Coile
Lernout & Hauspie Speech Products
Rozendaalstraat 14, 8900 Ieper, Belgium
&
University of Gent
Dpt. of Electronics and Information Systems,
St. Pietersnieuwstraat 41, 9000 Gent, Belgium

ABSTRACT

An experiment was carried out to investigate the effect of varying the location of the (unique) sentence accent on the concrete melodic behavior of that accent. More specifically, we measured the onset, the duration and the size of both the rise and the fall on the so-called 'pointed hat' on the accent. The main finding was that accents are getting smaller the later they occur in the sentence.

INTRODUCTION

In natural speech, pitch accents can be realized in many different ways. In Dutch, a common realization of a pitch accent is the so-called *pointed hat*: a sudden pitch rise on the accented syllable, immediately followed by a sudden fall ('t Hart et al., 1990). Yet, this pointed hat can vary a lot with respect to the range and the duration of both pitch movements, as well as with respect to its location in the syllable. The present study has investigated the melodic behavior of single pitch accent in short, monophrasal Dutch utterances. Previous studies have pinpointed the notion of *converging declination lines* (Terken, 1993). This notion states that consecutive pitch movements become smaller the later they occur in the phrase, so that one could interpret them as being realized between two converging declining lines: the baseline and the topline. We were interested in looking if this notion also holds in utterances with but *one* accent. To do so, we measured the onset, duration and size of both the rise and the fall constituting the pointed hat used to mark the unique sentence accent.

METHODOLOGY

In order to check for the influence of the accent location on the concrete melodic realization of the pointed hat, a set of utterances was recorded, all containing only one sentence accent[1]. The data base consisted of several readings of an utterance

[1] The construction criteria for the materials were imposed by the COST-233 project 'Prosodics in Synthetic Speech'.

containing monosyllabic words ('Een laag dal zag een man' (A low valley saw a man)), and an utterance containing disyllabic words ('De lieve dame gooide een ballon' (The lovely lady threw a balloon)).

Accents could fall either in initial, medial or final position: on the first noun, on the verb or on the second noun. Two speakers were asked to utter the sentences: one male (SL) and one female (YH). Every utterance was spoken eight times. Totally this equals: 2 lengths (short, long) x 3 accent locations (I, M, F) x 2 speakers x 8 replications (eight recordings) = 96 recordings.

The readings with the desired accent structure were elicited by appropriate questions. E.g. to get the answer 'The lovely lady threw a *balloon*', the (oral) question was '*What* did the lovely lady throw?'.

The original pitch contours were replaced by so-called 'close copies' (i.e. straight-line approximations of the original pitch contour without any audible difference with the original ones, cf. 't Hart et al, 1990). On the 96 close copies, the numerical values of the pointed hats were measured. More exactly, we measured the onset (relative to vowel onset, in cs), duration (in cs) and excursion size (in semitones) of both rise and fall constituting the pointed hat.

RESULTS

Results for speaker 1 (SL)

The results for speaker SL are shown in the following table. The eight replications for every recording were pooled.

Table 1. *Speaker 1: mean onset, duration and excursion size for rise and fall in pointed hats as a function of the location of the accented syllabe (I(nitial), M(edial) or F(inal)).*

sent.	accent		rise			fall	
length	loc.	onset	dur	exc.size	onset	dur	exc.size
		(cs)	(cs)	(ST)	(cs)	(cs)	(ST)
short	I	-11.62	16.25	6.60	12.12	15.87	6.28
	M	- 9.75	16.62	5.35	8.37	17.12	5.65
	F	- 7.37	11.50	4.31	8.75	11.87	4.75
long	I	- 6.50	18.25	6.11	13.50	16.87	5.59
	M	- 7.87	16.50	5.32	13.62	14.12	5.56
	F	-10.12	14.37	4.52	9.00	13.00	4.99

The most striking effect in this table clearly is the reduction in excursion size as a function of the accent location, and this both for the rise and the fall. Accents realized in initial position (on the first noun) have larger excursions than accents in medial position (on the verb), which are in turn larger than those in final position (on the second noun). For mean values see also table 3. An ANOVA with sentence length and accent location as main effects show that these effects are statistically significant (for the exc.size of the rise: F $<$ .001, and for the exc.size of the fall: F = .002).

Since the phenomenon of decreasing excursion size as a function of accent location holds for both the rise and the fall, we can look at the correlation between these two parameters via a regression analysis. Following figure shows how these two parameters are correlated.

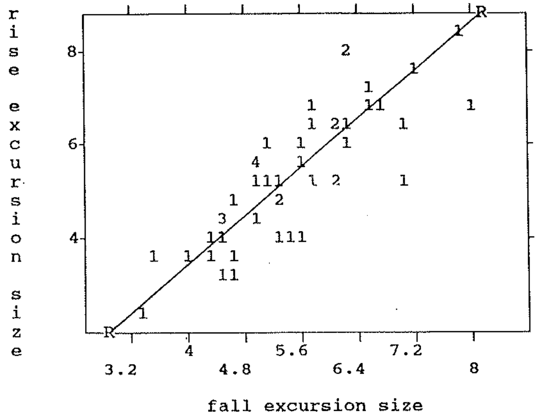


Figure 1. *Speaker 1: Correlation plot of excursion size rise with excursion rise fall (Correlation Coefficient = .817, R^2 = .668).*

A second observation for speaker SL concerns the duration of both the rise and the fall. Both rise and fall become shorter the later the accent falls. This can be seen in table 3 on the following page.

Results for speaker 2 (YH)

The results for speaker YH are shown in the following table. The eight replications for every recording were pooled.

Table 2. *Speaker 2: mean onset, duration and excursion size for rise and fall in pointed hats as a function of the location of the accented syllabe (I(nitial), M(edial) or F(inal)).*

sent. length	accent loc.	rise onset (cs)	rise dur (cs)	rise exc.size (ST)	fall onset (cs)	fall dur (cs)	fall exc.size (ST)
short	I	-14.87	23.87	6.07	13.87	26.75	7.67
	M	-16.75	26.50	4.82	13.25	15.62	6.10
	F	-10.00	14.62	4.51	6.50	12.00	5.57
long	I	-8.00	23.50	5.56	19.25	13.75	6.46
	M	-12.37	21.87	5.26	16.75	13.87	6.26
	F	-11.50	15.62	3.69	4.87	15.00	5.86

Just as in the measurements for SL, the most obvious effect of varying the accent location is the reduction in the excursion size of the movements. Both the rise and the fall get smaller excursions the later they occur in the sentence. Here as well, an ANOVA gives significant F-values for the excursion size of the rise and the fall (F $< .001$ for the first, and F $= .003$ for the latter).

For speaker YH, a regression analysis between the size of the rise and the size of the fall yields the following result: Correlation Coefficient = .753, $R^2 = .567$

Following table shows the mean excursion size and duration of both the rise and the fall for speaker SL, and the mean excursion size of both the rise and the fall for speaker YH. The data for the short and the long sentence are pooled, since the ANOVA shows that the only main effect influencing these variables is the accent location and there are no significant interactions.

Table 3. *Mean excursion size of rise and fall for both speakers, and mean duration of the rise and fall for speaker SL, pooled over the two sentences.*

accent loc.	speaker 1 (SL)				speaker 2 (YH)	
	rise exc.size	fall exc.size	rise dur.	fall dur.	rise exc.size	fall exc.size
I	6.36	5.94	17.25	16.37	5.82	7.07
M	5.34	5.61	16.56	15.62	5.04	6.18
F	4.42	4.87	12.94	12.44	4.10	5.72

CONCLUSION

For both speakers, the same main observation is found. The later the unique sentence accent occurs, the smaller both the rise and the fall constituting the pointed hat. This finding strengthens the observation in Terken, 1993. He has found that pitch accents tend to be realized between two converging declination lines: a baseline and a topline. In case of only *one* accent, it is problematic to draw a topline[2]. However, we have found that the principle of convergence still holds in this case. Unique pointed hats seem to be realized between two hypothetical declination lines as well.

This study was carried out within the framework of the European project COST-233 *'Prosodics in Synthetic Speech'*.

REFERENCES

Cohen, A., Collier, R., 't Hart, J. (1982), "Declination: construct or intrinsic feature of speech pitch?" *Phonetica*, 39, pp. 254-273.

't Hart, J., Collier, R., Cohen, A. (1990), *A perceptual study of intonation. An experimental-phonetic approach to speech melody* (Cambridge University Press, Cambridge).

Terken, J. (1993), "Synthesizing natural-sounding intonation for Dutch: rules and perceptual evaluation" *Computer Speech and Language*, 7, 1, pp. 27-49.

[2] Normally, base- and topline are drawn by interpolating the valley pitch values or the top pitch values, respectively. Cf. Cohen et al., 1982.

MÜSLI: A Classification Scheme For Laryngealizations

A. Batliner[1], S. Burger[1], B. Johne[1], A. Kießling[2]

[1] L.M.-Universität München, Institut für Deutsche Philologie,
Schellingstr. 3, 80799 München, F.R. of Germany

[2] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5),
Martensstr. 3, 91058 Erlangen, F.R. of Germany

ABSTRACT

We developed a classification scheme for laryngealizations that can be used to discriminate the many different shapes of laryngealizations with different feature values. Potential applications are phonetic transcription and automatic detection. The scheme was developed and tested with a database from 4 speakers that contains more than 1200 laryngealizations.

INTRODUCTION

The normal speech register (modal voice) comprises a F0-range from about 60 to 250 Hz for male speakers and a F0-range from about 120 to 550 Hz for female speakers. Below this register there is a special phonation type whose mechanisms of production are not totally understood yet and whose linguistic functions are little investigated until now. There is a variety of different terms for this phenomenon, which are used more or less synonymously: creak, vocal fry, creaky voice, pulse register, laryngealization, etc. We use "laryngealization" (henceforth LA) as a cover term for all these phenomena that show up as irregular, voiced stretches of speech. Normally, LAs do not disturb pitch perception but are perceived as suprasegmental irritations modulated onto the pitch curve. Although LAs can be found not only in pathological speech, but also in normal conversational speech, most of the time, they were not objects of investigation, but considered to be an irritating phenomenon that has to be discarded. On the other hand, recently the fact that LAs often occur at word or morpheme boundaries and thus could be used in speech recognition, has been realized. Efforts for their investigation and classification have been undertaken [2] [3]. In the time signal LAs can look quite different (cf. figure 1–6) and it is not far-fetched to claim that the only common denominator of the different types is their irregularity. LAs can be produced with different means and different states of the glottis but it is not clear yet whether there is a regular relationship between different production mechanisms and different types of LAs showing up in the time signal. In [2] four different types of LAs are characterized (cf. below). We will follow another approach and use non-binary features for our description scheme that can be used by different transcribers in a consistent way. An overspecification can be reduced in a second step. It should be possible to extract the features automatically with standard pattern recognition algorithms.

MATERIAL

We investigated a database of 1329 sentences from 4 speakers (3 female, 1 male; 30 minutes of speech in total). One third of the database consists of real spontaneous utterances gained from human-human clarification dialogues, the rest consists of the same utterances read by the same speakers nine months afterwards (own utterances and partners utterances). Recording conditions were comparable to a quiet office environment. The utterances were digitized with 12 Bit and 10 kHz; for more details cf. [1]. Two trained phoneticians classified the voiced passages as [+/– laryngealized] with the help of a segmentation program (time waveform presented on the screen and iterative listening to the segmented part). 4.8% of the speech in total (7.4% of the voiced parts) were laryngealized (henceforth la).

The mean duration of the LAs was 64.1 ms with a standard deviation of 35.1; minimum = 12.8 ms (1 frame), maximum = 332.8 ms (26 frames). 16% of the LAs extend through a phoneme boundary. The non-la passages will not be considered in this paper. The la parts were plotted with their non-la context and a constant resolution, and a group of 6 experts tried to cluster a subset of these plots manually using different criteria (the 4 classes in [2] as well as phoneme-, context-, and speaker- specific peculiarities). Based on the similarities between the tokens within the clusters and the dissimilarities between tokens of different clusters respectively, several features were chosen for characterizing the LAs adequately. Afterwards, a classification scheme was developed heuristically and subsequently tested and verified with the whole material.

THE CLASSIFICATION SCHEME FOR LARYNGEALIZATIONS

In MÜSLI (M*ünchner* S*chema für* L*aryngalisierungs*-I*dentifikation*) six different features in four different domains (cf. table 1) are used for describing LAs. The values of these features can be determined independently from each other and are coded with integers within the ranges from 1–3 or from 1–4. Thus, every LA is determined by a sextuple of integers. In this paper, we will deal only with these features and not with other, e.g. speaker- or context-specific phenomena. Due to the lack of space, not every feature value can be illustrated in figure 1–6, but some of the values can be seen in the captions. The features and their values that are given in brackets are described in the following. In parentheses, the percentage of cases of all LAs assigned to the specific value are given. Values that can probably be combined into one single value (i.e. reduction of overspecification), are given in curly brackets at the end of the description of each feature. Reasons for combining are: either one of the values - e.g. [3] in AMPSYN - occurs very seldom, or because the two values might possibly not be told apart with great certainty by e.g. an automatic classification. At the same time, the values do not discriminate different LA-types such as e.g. the values [1] and [2] in F0SYN and F0PAR, cf. table 1.

1. **NUMBER** = Number of glottal pulses: [1] many periods (83.5%); [2] two to three periods (8.8%); [3] one period (7.3%); {2 3}
2. **DAMPING** = Special form of the damped wave: [1] relatively normal damping (42.4%); [2] strong exponential decay of the amplitude (2.6%); [3] "delta-like", triangular damping (24.4%); [4] "unusual" damping (30.1%); {2 4}
3. **AMPSYN** = Amplitude compared with left and right context (syntagmatic aspect): [1] normal (76.2%); [2] lower (23.3%); [3] higher (0.4%); {1 3}
4. **AMPPAR** = Amplitude variations inside the LA (paradigmatic aspect): [1] regular envelope, no variations (23.3%); [2] slightly irregular envelope (45.3%); [3] "diplophonic", i.e. regular variation between high and low amplitude (17.8%); [4] break down of envelope (12.7%); {1 2}
5. **F0SYN** = F0 compared with context (syntagmatic aspect): [1] regular, no variations (38.0%); [2] slightly irregular (15.1%); [3] subharmonic (25.2%); [4] extremely long period(s) or pause (20.3%); {1 2}
6. **F0PAR** = F0 variations inside the LA (paradigmatic aspect): [1] regular, no variations (39.7%); [2] slightly irregular (28.1%); [3] strong variations (25.3%); [4] periods not detectable (6.7%); {1 2}

The feature value [1] is always the default value as it is found regularly in non-la speech as well. A value was determined if it showed up during more of half of the la passage. A "compound type" (56 occurrences in the database) was determined if the la passage consisted of two clearly distinct parts that could be classified on their own. These two parts were treated separately. In total 1251 LAs were labeled with MÜSLI.

RESULTS AND DISCUSSION

Out of all 1251 LAs 81% could be classified unequivocally and completely. 18% could also be classified, but with a disagreement in at least one feature value between the two phoneticians. In only 18 cases there was at least one feature value that could not be determined at all (feature value 0, cf. figure 6). The numbers given in the following always refer to all LAs except these 18 cases. For a grouping of the LAs into distinct LA-types, we first chose those combinations of feature values (sextuples) that occurred $\geq$ 10 times. These sextuples were grouped so that (near) default values were combined with as few as possible non-default values. We distinguish four different domains in the time signal: *Number*, *Damping*, *Amplitude*, and *Frequency*. In the following description, parentheses contain one or more of: 1. the relevant domains; 2. the number of the figure showing an example; 3. the terms used in [2] if they differ. Three LA-types could be differentiated with the help of **different domains**: GLOTTALIZATION (*Number* and *Frequency*, figure 1), DAMPING (*Damping*, figure 2, creak), DIPLOPHONIA (*Amplitude*, figure 3). Two LA-types could be differentiated **within one single domain**, namely SUBHARMONIC (figure 4, creak), and APERIODICITY (figure 5, creak or creaky voice) both having different values inside *Frequency* for F0SYN and F0PAR. In figure 6, the WASTE PAPER BASKET LA-type is illustrated with an example where two feature values (for AMPPAR and F0SYN) could not been defined. AMPSYN is no "distinctive feature" because it does not discriminate LA-types but it can characterize LAs in general. In figure 1–6 the sextuple of feature values is given in each caption in parentheses. Although a "standard" GLOTTALIZATION has only one period followed by a long pause, the example given in figure 1 represents roughly half of all the GLOTTALIZATIONs in our material.

Table 1: LA-*types and their characterization with MÜSLI*

LA-type (number of cases)	*Domains* & FEATURES					
	Number	*Damping*	*Amplitude*		*Frequency*	
	NUMBER	DAMPING	AMPSYN	AMPPAR	F0SYN	F0PAR
GLOTTALIZATION (61/114/116)	'[23]'	3 [1234]	1 [123]	[12]	'4'	[13] [123]
DAMPING (161/292/680)	1 [123]	'[34]' [234]	1 [123]	[12] [124]	1 [12]	[12] [12]
DIPLOPHONIA (122/166/222)	1	1 [1234]	1 [123]	'3'	[12]	[12]
SUBHARMONIC (109/157/190)	1	[134] [1234]	1 [123]	[12] [124]	'3'	'[12]'
APERIODICITY (158/242/384)	1	[134] [1234]	[12] [123]	2 [1234]	[34]	'[34]'

Table 1 shows the five LA-types and their characterization with special feature values. The columns can be interpreted as regular terms: between columns holds conjunction, within brackets holds disjunction. Combinatorically $3 \cdot 4 \cdot 3 \cdot 4 \cdot 4 \cdot 4 = 2304$ different sextuples can occur. In the first line of each LA-type the combinations are shown that entail $\geq$ 10 cases (narrow condition; 56 possible, 24 occurring sextuples). Weakening the conditions more cases can be classified; cf. the possible feature combinations in the second line of each LA-type (broad condition; 780 possible, 178 occurring sextuples). In the second line cells are left empty, whose terms do not differ from the corresponding terms in the first line. Cases that are comprised in line two are kept disjoint, i.e. there is no intersection of two LA-types. They represent so to speak **pure LA-types**. However, if we use as criterion only the "distinctive feature" values quoted in line one, i.e. for the other features all values are valid (very broad condition), we get 3552 possible and 247 occurring sextuples. 532 cases belong to more than one LA-type, 83% of them forming an intersection of DAMPING with other

LA-types. In the first column of table 1 the number of cases for narrow/broad/very broad conditions are given in parentheses below the name of each LA-type.

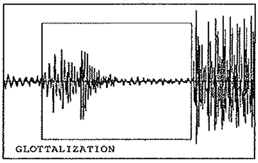


Figure 1 (231243)

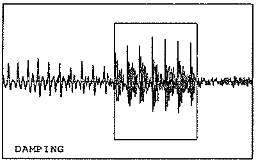


Figure 2 (141211)

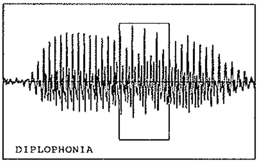


Figure 3 (111311)

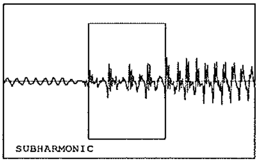


Figure 4 (111231)

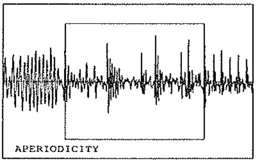


Figure 5 (131243)

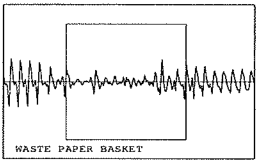


Figure 6 (111004)

FINAL REMARKS

It can be doubted that the features are distinctive phonologically but at least some of them might constitute allophones occurring in different contexts, while others might describe simply free variants. Yet, to our knowledge the feature matrix in table 1 is the first attempt to describe a large corpus of LAs systematically and exhaustively with a feature approach; it seems to work reasonably well. MÜSLI should, however, not be taken as **the** final classification scheme for LAs but rather as a starting point for further investigation. Other possible features as e.g. spectral tilt, breathiness or (partial) devoicing could be taken into consideration as well. The next step will be the automatic extraction of the different features and then hopefully a more straightforward but at the same time more robust feature description and a reduction of overspecification. It should be investigated further whether different LA-types can be discriminated perceptually, whether different LA-types have different functions such as e.g. boundary marking, and if the different LA-types are speaker-, language-, or register-specific.

Acknowledgements

This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research project ASL/VERBMOBIL and by the *Deutsche Forschungsgemeinschaft* (*DFG*). Only the authors are responsible for the contents of this paper.

References

[1] A. Batliner, C. Weiand, A. Kießling, and E. Nöth. *Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody.* In this volume.

[2] D. Huber. *Aspects of the Communicative Function of Voice in Text Intonation.* PhD thesis, Chalmers University, Göteborg/Lund, 1988.

[3] A. Kießling, R. Kompe, E. Nöth, and A. Batliner. *Irregularitäten im Sprachsignal — störend oder informativ?* In R. Hoffmann, editor, *Elektronische Signalverarbeitung*, volume 8 of *Studientexte zur Sprachkommunikation*, pages 104–108. TU Dresden, 1991.

Prosodic Modelling of Phrasing in Swedish

Gösta Bruce*, Björn Granström**, Kjell Gustafson** and David House*
(names in alphabetical order)
*Lund University, Dept. of Linguistics and Phonetics, Helgonab. 12, S-22362 Lund
**KTH, Dept. of Speech Comm. and Music Acoustics, Box 70014, S-10044 Stockholm

ABSTRACT

In this contribution we will give a summary report on a project on prosodic phrasing in Swedish. Examples of three different methodological approaches will be given: speech analysis, speech synthesis and prosodic parsing. The results include expansions to the model of Swedish prosody.

INTRODUCTION

The starting point for our research effort within phrasing and prosody of Swedish has been our judgement that we possess a fair amount of knowledge about prominence relations and accentuation but that we know relatively little about prosodic grouping and phrasing. It is widely recognised, however, that grouping – involving the double aspect of coherence (connective) signalling and boundary (demarcative) signalling – is one of the main functions of prosody.

This contribution summarizes the work within a recently completed co-operative project between Lund University and KTH. Our primary goal has been to attain new knowledge about phrasing and prosody in Swedish. Our focus of interest here is in particular the grouping of words into prosodic phrases and domains of similar size. The problems to tackle concern questions in both phonology and phonetics. The main phonological issue is to try to understand what structure could be assumed for prosodic phrasing, particularly what types of prosodic phrase can be identified as relevant domains between a 'prosodic word' and a 'prosodic utterance'. The basic phonetic issue is to find out what speech variables (F0, duration, intensity, phonation type, pausing, etc.) and combinations of them can be used to signal phrasing. Basically three different methods are being utilised within the project: speech analysis, speech synthesis and prosodic parsing.

SPEECH PRODUCTION/ANALYSIS

The first method is the collection and analysis of speech production data. It involved the construction and recording of specially designed test material produced by one Stockholm speaker, as well as the selection and recording of suitable read text passages by several speakers (laboratory speech). The first set consisted of 22 sentences repeated three times, typically occurring as minimal pairs, where the location of the sentence internal clause boundary (here represented by a comma) was varied. These sentences were, for the most part, syntactically ambiguous, and designed to give us an idea about phrasing strategies and to enable us to easily test these strategies in the text-to-speech framework. One such pair is presented below:

När pappa fiskar, stör Piper Putte. (When daddy is fishing, Piper disturbs Putte)
När pappa fiskar stör, piper Putte. (When daddy is fishing sturgeon, Putte peeps)

Considerable variation in the signalling of phrasing was observed (Bruce, Granström and House, 1992) including: 1) the use of duration only in clause/phrase boundary signalling, 2) signalling of coherence by deaccentuation, 3) coherence signalling through the use of a "hat pattern", and 4) in longer sentences with greater syntactic complexity, the combined use of duration and F0 cues, including the absence vs. presence of F0 downstepping.

An extension of the above speech material has also been investigated within the project (Bruce, Granström, Gustafson and House, 1991). The distinctive feature of these sentences (also minimal pairs, repeated ten times) is the absence vs. presence of an internal clause boundary. An example pair is:

Lärarna backar för pojkarnas sparkar. (The teachers back away from the boys' kicks)
Lärarna backar, för pojkarna sparkar. (The teachers back away because the boys kick)

While a connective F0 downstepping was a characteristic feature of a sentence without an internal boundary, a fair degree of variation was found in the production of sentences containing a boundary. This variation can be summarised by the following strategies: 1) a boundary cue comprising marked pre-boundary lengthening preceding a small physical pause, with F0 downstepping indicating coherence, 2) the addition of a focal accent to the accent before the boundary, accompanied by a terminal F0 fall and moderate pre-boundary lengthening, and 3) an upstepping F0 pattern initially — interpretable as an extra emphasis for contrast on the first accent — and a relatively wide F0 range on the accent before the boundary accompanying moderate pre-boundary lengthening.

In most of our material, phrasing and accentuation are partly interdependent, as deaccentuation is often used as a coherence cue for the division into phrases. In one type of sentence, however, accentuation stays the same, while phrasing is varied. The following sentence pair is used to illustrate this subset, where the characteristic difference is the location of the internal boundary, resulting in either a grouping of 2+3 accents or 3+2 accents:

B. Fast man offrade bonden, och löparen hälsade kungen.
(But we sacrificed the pawn, and the bishop greeted the king)
L. Fast man offrade bonden och löparen, hälsade kungen.
(Though we sacrificed the pawn and the bishop, the king greeted us)

These two sentences, as well as an ambiguous version of them, were recorded three times. One typical and clearly identified version of each sentence is illustrated in Figure 1. It is clear that both tonal and temporal cues are combined to signal the difference in phrasing. The notable F0 difference occurs after the 2nd accent 'bonden' as a deep vs. shallow F0 valley. Interestingly, there is no corresponding F0 difference after the 3rd accent 'löparen'. The main durational difference can be seen as a pre-boundary lengthening after the 2nd and 3rd accent respectively depending on the phrasing. This sentence type was also used in a formal perception test using text-to-speech synthesis (see next section).

SPEECH SYNTHESIS AND PERCEPTION

The second method employed was the use of text-to-speech synthesis for the testing of hypotheses about the signalling of prosodic phrasing. In the KTH text-to-speech system there are several ways of interacting with rules and parameters (Carlson, Granström and Hunnicutt, 1990). Text-to-speech synthesis was used to test the phrasing strategies observed in the production data above (Bruce, Granström and House, 1992). The default realisation of clause boundary signalling in the text-to-speech system, including a silent pause, appears to be unambiguous, but too strong and hence unnatural in many contexts. Several alternative boundary signalling strategies that also have the function of disambiguating the synthetic versions of the potentially ambiguous sentences in our speech material were identified and explored.

Some of the strategies concerning the relationship between tonal and durational cues were tested perceptually using the KTH rule synthesis where subjects can interactively vary duration and tonal parameters by moving a point on the computer screen. A sentence pair similar to sentences B and L above was used in a formal perception test where

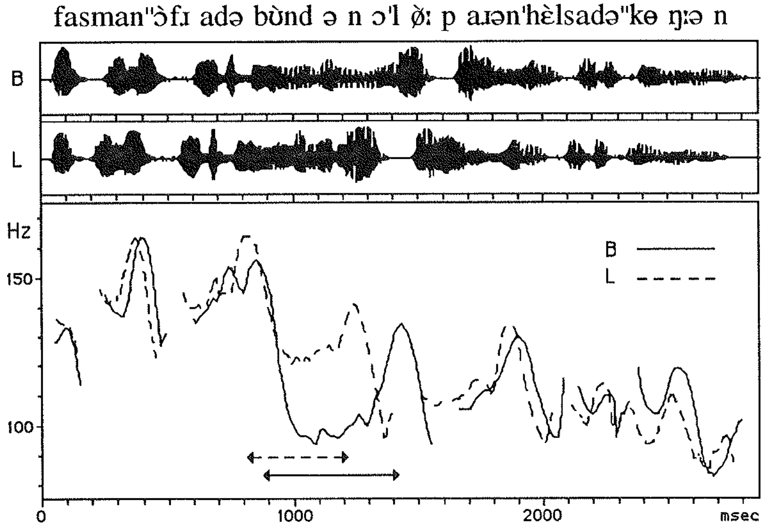


Figure 1. *Waveforms and fundamental frequency contours of the sentence "B" (solid line) and "L" (dashed line). Arrows indicate the domain used for parameter manipulation of synthetic versions in the perception experiment.*

subjects were asked to determine an optimal position for each interpretation and a line of ambiguity across the screen. Results of this test, presented in Bruce, Granström, Gustafson and House (1993a), clearly indicate that both duration and F0 are effective as phrasing signals. Although some of the individual listeners tended to have a cue preference (F0 or duration), the interaction between these two cues seems to be one of complementation rather than comprising a primary and a secondary cue. Furthermore, the durations and F0 values of the test word seem to be judged in relation to both what precedes and what follows for the same speech parameter in the test utterance. Thus the results are also consistent with our more specific observation from production data that a shallow F0 valley as part of an F0 downstepping pattern has a connective function signalling coherence within a prosodic phrase, while a deep F0 valley, as a break in the downstepping trend, has a demarcative function signalling a phrase boundary.

PROSODIC PARSING

The third method used in this project was prosodic parsing directed towards the recognition of phrasing. The first stage of the prosodic recognition method is the use of a human recogniser, an expert reader of an acoustic record of speech for the identification of potential phrases of an 'unknown' speech signal (cf. House and Bruce, 1990). Based on the knowledge used by the expert reader for prosodic parsing, the subsequent procedure is then to teach the computer to make an automatic analysis of prosodic phrases. Generally, we believe that the prosodic parser is particularly suitable for testing hypotheses about the interaction between different speech variables for the expression of prosodic phrasing.

The prosodic parsing experiment carried out in this project involved recording and analysing longer text passages read by two different speakers. In the experiments an expert reader is given the task of identifying prosodic phrases solely on the basis of a visual representation of the text showing the waveform, intensity and fundamental frequency. The results are then compared to two independent, auditively based transcriptions of the readings (Bruce, Granström, Gustafson and House, 1993b). There is fair agreement between the transcribers (~80%) and the expert is frequently able to identify the boundary. The mean absolute boundary locations identified across all experiments by the expert was slightly more than 70%, and by relaxing the criterion to +/- one word it increased to 85%. Sometimes the strength of the boundary is not agreed on. Results from these experiments will form the basis for the formulation of automatic recognition rules for phrases which can be integrated into an automatic prosodic parsing system.

CONCLUDING REMARKS

Our modelling of prosodic phrasing involves both coherence and boundary marking. We assume that successive prosodic words are typically grouped into prosodic phrases. This means that the prosodic phrase is related to the accentual structure. Thus specific combinations of tonal gestures for accentuation can signal coherence within a prosodic phrase. Boundary signals for a prosodic phrase can be either a separate gesture before the first accent (initial juncture) or after the last accent (terminal juncture), or they can coincide with an accentual gesture at the beginning or at the end of a phrase.

In the continuation of this project we want to widen the scope and direction of research in two different ways. Firstly we intend to cover not only the grouping function (e.g. phrasing) but also the prominence function of prosody (e.g. accentuation). We consider these to be the two main and basic functions of prosody. We specifically intend to study the interaction between phrasing and accentuation. These categories seem to be easy to separate from each other in theory but are frequently conflated in the practical situation. Secondly we would like to study how prosodic grouping and prominence are exploited in a dialogue context.

ACKNOWLEDGEMENT

The present work is in part carried out under a contract from the Swedish Language Technology Programme.

REFERENCES

G. Bruce, B. Granström and D. House (1992), "Prosodic phrasing in Swedish speech synthesis", in *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoît and T.R. Sawallis (eds.), pp. 113-125. Elsevier Science Publishers B.V., Amsterdam.

G. Bruce, B. Granström, K. Gustafson and D. House (1991), "On prosodic phrasing in Swedish", *PERILUS XIII*, pp. 35-38. Inst. of Linguistics, University of Stockholm.

G. Bruce, B. Granström, K. Gustafson and D. House (1993a), "Interaction of F0 and duration in the perception of prosodic phrasing in Swedish", in *Nordic Prosody VI*, B. Granström and L. Nord (eds.), Almquist and Wiksell, Stockholm.

G. Bruce, B. Granström, K. Gustafson and D. House (1993b), "Phrasing strategies in prosodic parsing and speech synthesis", in *Proc. Eurospeech '93, European Conf on Speech Comm and Technology*, Berlin, September 21-23, 1993.

R. Carlson, B. Granström and S. Hunnicutt (1990), "Multilingual text-to-speech development and applications", in *Advances in speech, hearing and language processing*, W. Ainsworth (ed.), pp. 269-296. JAI Press, London.

D. House and G. Bruce (1990), "Word and focal accents in Swedish from a recognition perspective", in *Nordic Prosody V*, eds. K. Wiik and I. Raimo, pp. 156-173. Turku University.

Automatic Detection of Prosodic Cues for Segmenting Continuous Speech into Supralexical Units

Noëlle Carbonell, Yves Laprie
CRIN-CNRS & INRIA-Lorraine
BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France

ABSTRACT

We are currently working on the automatic detection of syntagm and word boundaries in French, from the study of pitch and rhythm temporal evolution. After a brief presentation of our method, we discuss results obtained on various continuous speech corpora .

OBJECTIVES AND HYPOTHESES

Numerous experimental studies have brought out the useful contribution of prosody to speech perception and understanding (cf. for instance: Blesser, 1969 in Waibel, 1988; Grosjean, 1987). However, J. Vaissière and A. Waibel (Vaissière, 1988, Waibel, 1988) observe that few prototypes — among recent continuous speech recognition and understanding systems — take into account the linguistic information embedded in prosodic events.

The complex nature and multiple functions of prosody may account for this paradoxical situation, namely:

- the great intra- and inter-speaker variability that characterizes prosodic expression;
- the interactions between microprosody, which is deeply influenced by the phonetic contents of utterances, and macroprosody, which contributes to the interpretation of speech at various levels: lexical, syntactic and semantic (cf. Waibel, 1988, page 11);
- the complex relationships between macroprosody on the one hand, syntax and semantics on the other hand (Hirst et al., 1991).

We report a recent empirical study that we conducted on the contribution of French prosody to the automatic segmentation of continuous speech into lexical units. Our aim was to determine whether reliable predictions on word-boundary locations could be derived from the automatic analysis of pitch evolution and rhythm variations.

This issue has not yet been carefully investigated. In the few existing prototypes that involve prosodic data in the recognition/understanding process, prosody is assigned a limited function, that is: the validation of word hypotheses generated from acoustic-phonetic, syntactic or semantic evidence (Waibel, 1987).

But, in the case of French at least, the role of prosodic information should be carefully considered, since studies on French prosody (cf. in particular, Rossi, 1981; Di Cristo, 1981) indicate that: most often, main lexical accents are located on the last syllables of words, and last syllables of syntagms are associated with specific prosodic patterns.

Therefore, we contend — against (Vaissière, 1982) for instance — that the function of prosodic information in systems for the recogntion of spoken French should not, a priori, be restricted to the validation of word-boundary hypotheses generated by other information sources.

Our study is based on the following assumptions:

1. We hold, with M. Rossi and other phoneticians, that the contribution of energy to the expression of suprasegmental linguistic information is rather limited in French (Rossi, 1981), compared with the role that this prosodic parameter plays in other languages

(such as English, for instance); which explains why we have restricted the scope of our study to the analysis of pitch evolution and rhythm variations.

2. We assume, with I. Gaïtella (Gaïtella, 1991) that prosodic expression greatly varies according to the pragmatic context (or situation) in which speech is produced; therefore, our study includes the analysis of speech corpora recorded in various situations (reading, sentence repetition from memory, simulated human-computer dialogue).
3. We chose to use macroprosody as an independent source of information, in order to promote robust speech recognition and understanding; therefore, we decided to focus pitch analysis on the centre parts of vocalic nuclei exclusively, so as to eliminate most microprosodic effects; with regard to the study of rhythm, it is based on a rough estimate of the mean duration of vocalic nuclei per prosodic group.
4. Since phoneticians do not agree on a unique model for the description of French macroprosodic sentence patterns, we did not attempt to interpret prosodic cues in terms of comprehensive macroprosodic structures; instead, we focussed on the detection of prosodic cues that may contribute efficiently to lexical segmentation, that is: syntagm-boundary markers and chief lexical accents; with respect to the latter, we assumed that they could be reliably distinguished from secondary and emphatic accents which may occur on first or intermediate syllables in words (cf. polysyllables).

We briefly describe our method in the next paragraph. Then, results obtained on three different speech corpora are presented and discussed. Finally, we indicate how these results influence our current work on the contribution of prosody to speech recognition.

METHOD AND ALGORITHMS

Pitch evolution analysis

In order to get robust pitch data, we chose to implement the time-domain auto-correlation method for pitch detection proposed by M.M. Sondhi in 1968 and improved by L.R. Rabiner namely.

We developed an algorithm for detecting peaks on the F0 curve (in the time x frequency plan), since in French:

- F0 values are significantly higher on stressed syllables (lexical accent) than on unstressed syllables,
- last syllables in some syntagms correspond with prominent local maxima on the F0 curve (cf. the notion of 'continuation majeure' in Rossi, 1981).

Peaks are determined on the smoothest curve (natural cubic spline) resulting from interpolating selected F0 values: one per vocalic nucleus. With respect to syntagm-boundary and stress detection, results are best when the selected values are located in the centre parts of vocalic nuclei; which is in keeping with the observation that microprosody effects are reduced on vowels.

Vocalic nuclei are identified thanks to a speaker-independent algorithm (NOVOCA) (Fohr, 1989); decision criteria are based on a coarse analysis of the spectral distribution of energy.

Study of rhythm variations

Our goal was to detect rhythm variations that contribute to the marking of prosodic group endings and the expression of stress. In French, stressed syllables and last syllables in most syntagms (Di Cristo, 1981) are significantly lengthened.

Since present acoustic-phonetic decoders cannot reliably segment speech into syllables, we used vocalic nucleus duration as a basic unit for the study of rhythm, although vowel duration is greatly influenced by the nature of the vowel and by its phonetic context (1). Our choice is a compromise between practical considerations (i.e. the limits of present acoustic-phonetic decoders) and the constraint (cf. assumption 3 in the previous

(1) For instance, a nasal vowel is often significantly longer than an oral one. As an illustration of context effects, /a/ in syllables where it is followed by /R/ (cf. "part" for example) is longer than when it is followed by another consonant as in "chatte".

paragraph) that prosody should be used as an independent information source; thanks to the accuracy of NOVOCA (with respect to vocalic nucleus detection and bounding), this constraint is not infringed.

In order to determine significant vocalic nucleus lengthenings, we compare each vocalic nucleus in a speech fragment with the Mean Vocalic Duration (MVD) calculated over this speech fragment (2). If the length of a vocalic nucleus is superior to 1.5 x the MVD, we consider that it has undergone a significant lengthening.

PRESENTATION AND DISCUSSION OF RESULTS

Speech corpora

We have tested our algorithms on three multi-speakers speech corpora recorded in three different speech production pragmatic environments:

- corpus LABISE: reading of a short text in a sound-proof room (12 male speakers);
- corpus CMB: 5 male speakers in a quiet room were asked to say rapidly short sentences from memory (short-term memory);
- corpus METEO: simulated human-computer oral dialogues (on weather-forecast) in a realistic environment: each speaker (10 male speakers on the whole) interacted with a real microcomputer to which he could also 'talk' spontaneously.

Results

Results are summarized in the following table.

Table 1. *Automatic detection of macroprosodic cues (i.e. F0 peaks and vocalic nucleus lengthenings) for lexical segmentation — Three speech production situations are considered: reading, 'aloud recollection' of sentences, quasi-spontaneous dialogue.*

nb. syntagms: number of potential syntagmatic marks ('continuations majeures').
correct (F0): detected F0 peaks corresponding to syntagm endings.
lexical (F0): detected F0 peaks corresponding to word endings (inside syntagms).
errors (F0): detected F0 peaks located on first or intermediate syllables in words.
correct (D): detected vocalic nucleus lenghtenings associated with syntagm endings.
errors (D): detected voc. nucl. lengthenings on first or intermediate syllables in words.

Percentages are calculated in reference to:

- *the number of potential marks (first column), for the second and fifth columns;*
- *the number of correct lexical detections (syntagm+word marks), for columns 3, 4, 6.*

Corpus	nb. syntagms	correct (F0)	lexical (F0)	errors (F0)	correct (D)	errors (D)
LABISE	(479)	**78%**	19%	*9%*	**19%**	*6%*
CMB	(159)	**88%**	7%	*8%*	**30%**	*12%*
METEO	(1282)	**83%**	26%	*10%*	**25%**	*18%*

Discussion

Results in table 1 indicate that our algorithms are capable of detecting 4 out of 5 syntagm boundaries, thanks to the analysis of F0 variations only. Errors (i.e. erroneous word-boundary detection) are due to:

- optional secondary lexical accents in polysyllabic words with 3 or more syllables,
- and expressive/emphatic effects, which take the form of lexical accents but are located on first or intermediate syllables in words (Aouizerat and Lonchamp, 1991).

Since error rates are inferior or equal to 10% in the three conditions, word-boundary marks obtained from the analysis of F0 macrovariations may contribute efficiently to the validation of lexical hypotheses generated from other information sources in speech recognition systems. But better detection rates are needed, in order to ensure the success of

[2] We designate by "speech fragment" any sequence of speech sounds bounded by two successive pauses (i.e. silences of 250 ms or more). For any given speech fragment, the MVD is the median value in the set constituted by the durations of all vocalic nuclei included in this speech fragment.

attempts aiming at improving lexical identification thanks to the generation of word-boundary hypotheses from melodic cues.

Besides, syntagm-boundary detection is unreliable, especially regarding spontaneous speech and reading (cf. column 3 in table 1). Then, results from our algorithm should not be used at the syntactic level: their contribution might drastically reduce the accuracy and efficiency of syntactic analysis.

Results from the study of rhythm are rather disappointing. Error rates are high, especially for spontaneous speech. Moreover, relatively few syntagm boundaries are correctly detected: 1 out of 5 for reading, and 1 out of 4 for spontaneous speech. Besides, accuracy greatly varies from one pragmatic situation to another; differences between reading and spontaneous speech are particularly marked. Which suggests that, in French, intonation conveys more linguistic information than rhythm, and that rhythm is more influenced (than intonation) by other factors, such as pragmatic constraints on speech production.

Therefore, information supplied by the automatic analysis of rhythm cannot be involved in the interpretation of French spoken utterances, at least for the time being. Further studies are necessary, in order to determine the exact contribution of rhythm to speech interpretation.

CONCLUSION

The empirical study on French prosody that we have presented here points to the following conclusions. First, pitch analysis appears as a useful source of information for validating lexical segmentation hypotheses generated by word recognizers that operate on continuous speech, which confirms conclusions from previous studies. Secondly, the contribution of rhythm to lexical segmentation (more generally, to speech interpretation) is not clear, and requires further investigation.

Moreover, our results suggest two research directions that we are currently investigating:

- the definition and testing of criteria for assessing the reliability of prosodic analysis results, with a view to selecting the most robust cues for the generation of word-boundary predictions;
- the refinement of pitch and rhythm analysis, in order to improve syntagm-boundary detection; which implies the interpretation of F0 evolution and syllabic duration variations in terms of macroprosodic groups and patterns.

REFERENCES

Aouizerat H., Lonchamp F. (1992), "Description et génération par règles de l'intonation de la phrase énonciative lue en français", *Séminaire Prosodie*, Aix en Provence, Octobre 1992.

Di Cristo A. (1981), De la microprosodie à l'intonosyntaxe, Thèse d'Etat, Aix en Provence.

Fohr D., Carbonell N., Haton J.-P. (1989), "Phonetic Decoding of Continuous Speech with the APHODEX Expert System", in J.-P. Tubach, J.-J. Mariani (eds.), *Proc. EUROSPEECH 89*, Paris, Septembre 1989, Vol. 2, pp. 609-612.

Gaïtella A. (1991), Rythme et parole: comparaison critique du rythme de la lecture oralisée et de la parole spontanée, Thèse d'Université, Aix en Provence.

Grosjean F., Gee F. (1987), "Prosodic structure and spoken word recognition", *Cognition 25*, pp. 135-155.

Hirst D., Espesser R., Di Cristo A. (1991), "Constituants prosodiques et macro-segmentation du signal: méthodologie et critères d'évaluation", *Actes Séminaire Prosodie et Reconnaissance de la Parole*, Aix en Provence, Mars 1991, pp. 1-8.

Rossi M. (1981), *L'intonation - De l'acoustique à la sémantique*, Paris: Klincksieck.

Vaissière J. (1982), "A Suprasegmental Component in a French Speech Recognition System", *Recherches Acoustiques*, CNET, Vol. VII, pp. 109-125.

Vaissière J. (1988), "The use of Prosodic Parameters in ASR", in *Recent Advances in Speech Understanding and Dialog Systems*, Berlin: Springer-Verlag, pp. 71-99.

Waibel A. (1987), "Prosodic Knowledge Sources for Word Hypothesization in a Continuous Speech Recognition System", *Proc. IEEE ICASSP-87*, Dallas, Avril 1987, pp. 856-859.

Waibel A. (1988), *Prosody and speech recognition*, Londres: Pitman.

Perception of low-anchoring versus high-anchoring of Dutch accent-lending pitch rises

Johanneke Caspers & Vincent J. van Heuven
Holland Institute of Linguistics/Phonetics Laboratory, Leiden University
PO box 9515, 2300 RA Leiden, The Netherlands

ABSTRACT

A perception experiment was conducted to evaluate the hypothesis, derived from earlier production data, that the onset of the Dutch accent-lending pitch rise ('1') is attached to the syllable onset. The accent-lending rise, either anchored with its onset ('low-anchor') to the syllable onset, or with its peak ('high-anchor') at 50 ms after the vowel onset, was presented to listeners in a discrimination and a preference test. Results reveal that, counter to what the relevant literature suggests, there is a preference for the low-anchor, increasing with the magnitude of the shift between the two types of rise.

INTRODUCTION

In the Dutch Intonation Grammar ('t Hart et al. 1990), the *peak* of the standard accent-lending pitch rise is fixed at 50 ms after the vowel onset. This is in concurrence with what is generally assumed in the phonological school of intonation (Pierrehumbert, 1980; Gussenhoven, 1988): in a rising tonal accent, the pitch peak is called the 'target' and is associated with the accented syllable ('H*'). It seems reasonable to assume that this H-target is anchored in the segmental structure. Furthermore, it seems plausible that this important prosodic event coincides with the most salient part of the syllable, i.e. the CV interface (Ohala and Kawasaki, 1984). In production data, however (Caspers and Van Heuven 1992, 1993), we found that the onset of the Dutch accent-lending pitch rise ('1') is relatively fixed with respect to the onset of the syllable, whereas the distance between the peak of the rise and the vowel onset varies considerably under time pressure. We concluded that the onset of the Dutch accent-lending pitch rise is attached to the syllable onset.

We compared our anchor point for the onset of the rise ('low-anchor') with the anchor point used in the Dutch Intonation Grammar: synchronization of the offset of the rise at 50 ms after the vowel onset ('high-anchor'). A discrimination and a preference test were carried out. Assuming that mimicking actual speech production behaviour yields a perceptually adequate alignment of pitch movements, we predict that listeners will prefer our 'low-anchor' over a 'high-anchor' for the Dutch accent-lending pitch rise ('1'), in those cases where a difference can be perceived between both types of rise.

METHOD

Stylized pitch contours from the mentioned production experiment were used as basic material. Eight utterances were selected, containing either isolated accent lending pitch rises ('1'), or so called 'flat hat' intonation contours ('1∅A', i.e. an accent lending pitch rise, followed by an accent lending pitch fall in the second syllable after the syllable containing the rise). The accent lending rise was positioned on four different (CVC) target syllables. Target syllables started either with a relatively short consonant (/m/) or a longer consonant (/p/), followed by a low vowel and an /n/ (i.e. /mɑn/, /ma:n/, /pɑn/ and /pa:n/). The duration of the accent lending rise varied with contour type (220 ms for the isolated rise and 190 ms for the rise in a flat hat). The accent lending rise in each utterance was (i) attached with its onset to the syllable onset ('low-anchor'), and (ii) with its end at 50 ms after the vowel onset ('high-anchor'). After manipulation, the utterances were resynthesized to waveforms, using straightforward LPC resynthesis.

A stimulus consisted of a pair of utterances, separated by a 200 ms silent interval, **only** differing in alignment of the accent lending rise. All possible combinations of low-anchored and high-anchored versions of the accent lending rise were made (i.e. low/high, high/low, low/low and high/high), resulting in 32 stimuli for the discrimination test, 16 'different' and 16 'same'. For the preference test, the 16 different pairs of utterances were used.

Twenty-five naive and 25 experienced listeners (intonologists) participated in the experiment.

Listeners were seated in a sound-isolated booth and listened to the stimuli over good quality headphones. The listener's task was to listen to each of the 32 stimuli and to indicate whether he or she perceived a difference between both utterances or not (discrimination test). In the preference test, listeners indicated for each of the 16 stimuli which of the two utterances they preferred. An ordered list of stimuli was presented online to the subjects (12 kHz, 12 bits, 4.5 kHz LP, 96 dB/oct). They had to press keys to make a pair of utterances audible, and to mark their judgements. Within one trial, subjects could listen to each stimulus as often as they felt necessary.

RESULTS

Overall discrimination and preference responses

In table 1 the overall discrimination and preference responses are presented.

Table 1. *Discrimination responses (absolute and relative frequency of 'different' and 'same' responses) for different and same stimuli (correct responses in shaded cells), and preference responses (absolute and relative frequency of 'pro low-anchor' and 'pro high-anchor' responses).*

	discrimination responses		preference responses	
stimulus	different	same	pro low-anchor	pro high-anchor
different	487 (61%)	313 (39%)	520 (65%)	279 (35%)
same	137 (17%)	663 (83%)		
total	624 (39%)	976 (61%)		

On average, 72% of the utterance pairs were discriminated correctly (shaded cells), and in 65% of all cases the low-anchored rise was preferred to the high-anchored rise. A binomial test showed both frequency distributions to be different from chance (z= 53.08, p<<.001 and z= 8.49, p<<.001). These results indicate that listeners are able to hear the difference between the two ways of timing the accent-lending pitch rise, and that a preference exists for the alignment of the onset of the rise with the syllable onset (low-anchor) to an alignment of the end of the rise at 50 ms after the vowel onset (high-anchor).

The experienced listeners discriminated different stimuli better than naive listeners (70 vs 52% correct responses, $\chi^2 = 27.97$, df = 1, p<<.001). For the class of same stimuli, no effect of experience is found (84 vs 80% correct responses, $\chi^2 = 1.49$, df = 1, ins.). As opposed to the discrimination data, there is no association between experience and the preference responses ($\chi^2 = 0.88$, df = 1, ins.). Both groups of listeners prefer the low alignment of the rise in ca. 65% of the cases.

Post hoc analysis: effect of 'shift size'

The shift in alignment between the low-anchored and high-anchored rises relative to the segmental structure, is influenced by the duration of the rise and the duration of the initial consonant, and has a magnitude of 40, 60, 70, 80 or 100 ms. In a post hoc analysis, the effect of the magnitude of this 'shift size' on the discrimination and preference responses was examined directly. We hypothesized that larger shift sizes will be easier to perceive, and therefore will produce clearer preferences than smaller shifts. The results are presented in figure 1a and b.

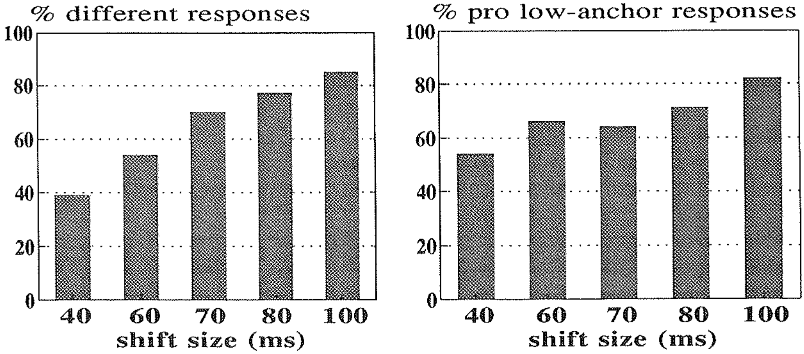


Figure 1. *Percentage of 'different' (a) and 'pro low-anchor' (b) responses to different stimuli, broken down by shift size in ms.*

It is clear from figure 1a, that discrimination improves with the shift size: the larger the shift, the more correct responses are given. The effect of shift size is significant ($\chi^2 = 87.10$, df = 4, p<<.001). A comparable effect of shift size is found for the preference responses ($\chi^2 = 25.16$, df = 4, p<<.001). This means that the preference for the low-anchor grows with the shift size, as expected.

CONCLUSION AND DISCUSSION

We found that in 65% of the cases, the low-anchor was preferred to the high-anchor, which largely confirms our hypothesis. There was no effect of the experience of the listener on the preference responses, which means that to experienced as well as naive listeners an anchoring of the onset of the rise is more acceptable than an anchoring of the offset of the rise. The magnitude of the shift between low-anchored and high-anchored rises affects the preference responses considerably: the larger the difference in timing, the stronger the preference for the low-anchored version of the rise is, resulting in over 80% pro high-anchor judgements for the largest shift size (100 ms). The results of this perception experiment lend more credibility to our claim that the onset of the Dutch accent-lending pitch rise is attached to the syllable onset rather than with its offset to the vowel onset.

It is possible that a connection exists between the preference for a low-anchored rise and the P-centre phenomenon (psychological moment of occurrence of a syllable). The location of the perceptual centre is strongly correlated with the duration of the initial consonant(s) (Pompino-Marschall, 1990). Further research into the relationship between P-centres and timing of pitch movements is required.

ACKNOWLEDGEMENT

This research was supported by the Linguistic Research Foundation, which is funded by the Netherlands Organization for Scientific Research, NWO, under project # 300-173-005.

REFERENCES

J. Caspers and V.J. van Heuven (1992), "Phonetic properties of Dutch accent lending pitch movements under time pressure", *Proc. Internat. Conf. Spoken Lang. Proc., Banff, 12-16 October 1992*, Vol. 1, pp. 731-734.

J. Caspers and V.J. van Heuven (1993), "Effects of time pressure on the phonetic realisation of the Dutch accent lending pitch rise and fall", in *Phonetica* (to appear).

C. Gussenhoven (1988), "Adequacy in intonation analysis: The case of Dutch", in *Autosegmental studies on pitch accent*, ed. by H. van der Hulst and N. Smith (Foris, Dordrecht), pp. 95-121.

J. 't Hart, R. Collier and A. Cohen (1990), *A perceptual study of intonation* (Cambridge University Press, Cambridge).

J.J. Ohala and H. Kawasaki (1984), "Prosodic phonology and phonetics", in *Phonology Yearbook*, Vol. 1, pp. 113-127.

J.B. Pierrehumbert (1980), *The phonology and phonetics of English intonation* (PhD Dissertation, Massachusetts Institute of Technology).

B. Pompino-Marschall (1990), *Die Silbenprosodie* (Tübingen, Niemeyer).

A Cognitive Approach to Planning and Representation of Prosodic Features in a Concept-to-Speech System

Carsten Günther
Universität Hamburg
Fachbereich Informatik, AB Wissens- und Sprachverarbeitung (WSV)
Bodenstedtstr. 16, D-22765 Hamburg
e-mail: guenther@informatik.uni-hamburg.de

ABSTRACT

This paper presents a cognitive approach to prosodic planning in a language generation system. The macro- and microprosodic planning component is part of a computational modeling of main processing stages of the human language production process. The architecture of the system which is motivated by appropriate psycholinguistic insights, and some representational formats of prosodic knowledge are introduced.

INTRODUCTION

This paper is concerned with the phonological and phonetic planning component of the SYNPHONICS (Syntactic and Phonological Realization of Incrementally Generated Conceptual Structures) Formulator, and in particular with the representation of phonological and phonetic knowledge. The SYNPHONICS approach to the computational modeling of natural language production takes into consideration results from psycholinguistic research about the time course of the human language production process as well as recent developments in theoretical linguistics and phonetics concerning the representation of syntactic, phonological, and phonetic knowledge. The crucial point of linguistic investigation lies in the analysis and modeling of the syntactic and prosodic realization of different information structures (e.g. focus-background structure) in accordance with conceptual and contextual variations. The SYNPHONICS Formulator is the central part of the SYNPHONICS System[1], which is at present in a conceptual stage and will comprise the whole generative processing of utterances from pre-linguistic conceptual structures over complex semantic/syntactic/phonological structures onto acoustic parameter sets for controlling a speech synthesizer.

In our approach, language production is seen as an incremental process (Levelt 1989) which combines parallel and serial processing. Therefore, the planning processes must be hold local and must act over incomplete structures (e.g. there will be no preplanning of metrical trees or of complete intonation contours over whole utterances). This assumption about processing properties adheres to a special relational account of linguistic structures. In this account we assume that semantic, syntactic, and phonological information can be linked to each other, building a complex sign with inherent constraints. In abandoning a strictly functional dependency of phonological structure on syntactic structure we assume a direct interrelationship between semantic and phonological structure[2]. Such a view directly influences the organization of the prosodic planning processes and also the structure of the processing units (increments). In the next section, the architecture of the phonological and phonetic encoder within the SYNPHONICS Formulator will be described.

[1] For a detailed description of the SYNPHONICS system see Herweg (1992) and Schopp (1993).

[2] Evidence for the relevance of such a direct relationship between semantic and prosodic structure is shown by means of examples of focus/background structuring in Günther et al. (1993).

THE MODEL OF THE PHONOLOGICAL AND PHONETIC ENCODER

The architecture model of the phonological and phonetic encoder (Figure 1) with its processing steps of phonological, phonetic-articulatory, and acoustic encoding shows a clear separation of declarative knowledge components from procedural control components. This is due to the integration of a declarative grammar component (a variant of a HPSG (Pollard&Sag 1992) for German) in a procedural control structure. The grammar component is expanded by abstract semantics/phonology schemes (e.g., a Focus-Accent Scheme), a detailed phonological lexicon , and prosodic principles.

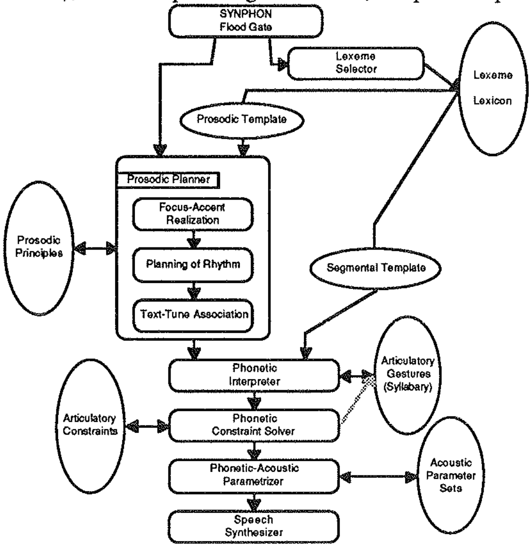


Figure 1. The architecture of the phonological and phonetic encoder.

Within the SYNPHONICS Formulator, we will represent phonological and phonetic information according to the event-based paradigm (Bird&Klein 1989). But the type of events and their properties will be separated into phonological and phonetic ones, reflecting the peculiarities of phonological and phonetic processes. During phonological encoding, we use a type hierarchy based on autosegmental phonology, whereas during phonetic encoding, we prefer a type hierarchy and property assortment that takes its bearings from articulatory processes (Browman and Goldstein 1989). A separation of phonological and phonetic encoding processes with adjusted structures increases the modularity of speech production models and applies to the research topic of current language generation investigation: Each decision or evaluation has to be carried out on its hereditary processing stage.

The *SYNPHON Flood Gate* forms the interface from semantic and syntactic to phonological planning. This module ensures the incremental subsequent treatment of already semantic and syntactic specified utterance fragments. The *Flood Gate* selects structure units which meet the inherent needs of phonological processes. According to psycholinguistic investigations (Levelt 1989), *phonological phrases* or *accent domains* (Gussenhoven 1983, Ladd 1983) - a semantic pendant to the well known syntactically defined phonological phrase - are conceivable to be such incremental units.

The *Lexeme Selector* selects the corresponding lexemes from the lexicon by dereferencing the lexeme pointer (an abstract address determined during lemma selection) and using syntactic agreement information as well as case information. Only this second lexicon access (after lemma selection) makes available the concrete word form information (Levelt 1992a). In the *Lexeme Lexicon*, morphological, metrical and segmental information are stored. This information is specified during lexical-phonological spellout processes. Figure 2 shows a prosodic specified lexical entry of the proper name *Hans* in an HPSG-like style (semantic, syntactical, morphological, and concrete subsegmental event information is omitted).

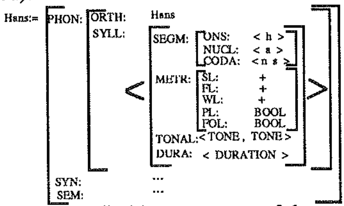


Figure 2. A prosodic specified lexeme entry of the proper name *Hans*.

The feature SYLLabic is a list of syllable templates with syllabic, metrical, tonal, and durational information. A complex type hierarchy of METRical types permits prominence rules to interpret feature values (of type BOOLean) as metrical grid positions. The number of grid levels is limited to five prosodic structure levels (syllable, foot, phonological word, accent domain, intonational phrase). The TONAL and DURAtional features are not specified till application of postlexical prosodic rules. The feature TONAL has two tone place holders because some syllables can carry an accent tone and a boundary tone (e.g., focused monosyllabic words at the end of a sentence).

The *Prosodic Planner* obtains results from the lexical-phonological spellout processes. The system's architecture allows a metrical and a tonal planning which is independent of the concrete segmental spellout. Such an architecture corresponds to the findings in psycholinguistic priming experiments of speech production (Levelt 1992b). Within this module, the macroprosodic planning takes place. The *Prosodic Planner* derives an abstract prosodic structure[3] of the utterance from semantic, syntactical and lexical information according to prosodic principles. Prosodic planning involves the projection of the focus structure onto prominence structure (specifying the feature METRic of the syllable template), the rhythmical planning (e.g., in order to avoid stress clashes) and the text-tune association (specifying the feature TONAL for representing pitch accent or boundary tone). The internal ordering of the prosodic planning steps has to follow prosodic structure building constraints, e.g. in accordance with mutual dependencies between metrical structures, boundary tones, and accent tones. Figure 3 shows an HPSG-like rule which ensures that the word accent bearing syllable of a focused word will become the most prominent one within an intonational phrase and carry a high accent tone.

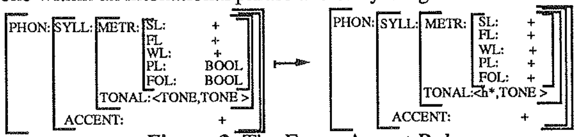


Figure 3. The Focus-Accent Rule.

The next processing stage is the *Phonetic Interpreter*, which forms our interface between phonology and phonetics and deduces a phonetic-articulatory event structure from abstract prosodic and segmental information by paying attention to segmental phonetic parameters. An interplay between global abstractly planned prosodic features and segment specific parameters takes place determining the concrete phonetic events which realize the prosodic

[3] The psychological reality of an abstract prosodic structure representation during sentence production was recently demonstrated by Ferreira (1993).

features. The standard articulator hierarchy of Browman and Goldstein's proposal (1989) is expanded by the articulator *jaw*, which is necessary in order to plan correctly the co-articulation effects and formant transitions of vowels. The phonetic interpretation of sub- and suprasegmental information relies upon a declarative knowledge base, the syllabary (Levelt 1992b). According to psycholinguistic investigations, the eventual increments which will be handed over from phonological encoding to phonetic interpretation are metrically structured, subsegmentally underspecified syllables or phonological words. These structures serve as the access code to the appropriate gestural score.[4] In order to model the cognitive language production process, we represent articulatory gestures as well as articulatory scores within our syllabary. Articulatory plans of syllables are already fully specified temporally and only the syllable environment has to be taken into account. But in case of assembling an articulatory unit from single gestures an articulatory constraint satisfaction process must be performed. At this point it is possible to implement a learning process which enlarges the syllabary in the case of frequently appearing articulatory scores.

The output of the phonetic constraint solver controls a speech synthesizer. Because of using an acoustic synthesizer, namely a Klatt-based formant synthesizer[5], a phonetic-acoustic interface is required. This module calculates the acoustic control parameters in accordance with the articulatory targets on the different articulatory event tiers.

ACKNOWLEDGMENTS

The research reported in this paper is carried out in a research project which is funded by the German Science Foundation (DFG) under grant no. Ha 1237/4-1.

REFERENCES

Bird,S. & E. Klein (1989): "*Phonological Events*", Research Paper EUCCS/RP-24. Centre for Cognitive Science, U Edinburgh.

Browman, C.P. and L. Goldstein (1989), "Articulatory gestures as phonological units", *Phonology*, Vol. 6, pp. 201-251.

Ferreira, F. (1993), "Creation of Prosody During Sentence Production", *Psychological Review*, Vol. 100, pp. 233-253.

Günther, C., Ch. Habel, C. Maienborn and A. Schopp (1992), "What´s up with the printer? - Context relative presentation of conceptual structure and its prosodic realization in a language production system", in Schopp (1993), pp. 5-16.

Gussenhoven, C. (1983), "Focus, mode and the nucleus", *J. of Linguistics*, Vol. 19, pp. 377 – 417.

Herweg, M. (ed., 1992), *Hamburger Arbeitspapiere zur Sprachproduktion - I*, GK-Kognitionswissenschaft, AP 9, Uni Hamburg.

Ladd, D.R. (1983), "Even, focus and normal stress", *J. of Semantics*, Vol. 2, pp. 257-270.

Levelt, W.J. (1989), *Speaking: From Intention to Articulation,*. Cambr., Mass., MIT Press.

Levelt, W.J. (1992a), "Accessing word in speech production: Stages, processes and representations", *Cognition*, Vol. 42, pp. 1-22.

Levelt, W. J. (1992b), *Timing in Speech Production with Special Reference to Word Form Encoding*, Ms. MPI Nijmegen.

Pollard, C. & I. Sag (1992), *Head-driven Phrase Structure Grammar,* Ms., April 1992, Stanford, CSLI (to appear by CSLI).

Schopp, A. (ed., 1993), *Hamburg Working Papers on Language Production - II*, GK-Kognitionswissenschaft, AP 13, Uni Hamburg.

[4] If articulatory scores are stored syllable-sized in the syllabary then the retrievel should reveal a frequency effect. And indeed, low-frequently syllables are harder to access as high-frequently ones (Levelt 1992b).

[5] For synthesizing speech, a Klatt-based synthesizer was kindly made available to us by the Institute for Technical Acoustics of the Technical University of Dresden.

A Transdisciplinary Model for Prosody Applied to the Teaching of Clinical Populations

Patricia M. Hargrove, Ph.D
Department of Communication Disorders
Mankato State University, Mankato MN, USA 56002
and
Nancy S. McGarr, Ph.D.
Department of Speech, Communication Sciences and Theatre
St. John's University, Jamaica NY, USA 11438
and
Haskins Laboratories, New Haven CT, USA 06511

ABSTRACT

This presentation describes the Prosodic Teaching Model which is an integrated approach to viewing the prosody of speech and language impaired persons. The Prosodic Teaching Model consists of two major divisions: prosodic features and prosodic components. The prosodic features include pitch, loudness, duration, and pause; the prosodic components comprise tempo, intonation, stress, and rhythm.

BACKGROUND

The purpose of this presentation is to explain the Prosodic Teaching Model, an organizational framework for teaching prosody and for using prosody to facilitate communication with speech, language, or hearing impaired persons. As an applied, transdisciplinary model, the Prosodic Teaching Model is derived from theoretical and research literature, from clinical experience, and from clinical research. The model is specifically concerned with teachability issues and attempts to account for both empirical and anecdotal information about the teaching of English prosody from a variety of fields including articulation, phonology, fluency disorders, neurogenics, developmental delay, voice problems, child language impairment, hearing impairment, English as a second language, special education, learning disabilities, and music therapy.

In many instances, clinical approaches adopted for speech and/or language training in each of the aforementioned areas are fairly parochial and thus important information from other disciplines or disorders may be inadvertently overlooked. The use of an organized system that addresses this entire literaturé, such as the Prosodic Teaching Model, may improve clinical practice by assisting the clinician in identifying different aspects of prosody that should be treated and in ordering treatment objectives. The use of such a model to set priorities and to organize treatment is critical given the lack of empirical data regarding prosody in the clinical literature. Such an approach encourages more rigorous research and questioning of current pedagogical protocols.

DESCRIPTION

The Prosodic Teaching Model advocates a traditional division of prosody into two areas which may be particularly helpful in facilitating speech and language training: (1) prosodic features and (2) prosodic components. Prosodic features, which may be considered the building blocks for the prosodic components, include pitch, loudness, duration and pause. Combined, these prosodic features are implemented (realized as) the varying prosodic components of tempo, intonation, stress and rhythm. For the clinically impaired client to actualize the linguistic message, he or she must be able to perceive and to produce these prosodic features. A failure to marshal one or more features is a hallmark of many disordered groups. Indeed, some might argue that the clustering of specific error types perceptually distinguishes one disordered group from another and may be diagnostically significant.

Prosodic Features

Prosodic features (pitch, loudness, duration, and pause) are building blocks of the prosodic components and have been described in terms of their perceptual and productive characteristics. Analyses of both normal and disordered speakers have considered prosodic features from descriptive, acoustic, and physiological perspectives. Linguistic models, especially of normal speakers, often take these features as a given. However, in work with disordered speakers, they must receive prime attention. The inability of many disordered speakers to perceive or to produce one or more of these features, may be paramount to failure to achieve prosodic components. A significant literature exists that details aspects of the prosodic features pitch, loudness, duration, and pause in a variety of disordered groups and therapeutic protocols specific to each group have been attempted.

Clients need not use an "idealized" pattern of prosodic features to produce a particular meaning. If control of "typical" prosodic features is beyond their capabilities, impaired speakers may use only one of the several prosodic features usually used to produce a particular prosodic component or they may develop compensatory patterns to produce the targeted prosodic component. Alternately, if clients have sufficient control of the prosodic features, clinicians may consider using them to improve other aspects of communication (e.g., increased duration may be used to improve fluency).

The four prosodic features are described briefly and their categories are listed below:

A. Pitch is the auditory perception primarily associated with the acoustic dimension of frequency. The categories of pitch are

1. Pitch height (average vocal pitch)

2. Pitch slope and declination (time element involved in increases or decreases in F_0 at the syllable, word, phrase, or sentence level. Declination refers to one type of decreasing pitch slope--the pitch slope from the onset of an utterance to the end of the utterance.)

3. Pitch direction (perception of pitch change)

4. Pitch variation (pitch range or pitch width)

B. Loudness is the auditory perception most often associated with the acoustic dimension of amplitude or intensity. The categories of loudness are

1. Loudness level (magnitude of the excursion from an arbitrary reference point)

2. Loudness variation (range of loudness levels)

C. Duration is the auditory perception associated with the acoustic dimension of time. The categories of duration include

1. Inherent duration (length modifications pertaining to speech sounds)

2. Prosodic duration (changes in length that are not related to differences among speech sounds but, nevertheless, having communicative value). Changes in prosodic duration are concerned with the production of tempo, intonation, stress, and rhythm.

D Pause is the auditory perception primarily associated with silence. The categories of pause are

1. Intraturn pauses (gaps in vocalization that occur within a single speaker's turn). There are two variations of intraturn pauses: (a) lexical and (b) phrasal.

2. Interturn pauses (pauses that occur at the end of a speaker's turn)

Prosodic Components

Prosodic components are the elements that constitute prosody and include tempo, intonation, stress, and rhythm. The use of these components have linguistic significance. In the

Prosodic Teaching Model, clinicians must determine how a failure to achieve a prosodic feature may result in an error in a prosodic component and attempt to augment feature cues or provide clients with alternate strategies. Alternately, if clients possess sufficient control of their prosodic features, clinicians must determine if they express sufficient and appropriate meanings using prosodic components. If clients' prosodic components are sufficient and appropriate, clinicians may consider using prosodic components to facilitate other aspects of communication (e.g., changes in stress patterns may be used to promote intelligibility). Again, clinical resources address these prosodic components, but there is little cross referencing from one disorder to the next. The following briefly describes each of the four prosodic components.

A. Tempo involves use of timing elements, such as rate of speech, to impart meaning. The categories of tempo include

1. Speaking rate (number of syllables produced over a given amount of time). Rate changes are accomplished by (a) duration changes, (b) pause, and/or (c) phonemic changes.

2. Concordance (the movement from one element in an utterance to the next element). There are two factors that contribute to the quality of concordance: (a) insertion or elimination of pauses and (b) phonetic blending.

3. Phrasing (speaker's marking of the beginning or the ending of phrases)

B. Intonation is the communicative use of pitch. The categories of intonation are divided into two major subdivisions.

1. The internal organizational level of intonation:

a. Onset (pitch height of the first full syllable in an utterance)

b. Nucleus (most prominent syllable)

c. Terminal contour (final pitch direction of the last syllables)

d. Overall contour (holistic representation the pitch configuration from the initiation of the utterance to its end)

2. The external organizational level of intonation:

a. Cohesive devices (stretches of identical or related intonation patterns that extend across utterance boundaries)

b. Pitch Agreement (degree of concordance or agreement in pitch height between the end of one utterance and the beginning of the next).

C. Stress is the use of prominence for purposes of communication. The three stress categories include

1. Lexical stress (pattern of the stressed and unstressed syllables at the word level). Lexical stress occurs in (a) multisyllabic words, (b) weak/strong forms, and (c) stress shifting.

2. Phrasal stress (most prominent syllable in a phrase or sentence)

3. Emphatic stress (strong level of prominence)

D. Rhythm is concerned with the use of sequences of stresses and the flow of speech during communication. The categories of rhythm are

1. Stress sequences (stressed syllables at regularly perceived intervals)

2. Alterations (changes in tempo, intonation, and/or articulation)

3. Continuity (the ability to maintain an uninterrupted flow of speech)

SUMMARY

This presentation describes an integrated approach to organizing information about treating prosody (i.e., the Prosodic Teaching Model). The analysis of treatment objectives from a variety of disciplines, from the theoretical literature, and from clinical studies serves as the basis for the model which attempts to provide clinicians with a firmer grasp of the nature of communication problems and ways to facilitate appropriate treatment.

Experimental Study on the Role of Prosodic Features in the Human Process of Spoken Word Perception

Keikichi Hirose, Nobuaki Minematsu and Mika Ito
Department of Electronic Engineering, Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

ABSTRACT

As a step toward the clarification and formulation of the human process of speech perception through prosodic features, perceptual experiments were conducted on the transmission of accent types as well as on the transmission of word meaning using synthetic speech of Japanese words of four morae. As for the transmission of word meaning, the role of prosodic features for word identification was found to be largest for type 1 accent. This is because the type 1 accent has a falling in the fundamental frequency contour at the beginning portion of the word and, therefore, the prosodic features can be utilized before the perception of whole segmental features. As for the transmission of accent types, several results were obtained implying the existence of the process for accent-type identification aside from the process for perceiving segmental features.

INTRODUCTION

Although the segmental features of speech acoustically play the dominant role in the human process of speech recognition, the prosodic features may also play an important role. This is because the prosodic features of speech are tightly related to the linguistic information of an utterance, such as the word meaning, the syntactic structure, and the focal condition. With the premise that the knowledge on the human process of speech perception should be incorporated in the machine systems for speech recognition to improve their current performance, we have been conducting several perceptual experiments to clarify and to formulate the process (Fujisaki, Hirose, Ohno and Minematsu 1990, Minematsu, Ohno, Hirose and Fujisaki 1992). These experiments, however, were restricted to the segmental features of speech and did not cover the prosodic features.

From the viewpoint above, as a first step toward the clarification and formulation of the process of speech recognition through prosodic features, we have conducted perceptual experiments on the transmission of accent types as well as on the transmission of word meaning using synthetic speech of 4-mora words of Japanese with their fundamental frequency contours manipulated. In this paper, each of both experiments is separately explained followed by the discussion on the results.

WORD ACCENT OF JAPANESE

For an n-mora word of Japanese, (n+1) accent types are possible in the Tokyo dialect. These are denoted by "type i" accents (i=0 to n). Each of type i accents has a rapid downfall in the fundamental frequency contour respectively at the end of ith mora except for the case of i=0 without apparent downfall. When uttered in isolation, type n accent has a fundamental frequency contour similar to that of type 0 accent. For the current perceptual experiments, isolated utterances of 4-mora words were used with type 4 accent excluded. Utterances of a male speaker were recorded and digitized with 10 kHz sampling frequency and with 12 bit accuracy for the further process. Stimuli for the perceptual experiments were generated by the PARCOR analysis-synthesis method with manipulations on the fundamental frequencies. For the current experiments, the manipulations were conducted only on the fundamental frequency contour and no on other prosodic parameters, such as the syllable duration and the source power. This is because of the priority of the fundamental frequency contour in the acoustic manifestation of the prosodic features of Japanese speech.

ROLE OF ACCENT TYPES FOR THE IDENTIFICATION OF SPOKEN WORDS

Method of experiment

As shown in table 1, utterances of 12 nouns were selected for each of the accent types 0, 1, 2 and 3. Following three types of manipulation were conducted on the fundamental frequency contours during the process of PARCOR analysis–synthesis:

(Case 1) Keeping constant at 100 Hz,
(Case 2) alternating into other accent types,
(Case 3) with no modification.

Manipulation for case 2 was performed based on the model of fundamental frequency contour generation (Fujisaki and Hirose, 1984). Fundamental frequency contours of alternated accent types were generated by the model after shifting the onset and the offset of the accent command to their typical values. A band elimination of 0.5 kHz to 3.0 kHz was further performed for all of the synthetic speech samples so that the subjects of the perceptual experiment may perceive a sample as a whole. The syllable-based recognition using the segmental features is difficult for the band–eliminated speech stimuli. These stimuli were presented through headphones with 4 sec inter–stimulus interval to 10 male subjects of Japanese who were asked to reproduce the words orally. The experiments were conducted for the three cases shown above in the order of 2, 1 and 3.

Type 0 to Type 1:	"raion"	"akabou"	"ninjin"	"naiyou"
Type 0 to Type 2:	"shingou"	"omatsuri"	"yokujitsu"	"amerika"
Type 0 to Type 3:	"hiroshima"	"aimai"	"orugan"	"raihin"
Type 1 to Type 0:	"nekutai"	"koumori"	"wakuchin"	"randamu"
Type 1 to Type 2:	"naitaa"	"monoraru"	"amazon"	"unsei"
Type 1 to Type 3:	"kamakiri"	"unmei"	"ookami"	"enbun"
Type 2 to Type 0:	"imomushi"	"norimaki"	"omusubi"	"yononaka"
Type 2 to Type 1:	"mimizuku"	"katakori"	"onigiri"	"nodoame"
Type 2 to Type 3:	"aomori"	"toraburu"	"murasaki"	"origami"
Type 3 to Type 0:	"kamisori"	"tamanegi"	"nokogiri"	"machigai"
Type 3 to Type 1:	"kaminari"	"nissuu"	"nenryou"	"nonbiri"
Type 3 to Type 2:	"kaminoke"	"nakigoe"	"noumiso"	"tenkizu"

Table 1. *Four–mora words used for the experiment.*

Experimental results and considerations

Figures 1 and 2 show the word recognition rate separately for the three cases. The recognition rates of Fig. 1 are calculated for each original accent type, while those of Fig. 2 are calculated for each accent type after the case 2 alternation. Case 2' indicates the rates of accent–type recognition. In both figures, the recognition rate of each accent type has a similar value for case 3 samples without modification in fundamental fre–

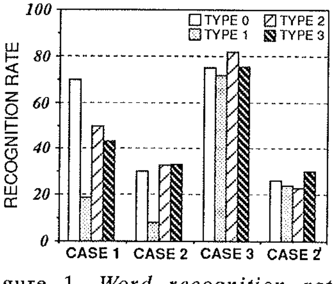


Figure 1. *Word recognition rates summarized separately for each original accent type.*

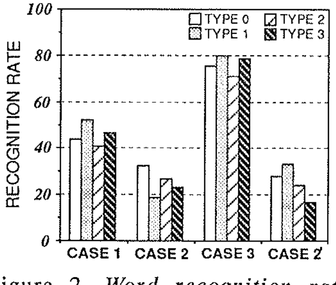


Figure 2. *Word recognition rates summarized separately for each accent type after alternation.*

quency contours. As clearly indicated in Fig. 1, the largest drop in the recognition rate due to the accent type alternation is observed for the words with type 1 accent. In Fig. 2, the largest drop is observed when non–type 1 accent is altered to type 1 accent. These results imply the greater role of prosodic features for the perception of words with type 1 accent. The recognition rate of accent types has the largest score for the samples with type 1 accent as shown in Fig. 2. This result may support the above findings on the role of type 1 accent for the word identification process. The reason of the findings is considered to be due to the fact that type 1 accent has a falling in the fundamental frequency contour at the beginning portion of the word and, therefore, the prosodic features can be utilized before the perception of whole segmental features. Supposing the importance of the prosodic features for the word recognition is affected only by the location of the downfall in the fundamental frequency contour, the role should decrease in the order of types 2, 3 and 0. No result, however, was found supporting this hypothesis. This implies the existence of the perceptual mechanism which makes possible to recognize the word only with the acoustic features of the first half of the utterance.

PERCEPTION OF WORD ACCENT

Method of experiment

Utterances of 4–mora words of type 3 accent were selected for the experiment on the transmission of accent types. Three Japanese words "aozora (blue sky)," "genshiro (atomic reactor)," "korigori (have had enough and never do it again)" were selected together with three nonsense words "imeyuro," "nemeira," "ominere." For each of these words, synthetic speech samples were prepared with accent types 0, 1, 2, 3 and 6 artificial accent types not found in the Tokyo dialect. As shown in Fig. 3, the same rising/falling pattern in fundamental frequency contour was realized at the boundary of 3rd and 4th morae for each sample. The following two types of gating techniques were then applied to the samples to produce stimuli for the perceptual experiment:

(Case A) retaining the initial portion of x msec and replacing the rest by silence,

(Case B) adding to case A, replacing the portion of first and second morae by silence.

These replacements are shown schematically in Fig. 4. Case B was planned to investigate the effect of initial part of an utterance for the accent–type identification.

The word stimuli were presented through headphones to the subjects in several sessions for each of 6 words. Each session includes 10 stimuli, viz., one for each of 10 accent types. The gating duration x for stimuli in the first session was set equal to the total duration of the first to the third morae and was increased in 5 msec steps. The experiment was conducted firstly for the stimuli of case A with an inter–stimulus interval of 6 sec, and was then conducted for those of case B with an inter–stimulus interval of 2 sec. The subjects were asked to reproduce the accent types by schematically drawing them on a sheet of paper specially prepared. Based on their answers, necessary duration x was decided for the identification of the accent types.

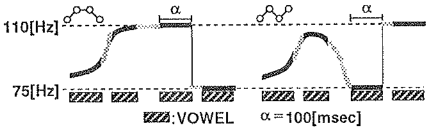


Figure 3. *Method for manipulating the fundamental frequency contours.*

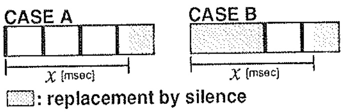


Figure 4. *Schematic illustration for the two cases of silence replacement.*

Experimental results and considerations

The results are shown in Figs. 5 and 6 separately for 3 Japanese words and 3 nonsense words. The ordinate denotes the length of 4th morae in the gating period. As for the case B, the accent types are identified with the shorter length for the words with rising fundamental frequency contour at the boundary of 3rd and 4th morae than those with falling fundamental frequency contour. This result indicates the rising is perceived faster than the falling. As for the case A, the length for accent type identification is shorter for the Japanese words than for the nonsense words. If we compare the results of the two cases for the words of accent type 3, the downfall in the fundamental frequency contour is shown to be perceived for shorter gating period in case A than in case B. This fact is found both for the Japanese words and the nonsense words. Assuming that human may have no inner dictionary for nonsense words, the above fact implies that an accent dictionary of known accent types exists aside from the ordinal word dictionary with information on part of speech, meaning and others, and that a process exists to perceive input spoken words as if they are accompanied by one of the known accent types. A pointer may exist from each item of the accent dictionary to each item of the inner word dictionary.

The above hypothesis was also supported by another preliminary experiment where the spoken word stimuli with original and alternated accent types were presented to the subjects in isolation and continuously. A larger drop in the rate of word identification due to the continuous presentation was observed for the stimuli with alternated accent types than for those with original accent types. Incorrect pointing to the inner word dictionary may occur in the case of alternated accent types and may largely degrade the performance of word identification based on the segmental features typically in the case of continuous presentation.

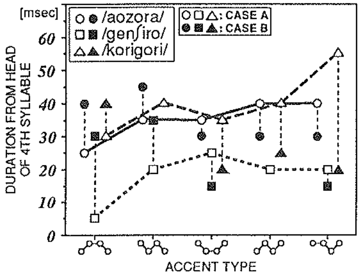


Figure 5. *Duration of 4th mora necessary for the identification of accent types for Japanese words "aozora," "genshiro," and "korigori."*

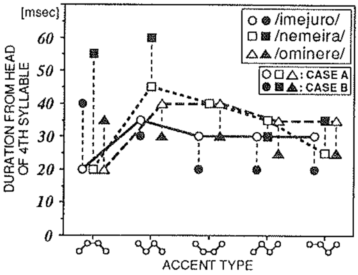


Figure 6. *Duration of 4th mora necessary for the identification of accent types for nonsense words "imeyuro," "nemeira," and "ominere."*

CONCLUSION

Perceptual experiments were conducted on the role of prosodic features for the identification of spoken words. It was found that the role is largest for type 1 accent. Several results were also obtained implying the existence of the process for accent-type identification. The above experiments, however, were restricted to word level information. Further experiments are necessary to examine the role of prosodic features in the process of perceiving higher-order linguistic information, such as syntactic structures.

REFERENCES

H. Fujisaki & K. Hirose (1984), "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn.*, Vol.5, No.4, pp.233-242.

H. Fujisaki, K. Hirose, Sumio Ohno & N. Minematsu (1990), "Influence of context and knowledge on the perception of continuous speech," *Proc. Int'l Conf. on Spoken Language Processing, Kobe, 18-22 November 1990*, 10.9, Vol.1, pp.417-420

N. Minematsu, S. Ohno, K. Hirose and H. Fujisaki (1992), "The influence of semantic and syntactic information on spoken sentence recognition," *Proc. Int'l Conf. on Spoken Language Processing, Banff, 12-16 October 1992*, Tu.fPM.4.5, Vol.1, pp.153-156

Multi-lingual modelling of intonation patterns

Daniel Hirst*, Albert Di Cristo*, Martine Le Besnerais**, Zohra Najim*, Pascale Nicolas* & Pascal Roméas*.
*Institut de Phonétique d'Aix, URA CNRS 261 Parole et Langage,
Université de Provence, Aix en Provence, France
**Universidad Autónoma de Barcelona, Spain

ABSTRACT

An account is given of an ongoing research programme developping a general method of investigation applicable to the intonation systems of different languages. At present, four languages in particular are being studied. : French, English, Spanish and Arabic. Four levels of prosodic representation are distinguished : underlying phonological, surface phonological, phonetic and acoustic. Each level is required to be interpretable in terms of the immediately superior and inferior levels. Automatic and semi-automatic procedures are being developped for this task.

INTRODUCTION

In this paper we present preliminary results of an ongoing research programme developping a general method of investigation applicable to the intonation systems of different languages. At present, four languages in particular are being studied. Three of these are Indo-European languages : French (Romeas 1991; Di Cristo in press; Nicolas 1992 in progress), English (Hirst in press), and Spanish (Alcoba et al. 1993; Murillo & Alcoba in press; Le Besnerais in progress), while one is a non Indo-European language : Arabic (Benkirane in press, Najim & Hirst in press; Najim in progress).

These four languages present interesting differences in their prosodic systems. From a phonological point of view it has been claimed that the intonation systems of English and French differ in at least two parameters.

P1. Whereas English accent groups are traditionally held to be organised into left-headed structures, with initial prominence (Pike 1945, Jassem 1952, Abercrombie 1964), accent groups in French are regularly structured the opposite way, into right-headed structures with final prominence (Wenk & Wioland1982; Hirst & Di Cristo 1984).

P2. A second parameter concerns the tonal sequence usually found on phrase internal Tonal Units (=accent groups) in the two languages. It has been argued (Hirst & Di Cristo 1984, in press) that while this tonal sequence is most commonly [H L] in unmarked declarative utterances in English, in similar French utterances it is usually [L H].

There is some evidence (Hirst & Di Cristo in press) that the first of these two parameters, P1, might in fact characterise a difference between Germanic and Romance languages in general. From this point of view a comparison of results obtained on English and French with those obtained on Spanish provides an interesting control of this prediction while the inclusion of a non Indo-European language should make it possible to see how far such a parameter can be applied to languages from other linguistic phyla.

Levels of representation and description.

At one extreme we may distinguish an abstract level of cognitive representation (phonological) and at the other the level of observation of physical data (acoustics, physiology etc.). It has been argued (Hirst 1992) that between these we should distinguish at least two intermediate levels, the level often referred to as "phonetic transcription", which is in fact a level of surface phonology, and the essentially hybrid level of phonetic representation, where *phonetics* is taken to constitute the interface between the cognitive (phonological) and the physical (acoustic) levels.

Each level of description is required to satisfy the *Interpretability Constraint* : each level i must be able to be interpreted on both levels $i+1$ and $i-1$ when such levels exist.

One of the most general questions we hope to be able to address in this research is to what extent the prosodic variability which is to be observed between different languages can be attributed to language specific parameters on each of the different levels.

Phonological representation of intonation patterns

While the exact nature of such a representation is of course unknown we assume,as in a number of recent models, that an intonation pattern can be derived from a phonological structure to which language specific templates associate appropriate tonal segments (for discussion cf Hirst & Di Cristo in press).

Phonetic modelling of F0 curves.

A number of different techniques have been developped in recent years for automatically generating fundamental frequency patterns of synthetic speech from symbolic input. Less research has been devoted to the inverse problem : the automatic coding of fundamental frequency patterns by symbolic output. There have been even fewer attempts to "model" such patterns where the output of the automatic coding is directly usable as input for the automatic synthesis system. Such an automatic modelling system (described in more detail in Hirst & Espesser in press) has recently been developped in our Institute. The output of the programme MOMEL is a sequence of target points <Hz, ms>, which we claim constitute an appropriate phonetic representation of the F0 curve. These target points can be used to generate a quadratic spline function giving a very close approximation to the smooth continuous F0 curve observed on fully sonorant segments of speech (Hirst 1980). The residual micromelodic profile can be stored separately and then added to the quadratic spline to obtain very high quality speech synthesis (Di Cristo & Hirst 1986).

Surface phonological modelling.

The target points obtained from the programme MOMEL are coded symbolically using the INTSINT transcription system (Hirst & Di Cristo in press). According to this system (used in half of the chapters in Hirst & Di Cristo (eds) in press), tonal segments are assumed to be of two types : absolute tones[T(op) M(id) B(ottom)], whose phonetic interpretation is assumed to be independent of the immediately preceding tone; and relative tones : [H(igher) L(ower) U(pstepped) D(ownstepped) S(ame)] whose phonetic interpretation is assume to be dependent on that of the immediately preceding tone.

A number of different options for automatically coding target points are being investigated. Among these are the use of a threshold for the distinction between absolute and relative tones; the use of syntagmatic constraints to ensure that only D and U are used for cases of iterative lowering or raising while T and H or B and L always correspond to peaks or valleys respectively. A further possibility is that of establishing an optimal coding of the target points, minimising the difference of variance between the model and the observed data.

The investigation of the association between the tonal segments and the segmental transcription raises a number of interesting issues and is at present in a very preliminary stage. Various models making use of different temporal parameters (word boundaries, syllable boundaries, vowel onset etc) are being explored.

Phonetic interpretation of surface phonology.

The phonetic interpretation of the symbolic coding of the F0 target points can be obtained by statistical analysis of the original dataset. Absolute tones are thus modelled as the mean value of the corresponding target points; relative tones are modelled by linear regression on the preceding target point, irrespective of its code. (Hirst, Nicolas & Espesser 1991). It remains to be seen what is the optimal domain for such interpretation.

Universal and language specific characteristics of the symbolic coding and the phonetic interpretation of such modelling will be investigated for the four languages.

Preliminary results

- Durational characteristics of Spanish

In order to test for evidence whether Spanish accent groups are organised into left-headed or right headed structures, measurements of syllable duration were made on three corpora : a list of declarative sentences, a portion of a continuous text and a portion of a continuous spontaneous monologue. Among the factors tested were position in the word, stressed/unstressed nature of syllable, number of syllables in the left-headed/right-headed foot, number of syllables in the left-headed/right-headed restricted foot (where a "restricted foot" is assumed not to cross word boundaries, cf the "narrow rhythm unit" of Jassem 1952, Jassem et al. 1984). Figure 1 shows the summed unexplained variance for four models : left-headed/right headed (LH/RH) unrestricted/restricted (/-R) accent groups. In all cases the variance of syllables not included in the relevant foot structures were added to the unexplained variance of the modelled structure so that the summed unexplained variances are comparable across models . The right-headed foot structure gave the best fit for each corpus, followed by the restricted right-headed foot-structure (Alcoba et al. 1992). The model based on the number of syllables in the word (not illustrated here) gave a worse fit for all three corpora.

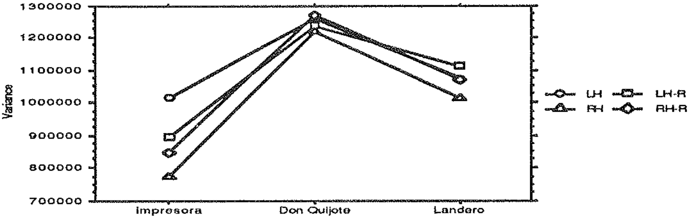


Figure 1 : Unexplained variance for four models of accent groups : Left/Right-headed [LH/RH], unrestricted/restricted [/-R] for three corpora in Spanish.

- Small corpus pilot study

A small corpus of 20 sentences was recorded for each of the four languages. The F0 curves of the sentences were modelled using MOMEL with manual correction when necessary. The target-points were then coded as INTSINT symbols. For temporal alignment of the tonal segments the onset of the stressed vowels was labelled as well as the word boundaries. It remains to be seen how far such a sparse labelling of the corpus, together with information concerning the number of syllables, is sufficient to account for the temporal localisation of the target points in the different languages. In particular the way in which these parameters can be related to the Left/Right-headed structures referred to above is of particular interest.

By way of illustration, the following figure shows the result of the procedure applied to one of the sentences from the Arabic corpus. :

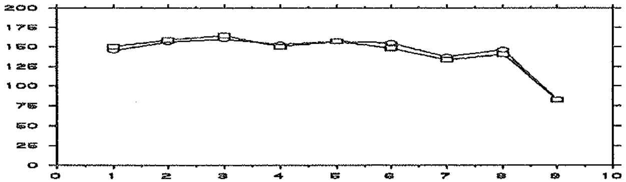


Figure 2 : Observed (circles) and predicted values (squares) for the Arabic sentence 'Kataba lwaladu alkabiiru risaalatan' (The big boy wrote a letter). The corresponding (automatically derived) INTSINT coding was [M U T L H D L H B].

The automatically derived quadratic spline targets (circles) were coded as the INTSINT sequence [M U T L H D L H B]. The statistical values of each tonal segment were then calculated on the complete corpus as described above and the predicted values were then derived (squares).

CONCLUSIONS

We have outlined a general methodology which is at present being developped for the comparative analysis of the intonation systems of different languages.

ACKNOWLEGEMENTS
This research was partly financed by the Franco-Spanish *Programme D'Action Intégrées* as part of a three-year project between the Universities of Provence and Barcelona entitled *Prosodie comparée du français et de l'espagnol contemporains.*

REFERENCES

Alcoba, S.; Di Cristo, A.; Hirst, D.J.; Le Besnerais, M.; Murillo, J. & Roméas P. (1993) *Rapport Final sur l'Action Intégrée Prosodie comparée du français de de l'espagnol contemporains.* (Unpublished report, Université de Provence).

Di Cristo (in press) "Intonation in French." in Hirst & Di Cristo (eds) in press.

Di Cristo, A. & Hirst, D.J. (1986) "Modelling French micromelody : analysis and synthesis." *Phonetica* 43, 11-30

Hirst, D.J. & Di Cristo A. (eds.) (in press) *Intonation Systems a Survey of Twenty Languages.* (Cambridge University Press; Cambridge)

Hirst, D.J. & Di Cristo A. (in press) "A survey of intonation systems" in Hirst & Di Cristo (eds) (in press).

Hirst D.J. & Espesser R. (in press) "Automatic modelling of fundamental frequency." *Travaux de l'Institut de Phonétique d'Aix*, 15.

Hirst, D.J. (1980) "Un modèle de production de l'intonation." *Travaux de l'Institut de Phonétique d'Aix* 7, 297-311.

Hirst, D.J. (1992) "Prediction of prosody : an overview." in G.Bailly & C.Benoît (eds) *Talking Machines : Theories, Models and Applications.* (Elsevier Science Publishers)

Hirst, D.J. (in press) "Intonation in British English" in Hirst & Di Cristo (eds) in press.

Hirst, D.J.; Nicolas, P. & Espesser, R. (1991) "Coding the F0 of a continuous text in French : an Experimental Approach." *Proc. ICPhS 12* (Aix), 5, 234-237.

Jassem, W. (1952) *Intonation of colloquial English.* (Panstowe Wydawnictwo Naukowe; Warswawa).

Jassem, W.; Hill, D. & Witten, I.H. (1984) "Isochrony in English speech : its statistical validity and linguistic significance." in Gibbon & Richter (eds) Intonation : Accent and Rhythm. (de Gruyter, Berlin)

Le Besnerais, M. (in progress) *Parámetros rítmicos para el estudio contrastado del francés y del español contemporáneos.* (Doctoral thesis, University of Barcelona).

Najim, Z. & Hirst, D.J. (in press) "Codage prosodique d'un corpus d'arabe littéral lu par un locuteur marocain." *Travaux de l'Institut de Phonétique d'Aix*, 15.

Najim, Z. (in progress) L'intonation de l'arabe littéral parlé au Maroc : analyse historique et expériementale. (Doctoral thesis, Université de Provence).

Nicolas, P. (1992) "Emergence de la structure intonative du texte lu en français." Actes du Séminaire Prosodie (La Baume lès Aix, October 1992), 103-112.

Nicolas, P. (in progress), *Apport de la prosodie à la parole de synthèse : cas du texte lu en français.* (Doctoral thesis, Université de Provence)

Romeas, P. (1991) *L'organisation prosodique des énoncés en situation de dialogue homme-machine simulé : théorie et données.* (Doctoral thesis, Université de Provence).

Wenk, B.J.. & Wioland, F. (1982) 'Is French really syllable-timed?' *Journal of Phonetics* 10 (2), 193-216.

Improving the Prosody in TTS Systems: Morphological and Lexical-Semantic Methods for Tracking 'New' vs. 'Given' Information

Merle Horne*, Marcus Filipsson*, Christer Johansson*, Mats Ljungqvist‡ and Anders Lindström‡
*Dept. of Linguistics and Phonetics, University of Lund, Helgonabacken 12, S-223 62 Lund, Sweden
‡INFOVOX, Box 2069, S-171 02 Solna, Sweden

ABSTRACT

The design of an algorithm for referent tracking in a restricted domain is presented. The algorithm allows one to preprocess a text and automatically tag words as either contextually 'New' or 'Given'. The procedure involves computational modelling of lexical semantic identity of sense relations as well as information on inflexional/derivational morphology and compounding. Referent identity is defined on head-word representations derived from the text input on the basis of the inflexional expansion rules contained in a lemmatized lexicon of Swedish. Information on the New/Given status of words can subsequently be used in the F_0-generating component of the text-to-speech system to trigger the assignment of focal vs non-focal word accents.

INTRODUCTION

A major goal in current research in text-to-speech has involved improving the prosody component by developing interfaces which extract contextual and syntactic information that condition pitch accent-type as well as level of accentual prominence (Hirschberg 1990, Youd and House 1991, Horne and Johansson 1991, in press, Horne, Filipsson, Ljungqvist and Lindström 1993, Monaghan 1990). In most commercial text-to-speech systems, the prosodic component does not have access to higher level syntactic and semantic information and is therefore able to generate only a very restricted number of intonation patterns. In the case of Swedish (Carlson and Granström 1973, Bruce and Granström 1989), the commonly used method is to assign a focal accent to the last content word in an intonational phrase. This focal accent pattern leads to the interpretation of all phrase-final words as 'new' information. It has not been possible to *automatically* assign phrase-final content words non-focal accents which are associated with 'given' information. In order to enable automatic assignment of the proper accent-type, it is necessary to process the input text with respect to the information status associated with the 'content' (lexical) words. We have currently been involved in developing such a linguistic preprocessor which models and tracks morphological and lexical-semantic coreferential relationships between content words (Horne, Filipsson, Ljungqvist and Lindström 1993). In what follows, we will describe how the processor works.

METHODOLOGY

In the modelling of the identity relationships and the development of the coreferent tracking algorithm, we have been currently exploiting the information contained in a computerized lexicon of Swedish (Hedelin, Jonsson and Lindblad 1987). The lexicon, which is lemmatized, contains approximately 116,000 headwords, each one listed with its part-of-speech specification, inflection code, and phonetic transcription. It also includes information on the morphological status of derivations and can handle the analysis of compound-words, either by explicit listing or by algorithmic generation.

A general feature of a lemmatized lexicon is the inherent relationships between the head-words and their inflected forms. As an example, the word *tända* 'to light' can be related to its paradigmatic forms: *tända* (inf.), *tänder* (pres.), *tände* (pret.), *tänt* (past part.), *tänd* (supine), *tändas* (inf. passive), etc. This is important in the present application since the identity relations are defined over the stems or head-words. The lexical structure has furthermore been amended with domain-specific knowledge describing semantic hierarchies (hyponymy, part/whole), synonymy relationships as well as pragmatically/situationally Given terms.

In the initial stages of the development of the algorithm, we did not have recourse to a lexicon (Horne and Johansson 1991, in press) and thus the tracking process was more limited as regards the number of different kinds of coreference relations that could be identified. Stem identification was achieved by means of a morphological truncation procedure (due to B. Brodda). This process applied each time a word was compared with each preceding word when determining its coreferential status. Morphological truncation is based on graphic information and searches for identical strings of letters in two words. If these are found and if the remaining strings in the two words are existing endings in the language, then the words are classified as coreferential. This method suffers, however, from the fact that it can only handle non-suppletive paradigms. Stems of morphs in suppletive paradigms such as *falla /föll* 'to fall/fell' cannot be found using the truncation procedure. Analysis of compounds is another problem area which was not solved using this non-lexical method.

DESIGN OF THE ALGORITHM

In the analysis of an input text, the following steps are currently involved. First, the lexicon is used to analyse the words and decompose them into morphs (see Figure 1). As mentioned earlier, the lexicon handles inflexions, derivations, and compound words. The treatment of compounds is an important feature, since the referent-tracking processes must apply not only to the compound as a whole, but to the component morphs. This process is complicated in Swedish by the fact that compounds, as in German, are written as undivided words, without hyphens or spaces between the component morphemes, e.g. *fondbörs* 'stock-exchange' consists of the morphemes *fond* and *börs*. Since the lexicon contains information on the internal structure of compounds, the tracking procedure can apply to the individual component morphemes. In Figure 2 can be seen an example of the decomposition of words into morphs.

After analysing each word in the text into its basic morph(s), the referent tracking procedure can apply to the text. Each word is then checked for its possible coreference with any previously mentioned word within an adjustable window. The window used in the examples in this paper has somewhat arbitrarily been chosen to be 60 words, but other domains could also be considered, e.g. the paragraph (Hirschberg 1990). The referent-tracking algorithm consists of four parts, as shown in Figure 1 (a-d):

The first one tracks and marks words that are situationally/pragmatically 'given', such as *börs* 'stock-market', and *krona* 'crown' in the domain we have studied (Swedish stock-market). The second stage involves identifying cases of coreference due to reiteration of root morphs (or stems), as obtained from the initial analysis using the lexicon. The third part uses 'domain-specific' synonymy relations to identify cases of coreference. Examples from the stock-market domain are: *kurs–nivå* 'rate' and *aktie–papper* 'share'. The fourth and final stage attempts to track words that are involved in hierarchical identity relations (hyponymy, part/whole relationships). In order to do this, superordinate relations have been modelled using 'is an example of' or 'is a part of' pointers to establish the relation between pairs of lemmas thus building up a forest of hierarchical, multi-branch trees. Cases of multiple inheritance, i.e. where a lemma has more than one parent (e.g. both *vardag* 'week-day' and *arbetsdag* 'work-day' are parents of *måndag* 'Monday') do not pose a problem, since the algorithm searches only among 'daughters' of the currently analyzed word, i.e. the tracking in semantic hierarchies is unidirectional, from the more general anaphor to the more specific antecedent. Output from the text processing scheme as described above is shown in Figure 3. Each word is marked as either 'new' (N) or 'given' (G).

DISCUSSION

The algorithm described in the present paper analyses the input text and annotates it with respect to 'new' and 'given' information, thereby allowing the TTS system to generate synthetic prosody of considerably improved naturalness as compared to the default behaviour, as verified by informal listening tests.The performance of the processing can, however, be improved by further analysis, such as identification of syntactic phrase boundaries (Bruce, Granström, Gustafson and House 1992) as well as relative levels of pitch prominence. The lexicon can be exploited at this stage as well in order to extract information on e.g. word-class designation. Contrastive prominence is another phenomenon which requires lexico-syntactic information. Thus, an even more attractive approach is to integrate the linguistic/contextual processing into the text-to-speech system (Lindström, Ljungqvist, and Gustafson 1993), thereby allowing it to exchange information with other knowledge sources of the TTS system, such as the syntactic parser, punctuation ambiguity resolution, treatment of numbers and abbreviations, etc.

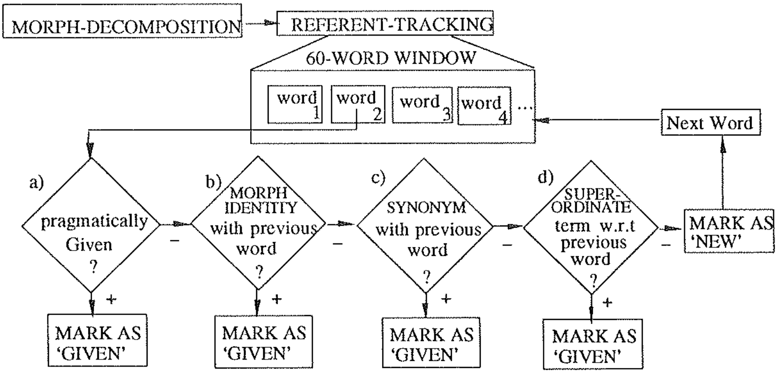


Figure 1. *Flow-diagram illustrating the different stages of the analysis in the lexical processor: MORPH-DECOMPOSITION first analyses each word into its content morphs. In the second stage, the REFERENT-TRACKING algorithm searches through a 60-word window for coreferents.*

:Stockholms	'Stockholm's'
Stockholm	'Stockholm'
:fondbörs	'stock-exchange'
fond	'stock'
börs	'exchange'
:generalindex	'general index'
general	'general'
index	'index'
:slutade	'closed'
sluta	'close (inf.)'
:på	'on'
på	'on'
:torsdagen	'Thursday+def.art'
torsdag	'Thursday'
:på	'at'
på	'at'
:858.8	'858.8'
858.8*	'858.8'

Figure 2. *Example of decomposition of words into morphs using the lemmatized lexicon. Analysed words are preceded by a colon (:). Under each of these words is/are the component root-morph(s). Notice that for verb-forms, the infinitive form is taken to be the root. An asterisk following a form indicates that there is no corresponding lexical entry.*

```
 1N    Stockholms   12N     marginella  23N     som            34Gw  och
 2Gg   fondbörs     13N     0.02        24N     oregelbunden   35G29 läkemedel
 3N    generalindex 14N     procent     25Gs 10 Kursstegringar 36Gw  att
 4N    slutade      15N     jämfört     26N     i              37N   bli
 5N    på           16Gi 11 med         27N     AGA            38N   bästa
 6N    torsdagen    17N     onsdagens   28Gw    och            39Gi 31bransch
 7Gi 5 på           18Gi  3 slutindex   29N     Astra
 8N    858.8        19N     Kursutveckl.30N     fick
 9N    en           20N     över        31Gi  3 branschindex
10N    uppgång      21Gh  6 dagen       32N     för
11N    med          22N     betecknades 33Gh 27 kemi
```

Figure 3. *An example of the output from the referent-tracking algorithm. Words are marked either as N for 'new' or G for 'given'. If a word is marked as G, the output shows in addition the kind of givenness that is present: Gg stands for pragmatically given, Gi, for 'given' due to morphological identity. The number directly preceding the word refers to the number of the word to which the word is construed as coreferent. Gh stands for 'given' due to a hierarchical relationship, e.g. word 21, dagen is marked as Gh since it refers back to word 6, torsdagen. Gs refers to coreference due to synonymy, e.g. the component stegring 'rise' in the compound word 25 kursstegringar 'rate-increases' is a synonym to word 10, uppgång, thus triggering a non-focal accent on the compound. Gw stands for grammatical words that are not eligible to receive focal accents.*

ACKNOWLEDGEMENTS

This research has been supported by a grant from the Swedish HSFR/NUTEK Language Technology Programme.

REFERENCES

G. Bruce and B. Granström (1989), "Modelling Swedish intonation in a text-to-speech system", *STL-QPSR*, Vol. 1, pp. 31-36.

G. Bruce, B. Granström, K. Gustafson and D. House (1992), "Aspects of prosodic phrasing in Swedish", *Proc. ICSLP, Banff, Canada,* Vol. 1, pp. 109-112.

R. Carlson and B. Granström (1973), "Word accent, emphatic stress, and syntax in a text-to-speech system, *STL-QPSR*, Vol. 2-3, pp. 31-36.

P. Hedelin, A. Jonsson and P. Lindblad (1987), *Svenskt uttalslexikon: 3 ed. Tech. Report, Chalmer's Univ. of Technology.*

J. Hirschberg (1990), "Using discourse context to guide pitch accent decisions in synthetic speech", *Proc. ESCA Workshop on Speech Synthesis, Autrans, France,* pp. 181-184, .

M. Horne and C. Johansson (1991), "Lexical structure and accenting in English and Swedish restricted texts", *Working Papers (Dept. Ling., U. of Lund),* Vol. 38, pp. 97-114.

M. Horne and C. Johansson (in press), "Computational tracking of 'New' vs 'Given' information: implications for synthesis of intonation", *Proc. Nordic Prosody VI, K.T.H., Stockholm, August 12-14, 1992.*

M. Horne, M. Filipsson, M. Ljungqvist and A. Lindström (1993), "Referent tracking in restricted texts using a lemmatized lexicon: implications for generation of intonation", *Proc. Eurospeech '93, Berlin, 21-23 September, 1993.*

A. Lindström, M. Ljungqvist and K. Gustafson (1993), "A modular architecture supporting multiple hypothesis for conversion of text to phonetic and linguistic entities". *Proc. Eurospeech '93, Berlin, 21-23 September, 1993.*

A. Monaghan (1990), "Treating anaphora in the CSTR text-to-speech system". *Proc. ESCA Workshop on Speech Synthesis, Autrans, France,* pp. 113-116.

N. Youd and J. House (1991), "Generating intonation in a voice dialogue system", *Proc. Eurospeech 91, Genua, Italy,* pp. 1287-1290.

The Prediction of Prosodic Timing: Rules for Final Syllable Lengthening in French

Eric Keller, Brigitte Zellner, Stefan Werner, and Nicole Blanchoud
LAIP - Lettres, Université de Lausanne
CH-1015 Lausanne, Switzerland
Fax: +41 21 692 4639, Email: Eric.Keller@imm.unil.ch

ABSTRACT

Timing is an essential part of prosody, since it contributes to the semantic and syntactic modulations of speech conveyed by accent and intonation. Theoretical and empirical considerations suggest that temporal organization takes two main forms. First, it is a necessary corollary of accent placement and intonational modification. Second, it manifests itself in prolongations, pauses and hesitations related to sentence structure. A set of rules for the second set of temporal modulations is presented. In contrast to previous work, such rules depend only minimally on syntactic structures and can be formulated nearly entirely in simple phonological terms.

Rules for the prediction of the temporal structure of speech are important for the development of text-to-speech systems. Furthermore, such rules are of interest to our understanding of human linguistic functioning, particularly if they capture general principles of psycholinguistic operation. Previous work in this area has suggested that syntactic structures derived from psycholinguistic evidence (so-called "performance structures") could successfully predict the durations of *pauses* in speech (Gee & Grosjean, 1983). Specifically, pauses at major syntactic boundaries of this type tend to be longer than those at minor boundaries. Furthermore, *final syllable durations* adjoining such boundaries show similar correlations with syntactic boundary types. Psycholinguistic processing may thus employ "final syllable+adjoining pause duration" as a basic prosodic vehicle for marking certain hierarchical structures in speech.

The Monnin-Grosjean Rules

Several sets of rules of this type have been proposed for English, and an adaptation of these rules has recently been prepared for French (Monnin & Grosjean, in press). Fundamentally, the Monnin & Grosjean rules proceed as follows:

(1) *Nuclei* of prosodic constituents are identified from left to right: nouns, verbs and post-posed adjectives.

(2) *Prosodic constituents* are created by grouping words around the nucleus. Words are attached one by one. Special conditions determine whether words are attached to the right- or the left-lying nucleus.

(3) Word boundaries in the prosodic constituent are *indexed* to provide a measure of the *strength* or *importance* of boundaries between them. Basically, this is a count of the attachments (nodes) separating two words. However, various adjustments may intervene to handle special cases.

(4) *Higher prosodic constituents* are created to form a constituent hierarchy. This hierarchy is different from traditional syntactic hierarchies, since various "weight" parameters are taken into consideration, such as number of branching nodes.

(5) Higher prosodic constituents are indexed. The index count between two constituents is based on the number of nodes required to connect the constituents.

(6) *Further adjustments* may be required, depending on constituent and word length.

(7) Finally, by multiplication with a simple constant, index counts can be translated into *durations* of final syllable+pause segments.

A Verification of the Monnin-Grosjean Rules

As a first step, these rules were verified with respect to both the Monnin-Grosjean corpus and a new corpus. The verification of the original corpus permitted us to check our understanding of the rules and to examine results in detail. The new corpus[1] represents three readings at different speech rates of three sentences by 12 speakers, 6 male, 6 female, 6 Parisian French, 6 méridional French (Toulouse) (7200 phonemes). The first reading was a practiced reading at normal speech rate, the second was a slow, deliberate reading, and the third was extra-deliberate (not used). The first sentence is syntactically and semantically complex, while the other two are quite simple. The corpus was manually labeled at the subphonemic level. Fo, energy and durations were measured at 10 ms-intervals.

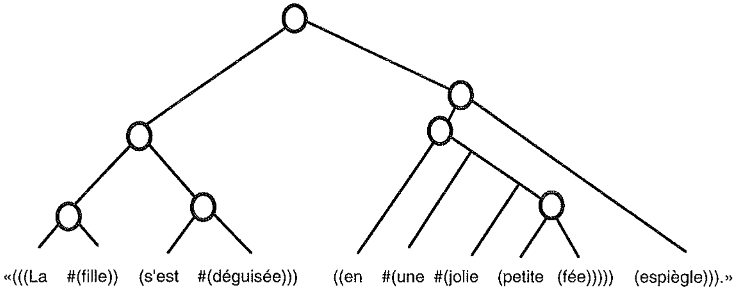

«(((La #(fille)) (s'est #(déguisée))) ((en #(une #(jolie (petite (fée))))) (espiègle))).»

Figure 1. A performance structure tree according to the Monnin & Grosjean rules. A single node separates «petite» and «fée», which predicts a short final syllable and adjoining pause for «petite». By contrast, five nodes separate «déguisée» and «en», which predicts a much larger final syllable and pause duration at this major juncture. The #-mark specifies an attachment of grammatical to lexical words.

Results of the application to the Monnin-Grosjean rules were uneven (Table 1). While sentences 1 and 3 of the Caelen-Haumont data showed acceptable predictions in the .7 - .9 correlation range, sentence 2 showed particularly low predictions. With respect to Monnin and Grosjean's own data set, predictions for slow productions were found to be less successful than those for productions at a normal speech rate. This led to the following considerations:

(1) Monnin & Grosjean analyzed simple sentences that pose few problems of *hierarchical structure*. They suggest using general syntactic principles for *more complex structures*. (Our prediction for Caelen-Haumont's complex sentence 1 is in fact based on syntactic theory). However, reference to syntactic theory destroys the computational simplicity of the algorithm. Are complex structures really required?

(2) The application of the index count is *quite complex*. In particular, a number of minor adjustments are required to handle constituents of various lengths. Are all of these necessary? Are all of these adjustments "psycholinguistically real"?

The Keller-Zellner Rules

In view of the considerable existing literature and much experimentation performed on the two data sets, the following principles were formulated, and an algorithm was defined. The resulting rules satisfy our criteria of *simplicity*, respect of *psycholinguistic principles*, and *high predictive capacity* for the data sets at hand.

(1) *Prosodic constituents are formed on the basis of simple proximal syntax.* No syntactic structures more complex than those applying to a single phrase are required. This is considered to be "psycholinguistically simple" in the sense that children around 4-5 years show satisfactory prosodic grouping, an age at which they show an insufficient command of complex syntactic structures. Prosodic groups can be identified by the application of steps 1 and 2 of the Monnin-Grosjean rules.

[1] Kindly made available to us by Geneviève Caelen-Haumont, ICP/INPG, Université Stendhal, Grenoble, France.

(2) *Final syllable+pause durations increase in duration as the constituent proceeds.* The increase proceeds from an empirical minimum to an empirical maximum. The initial hypothesis calls for equal steps. Increased durations correspond to a slowing down, which is a commonly observed phenomenon in speech.

(3) *Rhythmic alternance* was observed for two locations: post-verbally and in the middle of 4-6 word constituents. Rhythmic alternance occurs when one element is lengthened more than strictly required. As a consequence, the following element must be shortened "in order to conclude the constituent in time". Concretely, this amounts to postulating an *inversion of durations* for the word pair involved in the alternance.

The resulting algorithm is quite simple and is fully reproduced at the end of this paper. Correlations with the Caelen-Haumont and the Monnin-Grosjean data sets are reported in Table 1. It is found that correlations are quite regular. They never dip below a linear correlation of .7, and generally tend to be found in the .8 range.

An inspection of the evolution of Fo and energy values at the end of prosodic constituents postulated here shows some regularities. Fo values rise at the end of each constituent, except for the sentence-final constituent. Energy values fall regularly at the end of each constituent. This suggests that the temporal structure characterized here interacts directly with control over Fo and energy.

Table 1: Linear Correlations Between Predicted and Measured Final Syllable+Pause Durations According to Two Sets of Rules

	Monnin-Grosjean		**Keller-Zellner**	
Caelen-Haumont Data Set	**Normal**	**Slow**	**Normal**	**Slow**
Sentence 1	.786	.895	.862	.845
Sentence 2	.289	.375	.811	.829
Sentence 3	.925	.808	.878	.751
Mean	**.667**	**.693**	**.850**	**.808**
Monnin-Grosjean Data Set	**Normal**	**Slow**	**Normal**	**Slow**
Sentence 1	.890	.674	.873	.835
Sentence 2	.914	.796	.886	.954
Sentence 3	.981	.886	.773	.892
Sentence 4	.961	.826	.798	.850
Sentence 5	.947	.736	.827	.872
Sentence 6	.984	.711	.812	.835
Sentence 7	.931	.841	.754	.906
Sentence 8	.940	.585	.870	.809
Sentence 9	.968	.808	.701	.818
Mean	**.946**	**.763**	**.810**	**.863**

Conclusion

The performance of the new and simplified Keller-Zellner algorithm is encouraging. Proximal syntax can be used to create prosodic constituents, and final syllable+pause durations can be calculated using a simple set of rules. Text-to-speech systems could quite easily use these rules in conjunction with statistically determined values for non-final syllables (such as those proposed in O'Shaughnessy, 1984). It remains that the data considered here is limited. Only few sentences and speakers have been examined, and only read speech has been considered. Future research will automatize these rules and will examine predictions for larger and more varied data sets.

Acknowledgements

Grateful acknowledgements go to G. Caelen-Haumont and F. Grosjean for making their corpora available and for offering their comments. Supported by the Office fédéral de l'éducation et de la science, Berne, for the ESPRIT project "Speech Maps" and the COST project 233 "Prosodics of synthetic speech".

The Keller-Zellner Algorithm

(1) Identification, from left to right, of the *nuclei* of the prosodic constituents: nouns, verbs and free-standing adjectives, adverbs and pronouns (such as "La chemise est *sale*", "c'est *bien*", "pense à *ça*").

(2) Creation of the *prosodic constituents* by grouping the words around the nucleus. All words to the left of the nucleus are attached to the right-lying nucleus, except for post-posed adjectives and post-posed pronouns which are attached to the left-lying nucleus ("la chemise *blanche*", "donne-*lui*").

(3) Calculation of *predictions for final syllable+pause durations.* Within each prosodic constituent, durations increase from a minimum to a maximum duration. Initially, the increase is assumed to occur in equal steps. (The minimum and maximum are assumed to be 50 and 350 ms in normal speech, 50 and 525 ms in slow speech.) The first final syllable in a constituent has a duration of minimum+step size ms.

(4) *Rhythmic tradeoffs:*

1. *Post-verbal trade-off:* When a constituent follows a verb and there are at least two words prior to the nucleus, the final syllable duration of the first word is lengthened with respect to that of the second word. (Exchange durations for words 1 and 2.)

2a. *Rhythmic alternance:* If a constituent is 4 or more words long, and if word 3 is 2 or more syllables long, word 2 is lengthened with respect to word 3. (Exchange durations for words 2 and 3.)

2b. *If rule 1 has already applied:* If a constituent is 4 or more words long, and if word 4 is 2 or more syllables long, word 3 is lengthened with respect to word 4. (Exchange durations for words 3 and 4.)

3. *Single-word constituents:* Constituents containing a single word show reduced final syllable durations. (Reduce durations for single word constituents by 50 ms.)

(5) *Measure of final syllable+pause.* The measure begins with the vowel of the final syllable and ends at the end of the pause. It includes whatever intervening consonant may occur, but it excludes the characteristic optional schwa of French méridional speakers (as in «biologiste»). Excluding the optional schwa permitted us to make direct comparisons of data sets from northern and méridional speakers. Resulting time measures were very similar. For a limited data set, the intervening consonant was suppressed. However, resulting durations were found to show greater variability than those that included the consonant. Measures for sentence-final words were only known for a few sentences and were thus set to 0 in all cases for statistical purposes.

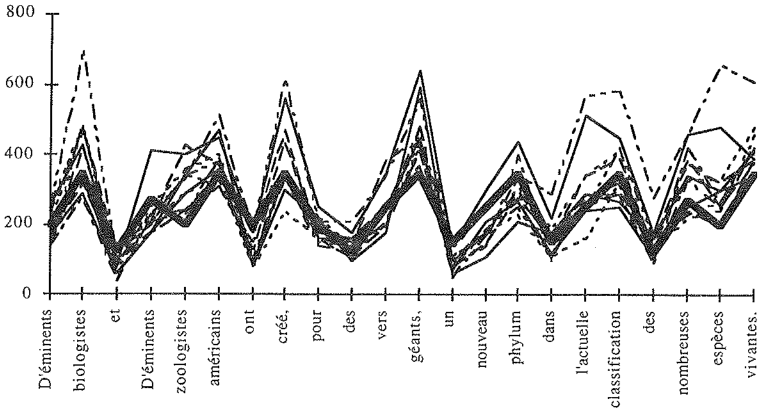


Figure 2. The prediction of the Keller-Zeller algorithm (thick line) for Caelen-Haumont's sentence 1 (thin lines: 12 speakers at normal speech rate).

References

Gee, J.-P., & Grosjean, F. (1981). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology, 15,* 411-458.

Monnin, P., & Grosjean, F. (in press). Les structures de performance en français: caractérisation et prédiction. *L'Année Psychologique.*

O'Shaughnessy, D. (1984). A multispeaker analysis of durations in read French paragraphs. *Journal of the Acoustical Society of America, 76,* 6.

Perceptual evaluation of rule-generated intonation contours for German interrogatives

Bernd Möbius
Institut für Kommunikationsforschung und Phonetik, Universität Bonn
Poppelsdorfer Allee 47, D-53115 Bonn, Germany

ABSTRACT

This paper presents the results of a perceptual experiment that was carried out in order to evaluate naturalness and adequacy of artificial, rule-generated intonation contours for German interrogatives. In previously reported experiments, the ratings for artificial intonation patterns of declaratives were satisfactory while those for interrogatives were not. Therefore, the present study aimed at improving the rules for echo, yes/no, and wh-questions. The integration of several additional features contributed to a more adequate generation of German interrogative intonation contours.

INTRODUCTION

In this paper, the results of a perceptual experiment are presented that was carried out in order to evaluate naturalness and adequacy of artificial, rule-generated intonation contours for German interrogatives. In previously reported perceptual experiments (Möbius and Pätzold 1992), the acceptability of rule-generated intonation patterns as well as the adequate modelling of prosodic properties were critically examined by expert and 'naive' listeners. The results suggested that rule-generated F_0 contours for declaratives were nearly as acceptable as close approximations of the original contours, while the ratings for interrogatives were significantly lower. Furthermore, detailed judgements concerning the realization of word accents and sentence mode were obtained.

The study presented here aimed at improving the rules for three types of interrogative sentences (echo, yes/no, and wh-questions) by extending the speech materials and by taking into account linguistic factors that were omitted in the previous investigations.

GENERATING F_0 CONTOURS BY RULE

General procedure

The rules that generate an artificial intonation pattern for a given utterance are based on the analysis of naturally produced F_0 contours by means of the quantitative model proposed by Fujisaki (1983, 1988). The model aims at a functional representation of the production of F_0 contours by a human speaker and has been successfully adapted to German (Möbius et al., 1991; Möbius, 1993). Using an analysis-by-synthesis procedure, the complex F_0 contour of a given utterance can be decomposed into the components of the model. This is achieved by successively optimizing the parameter values which leads to a close approximation of the original F_0 course. Thus, the model provides a parametric representation of intonation contours.

The potential sources of variation of the parameter values were explored using statistical methods. Standard values were derived on the basis of the statistically significant factors. A set of rules was formulated that control the adjustment of the parameters (see Möbius 1993 for details). The rules capture speaker-dependent as well as linguistic features such as sentence mode, sentence accent, phrase boundary signals, and word accent, and generate an artificial intonation pattern for any given target utterance. The input information needed for generating an F_0 contour by rule is the temporal position of accented syllables. At present, the rules are confined to rather short isolated utterances containing not more than two prosodic phrases.

Modifications

Taking the results of the previous experiments (Möbius and Pätzold 1992) as a starting-point, the major purpose of the study presented here was to improve the rules for generating intonation contours for interrogative sentences. This was achieved by extending the speech materials and by integrating prosodic features that were obviously omitted in the previous investigations. For instance, accentuation and signalling of interrogative sentence mode superposed and compressed on the utterance-final syllable was not sufficiently modelled by earlier versions of the rules. Furthermore, their were no rules for deaccentuation in compounds and in the case of two adjacent accented syllables. These features are now incorporated into the set of rules.

The speech material under investigation may be characterized as typical 'laboratory' speech. It covers three types of interrogative sentences, i.e. echo questions, yes/no questions, and wh-questions. Recordings were made in an anechoic chamber for two female and two male speakers who read the orthographically presented sentences aloud.

PERCEPTUAL EVALUATION

Method

The aim of the perceptual experiment was to evaluate the acceptability of rule-generated contours for German interrogatives in comparison with the acceptability of their original counterparts. More specifically, the subjects were asked to judge the melodic and stress features of the stimuli with respect to 'naturalness' and linguistic 'adequacy'. As an illustration of the two criteria, the listeners were given the following examples: a) Speech melody of an utterance may sound natural although sentence mode may not be clearly or adequately signalled; b) accented syllables may stand out clearly in the course of the utterance (linguistically 'adequate') but possibly by means that sound unnatural. For each of the two criteria, the subjects expressed their ratings on a seven-point scale ranging from -3 to +3. In individual sessions, the stimuli were presented by headphone to 19 prosodically 'naive' listeners who were payed for participating in the experiment.

Since there are several speaker dependent features that are covered by the rules, one female and one male 'voice' were used in the experiment. For each speaker, two versions of each test sentence were presented to the listeners, i.e. one version with a rule-generated intonation contour and one version with the original F_0 information as extracted by a pitch determination algorithm. In order to avoid, as far as possible, any differences in the sound quality of the stimuli that may affect listeners' judgements of prosodic properties, both kinds of stimuli were manipulated by means of the PSOLA algorithm (Moulines and Charpentier 1990).

Results

Not very surprisingly, there is a marked relationship between the ratings for the criteria 'naturalness' and 'adequacy' as expressed by the high value of the Pearson product-moment correlation coefficient (r = 0.81; p < 0.001). The value of the coefficient of determination (r^2 = 0.66) indicates that both variables have a considerable proportion (66%) of their respective variances in common.

The subjective ratings for 'naturalness' and 'adequacy' of the original and the rule-generated intonation contours, both for the female and the male voice, are shown in figures 1 and 2, respectively. The results are given for all utterances irrespective of sentence mode.

The ratings for the original contours are significantly higher than those for the rule-generated ones ($F_{1,\,82}$ = 67.6, p < 0.001 for 'naturalness'; $F_{1,\,82}$ = 24.7, p < 0.001 for 'adequacy'). There is an overall difference of about one point on the scale between the two types of stimuli for both 'naturalness' and 'adequacy'. In general, the female voice is rated higher than the male one, but this difference is statistically insignificant, and the relation between the two types of stimuli is consistent irrespective of speaker.

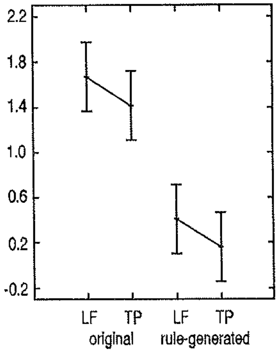


Figure 1. *Mean ratings and 95% confidence intervals for 'naturalness' of original and rule-generated intonation contours for the female (LF) and the male voice (TP).*

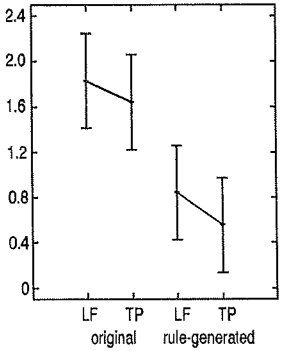


Figure 2. *Mean ratings and 95% confidence intervals for 'adequacy' of original and rule-generated intonation contours for the female (LF) and the male voice (TP).*

Discussion

The version of the intonation contour, i.e. original *vs*. rule-generated, seems to be the only factor exerting a significant influence upon the variance of the 'naturalness' and 'adequacy' judgements. Other potential factors, such as speaker or sentence mode, are statistically insignificant. The mean ratings for 'naturalness' as well as for 'adequacy' of the rule-generated contours are about one point of the seven-point scale lower than those for the original contours. This difference turns out to be statistically significant and consistent for both speakers and the three different interrogative modes.

Higher ratings for the original contours compared to the rule-generated ones meet the expectations if the procedure of developping the rules is considered. While it is true that the rules generate F_0 contours that are, in a qualitative way, representative for certain speakers or types of speakers, the standardized parameter values gained by statistical analysis can only be interpreted as averages. They generate 'prototypical' F_0 contours that will never be produced by any speaker in exactly the same way. Nevertheless, the rule-generated intonation contours seem to be not much less acceptable than the respective original ones.

CONCLUSION

The purpose of the study presented here was to improve the rules for generating intonation contours for interrogative sentences. This was mainly achieved by extending the speech materials and by providing rules for deaccentuation in compounds and in the case of two adjacent accented syllables. Furthermore, the combined effect of accentuation and signalling of sentence mode superposed in utterance-final position is now sufficiently modelled. The results of the perceptual experiment suggest that the integration of these features contributed to a more adequate generation of German interrogative intonation contours.

ACKNOWLEDGEMENTS

The author wishes to thank Ansgar Rinscheid for his pitch mark algorithm, Thomas Portele for implementation of the PSOLA algorithm and for assistance with the perceptual experiment, and Matthias Pätzold for his approximation algorithm. This study was partly supported by a grant from the German Federal Ministry of Research and Technology (BMFT).

REFERENCES

H. Fujisaki (1983), "Dynamic characteristics of voice fundamental frequency in speech and singing", in *The production of speech*, ed. by P.F. MacNeilage (Springer, New York), pp. 39-55.

H. Fujisaki (1988), "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour", in *Vocal physiology: voice production, mechanisms and functions*, ed. by O. Fujimura (Raven, New York), pp. 347-355.

B. Möbius, G. Demenko and M. Pätzold (1991), "Parametric description of German fundamental frequency contours", *Proc. 12th Internat. Cong. Phon. Sc., Aix-en-Provence, 19-24 August 1991*, Vol. 5, pp. 222-225.

B. Möbius and M. Pätzold (1992), "F_0 synthesis based on a quantitative model of German intonation", *Proc. Internat. Conf. Spoken Language Processing, Banff, Alberta, Canada, 12-16 October 1992*, Vol. 1, pp. 361-364.

B. Möbius (1993), *Ein quantitatives Modell der deutschen Intonation - Analyse und Synthese von Grundfrequenzverläufen* (Niemeyer, Tübingen).

E. Moulines and F. Charpentier (1990), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, Vol. 9, pp. 453-467.

Perceptual Evidence for Separate Processing of Stress Pattern and Phonemic String

Valérie Pasdeloup, José Morais and Régine Kolinsky
Laboratoire de Psychologie Expérimentale, Université Libre de Bruxelles,
117 av. Adolphe Buyl, 1050 Brussels, Belgium

ABSTRACT

In order to determine if lexical stress pattern and phonemic string are represented separately at some level of processing, a target word recognition task was used under dichotic listening presentation. Two stimuli were presented simultaneously, one to one ear and one to the other ear. The stimuli are constructed in such a way that, if listeners combine the phonemic string from the stimulus presented to one ear with the stress pattern from the stimulus presented to the other ear, they experience the illusory perception of a word which corresponds to the prespecified target. The results suggest that stress pattern and segmental string can be represented separately at least at a prelexical level.

INTRODUCTION

In nonlinear phonological theories (metrical and autosegmental), suprasegmentals are independent of the segmental strings to which they are associated. In the present experiment, we investigated whether in Portuguese, a lexical-stress language, the stress pattern and the phonemic string are represented separately at some processing level. We used a target detection task under dichotic listening presentation (two different stimuli are presented simultaneously, one to one ear and one to the other ear). We hypothezed that stress pattern can be analyzed separately of phonemic string. We predicted that an illusory word could be obtained by combining the phonemic string from the stimulus presented to one ear with the stress pattern from the stimulus presented simultaneously to the other ear. In Portuguese, stress syllables have a full vowel, a longer duration, and a stronger intensity while most unstressed syllables have a reduced vowel, a shorter duration, and a weaker intensity. Still, in formal register or when intrinsically heavy, vowels may not be reduced when unstressed. Since we wanted stress contrasts to be expressed only in the prosodic dimension, we chose a set of items whose vowels are not reduced when unstressed.

EXPERIMENT

The purpose of this perceptual experiment is to establish if subjects can experience in a dichotic listening situation the illusion of hearing a word, although its phonemic information is given to one ear and its lexical stress information to the other ear. Let us take the Portuguese word *MINgua* (upper case will be used to represent a stressed syllable). It shares the phonemic constituents (but not the stress pattern) with the word *minGUa* presented to one ear, and the stress pattern (but not the phonemic string) with the word *VIcio* presented to the other ear. If, when presented with the dichotic pair minGUa-VIcio (called experimental (E) trial), subjects experience sometimes the illusion of hearing the target *MINgua*, it could support the hypothesis that phonemic string and stress pattern are processed separately and recombined erroneously.

Of course, not all false detections (FDs) of the target MINgua observed in this E trial could safely be attributed to such a false recombination. The listener could interpret incorrectly the stress of minGUa, which shared the segmental string with the target MINgua. Consequently, we designed a control (C) trial where none of the items presented dichotically has the stress pattern of the target: the target word *MINgua* shares the phonemic constituents with the word *minGUa* presented to one ear (as in the E trial), but does not share the stress pattern with the word *viCIo* presented to the other ear. In order to establish the occurrence of false recombinations, FDs of the target MINgua in E trials should be higher than Fds of this target in C trials. In addition to the E and C trials described above which are target-absent trials, two target-present trials were designed. E and C trials obeyed the same rule in both target-absent and target-present situations (cf. Table 1).

Table 1. *Set of trials for the target word MINgua*

	Experimental trial	Control trial
Target-absent	minGUa - VIcio	minGUa - viCIo
Target-present	MINgua - VIcio	MINgua - viCIo

Method

Materials. We chose bi- and trisyllabic stress-pairs which differ only in the position of lexical stress, on the first or second syllable. Four groups of stimuli were built, each containing four word targets, presented with their corresponding set of stimuli. The four groups are defined as follows: (1) pairs of trisyllabic word stimuli which do not share any phoneme (as in VIcio-minGUa), (2) pairs of bisyllabic word stimuli which share the second syllable (as in CANsam-PENsam), (3) pairs of trisyllabic pseudoword stimuli which do not share any phoneme (as in siRIo-VINgua), (4) pairs of bisyllabic pseudoword stimuli which share the second vowel (as in FUNçam-VENgam). The pairs of stimuli sharing phoneme are called stimuli S and those which do not share any phoneme NS. Sixteen stress-pairs consisting of words (W), pseudowords (PW), or both W and PW were selected in order to build the trials for these four groups. W and PW were recorded by a native male speaker of Portugal. In order to synchronize, for each pair, the dichotic stimuli onset and offset, we used a speech compression algorithm (from Digidesign's Sound Tool II). Synchronization was made on the basis of the average duration of dichotic stimuli.

Subjects. Fourteen native speakers of Portugal from University of Lisbon took part in the experiment.

Procedure. Subjects were tested individually and heard the dichotic stimuli over headphones. They were instructed to listen carefully to each pair and to respond if the target word written in a booklet has been presented or not. They heard a pair of stimuli every 3 seconds. Half of subjects received the W stimuli trials in a first session and the PW stimuli trials in a second one; the other half received them in the reverse order. Each target-absent trial corresponding to one target word was presented six times and each target-present trial two times. Trials were presented in a pseudo-random way. Each session was preceded by a training phase of 16 trials. Items of each type of pair were counterbalanced for ear assignment.

RESULTS

Results for target-absent trials (see Table 2)

In order to evaluate the false recombination of stress pattern and segmental string, an illusion rate was calculated by dividing the number of Fds on E trials by the total sum of Fds on E and C trials. A rate superior to 0.5 signals the occurrence of false recombination. The mean percentage of Fds is higher in the E than in the C trial. Thus, subjects more often get the illusion of hearing the target word when it shares its stress pattern with one of the stimulus of the pair than when it does not. The illusion rate is higher than 0.5 in all groups, except for the group of W NS stimuli which share no phoneme. Moreover, it seems that sharing phonemic units strongly favours the recombination phenomenon within the groups of W stimuli, whereas no such effect is apparent for the groups of PW stimuli.

Variance analyses (Trial [E vs. C] X Group of stimuli) by subjects and by target word were conducted. The analyses revealed highly significant effects of Trial ($p < .0005$) and Group ($p < .0005$ and $p < .025$, respectively). The interaction is also significant ($p < .0005$ and $p < .01$, respectively). Trial is not significant for the W NS stimuli, but is highly significant for the W S stimuli sharing the second syllable ($p < .0005$ and $p < .0005$, respectively). Trial is significant for both groups of PW stimuli, but only in by-subjects ($p < .01$ and $p < .005$, for NS and S, respectively). The suggestions that can be drawn are the following: first, stress pattern and segmental string can recombine erroneously, and second, this recombination may depend on both lexicality of the stimuli and the fact that they share or not phonemic units.

Table 2. *Percentages of false detections and illusion rates for target-absent trials*

Stimuli	Experimental trial	Control trial	Illusion rate
Word NS	25.1%	23.2%	0.519
Word S	54.2%	25.9%	0.677
Pseudoword NS	30.1%	21.4%	0.584
Pseudoword S	57.4%	43.2%	0.571
Mean	41.7%	28.4%	0.595

Results for target-present trials (see Table 3)

Subjects correctly detect the target word when it is present in the trial. The rate of correct detections is slightly higher in the E trial, especially in the groups of PW stimuli. In these trials, the distractor (i. e., the stimulus paired with the target) shares the stress pattern with the target, whereas it does not in the C trials. Thus, the detection of the target word was presumably facilitated when it was paired with a stimulus that carried the same stress pattern. These results could be interpreted as a "stress sharing advantage"; the so-called "feature sharing advantage" describes a similar effect with segmental features such as place and voicing. The variance analyses revealed that Trial is significant by-target ($p < .05$), and reaches only the 0.065 level of significance by-subjects ($p = .065$). Group is highly significant ($p < .0005$ and $p < .001$, respectively). The interaction is not significant.

Table 3. *Percentages of correct detections for target-present trials*

Stimuli	Experimental trial	Control trial
Word NS	92%	92.9%
Word S	72.3%	68.8%
Pseudoword NS	97.3%	89.3%
Pseudoword S	96.4%	89.3%
Mean	89.5%	85%

Results of d'scores (see Table 4)

One problem for the interpretation of illusion rates calculated only on the basis of Fds on target-absent trials is that Fds could occur more frequently in E than in C trials because the stress pattern corresponding to the one of the target can only be perceived in E trials. It might have some biasing effect at the response level. It was important to assess whether false recombinations of stress pattern and segmental string actually occurred at the *discriminability* level. For this reason, we calculated a further illusion rate on the basis of d' score, from the Signal Detection Theory. d' score is more reliable than a score based only on Fds, since d' gives a response bias-free index of discriminability. It depends on both Fds on target-absent trials and correct detections on target-present trials. An Ed' and a Cd' were calculated. Illusion rate on d' is calculated by dividing Ed' scores by both Ed' and Cd' scores. The hypothesis that stress pattern can recombine erroneously with segmental string predicts a situation of lower discriminability in the E trial than in the C trial (Ed' < Cd'): an illusion rate on d' < 0.5 signals the occurrence of those recombinations.

The illusion rate is lower than 0.5 only for the W S stimuli sharing the final syllable. Variance analyses show that Trial is not significant, but Group is highly significant (p < .0005 and p < .025, respectively). The interaction approaches significance, both by subject and by target (p = .054 and p = .07, respectively). Trial is significant only for W S stimuli (p < .005). Thus, on the basis of d' scores, it seems that stress pattern can be analyzed separately from segmental string.

Table 4. *Illusion rates on d' and standard deviation (in parentheses)*

Stimuli	Illusion rate on d'	
Word NS	0.490	(0.089)
Word S	0.258	(0.189)
Pseudoword NS	0.542	(0.043)
Pseudoword S	0.538	(0.088)
Mean	0.457	

DISCUSSION

The results suggest that, under certain conditions, subjects can experience the illusion of hearing a word, when the phonemic information is given to one ear and the lexical stress information to the other ear. Thus, it seems that before target recognition, in other words at some prelexical level, stress pattern and segmental string can be represented separately, otherwise they could not recombine erroneously. This outcome is not inconsistent with the hypothesis postulated by plurilinear phonological theories that stress pattern and segmental string to which it is associated form separate representations at the mental lexicon.

Illusion rates calculated on the basis of *d' scores* are significant in only one case, namely when the stimuli are words sharing the final syllable. Thus, false recombinations of stress pattern and segmental string can apparently occur at a discriminability level, which means that these phonological properties were, at some time during the recognition process, represented separately. Given the suggestion that separate representation of segmental string and stress pattern only occurs for syllable-sharing stimuli, it seems that separate representation does not mean independent processing. Whether stress pattern separates or not from, and recombines or not with, a segmental string would depend on the relation between the segmental strings themselves. Separation would not occur, anyway, before the dichotic stimuli interact with one another. Illusions might be related to constraints at the recombination stage. Further study of these issues is obviously required.

Acknowledgements: This research was supported by the Programme Cognisciences (CNRS, France), through a post-doctoral grant accorded to the first author, and by the Human Frontier Science Program.

Judgement on quality and diagnostic evaluation of synthetic prosody

Serge Santi & Isabelle Guaïtella
Institut de Phonétique d'Aix-en-Provence
Laboratoire Parole et Langage URA CNRS 261

ABSTRACT

Two original methods of evaluation of synthetic prosody have been developped and applied to three prosodic rule generators using four French dialogue oriented applications (plane tickets reservation, remote inquiry of databases, etc.). Our data show a good correspondance between the results of two tests. They are both capable of discriminating between algorithms according to each type of application. Each method seem to meet our expectations, i.e., evaluation of quality only vs diagnostic evaluation. The question of what is to be evaluated must be answered before determining how to evaluate.

INTRODUCTION

Prosody can be considered an independant module of the speech synthesis process. As a consequence, a separate evaluation of both prosodic and segmental rules is needed. For prosody -and even for the segmental level-, despite the interest of the speech technology community (see, for instance, Grice et al., 1991), no standard evaluation method seems to be available yet (Santi, 1992).

The principle according to which the goals of the experiment must determine the choice of the method, seems to be unequivocal. For instance, if different systems (or algorithms) are to be compared, a quality test based on satisfaction scores (see, for instance, Pavlovic et al., 1990) or a pair comparison procedure have to be conducted. On the other hand, if evaluation is carried out in order to provide information about eventual defaults of the rules, then a diagnostic evaluation is needed.

Two original methods of evaluation of synthetic French prosody (Test 1: Localisation of prosodic defaults by underlying misfunctionning sequences, Test 2: Evaluation of quality based on the satisfaction criterion) have been developped and applied to three prosodic rule generators using four dialogue oriented applications. These methods and their main results will be briefly presented here.

METHOD

Speech material

Four representative dialogues corresponding to four applications were considered: "Route" (Road): delivery of messages about road traffic (no human speaking intervention); "Camif": shopping orders by telephone (no human speaking intervention); "SNCF" (Train): train ticket reservation (with human speaking intervention); "Avion" (Plane): plane ticket reservation (with human speaking intervention).

Three prosodic rule generators (also called prosodic "styles") were tested: Multivoc, Cnetvox-lecture, Cnetvox-dialogue. The segmental continuum (male voice) was synthesised by a TTS rules system develloped by the French Telecom (CNET at Lannion, France). Recordings of the user -for the two last applications- were recorded in Aix-en-Provence by a female speaker in order to avoid possible ambiguity between the human speaker and the machine.

All dialogues were segmented into pragmatic units called "blocks". All blocks have semantic, syntactic and pragmatic (interactional) coherence. As a consequence the size of the block is highly variable (from 8 words to 84 words, from a single sentence to 8 speech turns).

Example:

block Avion 3

Machine:	quel jour desirez-vous partir
Human:	*le cinq avril en fin d'après-midi*
Machine:	le cinq avril vers quelle heure
Human:	*à partir de dix-sept heures*

The method used in test 1required a written version of the dialogues. Because of the close relationship existing between punctuation and prosody (Guaïtella & Santi, 1992), no punctuation marks were used.

Subjects and hardware

A first group of thirty listeners took part in test 1 and fifteen others in test 2. They were all native French students (male and female in equal proportion) and were paid for the test.

Speech material was presented to listeners by means of headphones (AKG, K240). Storage and restitution of audio stimuli was completed with an Intel 386/25 PC micro-computer. The human voice was recorded on a SONY Digital Audio Tape. Tests 1 and 2 were performed individually in a sound-proof chamber.

Test 1: Diagnostic procedure of localisation of misfunctionning sequences

Subjects had to listen carefully to each block and to concentrate on the prosody. In a second hearing they were asked to to underline all eventual sequences they judged to be not or hardly acceptable. Written and oral explanations and a pilot test were proposed to the subjects before the test. All subjects listened to all blocks in a specific random order but a single prosodic version was presented to each subject. The duration of test 1 was about 25 minutes. This method is similar to that used by Hirst et al. (1991)

Test 2: Evaluation of quality based on satisfaction scores

This method is based on methodology proposed by the SAM (Speech Assessment Methodology) working group on prosody evaluation (Grice & Hirst, 1991). We also used the SOAP software to pilot the test (Howard-Jones et al., 1991).

Listeners had also to listen carefully to each block and to concentrate on prosody. After each block they were asked to give a satisfaction score based on prosody only (they were asked not to judge segmentals). A scaling method was prefered to magnitude estimation (see Pavlovic et al., 1990). All other elements were similar to test 1 exept the fact that subject listened to the three prosodic versions according to the Latin Square method.The duration of test 2 was about 25 minutes.

RESULTS

Test 1

Due to the nature of the data, a manual exploitation of the answers was carried out. For each block two scoring methods were used: 1- Number of underlinings, 2- length of underlining distance, in mm (see Hirst et al., 1991). Because of the non-homogenous sizes of the blocks relative values were calculated (percentage of underlined text -distance-, or percentage of underlined words).

Two analyses of variance were carried out:

ANOVA 1- factors: prosodic style and application, response: number of underlinings.

ANOVA 2- factors: prosodic style and application, response: percentage of underlined text.

For ANOVA 1 both factors are highly significant (prosodic style: $p < .0003$; application: $p < .0001$) but the interaction is not ($p <.1222$). For ANOVA 2 both factors and interaction are significant (prosodic style: $p < .0001$; application: $p < .0001$; interaction: $p <.0232$). The incidence table of ANOVA 2 shows that some prosodic styles seem to be better adapted to some applications (for instance: Route and Cnetvox-lecture) (table 1)

Table 1. *Incidence table of ANOVA 2: interaction between "prosody" and "application" factors (percentage of underlying text is taken into account).Up: population, low: percentage.*

	application:	avion	route	sncf	camif	Totals :
prosody	Multivoc	100 21,44	90 11,53	100 19,24	120 9,2	410 15,16
	L-cnetvox	100 13,34	90 7,9	100 12,4	120 9,13	410 10,68
	D-cnetvox	100 13,33	90 12,5	100 12,46	120 7,65	410 11,27
	Totals :	300 16,04	270 10,64	300 14,7	360 8,66	1230 12,37

A significant but weak correlation between the two types of scores (number of underlinings and percentage of underlined text) does not allow us to consider only one of these scores. We also noticed that individual variability was important. Listeners used different individual strategies in underlining. For more details concerning the results of test 1 see Santi & Guaïtella (1992a).

Test 2

Here too two analyses of variance were computed.

ANOVA 3- factors: prosodic style and application, response: scores.

ANOVA 4- factors: prosodic style and application type, response: scores.

The "application type" factor is a combination of the applications. The applications Route and Camif where no human intervention by voice is involved were labelled as *monologue*. The two others application (Train and Plane) are grouped into the *dialogue* category.

For ANOVA 3 both factors and interaction are significant (prosodic style: $p < .0001$; application: $p < .0229$; interaction: $p < .0164$) (Table 2). ANOVA 4 also shows a signicant effect of both "prosody" and "application type" factors but the interaction is not significant ($p < .4969$).

Table 2. *Incidence table of ANOVA 3: interaction between "prosody" and "application" factors (scores).Up: population, low: percentage.*

	application:	route	camif	sncf	avion	Totals:
prosody	Multivoc	45 9,733	60 8,567	50 8,64	50 10,46	205 9,302
	L-Cnetvox	45 12,733	60 11,95	50 12,96	50 12,4	205 12,478
	D-Cnetvox	45 11,356	60 12,6	50 12,8	50 13,56	205 12,61
	Totals:	135 11,274	180 11,039	150 11,467	150 12,14	615 11,463

The results of test 2 clearly show that Cnetvox-lecture and Cnetvox-dialogue are judged better by listeners than Multivoc. Hierachy between Cnetvox-lecture and Cnetvox-dialogue is harder to establish but we can notice that Cnetvox-dialogue is prefered for dialogue applications and Cnetvox-lecture best evaluated for monolog applications. This result can be explained by the fact that Cnetvox-dialogue takes into account some dialogic aspects of intonation. On the contrary Cnetvox-lecture is the standard prosodic output of the TTS system and is based on a "reading" model of intonation.

DISCUSSION

Results of test 1 are consistent with the results of test 2. While the two methods are quite different, the results often lead to the same conclusions:

- performance of Cnetvox-lecture and Cnetvox-dialogue are always superior to Multivoc.
- these prosodic styles are hardly discriminable and are dependent on the application in which they are used.
- the coherence between the results of test 1 and test 2 validate both methodologies...

In fact, the diagnostic capabilities of test 1 have been exploited but are not described here (see Santi &Guaïtella, 1992b). Test 1 is far more difficult to carry out than test 2 but is not suited to the same goals. If only subjective evaluation in order to discriminate among different algorithms is needed test 2 is sufficient and more efficient. However, if diagnostic information about misfunctionnings of the algorithms are of interest, then test 1 can be quite useful. No evaluation method can tell directly what to do but a well chosen methodology may be capable of telling you on what aspects of the system have to be improved. Concerning methodology itself we claim that the best method should be that closest to the real situation of communication. Even if in a test situation the listener cannot be considered as a user but rather as an observer, the coherence of what he observes is fundamental. As a consequence we maintain that isolated sentences or repetitive speech turns (i.e without any pragmatic context nor semantic coherence) do not constitute a good material for evaluation tests. Larger units such as the blocks used in our tests are certainly much more suited to keep the listener's attention on *what* is said even if he has to concentrate on *how* it is said.

AKNOWLEDGEMENTS

This research was financed by France Télécom (CNET, Lannion; Convention CNET/IPA CB 500 437). Thanks to D. J. Hirst, R. Espesser and C. Pavlovic.

REFERENCES

Grice M., Hirst D.J., 1991, "The evaluation of prosody in text-to-speech systems in a number of languages", *SAM Internal Repport*, Esprit Project 2589, So.2 (part 1).

Grice M., Vagges K., Hirst D.J., 1991, "Assessment of intonation in text-to-speech synthesis systems - A pilot test in english and italian", *Proceedings of the Eurospeech Conference*, Genova, vol.2, 879-882.

Guaïtella I, Santi S., 1992, "The punctuation and perception of read and spontaneous prosody: an application to speech synthesis", in: G. Bailly, C. Benoit & T. Sawallis (eds), *Talking machines: Theories, models and designs*, Elsevier, North Holland, 351-366.

Hirst D.J., Nicolas P., Espesser R., 1991, "Coding the f0 of a continuous text in French: an experimental approach", *Proc. of the XIIth I.C.Ph.S.*, Aix-en-Provence, 234-237.

Howard-Jones P.A. & the SAM Partnership, 1991, "'SOAP' - A speech output assessment package for controlled multilingual evaluation of synthetic speech", *Proceedings of the Eurospeech Conference,* Genova, vol.1, 281-283.

Pavlovic C.V., Rossi M., Espesser R., 1990, "Use of magnitude estimation technique for assessing the performance of text-to-speech synthesis systems", *J. Acoust. Soc. Am.*, 87, 373-382.

Santi S., 1992, "Methodes d'evaluation subjective de la composante prosodique en synthèse vocale", *Prépublication des actes du Séminaire Prosodie*, Aix-en-Provence, 36-46.

Santi S., Guaïtella I., 1992a, "Evaluation de la qualité de la prosodie de synthèse en situations de dialogue", *Document CNET/IPA 92-06.*

Santi S., Guaïtella I., 1992b, "Localisation et analyse de défauts prosodiques en synthèse du dialogue", *Document CNET/IPA 92-11.*

Issues in the Perception of Prosody

Jacques Terken
Institute for Perception Research
P.O.Box 513, 5600 MB Eindhoven, The Netherlands

ABSTRACT

Three issues in the perception of prosody are discussed: the perception of pitch variation, of prominence and of phrase boundaries. Also, the relation of these issues to the transcription of prosody is treated, and a number of remaining questions are addressed.

INTRODUCTION

According to Crystal (1969) one of "the main problems that have yet to be faced" in prosody research is "the development of a more immediately meaningful system of notation than has been hitherto available". The quotation might equally well have been taken from a much more recent source, as is evident from the current concern with transcription systems for prosody (Bruce, 1989; Silverman et al., 1992). In this contribution I will argue that knowledge about the perception of prosody may provide guidelines as to what kinds of information should be transcribed and for the design of an appropriate coding scheme. Understanding the perception of prosody means that we understand how the listener interprets actual prosodic patterns in relation to his knowledge of phonological constraints as studied by prosodic phonology. Below, I will explore the perception of actual prosodic patterns in relation to the listener's abstract knowledge for three aspects of prosody.

THE PERCEPTION OF PITCH VARIATION

An important issue in the perception of pitch variation is how the listener extracts categorial tonal distinctions from the continuously varying pitch. There are two main approaches to the description of sentence melody: either by means of level tones or targets (High, Low), or by means of contour tones (Rising, Falling). Exponents of both approaches have claimed that their inventory of tones represents the perceptually and linguistically relevant units. This issue is not without importance because transcriptions are not easily translated from one approach to the other.

Instrumental analyses have helped little to clear up this issue. F0 contours are the result of many different contributions, both involuntary and voluntary ones (the latter due to the speaker's intention to produce certain melodic properties). Psycho-acoustic investigations of pitch perception have usually been restricted to very brief signals lacking the complexity of speech signals, so that their application to speech research is limited. Except for Thorsen (1979), who represents an early attempt to incorporate findings about pitch perception to the analysis of F0 contours in a principled way, most investigators have applied "trial-and-error" methods, based on an analysis-by-synthesis approach, to extract the information relevant to the perception of prosody (Bruce, 1977; Di Cristo, Espesser and Nishinuma, 1979; Fujisaki and Hirose, 1984; 't Hart, Collier and Cohen, 1990). For instance, by omitting increasingly greater variations from the F0 tracking (called "stylization"), making the result audible and comparing the synthetic contour to the original, 't Hart c.s. determined which variations were perceptually relevant (supposed to be intentional) and which were not. The "trial-and-error" methods give rise to stylized F0 trajectories, made up of sequences of (usually) connected straight lines or more sophisticated mathematical functions. The problem with such representations is that they *contain*, but need not *exactly represent* the perceptually relevant information. There may be alternative representations giving the same perceptual result.

Therefore, we need more insight into the perception of pitch in speech to differentiate between alternative representations. Recently, several studies on pitch perception in speech have been conducted in order to explore the perception of prosodic characteristics. House (1990) applies insights from psychoacoustic investigations to the properties of the speech signal. Stimuli which differ along several dimensions usually give higher sensitivity thresholds compared to stimuli which differ only along one dimension. The same applies to dynamic stimuli compared to static stimuli. E.g., the threshold for the perception of a pitch rise in speech-like stimuli is influenced by the rate of amplitude decrease (Van Der Horst, 1993). On the basis of such considerations, House argues that adequate models of pitch perception in speech must take into account the dynamic nature of speech: Speech exhibits continuous variation of spectral and amplitude properties, and areas showing fast spectral and amplitude changes such as CV transitions alternate with areas showing relatively little spectral and/or amplitude change such as the steady-state portions of vowels. In a series of ABX experiments he explored the influence of spectral variation on the perception of pitch. The results showed that the perception of pitch glide as such requires a spectrally stable portion of at least 100 ms, below which no pitch change will be perceived, but a succession of pitch levels. Although it is difficult to determine exactly which information listeners may have used as a basis for classification, his conclusions are confirmed by the outcome of an informal listening test which I conducted with stimuli constructed on the basis of his descriptions.

Further evidence for the perceptual reality of pitch levels in speech is provided by Hermes and Rump (this volume).

From the available data the following picture emerges. The perception of pitch change as such is impeded in portions with strong amplitude modulation and fast spectral changes such as consonants, and, most notably, consonant vowel transitions. Instead, pitch levels are perceived associated with the successive syllabic nuclei, or – at least – the most sonorant parts of the syllables. Actual pitch change will be perceived only if the F0 change starts in a spectrally stable portion of the speech signal, i.e., somewhere after the vowel onset and when the spectrally stable portion exceeds a certain minimum duration. It appears that in these cases the actual percept consists of a pitch level associated with the so-called *syllabic nucleus* followed by a pitch change. Categorial tonal distinctions appear to be dependent primarily on the pitch in successive syllabic nuclei and the presence or absence of an immediately following pitch change, at least for accented syllables. Near prosodic boundaries the situation seems different. An informal ABX test to assess the relative contributions of F0 in the vowel and in the post-vocalic consonant to perceived pitch suggested that in syllables preceding a silence the consonant F0 has a stronger influence on perceived pitch than the vowel F0.

There are a number of issues which remain to be cleared up. In the first place, the view proposed here implies that what counts is the pitch associated with the syllabic nucleus and the presence or absence of a pitch change following the nucleus, but that subtle differences in the realization of these pitch changes (e.g. having to do with the magnitude of a pitch change) will not play an important communicative role, because their perception is impeded by simultaneous spectral and amplitude variations. Secondly, the notion of "syllabic nucleus" remains to be defined. Generally, this is interpreted as a small region following the vowel onset, usually associated with the amplitude peak. It is clear, however, that High targets in pre-nuclear accented syllables may occur considerably later than this point (Silverman and Pierrehumbert, 1990), while is is evident that the details of their realization are communicatively relevant, e.g. for signalling relative prominence. In Hungarian, the height of pitch maxima following the syllabic nucleus is relevant for the distinction between declarative statements and yes/no questions (Gósy and Terken, 1993).

Thus, further investigations along the lines of those of House and Hermes and Rump are needed to determine how spectral and amplitude variations affect the perception of pitch variations before we can define the categorial distinctions in a principled way. Also, these insights will prove useful for the automatic extraction and labelling of melodic properties (see the contributions of d'Alessandro and Beaugendre, and Ten Bosch, this volume).

THE PERCEPTION OF PROMINENCE

Prominence has traditionally been related to three phonetic properties: loudness, perceived duration and pitch. It has been argued that pitch plays the most important role, because independent manipulation of the three phonetic properties showed that pitch cues could overrule the other cues. On the other hand, pitch variation doesn't appear to be essential to prominence perception: prominence can be perceived in the absence of pitch variation such as in whispered speech or in artificially monotonized speech (Cutler and Darwin, 1981). Nowadays, the interest has shifted towards investigating the combined effects of different properties on the perception of prominence, in particular in connection with the perception of degrees of prominence rather than the presence or absence of accent. It has been found, for instance, that perceived prominence varies as a function of the timing of pitch changes (Kohler and Gartenberg, 1991) and their phonetic realization (Repp, Rump and Terken, 1993). Also, the question has been addressed whether there is a trade-off between vowel duration and the timing of pitch changes (Rump, 1992; Fox, this volume). Finally, it has been suggested that conclusions with regard to the modest role of amplitude drawn from earlier research may be inadequate because the experiments involved inappropriate amplitude manipulations (Sluijter, this volume).

Still, it is evident that variations in the excursion size of pitch changes (i.e., variations in local pitch range) play an important role in the perception of degrees of prominence for focal accents, both for speakers and listeners. Speakers who are asked to speak a word with varying degrees of emphasis do so primarily by manipulating the local pitch range (Liberman and Pierrehumbert, 1984). In turn, listeners asked to judge the degree of prominence of an accented syllable do so very reliably as a function of variations in local pitch range (Gussenhoven and Rietveld, 1988). Therefore, models of prominence perception have in the first place aimed to capture the relation between pitch range variation and perceived prominence. One such model is proposed by Hermes and Rump (this volume), which was already briefly mentioned in the preceding section. It attempts to relate the paradigmatic aspect of prominence perception (i.e., having to do with intrinsic accent strength) to variations in local pitch range (excursion size).

Our understanding is still far from complete. In the first place, two accented syllables with the same excursion size measured on some appropriate dimension in the same phrase will not be perceived as equally prominent, due to the listeners' expectations about declination as the utterance proceeds (Pierrehumbert, 1979; Terken, 1991). The role of declination in the perception of relative prominence has been addressed by Terken (1991). Terken (1993) proposes a model for the perception of relative prominence, taking into account the observation that a given excursion size may result in varying degrees of prominence depending on how it is scaled in the overall pitch range of the speaker (see Figure 1). It does not yet incorporate the findings by Hermes and Rump (o.c.). Also, it is only tentative, and additional experiments are needed for evaluation. In fact, actual parameter values computed on the basis of earlier data did not give adequate predictions in two follow-up experiments (Repp et al., 1993), which means that either the parameter values may have to be adjusted or that the model is inadequate.

Secondly, it is not immediately obvious which phonological constraints are involved. Two influences besides declination need to be mentioned. Pitch maxima are varied in a gradient way to signal variations in prominence for paralinguistic reasons. In addition, the scaling of pitch maxima is also subject to downstep, a discrete phenomenon. However, the domain within which downstep operates, the conditions under which it may (if it is optional, i.e., contrastive) or must apply (if obligatory), and its relevance to everyday communication (as opposed to laboratory speech) are still unclear. As a consequence, investigating how the listener unravels the different influences is problematic.

Thirdly, the relation between paradigmatic and syntagmatic aspects of prominence perception (paradigmatic having to do with intrinsic accent strength and syntagmatic having to do with the strength of an accent relative to other accents) has not been cleared up yet. For instance, we don't know whether prominence gradations are perceived

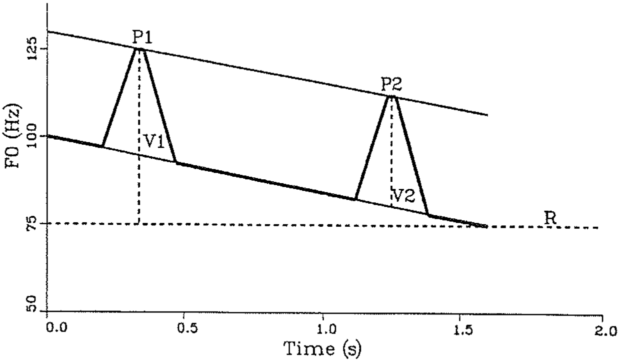


Figure 1: *Illustration of model relating judgments of relative prominence to pitch variation: accented syllables are perceived to be equally prominent if D2 = 0.9 (D1) + 0.23 (V1–R), with D: Distance between Peak P and Valley V, and R for a speaker-dependent fixed reference line. The values of 0.9 and 0.23 have been estimated on the basis of experimental data.*

primarily in a relative manner, due to the listener's ability to determine that one pitch change is perceptually more or less prominent than another, or whether listeners can also perceive gradations in perceptual prominence for individual pitch changes (and by implication may notice that one pitch change is not just more or less prominent but much more or much less prominent than another pitch change). In the latter case, the additional question is how such distinctions should be represented.

The frequently cited argument against a paradigmatic approach is that it implies categorial distinctions between different prominence levels primary, secondary etc.), and that such distinctions cannot be made very reliably which casts doubt on their categorial status. On the other hand, the experimental evidence from production and perception studies mentioned before favours the inclusion of paradigmatic aspects in the description. We can solve this problem by rejecting the assumption that prominence variations map onto a number of mutually exclusive categories such as ‘primary’ and ‘secondary’. Instead, we may treat prominence as a scalar feature with fuzzy category boundaries which may be employed to signal linguistic or paralinguistic distinctions. It seems likely that listeners have some mental representation of the probability density function for excursion sizes of pitch changes in speech which they employ to interpret an actual excursion size, in the same way as has been proposed for durations (cf. the notion of *normalized duration* proposed by Campbell, 1992; Wightman et al., 1992). Both the theoretical implications and the way in which listeners unravel linguistic and paralinguistic contributions to local pitch range variation remain to be investigated.

THE PERCEPTION OF PHRASE BOUNDARIES

In transcribing prosody in speech data bases one has to decide as to how phrase boundaries are to be coded. The particular decisions taken reflect assumptions about the distinctions that can be perceived by listeners. Often, a tripartition is made between strong, weak and no boundary, but experienced transcribers will immediately agree that there is much more variation that listeners might employ. Therefore, finer distinctions have been proposed, either within a strictly hierarchical framework (Price et al., 1991) or in a more recursive one (Ladd, 1992). The latter proposal appears particularly attractive, because it includes both paradigmatic aspects (having to do with the classification of boundaries as belonging to a certain type) and syntagmatic aspects (having to do with the grouping of units at the same level). However, the more refined coding scheme has been applied mainly by experienced transcribers, and therefore it should be explored whether untrained listeners can also make finer distinctions in a reliable way. Once this has been established, it can be investigated how the distinctions relate to phonetic properties.

De Pijper and Sanderman (1992) conducted an investigation to find out which distinctions can be made reliably by non-expert listeners. They presented listeners with a number of isolated sentences and asked them to assign a digit between 1 and 10 to each word boundary to indicate the perceived degree of separation of the two words, the idea being that a higher score would express a stronger prosodic boundary. It was found that listeners could reliably distinguish more than three levels, and that this was not due to the lexico-syntactic information, as the results correlated very well with those of a test in which the same sentences were presented with the segmental information distorted so as to make them unintelligible. Furthermore, the levels appeared to be overlapping, suggesting that there are no clear category boundaries but rather a more gradient scale. These results may be taken to reflect contributions both of categorial factors and gradient factors. For instance, while the presence of a pause (in combination with a tonal boundary marker) would always result in the perception of a strong prosodic boundary (e.g. a boundary between intonation phrases), it appeared that the perceived boundary strength increased as a function of pause duration.

Although a number of methodological issues have to be cleared up before firm conclusions can be drawn, the findings are compatible with Ladd's (1992) analysis. Listeners can make categorial distinctions between different types of prosodic boundaries (Price et al., 1991). At the same time they can make judgments as to the relative strength of two prosodic boundaries of the same type in order to infer hierarchical relations. This implies that the most informative transcription would represent both paradigmatic properties (concerning types of boundaries) and syntagmatic properties (concerning the hierarchical relations between constituents of the same type).

A related issue concerns the employment of different kinds of phonetic information. Beach (1991) has shown that there is a trade-off between tonal information, usually supposed to evoke a categorial distinction between different tones, and duration information, which is supposed to be much more gradient. Furthermore, there is individual variation in the extent to which listeners employ different sorts of information (Bruce et al., 1992). Simple decision models may be designed to account for the combined use of different kinds of information and the individual differences.

CONCLUSIONS

Considerable progress has been made in our understanding of how prosody functions in human communication. A major problem seems to be conceptual: to incorporate the insights into prosodic theory. Prosody is often conceived of as part of phonology, since phonology is concerned with characterizing the sound forms which may transmit certain distinctions in languages (i.e., distinguish between different readings of a sentence or a text). This implies that the description of prosodic phenomena is subject to the restrictions imposed by phonological representation. I have argued that if one does so, one runs into problems when trying to account for the full range of prosodic phenomena at different levels. It is not difficult, for instance, to design a decision model accounting for the relation between temporal and pitch variation in signalling phrase boundaries, but it is difficult to incorporate such a model into a phonological description. Therefore, the direction should be inverted: in order to understand the perception of prosody and the way prosody functions in human communication, the phonological constraints against which actual prosodic patterns are interpreted should be investigated as an intermediate step.

REFERENCES

Beach, C. (1991), "The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations", *Journal of Memory and Language*, Vol. 30, pp. 644-663.

Bruce, G. (1977), *Swedish word accents in sentence perspective* (Travaux de l'Institut de Linguistique de Lund, Lund).

Bruce, G. (1989), "Report from the IPA working group on suprasegmental categories", *Working papers*, Vol. 35, Lund University, Dept. of Linguistics, pp. 25-40.

Bruce, G., Granström, B., Gustafson, K. and House, D. (1992), "Aspects of prosodic phrasing in Swedish", *Proceedings of International Conference on Spoken Language Processing 1992*, (eds. J. Ohala et al.), University of Alberta, Canada, pp. 109-112.

Campbell, W.N. (1992), *Multi-level timing in speech* (Ph.D. Thesis, University of Sussex).

Cristo, Di, A., Espesser, R. and Nishinuma, Y. (1979), "Presentation d'une methode de stylisation prosodique", *Travaux de l'institut de phonetique d'Aix*, Vol. 6, pp. 125-146.

Crystal, D. (1969), *Prosodic systems and intonation in English* (Cambridge University Press, London).

Cutler, A. and Darwin, C.J. (1981), "Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency", *Perception & Psychophysics*, Vol. 29, pp. 217-224.

Fujisaki, H. and Hirose, K. (1984), "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *J. Acoust. Soc. Japan*, Vol. 5, pp. 233-242.

Gósy, M. and Terken, J. (1993), *Phonological relevance of timing and excursion size of pitch change: Evidence from Hungarian* (Manuscript, Institute for Perception Research, Eindhoven).

Gussenhoven, C. and Rietveld, A.C.M. (1988), "Fundamental frequency declination in Dutch: testing three hypotheses," *J. Phonetics*, Vol. 16, pp. 355-369.

't Hart, J., Collier, R. and Cohen, A. (1990), *A perceptual study of intonation*, (Cambridge University Press, Cambridge).

Hermes, D.J. and Rump, H.H. (1993), *Prominence in speech intonation induced by rising and falling pitch movements*, (Manuscript, Institute for Perception Research, Eindhoven).

Horst, Van der, R. (1993), *On the sensitivity for pitch movements in speech during amplitude changes*, (Manuscript, Institute for Perception Research, Eindhoven).

House, D. (1990), *Tonal perception in speech* (Lund University Press, Lund).

Kohler, K. J. and Gartenberg, R. (1991), "The perception of accents: F0 peak height vs. F0 peak position", *Arbeitsberichte Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel*, Vol. 25, pp. 219-241.

Ladd, D.R. (1992), *Compound prosodic domains* (Manuscript, University of Edinburgh).

Liberman, M. and Pierrehumbert, J. (1984), "Intonational invariance under changes in pitch range and length", in *Language sound and Structure*, ed. by M. Aronoff and R. Oehrle (MIT Press, Cambridge), pp. 157-233.

Pierrehumbert, J. (1979), "The perception of fundamental frequency declination," *J. Acoust. Soc. Am.*, Vol. 66, pp. 363-369.

Pijper, De, J.R., and Sanderman, A.A. (1993), "Prosodic cues to the perception of constituent boundaries", *Proceedings Eurospeech 93, Berlin, 21-23 September 1993*, to appear.

Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S. and Fong. C. (1991), "The use of prosody in syntactic disambiguation", *J. Acoust. Soc. Am.*, Vol. 90, pp. 2956-2970.

Repp, B.H., Rump, H.H. and Terken, J. (1993), *Relative perceptual prominence of fundamental frequency peaks in the presence of declination* (Manuscript, Institute for Perception Research, Eindhoven).

Rump, H.H. (1992), "Timing of pitch movements and perceived vowel duration", *Proceedings of International Conference on Spoken Language Processing 1992*, (eds. J. Ohala et al.), University of Alberta, Canada, pp. 1047-1050.

Silverman, K.E.A. and Pierrehumbert, J.B. (1990), "The timing of prenuclear high accents in English," in *Papers in Laboratory Phonology I: Between the grammar and the physics of speech*, ed. by J. Kingston and M. Beckman (Cambridge University Press, Cambridge), pp. 72-106.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992), "ToBI: A standard for labelling English prosody", *Proc. Internat. Conf. on Spoken Language Processing 1992, Banff, October 12–16 1992*, Vol.2, pp. 867-870.

Terken, J. (1991), "Fundamental frequency and perceived prominence of accented syllables", *J. Acoust. Soc. Am.*, Vol. 89, pp. 1768-1776.

Terken, J. (1993), "Baselines revisited. Reply to Ladd", *Language and Speech*, to appear.

Thorsen, N. (1979), "Interpreting raw fundamental frequency tracings of Danish", *Phonetica*, Vol. 36, pp. 57-78.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P. (1992), "Segmental durations in the vicinity of prosodic phrase boundaries", *J. Acoust. Soc. Am.*, Vol. 91, pp. 1707-1717.

A Numerical Model of Pitch Perception for Short-Duration Vocal Tones: Application to Intonation Analysis

Christophe d'Alessandro
LIMSI - CNRS
BP 133, F-91403 Orsay, France

ABSTRACT

An algorithm for automatic intonation analysis is described. It is based on a two-parameter model of weighted time-averaging with threshold for pitch perception. This model can be considered as a non-linear filter. In a first stage speech is decomposed into short-duration tonal segments using short-term energy. In a second stage these short-duration tones are analyzed using the numerical model. A set of short (static and dynamic) tones are then obtained, together with their (constant or time-varying) pitches. Stylized FØ contours are reconstructed from this set of tones. Stylized contours are resynthesized, and give synthetic sentences which are perceptually identical with natural sentences.

INTRODUCTION

This paper presents a numerical model of pitch perception for automatic intonation analysis. Automatic intonation analysis is an important challenge for both fundamental and applied prosodic studies. FØ, as it appears at the output of a pitch tracker, is generally difficult to interpret. A similar situation is encountered in singing: although the melody is precisely indicated in the vocal score, is precisely realized by the singer, and is precisely appreciated by the audience, there is no hope to read the melody directly on FØ curves. Perceptual constraints on melodic accuracy are more severe in singing than in speech, but in singing like in speech, many details of FØ are not perceived as significant variations, and many other details are not perceived at all. Clearly, a quantitative model of pitch perception is needed in order to fill the gap between the output of pitch trackers and the tonal decomposition described by linguistic analyses.

A quantitative model of tonal analysis for French has been presented by Mertens (1987). Both a linguistic description of French intonation and a system for automatic recognition of intonation have been developped. In the work by House (1990), qualitative results on the influence of spectral changes on intonation perception have been proposed, and a system for automatic recognition of intonation in Swedish has been described, along with an automatic FØ stylization procedure.

Compared to these works, we shall present herein a system using new psychoacoustic data on pitch perception for short tones. Our experimental data on pitch perception for short-duration tones with changing frequency were obtained in the context of a study on vibrato tones in singing. There is no room here to recall the experimental conditions of this study, and the reader is refered to d'Alessandro & Castellengo (1993) for details. We think that the model obtained for singing is fully applicable to intonation analysis in speech, because: 1) the durations, extents and FØ patterns used in our experiments are comparable to those observed in speech; 2) the psychological thresholds are probably higher for speech perception than for musical perception; 3) these thresholds are also higher for short-tones in isolation, compared to short-tones in context (see for instance Watson & al. (1990)).

TONAL SEGMENTATION

The main question in applying the results obtained for short tones to speech analysis is that, apparently, speech is not made of a succession of short tones. Nevertheless, it is generally accepted that a syllabic segmentation takes place in prosodic perception (at least for languages like French), and therefore one can consider that speech is perceived as a succession of tonal segments related to syllables.

An extensive review of syllabic segmentation, particularly in French, is reported in Mertens (1987): in this study the segmentation algorithms were mainly based on short-term energy, or on loudness. Another criterion for syllabic segmentation has been proposed by House (1990): the so-called "spectral constraint hypothesis". Nevertheless, in the computer implementation described by House, it appeared that the segmentation component used intensity measurements in much the same way as Mertens. As far as we know, no quantitative psychoacoustic data are yet available to give some evidence on optimal threshold values. In our system, tonal decomposition is also based on ad-hoc energy thresholds. Provided that speech is decomposed into short-duration tonal segments, the model of short-tone perception can be introduced.

STATIC AND DYNAMIC TONES

In d'Alessando & Castellengo (1993), the pitch perceived for short vocal vibrato tones has been measured using a method of adjustment. The stimuli were synthetic vocal tones, produced by a formant synthesizer. The parameters under study were the tone duration, the frequency extent, the vibrato rate, and the nominal frequency as a function of the fractional number of vibrato cycles.

All the subjects noted that for several patterns it was possible to perceive a dynamic tone (a glissando, or gliding tone), rather than a static tone.

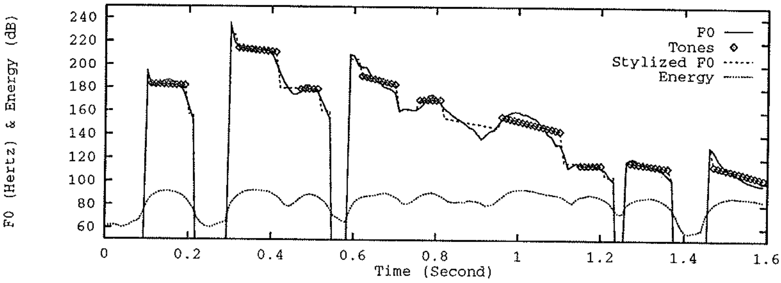


Figure 1: *Tones, FØ, stylized FØ and short-term energy.*

A dynamic tone is defined as a perceived tonal movement: the the FØ pattern (the physical variation of the fundamental frequency) results in a psychologically gliding tone (or glissando). For a static tone, the FØ pattern results in a single pitch percept.

A separate perception took place for the high and low parts of the vibrato cycles, for large frequency extents or slow vibrato rates (long duration). The same type of phenomenon was pointed out by Nabelek & al. (1970) in their study on pitch of tone bursts with changing frequencies. Our experimental data on fused (static tones) and separated (dynamic tones) pitch perception have been be consistently related to the glissando threshold.

Many psychoacoustic and psychophonetic data on the glissando threshold are available. One can find an interpolation procedure and a unified view of these data was presented in 'tHart and al. (1990). They studied the distribution of the glissando thresholds published in the literature and they showed that the glissando thresholds where distributed around a curve G_{tr} (expressed in Semi-Tones/Second) which approximately satisfies the equation:

$$\log(G_{tr}) \simeq -2.00 \times \log(T) - 1.83 \qquad (1)$$

where T is the duration of the tone. 'tHart et al. reported that more than 75 % of the data in the literature lie within a distance of a factor of two from Equation 1, i.e. within the interval $[\log(G_{tr}) - \log(2), \log(G_{tr}) + \log(2)]$, in the double logarithmic scale. The fusion/separation situations observed in our experiments are in good agreement with the glissando threshold. The glissando rates in case of separation are in the 75 % interval around the glissando threshold. In case of fusion, the glissando rates are all below the threshold.

FØ INTEGRATION

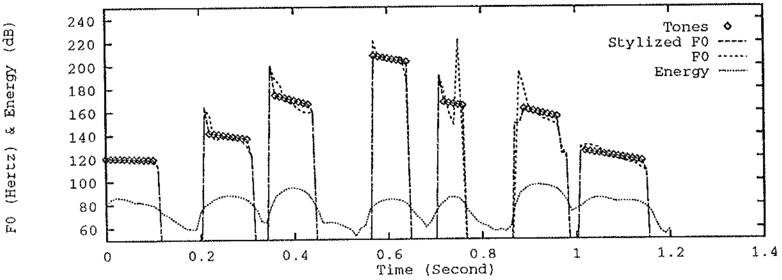


Figure 2: *Tones, FØ, stylized FØ and short-term energy.*

It appeared in the experiments that the final part of the tone had a larger weight on the pitch judgement than the initial one. The experimental results also suggested that the FØ patterns were time-averaged, at least in case of fusion. A quantitative model for such a process may be a time-average of the FØ pattern viewed through a data window. A simple model for the data window is a raised exponential memory function. Let $p(t)$ denote the pitch perceived at time t, f the time-varying FØ function, beginning at time 0, and let α, β be two constants, we have:

$$p(t) = \frac{\int_0^t (e^{-\alpha(t-\tau)} + \beta) f(\tau) d\tau}{\int_0^t (e^{-\alpha(t-\tau)} + \beta) d\tau} \tag{2}$$

In Equation 2, the constant β accounts for time averaging, and the constant α accounts for weighting of the past.

In the case of separation, the excitation patterns due to FØ extrema become more separated in time or in frequency, and these extrema are perceived as independant auditory events. It is reasonable to assume that if the glissando rate for a given condition is larger than the glissando threshold for the same condition, the amount of constant time-averaging represented by β is reduced. It does mean that in the separation case, the pitch judgement does not take into account the distant past. Therefore, Equation 2, with $\beta = 0$ is used when the glissando rate exceeds the glissando threshold.

The free parameters α and β have been estimated by minimizing the Root Mean Square distance between the model response and the experimental data. The optimal parameters obtained were $\alpha = -22$ and $\beta = 0.20$. The parameter β can be interpreted as the amount of long-term time-averaging (here $\beta = 20\%$). The parameter α represents the speed of decay of the exponential function. The bandwidth B of the corresponding low pass filter is: $B = \alpha/\pi = 7.00Hz$, and its time constant is $1/B = 0.14s$. Its seems that this optimal integration time constant can be consistently related to the auditory persistence.

APPLICATION

The model presented above was implemented in a computer program for tonal analysis. This progam contains several procedures : 1) pitch detection and voice/unvoiced decision. 2) short tone decomposition using short term energy. 3) glissando rate computation, and static/dynamic tone decision. 4) for static tones, the perceived pitch value is computed, according to Equation 2. For dynamic tones, two values are computed, at the beginning and end of the tone, according to Equation 2 with $\beta = 0$. These two target values are linearly interpolated.

This system was tested on a corpus of read French speech. It appeared that only two broad types of tone were necessary and sufficient to describe the intonation (static tones, and unidirectional dynamic tones). In order to test this stylization procedure, we synthesized the unprocessed and stylized tonal contours using LPC synthesis. The natural and synthetic sentences were generally perceptually indistinguishable. Errors occured only when the segmentation procedure failed (i.e. the tonal decomposition was wrong).

Figure 1 shows the application of the tonal analysis to the sentence “cessez de faire du bruit les enfants!” [sesedøfɛrdybrμilezɑ̃fɑ̃]. Figure 2 shows the sentence “ils sont partis pour Paris” [ilsõpartipurpari].

CONCLUSION

Further work is needed in order to understand the pitch of short tones. Particularly, no quantitative data and no model are yet available for describing the pitch of short-tones with joint time-varying energy and FØ patterns. Some work on glissando threshold in context is also needed : the glissando thresholds used in this study are probably too severe for continuous speech. It seems that the main problem remains tonal (or syllabic) decomposition.

The study presented herein concerned French, which is sometimes described as a “syllable-timed” language. For our corpus of read French, it appeared that only two types of short-tone patterns were sufficient: static tones and unidirectional dynamic tones. More complex patterns might be encountered in other languages or dialects (for instance in some dialects of popular French, rise+fall patterns can be found on a same syllable).

Our system describes intonation in terms of a succession of perceptually relevant tones. It provides a “tonal score” which can serve as a reliable basis for studying the melodic structures of speech.

REFERENCES

C. d’Alessandro & M. Castellengo, (1993). “The pitch of short-duration vibrato tones”, preprint LIMSI, NDL 93-04, to appear in J. Acoust. Soc. Am.

D. House, (1990). *Tonal Perception in Speech*, Lund University Press.

P. Mertens, (1987). *L’intonation du Francais. De la description linguistique à la reconnaissance automatique..* Unpublished doctoral dissertation, Catholic University of Leuven.

I. V. Nabelek, A. K. Nabelek & I. J. Hirsh (1970). "Pitch of tone bursts of changing frequency", J. Acoust. Soc. Am. **48**, 536-553.

J. ’tHart, R. Collier, and A. Cohen (1990). *A perceptual study of intonation*, (Cambridge Univ. Press, UK).

C. S. Watson, D. C. Foyle & G. R. Kidd (1990). “Limitted processing capacity for auditory pattern discrimination”. Acoust. Soc. Am. **88**, 2631-2638.

Experimental investigation on the perceptual consistency and the automatic recognition of prosodic units in spoken German

Matthias Reyelt
Institut für Nachrichtentechnik, Technische Universität Braunschweig
Schleinitzstr. 23, D-38092 Braunschweig, Germany
email: reyelt@ifn.ing.tu-bs.de

ABSTRACT

A corpus of about 500 sentences has been prosodically labelled by five students. They marked intonational phrase boundaries and accented syllables. The paper describes the inventory of prosodic labels that was used in the experiment and the resulting consistency of the parallel transcriptions. Also some preliminary results of the automatic recognition of these prosodic categories are presented.

INTRODUCTION

In the german compound project VERBMOBIL it is the task of the PHONDAT section to provide labelled speech data for training and evaluation purposes. At Braunschweig University the Institute for Communications Technology works at the development of a speech workstation for prosodic labelling. This workstation shall include software modules for speech signal analysis, linguistic analysis of the spoken text and a speech synthesis modul.

Additional research concerns appropriate labelling inventories and instructions and the achievable consistency of prosodic transcriptions.

In a pilot investigation a small basic inventory of prosodic labels has been considered. The labels were supposed to denote basic auditory units of prosody which should be perceiveable to human subjects after only a few simple instructions. On the other hand these units of course are also assumed to be linguistically relevant, and the instructions were such as to direct the attention of the subjects to an overall auditory impression rather than to certain specific features such as pitch or loudness. Using such a label inventory about 20% of the speech data recorded in PHONDAT were labelled prosodically. The labelling was done in parallel by five students. The parallel transcriptions not only provide the possibility of consistency investigations but can also be merged into a single less subjective reference transcription.

PROSODIC LABELS AND LABELLING INSTRUCTIONS

The prosodic categories under investigation refer to the *intonational phrase* domain. In particular, it was the task of the subjects to mark *phrase boundaries* and to assign to each syllable one of at most four levels of stress (or rather of *prominence*).

None of the students that took part in the experiments had any prior experience in labelling, either phonetic or prosodic. Therefore the description of the labels and the instructions had to be carefully chosen to be intuitively clear to the subjects.

For these reasons the labelling instructions were developped in a pilot study, in which 46 single sentences read by one speaker were labelled by two groups of two students each. The groups labelled the material several times, and after each session the results were evaluated and the instructions revised.

In the first test one group was instructed to assign to each syllable one of four *stress levels ("Betonungsstufen").* However, the resulting transkriptions revealed that this term is rather inconvenient for consistent labelling. The subjects were rather uncertain in their decisions, and frequently the labels reflected their impression of pitch contour rather than an impression of syllable prominence. The results for the phrase boundaries however were significantly better. The subjects seemed to have a clear idea about the category *intonational phrase.*

For the following session the instructions were revised. The subjects were instructed to label *primary accent ("Hauptakzent")* and *secondary accent("Nebenakzent")*. This time there was a rather clear correspondence between the transciptions of different subjects, although the number of marked accents still differed a lot due to the individual ideas about the degree of accentuation.

In these tests accented syllables and phrase boundaries were labelled separately. In the next session the label *phrase accent ("Phrasenakzent")* was defined as being the most prominent syllable in an intonational phrase. Additional *secondary accents* could be marked. In order to avoid training effects the test sentences were labelled this time by the second test group. Although these students were not used to the labelling the results of this test showed better consistency than the former and the students labelled faster and were more certain in their decisions.

PROSODIC LABELLING OF THE PHONDAT92-DATABASE

The PHONDAT92-database is a corpus of 200 sentences spoken by 15 speakers. From this corpus 60 sentences from 8 speakers were chosen. 5 students (different from the pilot test) labelled the 480 sentences. The speech signal was presented to the subjects on a computer screen. The subjects could mark or correct phrase boundaries with a mouse and they could play back the sentences completely or partially as often as they liked to.

Labels

Labels and instructions were similar to those in the final pilot test. The transcribers were instructed to denote intersections ("Einschnitte") between intonational units as *phrase boundaries (PB)* first, then to mark the most prominent syllable within each phrase as the *phrase accent (PA)*. Additional accented syllables could be marked as *secondary accent (SA)*. They were also allowed to use the label *emphasis ("Emphase")* instead of PA whenever they felt that a syllable was exceptionally prominent. The transcribers were also instructed not to pay attention to particular features of the speech signal (e.g. pitch contour or loudness) but only to their overall impression.

Labelling of distorted speech

In order to investigate how much the prosodic labelling is influenced by the linguistic sentence structure, the material of one speaker was distorted, so as to destroy the segmental structures whereas preserving the suprasegmental structures. To this aim the short time spectra were calculated and the magnitudes were clipped to a certain threshold level. These clipped spectra were multiplied with the speaker's long time spectrum and adjusted to their original loudness level. Thus the spoken text could not be understood any more, however the prosody was assumed to be the same as before. The distorted material was labelled by two of the transcribers.

Results

The students had no fundamental difficulties in perceiving the prosodic units described above. In table 1 the average number of syllables that were provided with certain labels is shown. At least for three of five transcribers the numbers of syllables marked PA and PB are quite similar.

The correspondence between two subjects for a specific label is calculated as follows:

$$corr_{1,2,label} = \frac{n_{corr(1,2),label}}{(n_{1,label} + n_{2,label})/2} \qquad (1)$$

where $n_{corr(1,2),label}$ is the number of syllables carrying the same specific label in

Table 1: *Number of labels as produced by five transcribers. The values are averages over eight speakers. The total number of syllables per speaker was 951*

subject	PA	SA	PB
CHR	157	114	90
KER	185	102	114
KAT	151	99	83
HEI	129	72	65
SEB	151	145	83

both the transcriptions of subject 1 and 2; $n_{1,label}$ is the total number of syllables carrying that label in the transciption of subject 1, and $n_{2,label}$ the total number in the transcription of subject 2. In table 2 the average correspondence between the five transcribers is illustrated for eight speakers. The consistency of the prosodic labelling matches that found for narrow phonetic labelling [1]. The correspondence for the SA is remarkably worse than for PA and PB.

Table 2: *Correspondence between the five subjects compared for eight speakers. The percentages are average values over the correspondences between two subjects*

speaker	AWE	KKO	KMA	RTD	MKN	HPT	CHK	WSE
phrase accent	66%	79%	75%	75%	71%	69%	72%	79%
secondary accent	32%	44%	44%	41%	39%	38%	42%	41%
phrase boundary	71%	75%	83%	75%	78%	67%	76%	84%

The transcribers who labelled the distorted speech had many technical difficulties. Although the syllable boundaries were displayed to the subjects on the screen, especially for short syllables it was very difficult to associate an acoustically percepted accent with a specific syllable. The comparison of this transcription with the original transcription of the same speaker and transcribent sometimes showed that the accent label had been assigned to the syllable just before or just after the syllable that carried the lexical stress. In spite of these errors the remaining correspondence with the original transcription was remarkably well: about 62% for PA and 45% for SA. Moreover the number of marked accents differed only slightly, 150/152 PA and 107/122 SA. Also the number of marked boundaries is quite similar (84/91), whereas the boundary positions are often different, and the correspondence is only 48%. This might be caused in part by the problems mentioned above; another problem might be that the loss of the segmental information makes it hard for the transcribers to recognize syllable lengthening.

AUTOMATIC RECOGNITION OF ACCENTED SYLLABLES

The prosodically labelled data were used for studies in the automatic recognition of the prosodic categories PA and SA . The 46 sentences from the pilot tests were taken as a sample set for classification. These sentences contained 505 syllables. For the sample fundamental frequency (F_0) and loudness were calculated.

The 505 syllables were classified by a *nearest neighbour classifier* using the *leave-one-out*-method (cf. [2]). The procedure is as follows:

The syllable to be classified is separated from the sample set. The remaining sample set is used as a reference for classification. The syllable is then classified and rejoined to the sample set. This is repeated for every syllable in the sample set.

The *nearest neighbour classifier* associates the *pattern* to be classified (e.g. F_0 or loudness contour) with the *class* (i.e. the prosodic category) of the least distant syllable in the sample set. The required distances were calculated from the F_0 or loudness contours by *dynamic time warping.*
Two tests were accomplished: In test 1 the reference transcription was produced by the author. This transcription contained 88 PA, 20 SA and 397 unaccented syllables. In the second test the sample set contained only those syllables which had been identically labelled by the author and the two subjects of the second test group (see above). The resulting sample set contained 367 unaccented syllables and only 44 and 3 syllables labelled PA and SA respectively. The tests were performed for F_0 and loudness.

Results

Since the sample set contains only a small number of accented syllables it may not representative for the conditional probability distributions of the features under consideration. Hence also the the recognition rates that are given in table 3 are not representative. Yet it can be concluded that for the automatic recognition of syllable stress F_0 is much more important than loudness. The results support the hypothesis that whenever the stress level as perceived by listeners is uncertain also automatic recognition becomes less certain.

Table 3: *Recognition rates for classification of prosodic labels*

	test 1		test 2	
prosodic label	SA	PA	SA	PA
loudness	0 %	19 %	–	–
pitch	20 %	40 %	0 %	43 %

CONCLUSION

As a result of this investigation a set of basic prosodic labels and and labelling instructions are defined that can be rather consistently labelled even by untrained listeners. With this label inventory a part of the PHONDAT-database is labelled. Further investigations will concern the consequences of this experiment for the design of the workstation for prosodic labelling and also the possibilities for automatic recognition of accents and phrase boundaries.

REFERENCES

[1] H.G. Tillmann, B. Eisen & Ch. Draxler. Consistency of judgements in manual labelling of phonetic segments: The distinction between clear and unclear cases. In *Proceedings ICSLP 92*, pages 871–874, 1992.

[2] M. Reyelt. Automatische Extraktion prosodischer Merkmale aus den Verläufen von Sprachgrundfrequenz und Lautheit. In G. Görz, editor, *KONVENS 92*, pages 385 – 389, Berlin, 1992. Springer.

This research was sponsored by the German Ministry of Research and Technology. Contract No. 01IV103C. The views and conclusions contained in this document are those of the author.

Algorithmic classification of pitch movements

Louis ten Bosch
Institute for Perception Research/IPO
P.O. Box 513, 5600 MB Eindhoven, the Netherlands.
e-mail: tenbosch@prl.philips.nl

ABSTRACT

In this paper, we discuss the construction of an algorithm that classifies pitch movements according to the IPO intonation labelling. The classification is performed by a feed-forward network, interpreted as a multi-linear classifier. In speaker-independent tests on a corpus of speech read by non-professionals, up to 81 % of the 279 pitch movements in the test corpus were correctly classified. These results are obtained by using information from the sampled speech data files only; a grammar will be used in the second stage of this study.

Keywords: Automatic classification, multi-linear classification, IPO-intonation system, speech recognition.

1. INTRODUCTION.

In this paper, an algorithm will be described aiming at the (semi-)automatic classification of pitch movements. The algorithm is trained and tested for Dutch. Its input is a sampled data file of an utterance; its output consists of a character string containing intonation transcriptions ('labelling'). Optimally, the algorithm should come up with a labelling that is indistinguishable from transcriptions produced by human intonation experts.

Algorithms that classify elementary patterns of speech melody are useful, e.g., the detection of phrase boundaries and accented syllables, the filtering of acoustically based hypotheses from an ASR-algorithm (Ostendorf, Wightman and Veilleux, 1991; Wightman and Ostendorf, 1992), and the labelling of large speech corpora.

In the present approach, the intonation labelling convention will be used which is known as the IPO-labelling. This labelling is chosen due to the relations posed in the theory between acoustic realization and perceptual labels ('t Hart, Collier, and Cohen (1990). This system defines ten labels (five different pitch rises labelled '1' to '5', five falls labelled 'A' to 'E'), with additional labels referring to a 'pointed hat' ('P'). Syllables bearing a perceptually relevant pitch movement can be labelled with at most one of these labels. Five labels are most common: '1', '2', 'A', 'B', and 'P' (also denoted '1&A'). The functional difference between the rises '1' and '2' and falls 'A' and 'B' corresponds to a phonetic difference with respect to the exact timing of the pitch movement: Accent-lending movements such as '1' and 'A' are generally earlier in the syllable than are the non-accent-lending movements '2' and 'B'.

The 'IPO-intonation-grammar' prescribes the permitted sequences of labels within one utterance ('t Hart *et al.*, 1990).

2. DESIGN OF THE ALGORITHM.

Several attempts have been made to come to a (semi)automatic classification of pitch movements, using, e.g., dynamic programming (Brew and Isard, 1990), or Hidden Markov Modelling (Butzberger, 1990). These approaches have proven to be fairly successful. We did not opt for these approaches, however, due to the difficulty of a proper interpretation of the many model parameters.

In this study, classification is based on multi-linear discrimination on features extracted from the sampled data file. The algorithm consists of two steps: (a) training, which is based on a labelled training corpus, and (b) classification, based on multi-linear discrimination.

(a) Training.

For the construction of a labelled database, a number of 800 Dutch sentences have been recorded. The average number of words per sentence was 7.4. These sentences (elicited speech) were spoken by over forty different speakers, male as well as female.

Sentences were manually labelled according to the IPO-system by four expert intonologists independently. A common subset was labelled, on which a consensus labelling has been defined. Eventually, a resulting total of 249 sentences were used in the test described below. The total number of labelled syllables was 817 (an average of 3.3 labelled syllables per sentence).

The distribution of the labels over the set of labelled syllables is '1' 31 %, '2' 14 %, 'A' 15 %, 'B' 14 %, 'P' 16 %, and other 10 %.

For each label, a corresponding class of acoustic realizations was constructed. The data space was constructed by feature representation in four steps (cf. Ten Bosch, 1993): (1) Pitch determination, (2) Correction of pitch measurements, (3) Determination of vowel onsets, and (4) Choice of pitch reference points.

Step (2) is included since the pitch determination algorithm usually returns the correct pitch as perceived on a (sub)syllabic scale. The actual pitch as it is perceived on the sentence scale (without gating) may deviate from the PDA outcome. A reinterpretation of the pitch contour results (figure 1).

In step (4), five pitch measurements per syllable were chosen as reference measurements: two measurements in the previous syllable, two measurements in the current syllable, and one measurement in the next syllable. These measurements were anchored at the moments of vowel onsets (Ten Bosch, 1993). The resulting data set (denoted $\mathcal{D}$) has dimension 5.

We make two observations about this representation. It is rather 'poor' in the sense that it does not make use of other spectral features. However, this 'poor' representation is sufficiently rich to cover the main distinctive features between the label classes (see below). Secondly, it deviates from the more standard representation spanned by 'excursion size', or moment of start and end of a pitch movement. The results of the training step allow these 'classical' features to be used in a description of class prototypes, but these features are certainly not unique.

(b) Classification.

The classification training was done on a subset of 65 % of the available set of labelled syllables. Most label classes in $\mathcal{D}$ are convex; they however do not necessarily obey a gaussian distribution (Ten Bosch, 1993). Consequently, the design of a Bayes classifier is not straightforward, and the recognition technique that is based on nearest prototypes may require more than one prototype per class (Ullmann, 1973, chapter 4). For the classification, a multi-linear discrimination was applied (cf. Fukunaga, 1972). The actual implementation of this optimization is done by a multi-layer classifier, i.e. a multi-layer perceptron (MLP), provided with a 5-n_h-n_y-topology, n_h ($2 \leq n_h \leq 5$) and n_y denoting the number of hidden units, and the number of output categories, respectively. For small-sized topologies, the MLP-results can be interpreted in a precise manner by relating them to a posteriori probabilities and CART-node questions (Richard and Lippmann, 1991; Breiman *et al.*, 1984).

In table 1, a summary is given of the results. The table shows results for several values of n_h and n_y. The normalized error (norm. error) denotes the mean error at an output unit. The column 'class. rate' denotes the fraction of correctly classified pitch movements. To cross validate the minimization, it was performed on 65 % of the available data, and tested on the remaining 35 % (279 syllables).

The results can be interpreted as follows. If the number of output categories n_y is clamped to 2, the best two 'class groupings' are {A, B} and {1, 2, P} (first row in table 1). Here, 'P' is more likely to belong to the group {'1', '2'}, rather than to {'A', 'B'}. An increase in the number of hidden units n_h, i.e. of the number of separating hyperplanes L_i used in the multi-linear discrimination in $\mathcal{D}$, shows an increasing classification rate (class assignment). If each class is to be labelled separately, an acceptable value of n_h is 5, as can be seen from the last four rows. As is suggested by the last row, it does not make sense to increase the discriminative power within $\mathcal{D}$ in order to optimize the classification rate substantially.

The result presented in the last row is tentative. It is the best result among 35 minimizations with substantially different initializations for the positions of L_i.

3. DISCUSSION.

In this paper, an attempt has been made to classify pitch movements by multi-linear discrimination. The main results are presented in table 1. These results have a unique interpretation from the point of view of technical optimization. The data obtained so far suggest that *unique* prototypical acoustical realizations of a class do not exist. In other words, prototypes do not specify the class topology on their own. Only under certain conditions, such as equal covariance matrices for each of the classes, the linear classification can be translated into a prototypical approach. These conditions are likely not to hold in $\mathcal{D}$. The present approach allows to look for distinctive features in the form of a set of hyperplanes L_i in $\mathcal{D}$, each hyperplane representing a specific 'property', i.e., a linear combination of the input features.

The question of how the classification results can be lined up with the 'classical' results in 't Hart *et al.* (1990) is solved by a close examination of the resulting MLP-weights. This shows that the difference between, e.g., '1' and '2' is mainly due to the value of the syllable-initial pitch in the current syllable relative to the syllable-final pitch in the previous and the current syllable. The behaviour of the 'classical' parameters that were known to be class-specific (e.g. timing and excursion differences, see 't Hart *et al.*, 1990) could be traced back in the test data as a trend only. This suggests that 'higher order' prosodic information (accents, grammar) must be used to further disambiguate between '1' and '2' or between 'A' and 'B'.

A final remark deals with the use of an intonation grammar. The disambiguation capability of the grammar presented in 't Hart *et al.* (1990) is, on the basis of the database presently used, for {1, 2} and {A, B} estimated to be 0.3. This means that the grammar disambiguates in the questions of the tree nodes {1, 2} and {A, B} in about one third of these cases.

ACKNOWLEDGEMENTS:

This research is supported by the Dutch Foundation for Scientific Research NWO, project no. 030-173-019.

REFERENCES:

Bosch, L.F.M. ten (1993). 'On the automatic classification of pitch movements.' Proceedings of the Eurospeech conference, Berlin, Germany.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and regression trees.* Wadsworth, Belmont, CA.

Brew, C., and Isard, S. (1990). 'Principles of contour labelling.' DYANA: Dynamic interpretation of natural language. CSTR-report, Edinburgh, UK.

Butzberger, J.W. (1990). *Statistical methods for analysis and recognition of intonation patterns in speech.* Thesis, Boston university.

Fukunaga, K. (1972). *Introduction to statistical pattern recognition.* Academic Press, NY.

't Hart, J., Collier, R., and Cohen, A. (1990). *A perceptual study of intonation. An experimental-phonetic approach to speech melody.* Cambridge University Press, Cambridge.

Ostendorf, M., Wightman, C.W., and Veilleux, N.M. (1991). 'Parse scoring with prosodic information: An analysis-by-synthesis approach.' Manuscript, Boston University.

Richard, M.D., and Lippmann, R.P. (1991). 'Neural network classifiers estimate Bayesian *a posteriori* probabilities.' Neural Computation 3, p. 461 – 483.

Ullmann, J.R. (1973). *Pattern recognition techniques.* Butterworth, London, UK.

Wightman, C.W., and Ostendorf, M. (1992). 'Automatic recognition of intonational features.' Proceedings 1992 International Conference on Acoustics, Speech, and Signal Processing, San Francisco. 221–224.

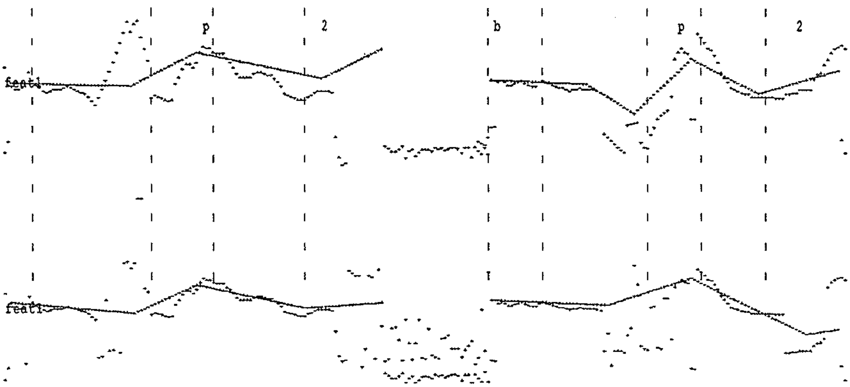


Figure 1. An example of correction and stylization of measured pitch. Along the abscissa, time is plotted. The pitch (scaled) is indicated along the ordinate. Correction and stylization remove perceptually irrelevant pitch jumps (e.g., octave errors).
Bottom: the original pitch measurement and its stylization.
Top: corrected pitch measurement and its stylization.

Table 1: *Results of the classification as a function of the two parameters n_h and n_y. For an explanation see the text.*

n_h	n_y	norm. error	class. rate	classes
2	2	0.81	0.79	$\{1, 2\} \cup \{P, A, B\}$
2	2	0.54	0.85	$\{1, 2, P\} \cup \{A, B\}$
3	2	0.59	0.83	$\{1, 2\} \cup \{P, A, B\}$
3	2	0.33	0.89	$\{1, 2, P\} \cup \{A, B\}$
3	3	0.41	0.81	$\{1, 2\} \cup \{P\} \cup \{A, B\}$
4	3	0.22	0.92	$\{1, 2\} \cup \{P\} \cup \{A, B\}$
4	5	0.42	0.54	$\{1\} \cup \{2\} \cup \{P\} \cup \{A\} \cup \{B\}$
5	5	0.27	0.81	$\{1\} \cup \{2\} \cup \{P\} \cup \{A\} \cup \{B\}$
6	5	0.23	0.83	$\{1\} \cup \{2\} \cup \{P\} \cup \{A\} \cup \{B\}$
≥ 7	5	< 0.23	$0.83 <$ rate < 0.86 ?	$\{1\} \cup \{2\} \cup \{P\} \cup \{A\} \cup \{B\}$

Perceptual cues of linguistic stress: intensity revisited

Agaath M.C. Sluijter & Vincent J. van Heuven
Holland Institute of Generative Linguistics /
Phonetics Laboratory
PO Box 9515, 2300 RA Leiden, The Netherlands

ABSTRACT

The general claim that intensity is a weak cue in the perception of stress is reconsidered. This claim is based on perception experiments in which intensity was varied in a naive way: all parts of the spectrum were increased with the same amount of energy. However, stressed syllables are produced with more vocal effort. If a speaker produces more vocal effort, higher frequencies increase more than lower frequencies. We show that the intensity differences as a function of stress are mainly located above 0.5 kHz. Varying intensity in this way would be much more realistic.

INTRODUCTION

The acoustical correlates of stress are pitch, duration, loudness and vowel quality. Of these, pitch and duration have been found the most important perceptual cues; intensity and vowel quality are generally claimed to be of lesser importance (Beckman 1986 and references mentioned there).

Indeed, an appropriate pitch movement provides an overriding cue, but is present only when the stress coincides with an accent (marking focus). When words are spoken outside focus, the position of the stress has to be inferred from the remaining cues. In our research we set out to determine the relative importance of vowel duration and intensity as cues for stress for materials spoken in and outside focus.

Traditionally, the cue value of intensity has been examined in a relatively naive and unrealistic way by simply manipulating the overall volume of syllables (or vowels). However, stressed syllables are produced with greater vocal effort than unstressed syllables. If a human speaker expends more vocal effort, as is required for the realization of a stress, intensity does not change uniformly across the spectrum, but higher frequencies are increased more than lower frequencies. Gauffin and Sundberg (1989) investigated the spectral consequences of changes in SPL during vocalization. Results show that with increasing loudness the levels of the higher bands (1.0-2.0 and 2.0-4.0 kHz) of a sustained vowel /ae/ increase more than the energy in the lower bands (0-0.5 and 0.5-1.0 kHz). Brandt, Ruder and Shipp (1969) independently varied vocal effort and intensity of continuous speech stimuli; speech produced with greater effort was estimated louder, even when intensity was held constant. Glave & Rietveld (1989) also showed that greater vocal effort is related to greater perceived loudness. Consequently, we expect that the spectral tilt of a stressed syllable differs from its unstressed counterpart: the energy in the higher frequencies of the spectrum increases more than the energy in the lower frequencies as this stressed syllable is produced with more vocal effort. In a production experiment we measured the energy in four contiguous frequency bands of stressed and unstressed vowels spoken in and outside focus, using both lexical and reiterant tokens.

METHOD

We selected the minimal stress pair /ká:non/-/ka:nón/ (cannon - canon) differing in stress position only. The target words were embedded in a carrier sentence: *Wil je [target] zeggen* 'Will you [target] say'. Targets were spoken in and outside focus. The condition with the target outside focus was realized by placing an accent on the word *zeggen*. In the other focus condition an accent was placed on the stressed syllable of the target, placing the target in focus. To control for influences of syllable structure, we also used the reiterant version of this word pair (repetition of the same syllable), where each syllable was replaced by the syllable *na* yielding nonsense words: /ná:na:/-/na:ná:/. The vowel /a:/ was chosen because it is the most open, longest vowel in Dutch. The resulting four stimulus types (2 stress positions * 2 focus conditions) with their reiterant versions were read by four male and six female speakers of standard Dutch. The speakers were recorded individually in a sound insulated booth, using semi-professional equipment. The subject's head was strapped to the chair so as to insure a constant distance between mouth and microphone.

Stimulus sentences were presented in normal Dutch orthography on a computer monitor in 8 random counterbalanced orders. Subjects always produced lexical and reiterant versions of each stimulus in immediate succession before going on to the next stimulus. After each stimulus, whether lexical or reiterant a 5s. pause was observed, during which interval the subject was required to inhale prior to initiating the next utterance.

RESULTS

The 640 utterances (2 stress positions * 2 focus conditions * 2 versions, i.e. lexical & reiterant * 10 speakers * 8 repetitions) were digitized. We used four repetitions (orders 2, 3, 7 and 8) yielding 320 sentences.

F1-maxima of the vowel in each syllable were determined using smoothed (30 ms integration) resograms. It was not possible to determine these maxima adequately in the syllable *non*, so these syllables were not used for further research. We measured the energy in dB in four contiguous filter bands: 0-0.5, 0.5-1.0, 1.0-2.0 and 2.0-4.0 kHz of stressed and unstressed vowels. In all cases we performed one-way analyses of variance for each filter band separately, with stress as a fixed factor.

In Figure 1 the energy differences in the four contiguous filter bands between the stressed initial syllable of *kánon* and the unstressed initial syllable of *kanón* are presented, i.e. in paradigmatic comparison. The results are broken down by gender and focus condition.

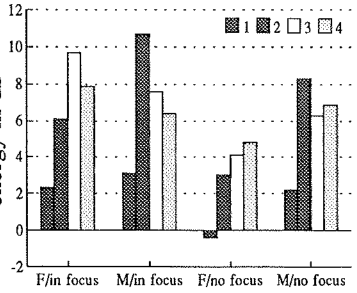


Figure 1. *Energy differences (in dB) in four contiguous filter bands (bar 1 = 0-0.5 kHz, bar 2: 0.5-1.0 kHz, bar 3: 1.0-2.0 kHz, bar 4: 2.0-4.0 kHz) between the stressed syllable 'ka' in 'kánon' and the unstressed syllable 'ka' in 'kanón'. The results are broken down by gender (M=male, F=female) and by focus condition (focus and no focus).*

As can be seen in Figure 1, there is an energy difference between stressed and unstressed vowels of about 3 to 11 dB in the highest three filter bands [all cases: p<.01], whereas there is only a slight difference in the base band [all cases: n.s., except the difference of 3.1 dB for male subjects in focus: $F_{1,31}$= 9.4, p=.005]. Male subjects have the largest energy difference between 0.5 and 1.0 kHz, whereas female subjects have the largest difference in the third filter band in words spoken in focus and in the fourth filter band in words spoken outside focus. This difference is probably due to the fact that female speakers usually have their formants at higher frequencies than males.

In Figure 2 the energy differences in the four contiguous filter bands between the stressed initial syllable of the reiterant version *nána* (replacing *kánon*) and the unstressed initial syllable of *naná* (replacing *kanón*) are presented in the left panel of the figure. The energy differences between the final syllables of the *nana* versions of the target words are presented in the right panel of the figure. As in figure 1, results are broken down by gender and focus condition.

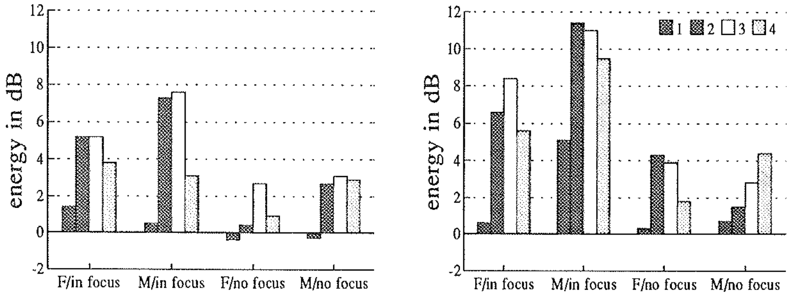


Figure 2. *Energy differences (in dB) in four filter bands between stressed and unstressed syllables 'na' (reiterant speech). In the left panel of the figure the initial syllables are presented, in the right panel the final syllables. Further see caption figure 1.*

The results for the initial syllables of the reiterant word pair (left-hand panel of figure 2), show that energy differences between stressed and unstressed vowels are mainly concentrated in the highest three filter bands with stronger effects in focus than outside focus. Outside focus female subjects only make a significant difference in energy between stressed and unstressed vowels between 1-2 kHz [$F_{1,47}$=5.3, p=.025, other cases F<1]. Male subjects only realize significant differences for targets spoken outside focus above 1 kHz [0-0.5 kHz: F<1; 0.5-1 kHz: $F_{1,31}$=3.2, n.s.; 1-2 kHz: $F_{1,31}$=6.7, p=.014; 2-4 kHz: $F_{1,31}$=7.1, p=.013]. The base band is hardly affected [all cases: n.s.].

The results for the final syllables of the reiterant word pair (right-hand panel of figure 2), show that for final syllables the effects are stronger, however comparable to the results for initial syllables (both reiterant and lexical).

In figure 3, the energy differences between the stressed and unstressed syllable of each member of the reiterant word pair are given separately. In the left panel of figure 3, the stressed and unstressed *na* of *kánon* are compared, in the right panel of *kanón* (syntagmatic comparison).

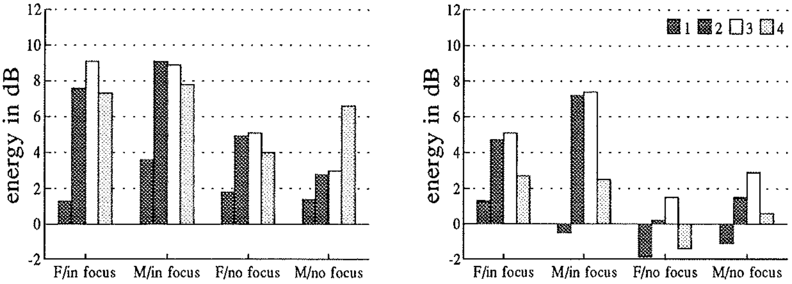


Figure 3. *Energy differences (in dB) of stressed and unstressed syllables within words. The left panel presents nána (for kánon), the right panel naná (for kanón). Further see caption figure 1.*

As can be seen in figure 3, the energy differences between the stressed and unstressed part of an initially stressed word are considerably larger than the differences in a finally stressed word, especially outside focus. Overall, the results are largely comparable to the results presented above.

CONCLUSION

We conclude that the intensity of the base band (0-0,5 kHz) was hardly affected by stress; however, intensity in the higher bands (0,5-1, 1-2, and 2-4 kHz) increased in stressed syllables by 5-10 dB, with stronger effects for accented than for unaccented words.

We are about to perform a subsequent perception experiment in which we shall examine the perception of stress position by manipulating vowel duration and intensity, the latter both in the classic way (i.e. uniform intensity differences) and in the more realistic way suggested by our production data (i.e. differences in higher bands only). Stimuli will be presented outside focus (without a pitch movement on the target) because of the fact that a pitch movement will always be the overriding cue in stress perception. We expect that realistic intensity manipulations (i.e., concentrated in the higher frequency bands) will provide stronger stress cues than can traditional intensity differences, and will be close in strength to duration differences. This finding then will rehabilitate the traditional claim that languages such as Dutch and English have dynamic (rather than melodic or temporal) stress.

REFERENCES

M.E. Beckman (1986), *Stress and Non-stress accent* (Foris, Dordrecht).

J.F. Brandt, K.P. Ruder and I. Shipp jr. (1969), "Vocal Loudness and Effort in Continuous Speech", *J. Acoust. Soc. Am.*, Vol. 46, pp. 1543-1548.

J. Gauffin and J. Sundberg (1989), "Spectral correlates of glottal voice source waveform characteristics", *J. Speech Hearing Res.*, Vol. 32, pp. 556-565.

R.D. Glave and A.C.M. Rietveld (1975), "Is the effort dependence of speech loudness explicable on the basis of acoustical cues", *J. Acoust. Soc. Am.*, Vol. 58, pp. 875-879.

Effect of Pitch Movement Timing on Perceived Duration and Prominence in Estonian and English Listeners

Robert Allen Fox, Department of Speech and Hearing Science
Ilse Lehiste, Department of Linguistics
The Ohio State University, Columbus, OH 43210

ABSTRACT

This study is one in a series in which we explore the possible influence of a listener's linguistic background on his/her perception of suprasegmental cues in the speech signal. Two experiments investigated the effect of the timing of a pitch rise on the perception of a medial syllable's perceived duration and/or prominence by native speakers of Estonian or American English. Results indicated that English subjects tended to perceive tokens with a later pitch rise as longer and more prominent than those with an early pitch rise. The reverse was true for the Estonian subjects in terms of perception of prominence but for these subjects, the timing of the pitch rise had little effect upon perceived duration.

INTRODUCTION

In the past two years we have examined the perception of "prominence" in sequences of both speech (using the nonsense token [bɑb]) and non-speech (signal-correlated noise) tokens by native Estonian and American English listeners while independently manipulating individual token duration and amplitude (Lehiste & Fox, 1992). As is well known, Estonian is a quantity language in which durational differences can distinguish between words. Duration in English, on the other hand, is not independently contrastive although it serves as one of the phonetic characteristics of stressed syllables. Our results have indicated that Estonian listeners are more sensitive to token duration in making their "prominence" decisions to both speech and non-speech stimuli than are English listeners and support the contention that the linguistic background of listeners has a potentially significant effect on the perception of prominence—particularly in terms of the utilization of the cues of amplitude and duration.

In these experiments we did not introduce any pitch variations. However, it is the case that variations in pitch patterns can influence both the perception of vowel duration as well as prominence, independent of changes in either duration or amplitude. For example, Lehiste (1976) demonstrated that the presence of a pitch change in a speech token could result in longer perceived duration. More recently, Rump (1992) demonstrated that the timing of accent-lending pitch movements had a significant effect both on perceived prominence and perceived vowel length in Dutch subjects. For example, he found that syllables with very early pitch rises were perceived as relatively more prominent than were syllables with later pitch rises. However, for Dutch speakers the prominence judgments did not always parallel perceived durations. In particular, there was a slight tendency for syllables with late pitch rises to be perceived as longer than those with early pitch rises (in opposition to the prominence decisions). The situation was somewhat different for pitch falls in that syllables with a late pitch were perceived as both relatively more prominent and longer. This suggests that, at least for Dutch subjects, the relationship between perceived prominence and duration is complex. This basic conclusion is supported by Hermes (1991) who argued that the timing of pitch movements, per se, may not be the most important factor determining prominence in Dutch, but rather this timing may determine the kind of "pitch movement with which a syllable is accentuated."

The present study compares the effect of the timing of pitch movements and language background on the perception of both syllable duration (Experiment 1) and prominence (Experiment 2) by Estonian and English listeners.

METHOD

Stimuli

The stimuli consisted of sequences of five monosyllables ([bɑb]) in which the third monosyllable had a pitch rise. The pitch pattern was superimposed on a slow F0 declination across the entire sequence of monosyllables as shown in Figure 1. The pitch rise started at 105 Hz and rose (over 120 ms) to 140 Hz; it then declined to 131 Hz by the end of the third monosyllable. The pitch rise began either at the onset of the vowel (0 ms onset) or 60 ms following vowel onset—this variation is similar to Rump's (1992) pitch movement onset factor. The experiment also manipulated the vowel duration of the third monosyllable as well as the timing of the pitch rise. In particular, vowel duration could be either 400, 425, 450 or 475 ms. Please note the these 5-syllable sequences differed only in terms of the third syllable.

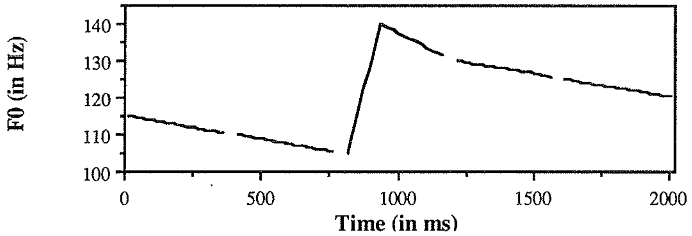


Figure 1. *Schematic of the basic pitch contour used (onset differences not shown).*

Procedure

In each experiment, subjects heard experimental trials which consisted an *anchor sequence* and a *comparison sequence* separated by a 500 ms interstimulus interval. There were two different anchors. In one anchor, the third syllable was 400 ms in duration with a pitch rise that began at vowel onset (0 ms). The third syllable of the second anchor was also 400 ms in duration, but had delayed pitch rise (beginning 60 ms after vowel onset) The comparison sequences represented either one of the anchors, or any of the other sequences (each of which had physically longer third syllables with a 0 or 60 ms pitch rise). The same set of comparison sequences were paired with each anchor sequence. In Experiment 1, subjects were required to indicate which sequence in a trial pair had the longer third syllable. In Experiment 2, subjects were required to indicate in which sequence the third syllable was "more prominent." Each possible pair of sequences was presented four times (twice with the anchor in first position, twice with the anchor in second position) in different random orders for the two experiments.

Subjects

There were two different subjects groups: twenty-three native speakers of English (living in Columbus OH) and thirty-three native speakers of Estonian (living in Tartu, Estonia). Subjects participated in both experiments (and always completed them in the order Experiment 1, Experiment 2).

RESULTS

As shown in Table 1, there was an overall tendency for both subject groups to perceive the second sequence as either longer or more prominent than the first sequence. This is very possibly due to a type of recency effect, making the relative "importance" of the second sequence (which was last heard) greater than that of the first sequence. However, since there was no significant difference between the two language groups in terms of this position effect (for either duration or prominence judgements), we will ignore positional variations in further descriptions of the data. Note, here and elsewhere, all significance tests were done using chi-square statistics.

Table 1. *Percentage of responses identifying either the first or second sequence as "longer" (in Experiment 1) or "more prominent" (in Experiment 2) across all experimental trials.*

	SEQUENCE POSITION	
	First	Second
Experiment 1		
Estonian	43.8	56.2
English	47.5	52.4
Experiment 2		
Estonian	42.8	57.7
English	44.9	55.1

One way to quickly determine the possible effect of the timing of the pitch rise upon perceived duration and prominence is to examine responses from those experimental trials which paired the different anchor sequences (a direct, "head-to-head" comparison). As shown in Table 2, the Estonian listeners judged the 0 ms anchor to be longer about as often as the 60 ms anchor. However, these same listeners judged the 0 ms anchor to be prominent 15% more often than the 60 ms anchor. The English subjects show a significantly different pattern for both duration and prominence judgements (at the .05 level). For the English responses, the 60 ms anchor is more often judged to be both longer and more prominent than the 0 ms anchor.

Table 2. *Percentage of responses identifying either the 0 ms or 60 ms pitch onset anchor tokens as "longer" (in Experiment 1) or "more prominent" (in Experiment 2) when the these two different anchors are compared directly ("head-to-head") in an experimental trial.*

	ANCHOR TOKENS	
	0 ms	60 ms
Experiment 1		
Estonian	51.6	48.4
English	41.3	58.7
Experiment 2		
Estonian	57.6	42.6
English	43.2	56.8

Shown in Table 3 is a breakdown of the responses when only one of the sequences in an experimental trial was an anchor. This table shows the percentage of time that the third syllable of the anchor sequence was identified as "longer" or "more prominent". Remember that in these experimental trials, the comparison sequence was always physically longer.

As expected, for both language groups, the number of times an anchor sequence was judged as being longer or more prominent decreased as the duration of the third syllable of the comparison sequence increased. In general, the Estonian subjects show little (non-significant) difference in duration judgments as a function of the timing of the pitch rise. The most critical comparison (highlighted in the table) is the mean response of 0 ms anchor vs. 60 ms comparison and 60 ms anchor vs. 0 ms comparison; that is, those trials in which the anchor token has a different pitch rise onset from the comparison sequence. For duration judgements, this comparison is almost identical for Estonian subjects (23.1% vs. 22.1%). This difference is only slightly larger (and only of borderline significance) in the prominence judgements (37.5% vs. 33.3%).

Again, however, the reverse is true of the English subjects, the 60 ms syllable is judged to be both longer and prominent more often than the 0 ms syllable. For example, in this critical comparison described above, English subjects show a significant difference both for the duration judgments (13.2% vs. 24.5%) and the prominence judgments (17.7% vs. 26.5%). It is interesting to note that for these data, English subjects seem to be more affected by the durational differences between the anchor syllable and the comparison syllable in making their prominence judgments than the Estonian subjects (i.e., the physically shorter anchor is identified more often as more prominent by Estonian subjects than by English subjects), contrary to the pattern that might be expected given our results in Lehiste & Fox (1992). However, one should note that the stimuli in this previous experiment on perception of prominence had no F0 variations.

Table 3. *Percentage of responses identifying the anchor stimuli as "longer" or "more prominent" when compared with the the longer comparison tokens. The critical comparison between anchor and comparison sequences that differ in pitch onset (are in bold italics).*

ANCHOR	COMPARISON TOKEN								
	0 ms Pitch Onset				60 ms Pitch Onset				*Overall Mean*
	425	450	475	*Mean*	425	450	475	*Mean*	
	Experiment 1 (Duration Comparison)								
Estonian Data									
0 ms	27.4	13.4	9.7	16.8	37.1	22.0	10.2	***23.1***	20.0
60 ms	32.3	19.4	14.5	***22.1***	35.5	14.0	13.4	21.0	21.6
Mean	29.9	16.4	12.1	19.5	36.3	18.0	11.8	22.1	
English Data									
0 ms	25.0	12.1	6.8	14.6	17.7	11.4	10.6	***13.2***	13.9
60 ms	42.4	22.0	9.1	***24.5***	28.0	11.4	7.6	15.7	20.1
Mean	33.7	17.1	8.0	19.6	22.9	11.4	9.1	14.5	
	Experiment 2 (Prominence Comparison)								
Estonian Data									
0 ms	34.1	33.3	31.7	33.0	35.0	41.4	36.0	***37.5***	35.3
60 ms	40.3	30.6	29.0	***33.3***	38.7	29.0	21.0	29.6	31.5
Mean	37.2	32.0	30.4	33.2	36.9	35.2	28.5	34.5	
English Data									
0 ms	24.2	27.3	18.9	23.5	22.0	15.2	15.9	***17.7***	20.6
60 ms	33.2	27.3	18.9	***26.5***	31.8	21.2	13.6	22.2	24.4
Mean	28.7	27.3	18.9	24.0	26.9	18.2	14.8	20.0	

Our data support the conclusion reached by Rump (1992) that differences in timing may produce differences in perceived prominence. However, there is a significant effect of language background. Unlike the Estonian listeners (or Rump's Dutch listeners), delaying the onset of the pitch rise tended to increase the perceived prominence of a syllable for English listeners. Our obtained increase in prominence when the pitch movement occurs later actually better matches the pattern obtained by Rump for a falling F0 contour. However, there are significant differences in the stimulus sets used in his study as opposed to ours (e.g., in his "early" timing condition the pitch rise occurs 120 ms before vowel onset rather than 0 ms in our study) and it is difficult to directly compare the two sets of prominence results.

ACKNOWLEDGMENTS

This work was partially support by a grant from NIA/NIH to the first author (AG08353) and by a grant from NSF to the second author (BNS-8719891). The help of Dr. Jüri Allik of Tartu University in running the Estonian listening tests is gratefully acknowledged.

REFERENCES

Hermes, D.J. (1991), "Prominence cased by rising and falling pitch movements with different positions in the syllable", In *IPO Annual Progress Report, Eindhoven,* Vol. 16, pp. 17-28.

Lehiste, I. (1976), "Influence of fundamental frequency pattern on the perception of duration", *Journal of Phonetics,* Vol. 4, pp. 113-117.

Lehiste, I. and Fox, R.A. (1992), "Perception of Prominence by Estonian and English Listeners", *Language and Speech,* Vol. 35, pp. 419-434.

Rump, H.H. (1992), "Timing of pitch movements and perceived vowel duration", in *Proc. of the International Conference on Spoken Language Processing,* Vol. 2, pp. 1047-1050.

Analysis and Perception of Intonation Expressing Paralinguistic Information in Spoken Japanese

Hiroya Fujisaki
Dept. of Applied Electronics, Science University of Tokyo
2641 Yamazaki, Noda, 278 Japan

Keikichi Hirose
Dept. of Electronic Engineering, University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

ABSTRACT
In addition to the linguistic information, prosody conveys para- and non-linguistic information. The present paper deals specifically with the role of intonation in transmitting the speaker's attitude/intention in spoken Japanese. Short declarative sentences, uttered with various attitudes/intentions, were analyzed to find their acoustic correlates. Perceptual experiments were also conducted using the same utterances as stimuli to find out the accuracy/reliability of transmission.

1. INTRODUCTION

Recent studies by the present authors and others on the prosodic features of the spoken Japanese have contributed much to the elucidation of the role of intonation in conveying linguistic information concerning such factors as lexical word accent, syntactic structure and discourse focus (Fujisaki and Kawai 1988). In the present paper, we define the linguistic information as the information that is explicit in or almost uniquely inferable from the written message.

In addition to linguistic information, however, intonation also conveys para- and non-linguistic information. Here we define paralinguistic information as the information that is not inferable from the written message but is added by the speaker to modify or complement the linguistic information. For instance, a written message can be uttered with various intonational patterns to express different intentions, attitudes and speaking styles which can be controlled by the speaker. On the other hand, nonlinguistic information concerns such factors as the age, gender, idiosyncrasy, physical and emotional conditions of the speakers which are not directly related to the linguistic or paralinguistic contents of the utterance and generally cannot be controlled by the speaker. It is, however, possible for the speaker to control the prosody of utterance to convey his/her emotion.

In comparison with the studies on the linguistic aspects of intonation, studies on the para- and non-linguistic aspects of intonation have been rather limited. However, studies on these aspects are no less important both for the basic understanding of human communication and for the realization of a high-quality man-machine communication through the spoken language. The present paper describes our initial effort toward the elucidation of these aspects of intonation and deals only with the paralinguistic aspects.

2. EXPRESSION OF SPEAKER'S ATTITUDE/INTENTION IN SPOKEN JAPANESE

In Japanese, as in many other languages, a sentence can be uttered with at least several different intonational patterns to express differences in the attitude/intention of the speaker. Let us take, for example, a positive declarative sentence such as "gakkō-e iku," meaning "go to the school," without specifying the subject. The sentence can be uttered at least with the following five different intonational patterns.

(1) Default intonation, indicating that the speaker is merely reporting the fact that someone (most commonly the speaker himself) goes to the school, without further attitudinal/intentional commitment.
(2) Assertive intonation, indicating that the speaker is definitely committed to the fact (i.e., "I am determined to go to the school.").
(3) Interrogative intonation, indicating that the speaker is addressing a question to a

second person (i.e., "Do you go to the school (now)?").

(4) Exhortative intonation, indicating that the speaker is addressing an invitation to a second person (i.e., "Shall we (now) go to the school?").

(5) Hesitative intonation, which indicates an interrogation to which the speaker is reluctant to accept a positive response, or expecting a negative response (e.g., "Do you (still) go to the school (in spite of this bad snowstorm, etc.)?").

On the other hand, some of these attitudes/intentions can also be expressed linguistically by adding a particle (or particles) to the verb which comes at the end of the sentence. For instance, the original sentence can be made into an assertion by adding the particle "-yo" to the end of the verb. Thus "gakkō-e iku-yo" is an assertion, while "gakkō-e iku-ka" is an interrogation. Some (but not all) of these final particles (or particle strings) and their functions are given below.

Table 1. *Some of the final particles and their default functions.*

particle(s)	function
-ka (-kai)	interrogation
-ne	confirmation
-yo	assertion
-ka-ne	interrogation
-nai	negation
-nai-ka	negative interrogation
-nai-ka-ne	negative interrogation

The addition of the particle(s), however, does not replace the role of intonation. Thus the sentence with the interrogative particle "-ka" is uttered still with an interrogative intonation. Furthermore, various modifications of their default linguistic meaning can be introduced by intonation, producing a variety of finer 'nuances.' For instance, "gakkō-e ikanai-ka," with a falling intonation indicates a strong suggestion rather than a negative interrogation, and thus can be considered to represent a directive attitude/intention.

The purpose of the present study is to find out the objective features that represent these differences on the other hand, and to find out to what extent these differences are perceived.

3. ANALYSIS OF INTONATION EXPRESSING PARALINGUISTIC INFORMATION

3.1 The speech material and the method of analysis

The sentences used for the current study has a very simple syntactic structure consisting of an object phrase and a verb phrase. The object phrase consists of a noun plus an accusative particle "o", while the verb phrase consists of a verb with or without being followed by a particle or a string of particles shown in Table 1, introducing various linguistic modifications of the original verb. Thus a total of eight sentence types (the original and its seven variants) are selected.

Since there exists an interaction between the intonation and the lexical accent of the constituent words, both 'accented' and 'unaccented' words were chosen for the noun and the verb; i.e., the accented "mame⌉" (bean(s)) and the unaccented "ame" (candy) for the noun, and the accented "mi⌉ru" (to look at) and the unaccented "niru" (to cook). Combination of these sentence types and word accent types produces a total of 32 sentences. The number is further doubled by adding the polite form of verb ending (e.g., 'mimasu' instead of 'miru'), resulting in a total of 64 sentences.

Each of these sentences was uttered with four or five different attitudes/intentions by four informants who were adult speakers of the common Japanese (i.e., the Tokyo dialect). At least three utterances were produced by a speaker for each sentence.

The speech material was digitized at 10 kHz with 12-bit precision. Fundamental frequencies were extracted by a modified autocorrelation method, and the F_0 contour was further analyzed using a model of the process of F_0 contour generation (Fujisaki and Hirose 1982). In addition to the phrase and accent components which constitute the F_0 contour of a declarative sentence with a default intonation, paralinguistic modification is often found to be expressed by another positive component which occurs toward the end of an utterance. Although this component may involve a mechanism other than that for the accent component, it was assumed in the present study that this component is generated by the accent control mechanism.

3.2 Results of F_0 contour analysis

Figure 1 shows the waveform, the F_0 contour and its closest approximation obtained by Analysis-by-Synthesis together with the extracted phrase component, and the underlying accent commands. The panels on the left-hand side are for the sentence "mame⌉-o mi⌉ru," and those on the right-hand side are for the sentence "mame⌉-o niru." The numbers on each panel indicate ① default, ② assertive, ③ interrogative, ④ exhortative, and ⑤ hesitative intonations, respectively. Comparison of these and other analysis results can be summarized as·follows.

(1) Of all the five different intonation patterns analyzed, ③, ④ and ⑤ are commonly characterized by a large utterance-final rising component, but its timing and magnitude are different among the three cases and differ also depending on the accent type of the verb. Significant differences also exist in the local and global tempo, especially for ② and ⑤.

(2) Compared with the default intonation ①, the assertive intonation ② is accompanied by a slightly faster overall speech rate and a longer accent command for the verb.

(3) Compared with the default intonation ①, the interrogative intonation ③ has the similar overall speech rate except for a longer final mora, accompanied by a significantly larger final rise command. When the verb is accented at the initial mora, this rise command appears as a separate command which is much larger than the accent commands of the 'accented' morae of the verb as well as for the object (noun), starting approximately at the segmental onset of the vowel of the final mora. When the verb is unaccented, however, this rise command coincides with the accent command for the second mora of the 'unaccented' verb, and starts approximately 30 msec prior to the segmental onset of the vowel of the final mora. Thus the timing of the final rise is siginificantly different depending on the accent type of the verb.

(4) Compared with the interrogative intonation ③, the exhortative intonation ④ is quite similar except that the magnitude of the final rise command is lower.

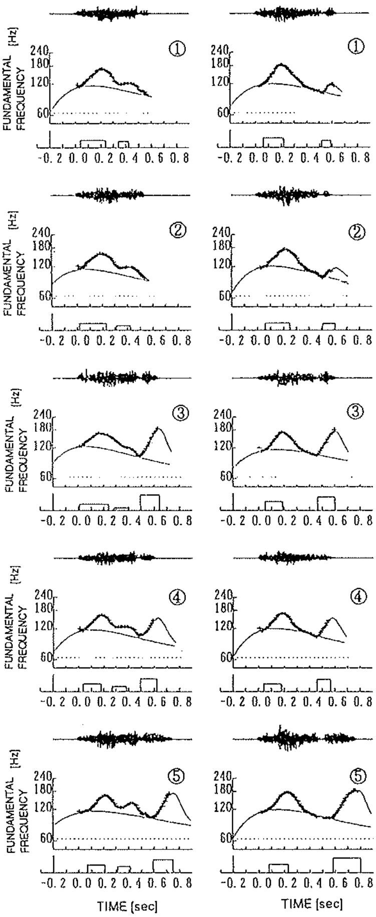


Figure 1. *Results of analysis of F_0 contours for the sentences "mame⌉-o mi⌉ru" (left) and "mame⌉-o niru" (right), with ① default, ② assertive, ③ interrogative, ④ exhortative, and ⑤ hesitative intonations.*

(5) Compared with the interrogative intonation ③, the hesitative intonation ⑤ is characterized by a marked change in the local tempo, i.e., the elongation of the final mora by a factor of two or more. This elongation is also accompanied by a delayed onset of the final rise command, starting at approximately 40 msec after the segmental onset of the vowel of the final mora. The magnitude of this rise command is similar to that for the interrogative intonation, but its duration is increased.

4. PERCEPTION OF SPEAKER'S ATTITUDE/INTENSION IN SPOKEN JAPANESE

The results of analysis mentioned in the foregoing section have indicated that differences in the attitude/intention are reflected in the acoustic characteristics, mainly in the F_0 contours and segmental durations. It remains to be investigated, however, whether or not the intended paralinguistic information is perceived as accurately as in the case of linguistic information such as the word accent type, etc. Perceptual experiments were thus conducted using both natural and synthetic speech sound stimuli.

4.1 Experiments using natural speech

One utterance each of the five variants of the sentence "mame⌉-o mi-mase⌉-n-ka," uttered by one male informant of the common Japanese, with five different attitudes/intentions was selected from the recorded speech samples. In this case, a default (neutral) intonation indicates a directive attitude and a falling intonation indicates a confirmative attitude, while interrogation, exhortation, and hesitation are expressed by more or less similar intonational patterns as in the case of "mame⌉-o mi⌉ru." They were arranged in random order with an inter-stimuli interval of 4 seconds and presented through headphones to the subjects, whose task was to identify the five attitudes/intentions.

Two normal-hearing subjects, who were both speakers of the common Japanese, took part in the experiment. Table 2 shows the averaged results of the two subjects in the form of a confusion matrix. The numbers indicate the percentages. It can be seen that the attitudes/intentions of the speaker are identified fairly accurately except for the confusion between interrogation and exhortation.

Table 2. *Results of a perceptual experiment on the accuracy of recognition of speaker's attitude/intention. The stimuli are natural utterances of "mame-o mima-sen-ka" uttered with five different attitudes/intentions. The numbers indicate the percentage.*

Stimulus	Response				
	① directive	② confirmative	③ interrogative	④ exhortative	⑤ hesitative
① directive	90			10	
② confirmative		97			3
③ interrogative	3		65	32	
④ exhortative	12		30	58	
⑤ hesitative					100

4.2 Experiments using synthetic speech

While perceptual experiments using natural speech confirmed that the paralinguistic information concerning the speaker's attitude/intention can be transmitted with a fair degree of accuracy in natural speech, further investigation was necessary to obtain a guideline for speech synthesis by rule. Thus another perceptual experiment was conducted using synthetic speech stimuli for the sentences "mame⌉-o mi⌉ru" and "ame-o niru."

Although the details cannot be given because of space limitations, the results of this perceptual experiment showed the range of parameter values for each intonational category and also indicated the influence of the word accent type of the verb on the timing of the commands for the terminal F_0 rise for interrogation, exhortation and hesitation.

REFERENCES

H. Fujisaki and H. Kawai (1988), "Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese," *Proc. 1988 Intl. Conf. on Acoust., Speech, and Signal Processing*, New York, 11–14 April 1988, vol. 1, pp.663–666.

H. Fujisaki and K. Hirose (1982), "Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation," *Preprints of the Working Group on Intonation, the 13th Intl. Congress of Linguists*, Tokyo, 31 August 1982, pp.57–70.

Modeling the Production of Prosody

Mary E. Beckman
Department of Linguistics, Ohio State University
222 Oxley Hall, 1712 Neil Ave., Columbus, OH 43210-1298, USA

ABSTRACT

This tutorial reviews some of the insights gained from fifteen years of research on the production of prosodic categories, concentrating primarily on modeling fundamental frequency patterns as reflections of intonation and phrasing. It closes with a hint of some at the directions in which this work is being extended to understanding the function of prosody in discourse and to the modeling of other phonetic dimensions.

INTRODUCTION

This tutorial was originally intended to introduce the "Production" of prosody. In the course of assigning papers to the different sections of this ESCA Workshop on Prosody, however, the organizers have expanded the scope of this section to cover "Models" and "Synthesis" as well. I can think of no more appropriate expansion. There is a particularly intimate and necessary connection between production data and modelling. Absent the kind of explicit laying out of theoretical assumptions afforded by a well-developed model, particularly by a model which has undergone that most rigorous test of being implemented in a synthesis system, production data are notoriously difficult to interpret. To be sure, this difficulty of interpretation is characteristic of all speech production data, but it is compounded for prosodic data, where there is not always the immediate theory of segmentation and contrast that the phonemic principle gives us. That is, for most aspects of prosodic structure, the contrasts are not the salient lexical contrasts that gave us the International Phonetic Alphabet for consonants and vowels. Rather, prosodic contrasts usually involve the pragmatic relationship of a syllable, word, or phrase to other elements in the utterance or discourse context, the sort of thing that is much harder to introspect about than the fact that one word is not another word. A speech scientist who wants to investigate some aspect of prosody must, therefore, first consciously decide to adopt some model to guide the design of the investigation and the interpretation of its results.

MODELING FUNDAMENTAL FREQUENCY PATTERNS

Nowhere are the connections between data and model building and between models and synthesis more fully exercised than in the investigation of phonologically contrastive intonational events as they are realized in fundamental frequency patterns. The location of this ESCA Workshop on Prosody in Lund is a felicitous reminder of the models of Swedish intonation developed here. The decade and a half since Bruce's (1977) seminal monograph has given us comprehensive descriptive models and systems for synthesizing fundamental frequency patterns in many other languages as well. Indeed for some of these languages, we have the luxury of two or more competing descriptions. For example, for English, we have Pierrehumbert's model implemented in the AT&T Bell Labs text to speech system (Anderson et al. 1984), Ladd's model implemented in the University of Edinburgh CSTR system (Ladd, 1987), and even an IPO-style system (Willems, Collier, and 't Hart, 1988). For Dutch, we have of course the IPO model itself ('t Hart and Collier, 1973), and a model implemented at Nijmegen (Gussenhoven and Rietveld, 1992). For Mandarin Chinese, there have been at least three different models (Shih, 1988; Gårding, 1987; Fujisaki, Hirose, Halle, and Lei, 1990), and likewise for Japanese (Fujisaki and Sudo, 1971; Sagisaka, 1990; Pierrehumbert & Beckman, 1988). The number of languages which have undergone the analysis and experiments necessary to construct synthesis models increases yearly. The system for German by Kohler (1991) and for Italian by Avesani (1990) are only two of many recent examples.

A notable result of this decade and a half of work is that we can now say with a fair degree of confidence which aspects of fundamental frequency patterns are likely to generalize across languages, and which aspects are likely to vary. One of the more important generalizations is that fundamental frequency patterns in all languages seem to be related in some way to discourse organization. In general, coherence among words or phrases can be signaled when each following F0 peak is systematically reduced relative to preceding peaks (as opposed to disjuncture, signaled when a following peak is clearly not in such a systematic relationship — e.g. by being produced in an expanded pitch range). In addition, some languages encode aspects of this general tendency in the phonology, by imposing strictly phonological constraints on the F0 downtrend across coherent sequences of phrases. For example, in both standard (Tokyo) Japanese and English, a large component to the downtrend is phonologically conditioned: each occurrence of a sequence of tones that is grouped together functionally into a "pitch accent" triggers a particularly large "downstep" of following peaks relative to utterances which do not have a downstep-triggering pitch accent in the same position. Moreover, in both languages, the downstep seems to be limited in application to a phonological constituent that is intermediate in the prosodic hierarchy between the prosodic word and the intonational phrase and a "reset" of the pitch range occurs at phrasal boundaries (Pierrehumbert and Beckman, 1988). This similarity is striking when we consider the very different distribution and function of the pitch accent in the two languages. In Japanese, the presence of a pitch accent on a particular syllable is part of the dictionary specification (accented words contrast lexically with unaccented), whereas in English, pitch accents are pragmatic markers associated to the most stressed syllable in words that are particularly salient in the current discourse segment. Yet the two languages are alike in that every bitonal pitch accent triggers downstep.

However, this kind of downstep is not a language universal. In some other languages, the trigger is different. For example, in Mandarin Chinese, downstep occurs each time there is a syllable bearing one of the three lexical tones other than the high level tone (Shih, 1988). In still other languages, there is no identifiable phonologically triggered component of the downtrend. For example, in Standard Danish (Thorsen, 1980) and Chonnam Korean (Jun, 1989), something like downstep seems to be triggered by each pitch accent, but there are no miminal pairs of accent type or of accented versus unaccented words to distinguish a potential phonological trigger from the more general pragmatic considerations of signaling discourse coherence. Bruce (1982) has shown that a similar reduction of successive accent peaks occurs in southern Swedish as well, but unlike in Danish and Korean, Swedish "downstep" is further limited to the region of the phrase after the nuclear stress (see below). Thus, by careful phonological and phonetic investigation of the intonational systems of many languages, we come across a broad generalization — that many languages use a reduction of later F0 peaks relative to earlier ones within a constituent at some level of the prosodic hierarchy to signal discourse coherence — but that languages differ in whether or how phonological contrasts condition the downtrend.

PITCH ACCENTS AND STRESS

Another notable result of this decade and a half of work on modeling F0 patterns in many languages is that we can now build more useful taxonomies. In comparing Swedish and Japanese with English and French in these F0 models, we see that the old impressionistic dichotomy between "pitch accent languages" and "intonation languages" is not so compelling. Whereas the older taxonomy classified Swedish and Japanese together in opposition to English and French because both have lexically determined pitch contrasts (accent 1 words versus accent 2 words in Swedish, accented versus unaccented words in Japanese), we now might note that all four languages have pitch patterns that fit the definition of "pitch accent" — namely, a tone or closely connected sequence of tones that is associated phonologically to some designated syllable within a word. Thus, they all are "pitch accent languages" in some sense. They all also are "intonation languages" in the sense of using tone patterns and pitch range relationships to group together words into prosodic phrases of various sizes: all four languages have tonally marked intonational phrases, and French and Japanese also have a smaller tonally delimited minor phrase (the French "rhythm group" and the Japanese "accentual phrase").

Given these similarities, the dimension of lexical contrast upon which the older taxonomy hinged now seems a grab bag of fundamentally unrelated characteristics. In Swedish, the lexical contrast between accent 1 and accent 2 is primarily one of pitch accent shape whereas in Japanese the lexical contrast involves pitch accent placement alone and shape is not distinctive. Both these types of lexical contrast now seem considerably less useful in categorizing the prosodic systems of these languages than the relationship of pitch accent to sentence rhythm. In both Swedish and English, a basic aspect of the rhythm of an utterance is the pattern of alternation between strong (stressed) and weak (unstressed) syllables, with pitch accent placement functioning to mark the strongest stressed syllables. They both contrast in this to French and Japanese, where the salient rhythm is instead a grouping of syllables into tonally marked minor phrases, with no very compelling relationship between pitch accent and syllable prominence.

Categorizing the languages on these lines leads, first of all, to a better understanding about the relationship between prosodic structure and segmental contrasts. Accented syllables (or accentable syllables) in Swedish and English are very different from unaccented syllables. The consonants in accented syllables are "stronger" (e.g., voiceless stops are strongly aspirated), vowels are longer and more fully realized (i.e., closer to the periphery of the vowel triangle), and, unlike the vowels in unstressed syllables, they cannot be reduced to the point of apparent deletion (cf. Fokes and Bond, 1993). De Jong (1991) characterizes these differences between accented and unaccented syllables in English as one of local "hyperarticulation": in languages such as English, accented syllables have special status in the conflict between the needs of the speaker to minimize effort and the needs of the hearer to maximize distinctive (Lindblom, 1990). In Japanese, by contrast, accented syllables are not different in length (see, e.g., Beckman, 1986) and seem hardly different in propensity to vowel "reduction". French is somewhat more akin to English and Swedish in that the (normally final) accented syllable is longer in duration than the syllables preceding it in each "rhythm group". However, closer inspection of the kinematic patterns involved shows that this lengthening is more like "pre-boundary" lengthening than it is like the accentual effect in English (Fletcher and Vatikiotis-Bateson, 1991). As Martin (this volume) puts it, the characteristics of "stress" in French are "particularly elusive" by comparison to those in English or Swedish.

Categorizing Swedish with English and French with Japanese on these grounds also leads to predictive insights about the ways in which such discourse categories as "broad versus narrow focus" will be realized in the prosody of naturally occurring utterances. In English and Swedish, the notion "nuclear stress" seems to be a useful concept in describing what happens when narrow focus is placed on a particular word or phrase. In both languages, focus is related to the placement of a phrasal tone (the "phrase accent" proper — see Bruce, 1977; Pierrehumbert, 1980) which is associated to the word with nuclear stress. In a context that puts broad focus on a sentence (with no single word or phrase particularly more salient in the discourse), the nuclear stress will be late in the utterance. Narrow focus on a word early in the utterance can be effected by associating the phrase accent to that word, thus effectively displacing the nuclear stress to a word other than the one that would normally bear it. Material after the early focus will then be differentiated from material before it, either by deaccenting all following words in the intermediate phrase in the case of English (Pierrehumbert, 1980), or by downstepping the accents of all following words, in the case of Swedish (Bruce, 1982). In terms of the discourse structure, we can think of this pattern as a strategy of increasing the relative stress of one word by reducing (or removing) the prominence of following accents.

This strategy differs markedly from that in Japanese and other languages that are prosodically like Japanese. In these languages, focus primarily involves patterns of phrasing and there is nothing like the notions "phrase accent" or "nuclear stress". The fundamental frequency modeling work described above shows that both Japanese and French have a tonally demarcated smaller prosodic phrase within the larger intonational constituent, and narrow focus seems to be realized primarily by deleting the tonal marks at following phrase boundaries. In Japanese, the deletion of later tones effectively groups everything after the focused word into a single accentual phrase together with the focused word (Pierrehumbert and Beckman, 1988). In French, on the other hand, the deletion of post-focus tones does not apply to the rise at the end of the focused word itself unless the word is very short. Unlike in Japanese, therefore, the deletion of tones after narrow focus groups the post-focus material separately from the focused word into a kind of postfocus "tail". This is somewhat oversimplified, of course, in that standard European French today is beginning to acquire an "extra" pitch accent — *l'accent d'insistence* — that is inserted toward the beginning of words with narrow focus (Touati, 1987). Still, in both languages the general strategy is one of reducing the salience of following material relative to the word with narrow focus by erasing "normal" tonally-marked minor phrase boundaries, where English (lacking this level of phrasing) reduces the relative salience of following material by erasing "normal" tonally-marked stresses. Thus, marking the edges of minor prosodic phrases in a language like Japanese is some ways functionally equivalent to marking stress by pitch accent placement iin a language like English. The distinction becomes particularly important when we try to extend the investigation of prosody and focus from the constructed material of lab speech to other, richer rhetorical styles or to spontaneous dialogue.

BEYOND SYNTAX

Indeed, it seems fair to say in general that the extensive gathering of production data on fundamental frequency and the modeling of fundamental frequency patterns in association with phonological description of intonational categories has now led us to a point where we are beginning to glean more useful insights into the functions of prosody in natural dialogue. Much earlier work on prosody concentrated on its relationship to syntactic structure. Phonologists have long attempted to predict the stress pattern of an English utterance from its syntactic organization. In the same vein, phoneticians have long investigated the role of prosody in disambiguating syntactically different but segmentally identical strings such as *Fast man offrade bonden, och löparen hälsade kungen* (Bruce, Granström, Gustafson, and House, 1992). Since the occasions must be extremely rare when comprehension hinges crucially on deciding between such contrastive readings, this research may seem irrelevant to the technology of speech synthesis or spoken language understanding. However, it would be a mistake for those concerned with technological applications to dismiss the results of such research. Even in the absence of two or more likely syntactic parses, prosody organizes speech in a way that is apparently critical for understanding. Pitch range relationships and accentuation patterns help listeners to parse topic structure and to resolve anaphoric reference (e.g., Hirschberg and Pierrehumbert, 1986). Work such as that of Silverman (this volume) demonstrates clearly that modeling such aspects of prosody is paramount in achieving natural and easily intelligible synthetic speech. Moreover, given the otherwise contrast-obliterating effect of prosodic position on segmental realization (e.g., Pierrehumbert and Talkin, 1992), a good understanding of intonation and prosody is also directly relevant for robust recognition of segments. The last decade and a half of research on prosody has thus taken us well beyond an inordinate emphasis on syntactic contrasts.

With this coming of age of our understanding of the role of prosody in recognizing segments in connected speech and in cueing discourse structure, there is also a very encouraging merging of research traditions. On the one hand, there is the detailed sorting out of prosodic categories proper in well-controlled phonetic experiments (the many references cited above), and on the other, there is the more impressionistic descriptions of larger speech corpora, including records of spontaneous speech (e.g., Altenberg, 1987). Researchers trained in one or the other of the sets of disciplines relevant to these two different traditions are now taking better advantage of each other's research in analyzing large corpora. This trend was well illustrated at the last International Conference on Spoken Language Processing by the many papers that combined the two approaches. As it becomes more feasible to gather and store ever larger corpora, the importance of well-controlled background work in the laboratory becomes more apparent. And, as Ostendorf, Price, and Shattuck-Hugnagel (this volume) point out, a good understanding of the relevant prosodic units (such as that which has been achieved for many languages through the detailed laboratory work of the last decade and a half) is prerequisite to the use of more general stochastic models in analyzing large spoken corpora. We see recognition of this especially in the emergence of cooperative efforts to build prosodically labelled databases, such as the development of the ToBI conventions for transcribing intonation and phrasing in English (Silverman et al. 1992).

BEYOND FUNDAMENTAL FREQUENCY

One thing that has also become clear with the development of these prosodic labelling systems, however, is how much room for basic research there still is in aspects of prosody other than fundamental frequency modeling. Work on articulatory correlates of rhythmic structure, such as De Jong (1991), make clear how poorly understood are the phonetic bases of local and global variations in speech timing. An equally large problem is the dearth of basic psychoacoustic research for relevant psychological correlates other than pitch. We know enough about pitch perception now that we can intelligently compare different phonetic representations (e.g., Hermes and van Gestel, 1991). By contrast to this, our understanding of the perception of spectral dynamics is still very limited and new, so new that we have advanced little beyond the guess that durational correlates of stress in languages such as English and Swedish might be related somehow to the temporal summation of loudness (e.g. Beckman, 1986).

However, here again, I see strong grounds for optimism. We are at least seeing renewed attention to aspects of the signal other than fundamental frequency (e.g. Bartkova, Haffner, and Larreur, this volume), and as our understanding of other aspects besides the fundamental frequency patterns improves, so should our prosodic models and synthesis. Also, as our understanding of speech timing and of more subtle spectral cues to voice source patterns improves, we should begin to be able to answer currently puzzling questions concerning the role of these other phonetic dimensions of prosody in differentiating pragmatic interpretations of the same intonation pattern (e.g., Hirschberg and Ward, 1992). Let us hope together for another fifteen years of productive research on these aspects of prosody.

REFERENCES

B. Altenberg (1987), *Prosodic Patterns in Spoken English* (Lund University Press).

M. J. Anderson, J. B. Pierrehumbert, and M. Y. Liberman (1984), "Synthesis by rule of English intonation patterns", *Proc. IEEE Internat. Conf. Acoustics, Speech and Signal Processing*, pp. 2.8.2-2.8.4.

C. Avesani (1990), "A contribution to the synthesis of Italian intonation", *Proc. Internat. Conf. Spoken Language Processing*, Vol. 1, pp. 833-836.

M. E. Beckman (1986) *Stress and Non-Stress Accent* (Foris, Dordrecht).

E. Gårding (1987), "Speech act and tonal pattern in Standard Chinese: constancy and variation", *Phonetica*, Vol. 44, pp. 13-29.

G. Bruce (1977), *Swedish Word Accents in Sentence Perspective* (Gleerup, Lund).

G. Bruce (1982), "Developing the Swedish intonation model", *Working Papers, Department of Linguistics, University of Lund*, No. 22, pp. 51-116.

G. Bruce, B. Granström, K. Gustafson, and D. House (1992), "Aspects of prosodic phrasing in Swedish", *Proc. Internat. Conf. Spoken Language Processing*, Vol. 1, pp. 109-112.

J. Fletcher and E. Vatikiotis-Bateson (1991), "Articulation of prosodic contrasts in French", *Proc. 12th Internat. Cong. Phon. Sc.*, Vol. 4, pp. 18-21.

J. Fokes and Z. S. Bond (1993), "The elusive/illusive syllable", *Phonetica*, Vol. 50, pp. 102-123.

H. Fujisaki, H. Hirose, P. Halle, and H. Lei (1990), "Analysis and modeling of tonal features in polysyllabic words and sentences of Standard Chinese", *Proc.Internat. Conf. Spoken Language Processing*, Vol. 1, pp. 841-844.

H. Fujisaki and H. Sudo (1971), "A generative model for the prosody of connected speech in Japanese", *Ann. Rep.Engineering Research Institute, University of Tokyo*, Vol. 30, pp. 75-80.

C. Gussenhoven and T. Rietveld (1992), "A target-interpolation model for the intonation of Dutch", *Proc. Internat. Conf. Spoken Language Processing*, Vol. 2, pp. 1235-1238.

J. 't Hart and R. Collier (1973), "Intonation by rule: a perceptual quest", *J. Phonetics*, Vol. 1, pp. 309-327.

D. J. Hermes and J. C. van Gestel (1991), "The frequency scale of speech intonation", *J. Acoustical Soc. America*, Vol. 90, pp. 97-102.

J. Hirschberg and J. Pierrehumbert (1986), "The intonational structuring of discourse", *Proc. 24th Assoc. Computational Linguistics*, pp. 136-144.

J. Hirschberg and G. Ward (1992), "The influence of pitch range, duration, amplitude, and spectral features on the interpretation of the rise-fall-rise contour in English", *J. Phonetics*, Vol. 20, 241-251.

K. De Jong (1991), *The Oral Articulation of English Stress Accent*, doctoral dissertation, Ohio State University.

S. Jun (1989), "The accentual pattern and prosody of Chonnam dialect of Korean", in *Harvard Studies in Korean Linguistics*, ed. by S. Kuno et al. (Harvard University Press, Cambridge), pp. 89-100.

K. Kohler (1991), "Prosody in speech synthesis: the interplay between basic research and TTS application", *J. Phonetics*, Vol. 19, pp. 121-138.

D. R. Ladd (1987), "A phonological model of intonation for use in speech synthesis by rule", *Proc. EuroSpeech*, pp. 21-24.

B. Lindblom (1990), "Explaining phonetic variation: a sketch of the H&H thoery", in *Speech Production and Speech Modeling*, ed. by H. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 403-440.

J. Pierrehumbert (1980), *The Phonetics and Phonology of English Intonation*, doctoral dissertation, MIT.

J. Pierrehumbert and M. E. Beckman (1988), *Japanese Tone Structure* (MIT Press, Cambridge).

J. Pierrehumbert and D. Talkin (1992), "Lenition of /h/ and glottal stop", in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, ed. by G. J. Docherty and D. R. Ladd (Cambridge University Press, Cambridge), pp. 90-117.

Y. Sagisaka (1990) "On the prediction of global F0 shape for Japanese text-to-speech", *Proc. Internat. Conf. Acoustics Speech and Signal Processing*, pp. 235-328.

C. Shih (1988), "Tone and intonation in Mandarin", *Working Papers, Cornell Phonetics Laboratory*, No. 3, pp. 83-109.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg (1992), "TOBI: a standard for labeling English prosody", *Proc. Internal. Conf. Spoken Language Processing*, Vol. 2, pp. 867-870.

N. Thorsen (1980), "Neutral stress, emphatic stress, and sentence intonation", *Ann. Rep., Institute of Phonetics, University of Copenhagen*, No. 14, pp. 121-205.

P. Touati (1987), *Structures prosodiques du suédois et du français* (Lund U. Press).

N. Willems, R. Collier, and J. 't Hart (1988), "A synthesis scheme for British English intonation", *J. Acoustical Soc. America*, Vol. 84, pp. 1250-1261.

Speaker specificity in prosodic parameters[1]

J. Kraayeveld, A.C.M. Rietveld and V.J. van Heuven[2]
Dept. of Language and Speech, Phonetics Section, Nijmegen University
P.O. Box 9103, NL-6500 HD, Nijmegen, The Netherlands

ABSTRACT

Ten time-integrated prosodic parameters were used to assign read-out and spontaneous speech fragments to the speakers that had produced them. Fifty speakers of standard Dutch participated in the experiment. They were stratified by gender and age. It was found, that the parameters are independent of each other in terms of speaker specificity. In combination, they could correctly assign 73 % of 500 fragments to the 50 speakers. As expected, differences in mean F_0 were important for speaker identification, but an analysis without mean F_0 still resulted in 56 % correct classification. Identification improves when spontaneous and read-out speech are analysed separately. Thus, speech type is an important factor to control for in speaker identification.

INTRODUCTION

Research on speaker specific variation is often directed at applications of speaker recognition (e.g. in forensic research) and speaker verification (e.g. for electronic access systems). However, it is also a necessary step in the process of separating the contrastive and linguistically meaningful properties from speaker-dependent and meaningless variation. In this contribution results are reported of a research project that aims at mapping out individual variation along *prosodic* parameters in Dutch.

First, in carefully selected sentences prosodic parameters were measured that are closely related to the linguistic and prosodic structure of the utterance, e.g. F_0 at specific turning points in an utterance, such as the top of a pointed hat-pattern. These measures we call *point* measures.

Not all prosodic parameters require strictly controlled utterances. By averaging over larger stretches of time and over many different segments, some can be made more or less text-independent. Examples of these *time-integrated* measures are mean F_0, certain temporal measures, and perturbation measures (measures for the instability of a speaker's frequency and amplitude).

In our research project the usefulness of both the point measures and the time-integrated measures were studied. Preliminary results on speaker dependent characteristics of such point measurements were reported earlier in Kraayeveld *et al.* (1991). In the present contribution, however, we examine the possibility of separating individuals using non-linguistic, time-integrated prosodic measures only.

In ten 15 s.-fragments of both read-out and spontaneous speech ten different time-

[1] This research was supported by the Linguistic Research Foundation (projectnr. 300-173-006), which is funded by the Netherlands organization for scientific research, NWO

[2] Dept. of Linguistics, Phonetics Laboratory, Leiden University, The Netherlands

integrated prosodic measures were used. Some of these measures have been found earlier to be very powerful tools in speaker identification. Especially mean F_0 and the standard deviation of F_0 appear to be very useful (e.g. Sambur, 1975).

From segmental studies we know that there are large differences between read-out (or premeditated) and spontaneous speech (eg. Van Bergem *et al.*, 1989). We anticipate similar, large effects of speech style on the use of prosodic parameters. Still we would like to know to what extent a single individual displays the same prosodic behaviour in both speech styles.

One, almost trivial, speaker difference that is reflected in prosody, specifically in mean F_0, is gender: female voices have about double the F_0 of males. Therefore our research was set up to examine speaker dependence of prosodic parameters both across genders and for male and female groups separately. We want to find out if individuals can be successfully discriminated when the sex of the speaker is partialled out, and on the basis of which parameters.

In the experimental design the factors gender, age[3] and speech type will be controlled.

METHOD

Speakers: A sample of 50 speakers of standard Dutch was selected that was stratified for gender and age: 25 male and 25 female speakers, belonging to the age groups 18-25, 26-35, 36-45, 46-55 and 56-65. Thus, each age groups consisted of five males and five females each.

Elicitation: In the first task speakers were interviewed on everyday issues. After editing out irrelevant material such as interviewer intrusions, an otherwise contiguous stretch of speech of 75 s. was selected from the end of each recording, and divided into five stretches of 15 s. each.

The second task was to read out a newspaper-like story. This story consisted of five paragraphs. Of each paragraph, the first 15 seconds were included. Thus, the read-out material roughly contains the same lexical material.

Analysis parameters: the following time-integrated parameters were determined:

1 F_0MEAN Mean F_0 in Hz.
2 F_0COV Coefficient of variation of F_0, the standard deviation of F_0 divided by its mean)
3 PPQ Pitch Perturbation Quotient, as defined by Davis (1976)
4 PZR Pitch period Zero-crossing Rate, the percentage of triplets of adjacent periods where duration does not increase or decrease monotonically
5 AMPCOV Coefficient of variation of the absolute peak-amplitude per period
6 APQ Amplitude Perturbation Quotient, analogous to PPQ
7 AZR Amplitude Zero-crossing Rate, analogous to PZR
8 VOICE Percentage of time the signal is considered Voiced
9 PAUSE Percentage of time the amplitude is below threshold
10 RATE Speaking Rate, number of syllables per fragment.

[3] There are no clear indications that age is related to speaking behaviour, at least not in the 18-65 yr. age bracket. Although our sample was stratified by age (see method) we shall not study the influence of age on the question of speaker separation.

RESULTS AND DISCUSSION

To determine how well these ten prosodic measures can be applied to the task of speaker identification, five discriminant analyses were carried out. Discriminant analysis is primarily a data reduction method, in which parameters are collapsed onto orthogonal discriminant functions such that the functions maximally separate the groups. Discriminant functions are linear combinations of variables in which the weights reflect the importance of the associated variables.

In the first analysis all 50 speakers functioned as 'groups', with 10 data points (five 15s.-fragments of two speech types) per group. Next, the analysis was repeated with parts of the data set: the read-out and the spontaneous fragments (50 groups, 5 data points each), and the sets of female and male speakers (25 groups per analysis, 10 data points).

All discriminant analyses resulted in 10 significant discriminant functions (the maximum possible number given 10 variables). To show the influence of reducing the number of dimensions from 10 down to one, the analyses were repeated, limiting the number of discriminant functions. Figure 1 displays the percentage speakers that are correctly classified as a function of the number of discriminant functions. Although the percentage of classified fragments does not improve substantially when more than seven functions are included, the further discriminant analyses will be based on all ten functions. This enables us to compare these results with analyses from which F_0 was excluded. Below, in Table 1, for all five analyses the variables are specified that have the highest correlation with the first three (Varimax-rotated) functions. From this table it becomes clear, that mean F_0 is the most important prosodic variable for speaker characterisation. Apparently this is not only the result of the obvious fact that men and women differ considerably on this parameter, since in separate analyses of men and women mean F_0 was the most important discriminating variable as well.

It also becomes clear, that it is not *only* F_0 that contributes to the classification of the individuals. If only F_0 is allowed as a variable in the analysis, the amount of correct classification is small, ranging from 9 % for males, to a value of 32 % for read-out speech (male and female speakers). Apparently, for F_0 there is an interaction between the factors speaker and speech style. The behaviour of speakers in the two speech styles is different, and can therefore be better classified if only one speech style is taken into consideration. Another way to study the role of F_0 is to exclude it from the analyses. If the maximally possible number of functions (nine) are allowed, the percentages of correct classification of the fragments for read out and spontaneous

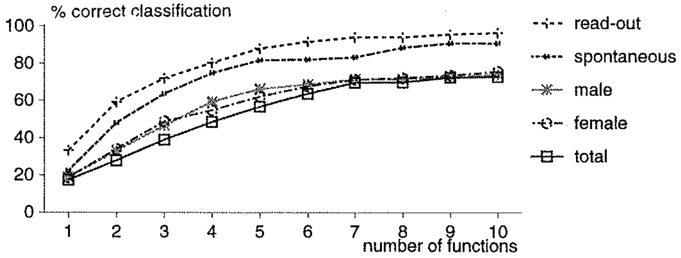


Figure 1: Percentage of fragments that is correctly classified in the *n*-dimensional space spanned by the discriminant functions

Table 1: Correlations between the three most important rotated discriminant functions and prosodic variables that exceed .50. Below the percentage correct classification for each analysis, the percentage correct classification is shown if F_0 is kept out of the analysis (resulting in an analysis with nine discriminant functions), and if only F_0 is allowed (resulting in one discriminant function):

	total	read-out	spontaneous	females	males
f. 1	F_0MEAN .87	F_0MEAN .97	F_0MEAN .95	F_0MEAN .87	F_0MEAN .84
f. 2	PZR .85	PZR .95	PAUSE .91	PZR .88	F_0COV .90
f. 3	F_0COV .92	F_0COV .95	PZR .92	F_0COV .93	PPQ .92
cor. class.	73 %	96 %	91 %	76 %	74 %
cor. class. without F0	56 %	93 %	82 %	58 %	62 %
cor. class., only F0	9 %	32 %	21 %	11 %	9 %

speech are somewhat lower than in the analysis with all variables. However, when we exclude mean F_0 from the separate analyses of male and female speakers, the percentage correct classification decreases considerably.

In summary, the differences between the speech types (i.e. read-out and spontaneous speech) appear to blur the speaker differences to some extent. If we compare the analysis of the total material with analyses of only parts of the data, we find that restricting the analysis to only one speech type improves the percentage correct classification more than analysing only one of the genders. Actually, the differences between the two speech types are so large, that in a discriminant analysis with the speech types as groups, 93.8 % of the fragments is correctly assigned to the speech types. The two parameters that correlate most with the only possible discriminant function are VOICE (.55) and AMPCOV (.52). Mean F_0 does not play any role in this function (-.10).

An analysis with the two genders as groups yields about the same percentage of correct classification as in the analysis of the speech styles: 98.4 % of the fragments was assigned to the correct gender.

REFERENCES

Bergem, D.R. van & Koopmans-van Beinum, F.J. (1989), "Vowel reduction in natural speech", *Proceedings of EUROSPEECH '89*, Paris, 285-288.

Davis, S.B. (1976), "Computer evaluation of laryngeal pathology based on inverse filtering of speech", *Speech Commun. Labs, Monogr.*, 13.

Kraayeveld, J., Rietveld, A.C.M. & Heuven, V.J. van (1991), "Speaker characterization in Dutch using prosodic parameters", *Proceedings of EUROSPEECH '91*, Genova, 427-430.

Sambur, M. (1975), "Selection of acoustic features for speaker identification", *IEEE Trans. ASSP*, ASSP-23, 176-182.

Prosodic regularities in the surface structure of French questions

Albert Di Cristo and Daniel Hirst
Institut de Phonétique, Laboratoire "Parole et Langage" URA 261
Université de Provence, Aix-en-Provence

ABSTRACT

This paper is concerned with the description of the tonal patterning of questions in French. The results are based both on the analysis of interviews recorded from broadcast programes and on an experimental study of simulated dialogues which constitute an iintermediate level between read-aloud decontextualised utterances and truly spontaneous speech. Special attention is paid to the recurrent downstepping pitch pattern which appears to constitute an important feature common to most of the basic types of questions analysed here.

INTRODUCTION

This study deals with the tonal organisation of questions in French. It is part of a larger project aiming at describing the prosodic pattern of General French within the framework of the parametric approach used in Aix. This approach has lead in the recent past to the development of two related models: a phonological model and a model of production.

The core system of the phonological model is formed by two hierarchical constituents: Intonation Units and Tonal Units to which language dependent tonal templates are assigned (Hirst, 1988, ; Hirst & Di Cristo, 1984).

The model of production is designed to model an Fo curve by selecting two parameters: local Fo targets (roughly but not entirely equivalent to the "turning points" of the Lund Model) and an appropriate interpolation function for connecting these targets (Hirst, 1981).

Thanks to the implementation of the model on a mini-computer (Hirst & Espesser, in press), it is now possible, among others, to factor out automatically an Fo curve into two components: a macroprosodic component and a microprosodic component (Di Cristo & Hirst, 1986), the former consisting of a smooth underlying curve. This curve is viewed in this approach as a phonetic level of interpretation of pitch variations which constitutes an interface between acoustic raw data and the higher levels of the phonological representation (Hirst & Di Cristo, 1992; Hirst & al., this volume).

MATERIAL AND METHODOLOGICAL PROCEDURE

The results which are presented in the next section of this paper are based on general observations arising from a preliminary study of interviews recorded from broadcast programs completed and illustrated by analysis of simulated short dialogues which were recorded by four subjects (two male and two female).

Turn-units in the dialogue consisted of short questions and answers, the former being of four types: a) yes-no Qs (*La voisine de la cousine de l'amie de Sylvie?*), b) elliptical Qs (*Et la copine de l'ami du mari de Corine?*), c) alternative Qs (*La voisine de la cousine de l'amie de Sylvie ou la copine de l'ami du mari de Corine?*), and d) left-dislocated Qs (*La voisine de la cousine de l'amie de Sylvie , elle vit aussi en Suisse?*). These turn-units have been chosen to be formed either by a single IU or by two IUs, each corresponding to an information-unit. As can be seen from the preceding examples and from the following dialogue, these IUs are nominal utterances including three to five syllable stress-groups.

- Tu penses qu'elle viendra demain?
- Qui ça?
- La voisine de la cousine de l'amie de Sylvie.
- La voisine de la cousine de l'amie de Sylvie ou la copine de l'ami du mari de Corine?
- Non, la voisine de la cousine de l'amie de Sylvie.
- La voisine de la cousine de l'amie de Sylvie, elle habite en Italie?
- Non, elle vit en Suisse.
- Et la copine de l'ami du mari de Corine, elle vit ausi en Suisse?
- La copine de l'ami du mari de Corine? Non, elle réside à Paris, la copine de l'ami du mari de Corine.

To describe the tonal patterns of questions embedded in the dialogue, a clear distinction is made between global and local characteristics affecting, respectively, the whole IU or simply a part of it as, for example, the pitch contour associated with the nucleus. Recurrent characteristics, i.e. iterative tonal patterns which are associated with smaller sequences such as stress-groups, are also taken into account since they also make up utterance intonation.

RESULTS AND DISCUSSION

We begin with some preliminary remarks about stress in French. Besides nuclear stress and emphatic stresses which are beyond the scope of this study, French possesses both primary and secondary stresses (Pasdeloup, 1990; Di Cristo, forthcoming).

Secondary stress, when present, generally affects the initial syllable of lexical units in, whereas primary stress occurs regularly on the last full syllable of a group of words, giving rise to a right-headed structure (Di Cristo & Hirst, in press). Since the main cue for rhythmic stresses is pitch prominence (Rossi & Di Cristo, 1980), we used the term Tonal Unit, (TU) to refer to what we have called the stress-group here (Hirst & Di Cristo, 1984) . Only TUs ending with a primary stress will be considered in this paper.

Let us consider first as a reference for a comparison the basic IU of a simple declarative utterance corresponding to a single information-unit. The tonal configuration of such an IU (Figure 1a) can be described as globally rising-falling. The initial rise which constitutes the onset pitch accent of the utterance is aligned with the first accented syllable and the pattern finishes on an extreme low pitch associated with the nucleus. A downtrend can also be observed joining the maxima of the utterance

As regards the recurrent pitch pattern, it can be accounted for by a general rule which specifies a rising pitch movement (from low to high) at the end of each TU, except the last one, so that the template of this iterative TU can be defined as [L H].

IUs in question structures can be classified into two types depending on the direction-rising or falling- of the final pitch contour. For yes-no Qs, which belong to the "final-rising" type and which are characterised by a globally rising-falling-rising pattern, it is possible to formulate the following rules.

If the IU contains one or two TUs, its pitch pattern, with the exception of the final rise, is similar to that of the declarative utterance. On the other hand, if the IU is formed by more than two TUs (Figure 1b), the recurrent pitch pattern of TUs between the first (which also exhibits an onset pitch accent similar to declaratives) and the last (which contains the final rise) is different from that of a corresponding declarative, consisting of a sequence of lowered pitches or downstepped tones , so that the template of such an UT can be defined as [D].

It is noteworthy that this description also applies to IUs of the "final-rising" type which are embedded in more complex question structures and which correspond, for example, to the first term of an alternative Q (such as: "*La voisine de la cousine de l'amie de Sylvie* ou la copine de l'ami du mari de Corine?") or to the rhematic part of a Left-dislocated Q (such as: "La voisine de la cousine de l'amie de Sylvie, *elle habite en Italie*?")

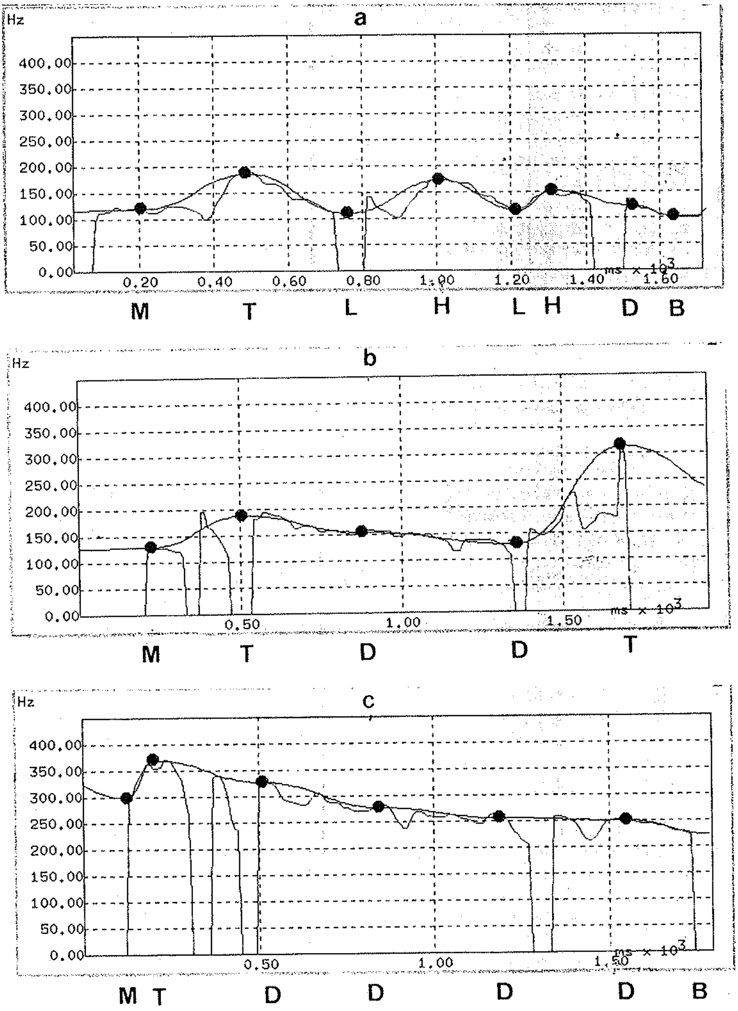


*Figure 1. Observed and modelled Fo curves with the INTSINT coding (Hirst & Di Cristo, fothcoming) of the target points: **M**id, **T**op, **B**ottom, **L**ower, **H**iger, **D**ownstepped and **S**ame. a) declarative utterance: "La voisine de la cousine de l'amie de Sylvie" (Sylvie's friend's cousin's neighbour), b) yes-no Q.: "La copine de l'ami du mari de Corine?" (Corine's husband's friend's girlfriend ?) and c) elliptical Q.: " Et la copine de l'ami du mari de Corine?" (What about Corine's husband's friend's girlfriend ?).*

This downstepping pattern in questions with a final rising contour has already been described in a previous study (Hirst & Di Cristo, 1984). Something similar seems to be implied by the decription proposed by Martin (1981) which accounts for a difference in the prenuclear pitch pattern of French utterances as conditioned by the form of the final falling or rising contour, appealing to a general principle of "contour opposition"

If this were the case, we should expect questions with a final falling contour to show recurrent rising prenuclear pitch movements, just like declaratives. This is not what we obseve however for IUs in question structures which belong to the "final-falling" type", such as elliptical Qs (e.g.: "Et la copine de l'ami du mari de Corine?"), the second term of alternative Qs (e.g.: "La voisine de la cousine de l'amie de Sylvie *ou la copine de l'ami du mari de corine*?") or the thematic part of left-dislocated Qs (e.g.: " *La voisine de la cousine de l'amie de Sylvie*, elle habite en Italie?") and which exhibit a global pitch pattern which is either rising-falling, or simply falling, depending on the place of the onset pitch accent. In fact, it is particularly interesting to observe (figure 1c) that these IUs, which all end either with a conclusive fall (mainly when they conclude a turn-unit) or with no final contour at all (when they are embedded in an utterance) are also characterised by a recurrent downstepped pitch pattern. The regularities which have been shown in the present paper seem consequently suggest that the downstepping pattern in fact constitutes a tonal feature of questions in French and not simply an anticipatory feature of a following rising contour.

REFERENCES

A. Di Cristo (forthcoming), "Intonation in French", in *Intonation Systems: A Survey of Twenty Languages*, ed. by DJ. Hirst and A. Di Cristo (Cambridge University Press, Cambridge).

A. Di Cristo and D.J. Hirst (1986), "Modelling French Micromelody: Analysis and Synthesis", *Phonetica*, Vol. 43, pp. 11-30.

A. Di Cristo and D.J. Hirst (in press), "Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français", *Trav. Inst. Phon. Aix*, Vol. 15.

D.J. Hirst (1981), "Un modèle de production de l'intonation", *Trav. Inst. Phon. Aix*, Vol. 7, pp. 297-311.

D.J. Hirst (1988), "Tonal units as phonological constituents: the evidence from French and English intonation", in *Autosegmental Studies in Pitch Accent*, ed by H. Van der Hulst and N. Smith (Foris, Dordrecht), pp. 151-165.

D.J. Hirst and A. Di Cristo (1984), "French intonation: a parametric approach", *Die Neueren Sprachen*, Vol. 83, pp. 554-569.

D.J. Hirst and A. Di Cristo (1992). "Niveaux de représentation, niveaux d'analyse et codage prosodique", *Communication au Séminaire Prosodie de La Beaume-lès -Aix.*

D.J.Hirst and A. Di Cristo (in press) "A survey of intonation systems." in Hirst & Di Cristo (eds) *Intonation Systems : a Survey of Twenty Languages*. (Cambridge University Press, Cambridge).

D.J. Hirst and R. Espesser (in press), "Automatic modelling of fundamental frequency", *Trav. Inst. Phon. Aix*, 15.

D.J. Hirst and A. Di Cristo (forthcoming), Intonation Systems: A Survey of Twenty Languages, Cambridge University Press.

Ph. Martin (1981), "Pour une théorie de l'inyonation", in *L'Intonation: de l'Acoustique à la Sémantique* edited by M. Rossi, A. Di Cristo, D.J. Hirst, Ph. Martin and Y. Nishinuma, Klincksieck, Paris, pp. 234-271.

V. Pasdeloup (1990), *Modèle de Règles Rythmiques du Français Appliqué à la Synthèse de la Parole*, Thèse de Doctorat, Un. de Provence.

M. Rossi et A. Di Cristo (1980), Un modèle de détection automatique des frontières intonatives, *Actes XIèmes Journées d'Etudes sur la Parole* (Strasbourg), pp. 141-164.

Combining Statistical and Linguistic Methods for Modeling Prosody[1]

M. Ostendorf* P. J. Price† S. Shattuck-Hufnagel‡
*ECS Engineering, Boston University, Boston, MA 02215 USA
†SRI International, Menlo Park, CA 94025 USA
‡Speech Communication, RLE, MIT, Cambridge, MA 02139 USA

ABSTRACT

We describe a general approach to computational modeling of prosody that combines statistical models with linguistic theory. Statistical models provide a mechanism for representing variability, for automatically training parameters, and for analyzing large corpora. Linguistic theory provides model structure and guides feature extraction. We illustrate the approach with examples from our own work and from the work of others.

INTRODUCTION

Recently, prosody research has seen increased use of corpus-based analyses and automatic learning techniques. In particular, statistical techniques have played an important role in advancing our understanding of prosody, as well as our ability to model prosody computationally for automatic speech processing. Still, these techniques, which are driven by the need to develop robust and portable modeling techniques, are currently underutilized because of cultural differences among linguists, computer scientists and engineers. Here we try to help bridge the multi-disciplinary gap by outlining some statistical methods and providing examples of their use when driven by linguistic theory.

Statistical techniques have long been used in speech research in the analysis of significance and in socio-linguistics for sampling of large corpora, but some researchers have been wary of more extensive use because the development of linguistic insights has often been ignored in statistical modeling. However, the use of prosody in automatic speech processing cannot ignore advances in speech recognition using statistical techniques, which have the advantages that they model variability (e.g. randomness due to incomplete knowledge of sources of variability) and that automatic training methods exist for porting or adapting the models to different speaking styles or domains. Further, in the scientific aim to understand human speech communication, statistical techniques are important in that they enable the use of large corpora, which is important because human intuitions can under-represent the full range of prosodic structure that can be uncovered through the analysis of large corpora. In addition, the use of large corpora can provide data representative of normal communication, while reducing the need to control context.

Of course, statistical techniques do not provide all the answers. Stochastic models can have burdensome data requirements unless constrained by linguistic structure, and analysis of large corpora is more informative if driven by questions raised in linguistic

[1] Research was jointly funded by NSF and DARPA, NSF grants IRI-8805680 and IRI-8905249.

theory. Thus, combining linguistic and statistical methods can provide insights and results beyond the reach of either approach alone. In the next sections, we further elaborate on this theme by describing general methods for combining statistics and linguistics, illustrating these with specific examples, and discussing future directions.

COMBINING LINGUISTICS AND STATISTICS

In the context of prosody modeling, we use "linguistics" to include both phonological models of abstract units (i.e. prosodic phrase constituents, prominence, and intonation markers) and phonetic hypotheses about their observed acoustic correlates (i.e. f0, duration and energy). By "statistics", we mean both statistical data analysis and modeling techniques, recognizing two roles for statistical techniques in prosody research: (1) to generate and test hypotheses about factors that influence the phonetics and phonology of prosodic patterns, as well as to assess our level of understanding of sources of variability, and (2) to model prosodic patterns for automatic speech processing.

In **data analysis**, linguistics can provide hypotheses to test statistically, and/or we can use distributional analyses to generate hypotheses to test with traditional perception and production experiments. For variables with different interrelated conditioning factors, multi-factor analysis techniques may provide more powerful tools than significance tests. Automatic clustering and techniques for estimating model order may help answer (or pose) questions related to the number of abstract units needed to represent different prosodic phenomena. Finally, multi-modal and hierarchical models may expose distribution differences not evident from mean computations.

In **computational modeling**, linguistic theory provides the model structure, reducing dimensionality to a practical size, and drives the signal processing or feature extraction. Examples are in the next section; here, we describe general statistical techniques appropriate for prosody modeling. Classification and regression trees (or, decision trees) are particularly useful for handling a combination of categorical and continuous variables, all of which are dependent, a common situation in prosody. Decision trees take a vector of features as input, and predict or estimate a variable after a series of binary questions about the features, modeling their dependence without making explicit prior assumptions. In prosody modeling, decision tree variables draw on features traditionally used in synthesis and/or recognition rules, but the sequence and number of rules (tree structure) and their threshold values are learned automatically. While decision trees are powerful, they can only predict a single variable or vector; other techniques are needed for handling sequences of variables (random processes). The most common techniques used are the Markov source model, and the hidden Markov model (HMM), which combines a Markov chain with a random observation model. Both types of process models can incorporate decision trees to handle non-homogeneous features.

PROSODY MODELING EXAMPLES

Data Analysis. In developing a model with many variables, tabulating results for all combinations of factors is impractical. Some alternative analysis methods are illustrated in [11] for duration modeling. Automatic clustering can also be a tool for data analysis, as in our experiments investigating questions about the categorical vs. gradient nature of acoustic differences among prominences. Decision tree design, a form of clustering, can yield insights into the relative importance of different variables, though in our work they have mainly served to confirm linguistic intuitions (e.g. hesitations are most likely to occur at a function-word/content-word boundary).

Our recent work on early accent placement within lexical items illustrates how corpus-based analysis can suggest new hypotheses. In a distributional analysis, we noticed that adjacent-stress words and alternating-stress words behaved differently with respect to within-word prominence, e.g. double accents were common only for alternating stress words. Although scattered clues might have suggested that adjacent-stress words form a special class, e.g. stress markings for these words are not consistent across dictionaries, the analysis of a large corpus made the systematic difference clearly visible.

Computational Models. As several research sites (particularly AT&T Bell Labs and ATR) have shown, corpus-based models can be powerful tools for *text-to-speech synthesis.* In prosody prediction, there are several models based on classification and regression trees, some using the predicted values associated with terminal nodes in the tree and others the probability distributions. Classification trees have proved useful for predicting abstract units [9, 8], e.g., prosodic phrase structure, pitch accent location and tone labels. Regression trees, used for estimating continuous variables, have been mainly applied to duration modeling, either for directly predicting segment duration [5, 6] or for deriving the terms in a parametric model [2].

Two aspects of prosody modeled in *speech recognition* systems are stress and duration. Several sites have used separate models for lexically stressed and unstressed vowels, though results have been mixed. Efforts in duration modeling, motivated by the fact that recognition errors often correspond to unlikely segment durations, use linguistic knowledge to define possible conditioning contexts for statistical models, e.g. [2, 5].

In *speech understanding,* where meaning of an utterance is extracted, prosody can provide information for determining the correct syntactic and semantic structure. For example, prosody has been used to reduce parsing ambiguity by automatically recognizing prosodic breaks using a decision tree and then using these breaks in a parser [4], or alternatively by scoring sentence parse hypotheses according to the likelihood of observed prosodic patterns [7]. Prosody can also provide information for speech understanding in semantic processing, since automatically detected phrasal prominence (e.g. [10, 1]) can provide clues to semantic focus. Finally, prosody can aid in detecting and correcting disfluencies such as word fragments, as in our work with decision trees.

A limitation of statistical modeling is the need for labeled training data. In speech recognition the labels are words, which can be hand-transcribed at a much lower cost than prosodic labels. Thus, it is critical to develop *automatic labeling* algorithms to assist this process. Our most successful efforts in automatic labeling have involved the use of decision trees, e.g. [10], which have outperformed HMMs.

CONCLUSIONS

Having argued the merits of combining statistics and linguistics in prosody research, we conclude with two areas ripe for further research. First, bidirectional models are important to study, since synthesis and recognition share many of the same problems. Since the speech synthesis and recognition communities have not intersected much, they have not benefited from combining their separate perspectives. There exist a few examples using the same model for both recognition and synthesis problems, e.g. [7] for prosodic constituents, but more work is needed. Second, although it is possible to use results of statistical analysis to help formulate linguistic hypotheses, there has actually been very little such work. Two possible tools are decision trees (e.g., for ranking prediction variables), and probabilistic information measures (e.g. entropy and mutual information) that may be used to assess our ability to account for observed acoustic or phonological variability. Perhaps the biggest barrier to be overcome is the amount of knowledge required from the different disciplines, but this simply argues for multi-disciplinary collaborations. To quote Ladefoged [3]: "We all have to rely on other people to fill in the gaps – the vast holes – in our knowledge. Any scientist today is part of a team that cannot hope to build a bridge to the future without a lot of help."

REFERENCES

[1] F. Chen & M. Withgott (1992), "The Use of Emphasis to Automatically Summarize a Spoken Discourse," *Proc. Inter. Conf. on Acoust., Speech and Signal Proc.,* I229-I232.

[2] C. Fong (1993), *Duration models for speech synthesis and recognition,* M.S. Thesis, Boston University.

[3] P. Ladefoged (1992), "Knowing Enough to Analyze Spoken Languages," *Proc. Inter. Conf. on Spoken Language Proc.,* pp. 1-4.

[4] M. Ostendorf, P. Price, J. Bear, & C. Wightman (1990), "The Use of Relative Duration in Syntactic Disambiguation," *Proc. DARPA Workshop on Speech and Natural Language,* pp. 26-31.

[5] J. Pitrelli (1990), *Hierarchical Modeling of Phoneme Duration: Application to Speech Recognition,* MIT PhD thesis.

[6] M. D. Riley (1992), "Tree-based Modeling of Segmental Durations," *Talking Machines: Theories, Models and Designs,* ed. by G. Bailly, C. Benoit and T. R. Sawalis, (Elsevier Science Publishers), pp. 265-274.

[7] N. Veilleux & M. Ostendorf (1993), "Probabilistic Parse Scoring with Prosodic Information," *Proc. Inter. Conf. on Acoust., Speech and Signal Proc.,* pp. II51-54.

[8] M. Ostendorf & N. Veilleux (1993), "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location," *J. Comp. Linguistics,* to appear.

[9] M. Wang & J. Hirschberg (1992), "Automatic classification of intonational phrase boundaries," *Computer Speech and Language,* Vol. 6, No. 2, pp. 175-196.

[10] C. Wightman & M. Ostendorf (1992), "Automatic Recognition of Intonation Features," *Proc. Inter. Conf. on Acoust., Speech and Signal Proc.,* pp. 221-224.

[11] J. van Santen and J. Olive (1990), "The analysis of contextual effects on segmental duration," *Computer Speech and Language,* Vol. 4, pp. 359-390.

Models and Reality: From Intonation Theory to Intonation Curves

Philippe Martin
Experimental Phonetics Laboratory
University of Toronto
21 Classic Ave. Toronto, Ont. Canada M5S 2X6

ABSTRACT

There are a few theories of intonation to account for various phonological facts in different languages such as English, French, Swedish, etc. Depending on their respective framework (glossematics, generative or linear phonology, etc.), these theories provide some understanding of observable phenomena of acoustic or auditory nature, such as the evolution of fundamental frequency in the sentence, the perception of syllable pitch and intensity, and so forth. Nevertheless, the apparent large number of discrepancies that appear between the theoretically predicted facts and the physical reality often baffle researchers dealing with the acoustical nature of the speech signal. In their eyes, any theoretical insight on prosody may seem useless. The object of this paper is to discuss some of the reasons for these discrepancies through examples, and to provide a better understanding of what an intonation theory delivers.

A THEORY: WHY?

As for other linguistic facts, prosodic descriptions come in three different flavors: empirical, formal and axiomatic.

1) Empirical: The data collected through experimental phonetic analysis are processed and organized to find interesting correlation. Ex.: Experimental data show a positive correlation between the amplitude of the fundamental frequency variation and the number of syllables of read sentences.

Graphs are commonly used to represent the results of this kind of analysis;

2) Formal: A common sense principle governs the presentation and the interpretation of the experimental data. Ex.: For possibly unknown reasons, speakers uttering a long sentence have a tendency to differentiate their word accents with larger variations of pitch. Therefore larger variations of fundamental frequency are to be expected in long sentences, as (clearly) shown in the data.

Models are often used to represent this analysis, together with appropriate mathematical formula. In a sense, formalism appears to be a variant of empirical analysis.

3) Axiomatic: A (common sense) principle controls the description process and is used consistently to organize and interpret the data. Ex.: A prosodic structure does exist in the sentence and prosodic contours indicate its characteristics. A direct consequence of their existence is found in the amplitude of variation of the fundamental frequency in sentences with a large number of syllables.

Somehow, only the axiomatic approach delivers a theory, as scientific knowledge produced by empiricism and formalism is in reality governed by the selection of experimental data. In other words, the starting principle (the axiom) determines a set of discovery principles that allow the classification and the interpretation of the data. Here a set of statements (theorems) summarizes the results of the analysis.

A THEORY: WHICH?

Let us consider an axiomatic theory of intonation applied to French. The axiom postulates that a prosodic structure organizes prosodic words in the sentence, as the syntactic structure organizes syntactic units (words). If it exists, this prosodic organization is indicated by some prosodic markers belonging to the prosodic words. To track those markers down, we will consider a sentence containing only the smallest unit, i.e., one syllable. This syllable corresponds to the smallest prosodic word, which must contain (at least) one prosodic marker: the word accent. Prosodic markers are therefore equated to word accents, revealed by pitch contours. The number of prosodic words in a sentence is thus equal to the number of word stresses (not emphatic or any other kind of stress), which must be in contrast to each other to properly encode the postulated prosodic structure. By describing these contrasts with an appropriate set of binary features (for example), the pitch contours lose their abstractness by a choice of characteristics such as the slope of fundamental frequency (rising vs. falling), duration, intensity, etc.

In order to force the prosodic markers to reveal their inner mechanism, a technique commonly used is to consider sentences with no other hierarchical organization than prosodic. Common examples are structured enumerations like telephone numbers, multiplication tables, etc. The prosodic grammar found at that stage can then be used to analyze more complex cases where the sentence is organized by other mechanisms (e.g., syntax) (See more details in Martin, 1987).

WHY IT WORKS

Such a hypothetico-deductive approach leads to the definition and the description of phonological prosodic contours and their relationships in the prosodic structure. Some of the immediate consequences are:

1) Fo variations: If the number of syllables in the sentence increases, so does the number of prosodic words. In the case where the prosodic structure acquires more levels, so does the number of contrasts between pitch accents. Contrasts using Fo will then have a tendency to use more of this material to encode the contrasts, and manifest larger Fo changes on the stressed syllables;

2) Penultimate prosodic contour: The prosodic grammar of French uses contrasts in Fo slope to mark a dependence to the right, which ensures that the penultimate prosodic word should always bear a rising contour (in the case of declarative sentences with a falling final contour). Being obligatory, this feature is thus neutralized, as shown in most experimental data;

3) First level contours: These contours often end a noun phrase and are rising only if preceded by at least another prosodic word. Indeed, only in this case a contrast falling/rising is at work to mark this part of the structure. If no prosodic contour precedes a noun phrase pitch accent, the rising Fo feature need not be present;

4) Prosodic features: Many acoustical, perceptual, or articulatory parameters are plausible candidates for manifesting the prosodic contrasts between contours predicted by grammar (A pure axiomatic attitude such as glossematics would even manipulate only abstract prosodic features). Although experimental data suggests that Fo slope, Fo swing and syllable duration are the most commonly used in read speech, other styles reveal the use of different sets of features (whispered speech constitutes an interesting case since it substitutes syllable pause duration for Fo contrasts).

WHY IT DOES NOT WORK

Experimental observations made on read sentences or on spontaneous speech are not always easy to relate or even to contradict theoretical predictions. Indeed, many processes other than just intonation grammar encoding the prosodic structure are involved in the production of prosodic parameters in a sentence. Some of the discrepancies commonly discussed are:

1) Prosody of competence vs. theory of performance: Syntactic descriptions normally use the competence of the speaker, whereas, curiously, most prosodic studies analyze only its performance. Except in the case of very specific studies, no grammatical description of a language is build from spontaneous speech data. Nevertheless, most researchers in prosody seem to rely only on those data. It is therefore no surprise that an intonation theory may appear inadequate when confronted with spontaneous speech data. Most syntactic descriptions would appear inadequate as well if tested in the same conditions;

2) Establishing the prosodic structure: Another source of problems stems from the fact that, except in rare cases, we do not have access to the prosodic structure other than the marking mechanism performed by the prosodic contours themselves. We can only postulate a certain homomorphy with the syntactic hierarchy, tempered by specific prosodic constraints (Such as the avoidance of stress clashes and the planarity of the prosodic structure);

3) Stressability, stress and destressing: A third reason for discrepancies pertains to the elusive nature of the stressed characteristics of a syllable in French. Whereas the stressability of a syllable can be clearly assessed, its effectively stressed or destressed nature is much more difficult to establish. Therefore, expected theoretical predictions on some contour features, say its rising Fo variation, rely on the effectively stressed nature of the corresponding syllable. This difficulty is somewhat alleviated by the (potentially dangerous) use of external insight provided by auditory of articulatory phonetics to help in establishing the stressed character of a syllable (see for example Di Cristo, 1980, on this subject).

4) Other prosodic features: Numerous other prosodic features are not taken into account by the theory: declination line, first syllable pitch movements, emphatic stress, etc. (See for instance Mertens, 1989, on this.) Again, this results from the choice of prosodic features that do participate to the marking of the structure.

APPLICATIONS

These considerations suggest that speech synthesis and language teaching are two typical applications that would benefit from a theoretical axiomatic approach. Text-to-speech synthesis obviously proceeds from the competence level, whereas speech recognition, dealing with spontaneous speech, can only make little or no use of a grammar of intonation.

CONCLUSION

By definition, an axiomatic approach sticks to its principles in its discovery process. This hopefully constitutes a guarantee that the results of the experimental analysis do reflect the starting hypothesis. All descriptions of prosodic contours, contrasts in Fo slope, syllable duration, etc., produced by the discovery process correlate with the existence of a prosodic structure. Obviously many other prosodic facts that are revealed experimentally and that do not participate to the assigned function are absent from the phonological description. In assessing the importance of those facts when considering specific applications leads to either change the theorems or to postulate other axioms.

REFERENCES

A. Di Cristo (1980), "Variabilité acoustique et intégration perceptive des cibles prosodiques", *Actes des XIèmes Journées d'Études sur la Parole* (Strasbourg), pp. 91-110.

Ph. Martin (1987), "Prosodic and Rhythmic Structures in French", *Linguistics,* Vol. 25, pp. 925-949.

P. Mertens (1989), *L'intonation du français, de la description linguistique à la reconnaissance automatique*, Doctorale dissertatie, Leuven Univ. , Dept. Linguistieck.

Intensity Prediction for Speech Synthesis in French

K. Bartkova - P. Haffner - D. Larreur
France Télécom - CNET Lannion A - TSS/RCP
Route de Trégastel - 22300 Lannion - France

ABSTRACT

The goal of the present study is to predict sound intensity for speech synthesis in French. In order to set up a model for intensity prediction, we first studied intensity variation in natural speech. The data base used was hand segmented and phonetically and syntactically labelled. The results of this part of the study were introduced into a rule-based model whose parameters were subsequently optimised using stochastic gradient procedure. Next, a neural network based model was developed and trained using part of the labelled data.

INTRODUCTION

Sound intensity is considered to be the least important of the three prosodic parameters for the perception of synthetic speech quality. Most of the time, researchers have settled for a decrease in sound intensity at the end of the sentence (Calliope 1989) while maintaining fairly constant intensity elsewhere. As far as diphone speech synthesis is concerned, the intensity of the basic units stored in the dictionary has been kept unchanged during speech generation. Some studies (Granström 1991), however, have highlighted the role of intensity in speech synthesis for modelling different styles of speech.

INTENSITY VARIATION IN NATURAL SPEECH

In order to set up a model for sound intensity prediction in French, we first investigated intensity variation in a corpus of about 1 hour of natural speech recorded by a male speaker. This corpus contained isolated sentences as well as 9 short texts. Its structures covered the majority of linguistic and prosodic possibilities in French. The intensity of the whole corpus was normalised (sentence by sentence) to its highest value. Intensity values were measured in dB in approximately the middle of the each phoneme thus avoiding accidental humps (if any) known to sometimes occur in unvoiced fricatives.

Even if the corpus used didn't contain sentences dedicated to the intrinsic intensity study, it afterwards became possible to bring out this value. Indeed, as the corpus was large enough, constraints could be imposed when sounds where chosen for intrinsic intensity calculation. Duration, pitch value and the left and right context of each sound were thus controlled. This way, for example, only vowels with a duration between 100 and 150 ms, pitch value between 70-120 Hz and left and right context belonging to the same consonantal classes were used for calculating intrinsic intensity. Since large variations can exist from one group to another[1], duration threshold was determined for each consonantal group. A distinction was made between the two allophones of /R/ : one was voiced (surrounded by voiced contexts) and the other voiceless (surrounded by unvoiced contexts). For stop consonants, voiceless stop intensity was measured during their burst whereas voiced stop intensity was measured during stop closure.

[1] The mean duration of unvoiced fricatives is much longer than that of semi-vowels.

Our findings on the intrinsic values of vowel and consonant intensity (Fig.1 & 2) are very similar to those observed in other studies (Di Cristo 1978).

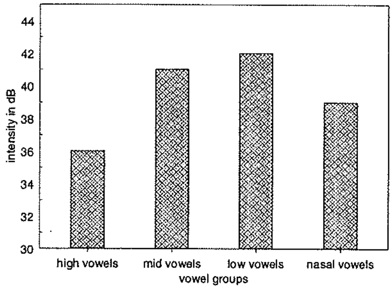


Figure 1. *Intrinsic intensity for vowels.*

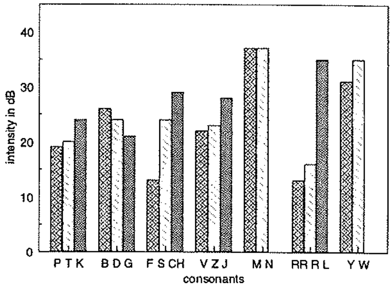


Figure 2. *Intrinsic intensity for consonants*[2].

The first part of this study revealed that duration, pitch, left and right contexts and sound position (with respect to the word and pause occurences) are among the most important parameters for detemining sound intensity. When the influence of one of these parameters on sound intensity was studied (duration, pitch, contexte...), the others were only allowed to vary between controlled thresholds. In this way, we hoped to clarify the relationship between sound intensity and other phonetic or syntactic events.

The number of sentences containing word focus was quite small in our corpus and therefore it would be dangerous to draw general conclusions. Nevertheless, it can be noted that increased sound intensity was observed in focus syllables.

RULE-BASED MODEL

The findings from the first part of our study were introduced into the rule-based model under different rules. The prediction formula took into account intrinsic sound intensity along with some additional coefficients expressing the influence of the previously discussed phonetic and syntactic parameters. The following formula was used for intensity prediction :

$$\mathit{Sound\_Intensity} = \mathit{intrinsic\_int.} + \mathit{coef}_{(\mathit{left\_context})} + \mathit{coef}_{(\mathit{right\_context})} \\ + \mathit{coef}_{(\mathit{duration})} + \mathit{coef}_{(\mathit{pitch})} + \mathit{coef}_{(\mathit{position})}.$$

An intrinsic intensity value was determined for each phoneme. The parameter coefficients relevant to the model (context, duration...) could be positive or negative. They were determined for a homogenous group of sound (ex: unvoiced plosives, voiced plosives, nasals, semi-vowels...).

Mean error prediction was calculated as the mean value between the predicted value and the target value (measured value). Figures 3 and 4 represent the histograms of prediction dispersion for vowels and consonants. The ***intensity prediction error rate*** for the ***whole corpus*** was ***4.3 dB for vowels and 4.9 dB for consonants (leading to an average error rate of 4.6 dB per sound).***

Coefficient and intrinsic intensity values were subsequently optimised using stochastic gradient procedure. The corpus was split into two parts. One part was used for "tuning" model values; the other for testing new values (data was split as in the NN-model). The prediction error rate was reduced by about 1 dB per sound (i.e. 20% reduction).

[2] **"J"** was used for the French fricative /ʒ/ as in "jour"; **"RR"** was used for the unvoiced /R/ allophone; **"Y"** was used for the semi-vowel /j/ as in "bien".

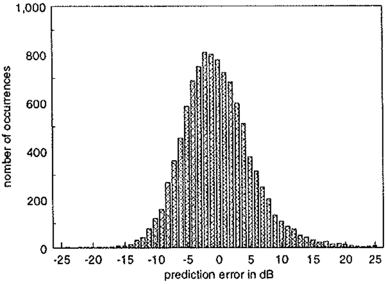


Figure 3. *Vowel prediction dispersion.*

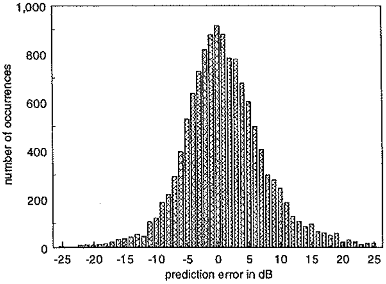


Figure 4. *Consonant prediction dispersion*

Table 1. *Intensity prediction error in dB for vowels and consonants provided by the rule-based model before and after optimisation by a stochastic gradient technique.*

	vowels		consonants		all sounds	
	training	**testing**	**training**	**testing**	**training**	**testing**
rule	**4.6**	**4.6**	**4.9**	**5.0**	**4.8**	**4.8**
stoch. gr.	**3.3**	**3.4**	**4.0**	**4.2**	**3.7**	**3.8**

NEURAL NETWORK MODEL

The Neural Network (NN) used in this study was developed with the Aspirin/Migraine software. A three layer NN was created in order to approximate the function:

Intensity_Code = f(Phoneme_Code, Left_cont_Code, Right_cont_Code, Position_Code, Pitch_Code, Duration_Code).

Phonetic knowledge was used to set up numeric codes for the relevant parameters, which were afterwards normalised to facilitate the training procedure convergence. The network architecture consisted of an input, a hidden and an output layer. The input layer gathered the cells coding for the input parameters cited above. The hidden layer was completely connected with the input layer. The output layer contained 36 cells which were arranged in the form of a thermometer. The maximum intensity was 36 dB with 18 above and below the mean value. Because of the considerable difference in vowel and consonant intensity, two separate networks were trained to predict them.

The NN was trained with the Gradient Back-Propagation procedure. Between 5 to 40 hidden cells were tested in the training procedure.

The ***mean error rates*** for ***vowel and consonant intensity*** prediction were ***3.6 dB and 4.5 dB respectively (on average, about 4 dB per phoneme)***. The number of hidden cells had little effect on results. The prediction error rate for consonants was the mean value provided by 8 networks trained for different groups of consonants.

PRELIMINARY AUDITORY TESTS USING NATURAL SPEECH

Phoneme intensity of 40 isolated sentences (belonging to our testing data) was modified according to the model parameters[3]. 20 listeners were asked to judge the quality of these sentences and in so doing compare altered and natural sentence intensity. The paired sentences (natural A and model-provided B) were presented in random order. For each

[3] All the other prosodic parameters (sound duration and pitch) were kept unchanged.

sentence pair listeners noted whether or not they had a preference for one of them. The following results were obtained :

Table 2. *Auditory test results in % for rule-based model and NN model.*

	no preference	pref. for nat. int.	pref. for mod. int.
rule-based mod.	**46.3**	**28.3**	**24.2**
NN model	**31.9**	**40.41**	**27.6**

Thus in **70.5%** of the time, the rule-based model provided intensity and in **59.5%** of the time, the NN model provided intensity was considered by listeners as fully natural.

We are well aware that natural signal can more readily undergo modifications without quality corruption than synthetic signal. Nevertheless, results obtained by the test are encouraging and it can be hoped that no major problems will be encountered when the intensity model is implemented to speech synthesis.

IMPLEMENTATION TO SPEECH SYNTHESIS

In this part of sudy we had to cope with the problem of sound level differences between the diphones and sounds of our corpus. This is why the absolute intensity value provided by the model was converted into a multiplicative coefficient expressing the slope movement of sentence intensity variation. As great care was taken to record diphones with a cosntant intensity; thus it was possible to consider diphone intensity as an intrinsic intensity value. The diphone intensity is then subject to modification depending on event at the sentence and sound level.

CONCLUSION

The aim of this study was to set up a model for sound intensity prediction for speech synthesis in French. We hope that controlling all three of the prosodic parameters will increase the resemblance of synthetic to natural speech. Accurate energy prediction at the sentence level will improve the perception of speech fluency by eliminating unpleasant, too salient sounds which occur in unexpected positions. Last but not least, controlling sound intensity will help to introduce the perception of sound depth into speech synthesis. It is true that sentence focus can be satisfactorily modelled by appropriate pitch movement and sound duration. But adding intensity will improve naturalness and the impression of text comprehension by the reading system.

ACKNOWLEDGEMENTS

The authors would like to thank F. Agashira and B. Cherbonnel, for implementing the signal processing part of intensity control, and for their valuable comments and great patience.

REFERENCES

Calliope (1989) *"La parole et son traitement automatique"*, Masson.

A. Di Cristo (1978) *"De la microprosodie à l'intonosyntaxe"*, Thèse d'Etat, Université de Provence, Aix-en-Provence.

B. Granström, & L. Nord (1991) "Neglected dimensions in speech synthesis", *Proc. of the ESCA workshop*, pp. 27-1 - 27-5; Barcelona, Spain.

INDEX OF AUTHORS