LUND UNIVERSITY
DEPARTMENT OF LINGUISTICS

General Linguistics
Phonetics

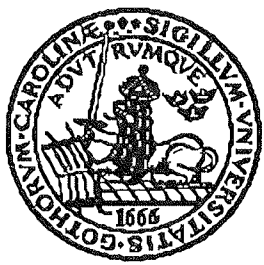
WORKING PAPERS

34. 1988

Papers from the
Second Swedish Phonetics Conference
held at Lund, May 5-6 1988

LUND UNIVERSITY
DEPARTMENT OF LINGUISTICS

General Linguistics
Phonetics



WORKING PAPERS

34. 1988

Papers from the
Second Swedish Phonetics Conference
held at Lund, May 5-6 1988

Wood, Sidney (editor).
Papers from the Second Swedish Phonetics Conference Held at Lund, 5-6 May 1988.

Working Papers
Department of Linguistics and Phonetics,
University of Lund,
Helgonabacken 12,
223 62 LUND, Sweden.

Earn/Bitnet:LINGLUND@SELDC52

Printed: Lund, 1988, viii+130 pp., ISSN-0280-526x.

A collection of 33 reports on current research on speech production, analysis and synthesis presented at the phonetics conference held at Lund in May 1988.

PREFACE

This number of Working Papers contains some 30 reports on current speech research that were presented at the 2nd Swedish Phonetics Conference held at the Department of Phonetics of the University of Lund, in May 1988. We were especially pleased to welcome phoneticians from the University of Copenhagen. The papers bear witness to the wide range of topics currently being worked on in Sweden and Denmark. One session, chaired by Jan-Olof Svantesson and introduced by Claes-Christian Elert, was reserved for a general discussion of the IPA alphabet for phonetic transcription. We, the organizers, heartily thank all participants for contributing to the success of this meeting. The next conference in the series is scheduled to be held at the Department of Speech Transmission and Music Acoustics at the Royal Institute of Technology, Stockholm. Publication of this volume was made possible by a generous grant from the Swedish Council for Research in the Humanities and Social Sciences.

Sidney Wood

PROGRAMME

SPEECH AFTER GLOSSECTOMY

Ann-Marie Almé & Olle Engstrand
Department of Linguistics
University of Stockholm

Introduction

Glossectomy is the surgical removal of all or part of the tongue, usually performed to treat carcinoma of the tongue. If larger parts of the tongue are missing, oral vegetative functions are severely impaired. A variety of symptoms arise which may interfere with speech production, intelligibility of speech, and swallowing.

Very few studies dealing with speech after glossectomy have been published (see however LaRiviere et al, 1975; Massengill et al, 1970; Morrish, 1984). We are not aware of any pre- and post treatment studies.

In a recently started research program "Speech after glossectomy", we intend to examine the phonetic characteristics of speech following total or partial resection of the tongue. Special attention will be paid to compensatory articulation in relation to the type and extent of tongue resection. This paper presents acoustic and perceptual data from a pilot study of two subjects.

Data Collection

Subjects were one normal speaker (OE) and one glossectomized speaker (PAT), both male and with the same dialect. PAT underwent radical glossectomy and neck dissection 6 years prior to this investigation.

Speech samples used in this pilot study are a word list and three short text passages. The word list is made up of 51 words with the structure /CV:l/. (C = all morpheme-initial phonemes; V = /i:/, /a:/ and /u:/).

Data Analysis

Wide band spectrograms were made of each CVC-word and the lexically stressed vowels /i:/, /a:/ and /u:/ in the text material. Based on this, the mean and standard deviation were calculated for the first two formants.

Perceptual Study. 14 students with normal hearing listened to the randomized CVC- syllables, their task being: a) to identify the first consonant and b) to identify the vowel. They heard the tape twice; half of the group was asked to identify the consonants first, the other half was asked to identify the vowels first.

Results and Discussion

Acoustical Analysis

Mean values for the first two formants of the vowels /i:/, /a:/ and /u:/ are plotted in figure 1. Looking at the F_1 versus F_2 plot, one can see that PAT's vowel structure is shifted to a more "neutral" position compared to the normal speaker's. Values for F_1 seem to be more stable than values for F_2.

For both speakers vowels in running speech are reduced compared to vowels in words in isolation.

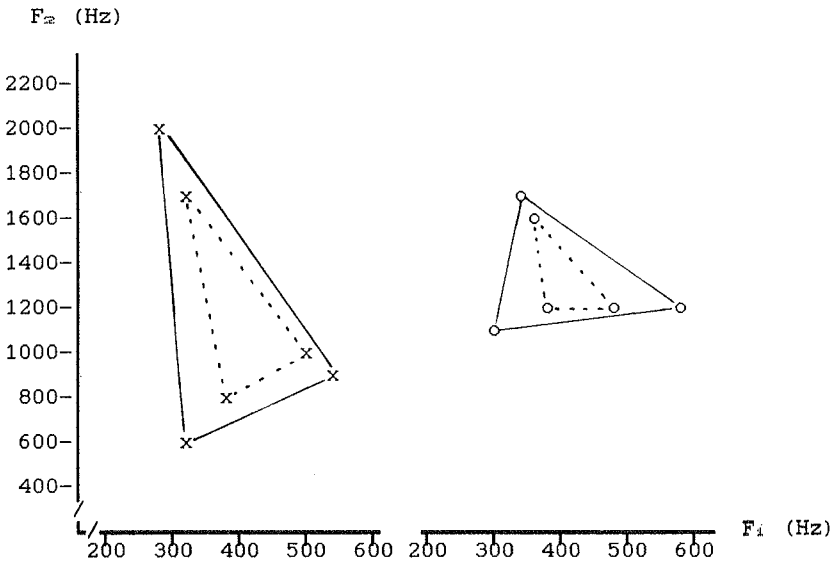


Fig. 1. F_1 versus F_2 plot for the vowels /i:/, /a:/ and /u:/ produced by a glossectomized speaker PAT (o) and a normal speaker OE (x). Mean values from word list (solid line) and running speech (dotted line).

The glossectomized subject seems to produce most consonants with the lips. Acoustic measurements reveal considerably less differentiation between consonants than the normal speaker. Formant transitions are rather flat. A representative example is given for /ba:/, /da:/, and /ga:/ in figure 2.

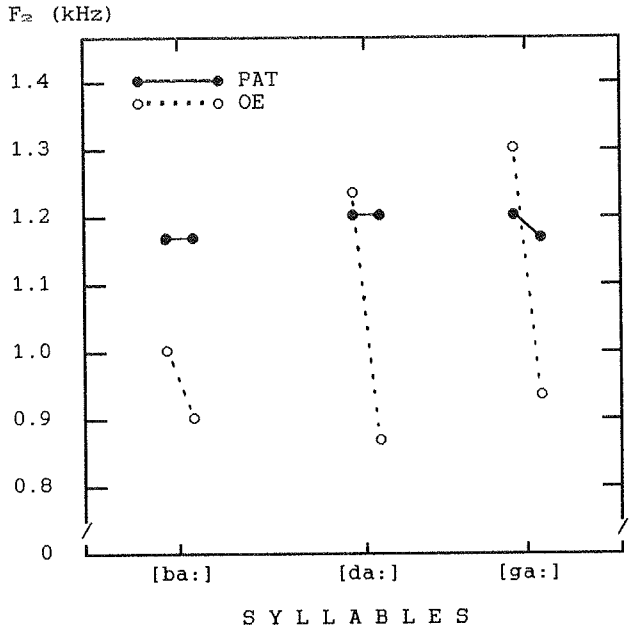


Fig. 2. Formant frequencies (F_2) for [ba: da: ga:], measured at the initial locus and target of the vowel. Normal speaker (OE) dotted line and glossectomized speaker (PAT) solid line.

Perceptual Analysis

Results from perception of consonants produced by PAT are summarized in a confusion matrix (Table 1). The figures repeat percent of listener responses (14 listeners). As expected, the labial sounds were well identified. Non-labial sounds were often perceived as labials, but to a lesser extent than expected. Apparently PAT compensates for the missing tongue articulation. The acoustic correlates of this effect will be examined in future work.

The vowels were identified to almost 100 %. This was on the other hand an easy task, since the listener panel was forced to choose between only three vowels. If the task had been to identify more vowels, the responses may have been different.

Table 1. Confusion matrix for consonants produced by the glossectomized speaker (PAT).

INTENDED CONSONANTS \ PERCEIVED CONSONANTS	p	t	k	b	d	g	f	v	s	ç	ɧ	h	j	l	r	m	n
p	93	5	2														
t	43	40	2				15										
k	50	26	7				15	2									
b				98	2												
d				83	15			2									
g				79	21												
f							98				2						
v					2			71					7	5	15		
s							67		5	2	26						
ç							93			7							
ɧ							86			2	10	2					
h												100					
j								12					17	28	43		
l					2			44					10	10	34		
r							2	14					12	27	45		
m																74	26
n																66	34

References

LaRiviere, C., Seilo, M.T., & Dimmick, K.C. (1975) Report on the speech intelligibility of a glossectomee: perceptual and acoustical observations. Folia Phoniat. 27: 201-214.

Massengill, R., Maxwell, S. & Pickrell, K. (1970) An analysis of articulation following partial and total glossectomy. J. Speech Hear. Disorders 35: 170-173.

Morrish, L. (1984) Plosive articulations of the glossectomee and some acoustic correlates. Working Papers in Linguistics and Phonetics, University of Leeds.

AUTOMATIC RECOGNITION OF FOCUS ACCENT IN GERMAN

Robert Bannert
Department of Linguistics and Phonetics
Lund University

Speech recognition systems of the future, to a growing extent, have to be able to process information of various kinds. In my paper I would like to present a new project[1] where an attempt is made to integrate the interplay of semantic and tonal information in a dialogue system. Here I will confine myself to the tonal component.

When a speaker and a hearer communicate by means of a dialogue, it is done in a given world and under certain given linguistic constraints. Both partners of the dialogue react according to certain rules of the game. Therefore it is assumed that, among other things, the semantic focus in a response can be predicted as a consequence of semantic and logic rules of the dialogue. Semantic focus is mainly signalled phonetically by a tonal peak, in German associated with the pitch accent of a word (or syllable).

From a semantic point of view, a sentence may have one focus or several foci and the domain of focus may vary, it may be wide or narrow. The interplay between the semantic aspect of focus and its tonal manifestion in German are not quite understood yet. However, it seems to be totally clear that narrow focus is to be found not only in normal, "neutral" intonation, - the one that is used when plain facts are communicated -, but also in contrast and emphasis. Each sentence contains one or more pitch accents that reflect semantic focus. These accents mark those words of the sentence that are most important semantically and I shall call them focus accents.

In our dialogue system to be modelled, the recognition of focus accent is one component out of three. It is the link between spectral speech recognition and the semantic rules of the dialogue game. The Fo-values of the sentences of the dialogue are processed using an algorithm in order to find the focus accent or accents. The sentences used in the dialogue show a prosodic structure

[1]The project is called "Modellbildung für die Auswertung der Fokusintonation im gesprochenen Dialog" and is carried out at the Research Group for Artificial Intelligence and Speech Signal Processing at the Department for Advanced Information and Communication Technologies, Fraunhofer Institute for Industrial Engineering (IAO) in Stuttgart.

corresponding to a prosodic utterance which may consist of one or several prosodic phrases. To start with, the semantic component will use simple statement sentences and their negations like "Johannes liebt Susanne (nicht)" (John loves Susanne/John does not love Susanne). The sentences of the dialogue are pronounced by a speaker, recorded, sampled and analysed for Fo.

The algorithm for recognizing focus accent consists of two steps. First all those changes of Fo are to be found that correspond to pitch accents. Fo-differences spanning a certain period of time are picked after having taken into consideration the total intonation of the whole prosodic utterance. In the second step, all the pitch accents that were picked are evaluated in order to find the focus accent or accents. At this level, the pitch accents are compared with each other using criterea in the total Fo-course of the prosodic utterance that lie in the time and Fo-dimension as well. To start with, the system will be constrained to speakers whose basic characteristics of their intonation are known. Later on these basic tonal facts will be established automatically by speaking some appropriate model sentences.

The projection of focus accent onto the time axis, i.e. its temporal position in the utterance, will be marked in an adequate way. This information about the temporal projection of focus accent or accents is then passed on to the dialogue system via the spectral speech recognition component which is not part and parcel of the project but is available in our department. The information about which word is in semantic focus acts as feed back and verification for the further semantic development of the dialogue.

References

Bannert, R. 1083. Some phonetic characteristics of a model for German prosody. Lund Univerity, Department of Linguistics and Phonetics, Working Papers 25, 1-34.

Bannert, R. 1985a. Towards a model for German prosody. Folia Linguistics XIX, 321-341.

Bannert, R. 1985b. Fokus, Kontrast und Phrasenintonation im Deutschen. Zeitschrift für Dialektologie und Linguistik 52, 289-305.

Lea, W.A., Medress, M. F. and Skinner T. E. 1975. A prosodically guided speech understanding strategy. IEEE Transactions on Acoustics, Speech, and Signal Processing. Vol. ASSP-23, 30-38.

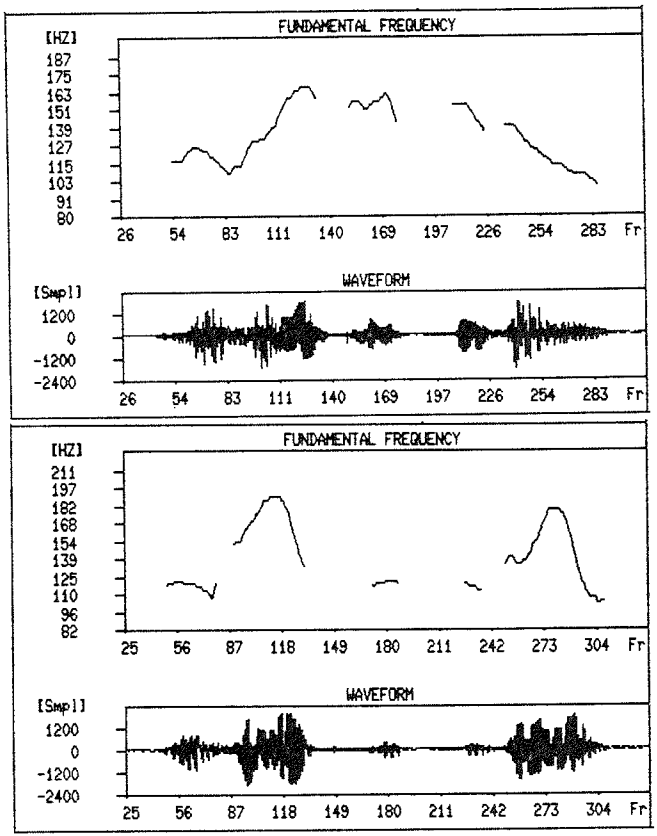


Fig. 1: Fo-curves of the utterance "Johannes liebt Susanne " containing two equal pitch accents. Hat pattern (Brücke) above, peak accents below.

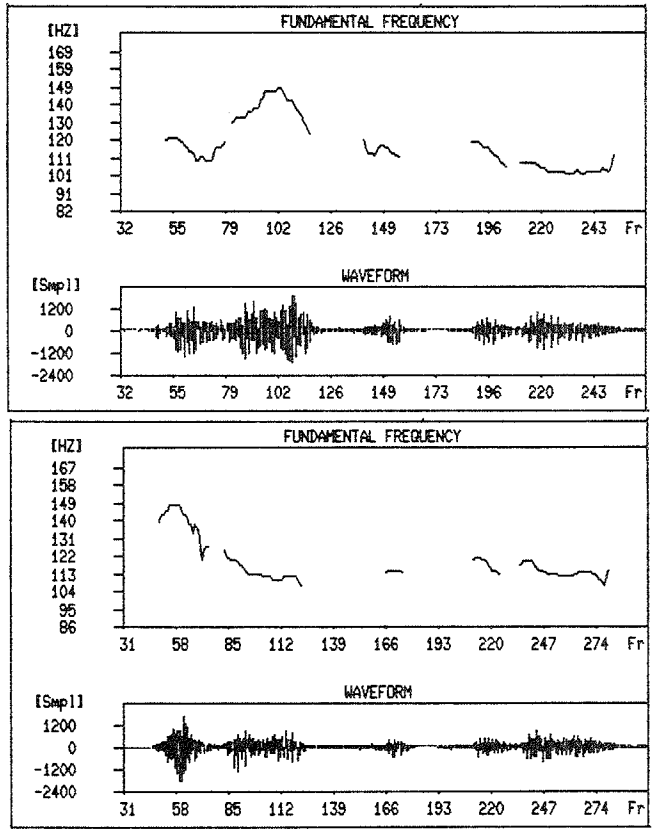


Fig. 2: Fo-curves of the utterance "Johannes liebt Susanne" above containing one accent (on Johannes). Peak accent above, low accent below.

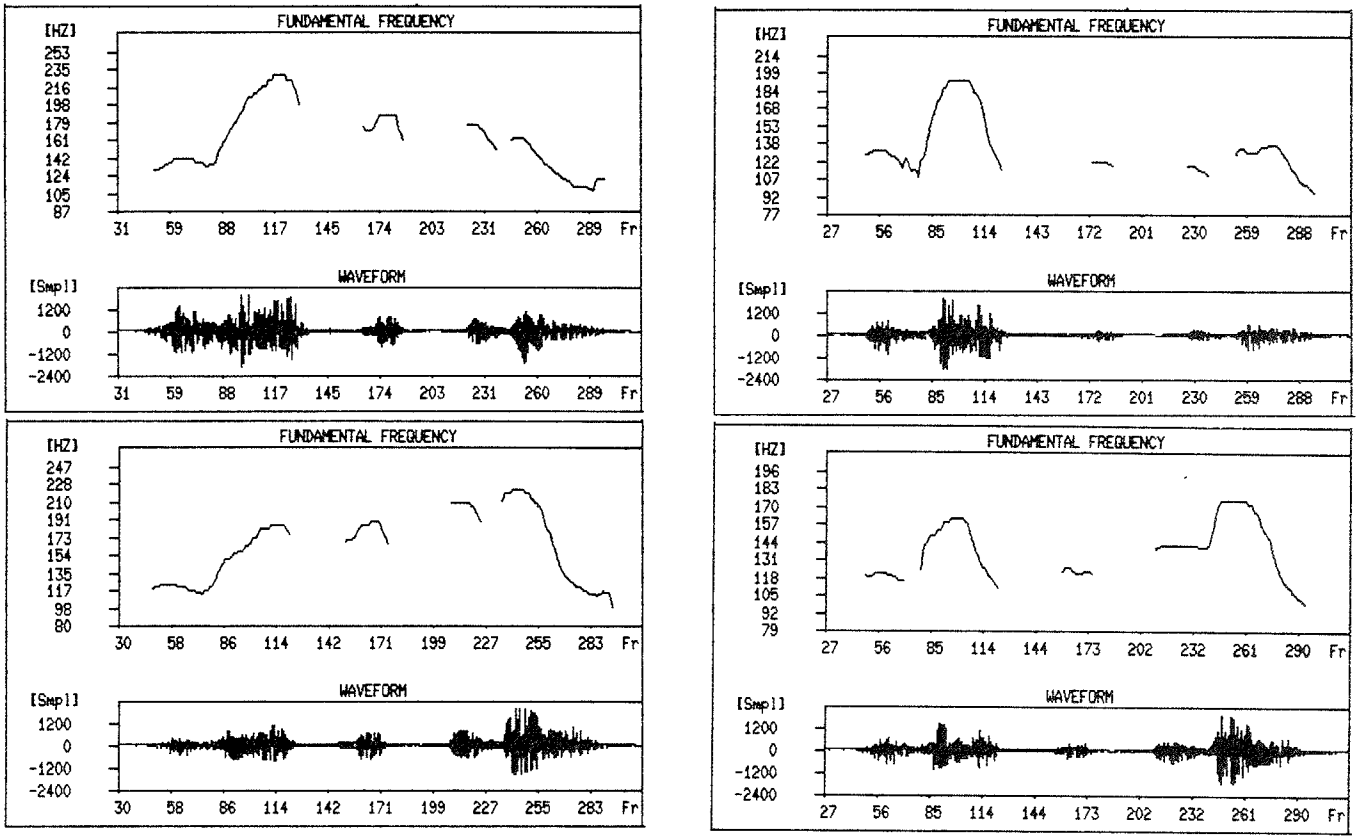


Fig. 3: Fo-curves of the utterance "Johannes liebt Susanne" with two unequal pitch accents involving contrast of emphasis on the strongest one (above on Johannes, beolw on Susanne). Left: hat pattern (Brücke), right: peaks.

SYNTHETIC PHONEME PROTOTYPES IN SPEECH RECOGNITION

Mats Blomberg
Department of Speech Communication and Music Acoustics
Royal Institute of Technology
Stockholm

ABSTRACT

A phonetic recognition system based on synthetic phoneme prototypes is described. The phoneme templates are specified in terms of synthesis parameters. They are transformed to the spectral domain in the form of a 16 channel filter bank section, the same way as the incoming speech is analysed. The vocabulary and grammar is described in a finite state network where each node contains a phoneme and its context-independent frequency spectrum. The recognition process is to find the phoneme sequence in the network that gives the lowest spectral distance to the input utterance. Preliminary results are given for recognition of connected Swedish digits. Future improvements include context-dependent phoneme spectra and speaker voice source adaptation. Another application for the program is automatic time alignment of phonetic transcription to speech data bases.

INTRODUCTION

Speech recognition of large vocabularies and speaker independent recognition are normally implemented using large amounts of training data. There is a wish to decrease the time spent during adaptation of a system to each speaker for speaker adaptive systems. There is also a need to decrease the very large number of speakers required to achieve high accuracy speaker independent recognition. Since it is unreasonable to train a system by repeating each word of a vocabulary consisting of several tens of thousands of words, training is often done on smaller units, like syllables, phonemes or allophones. Still, there is considerable amount of training needed to describe every unit in all its possible contexts. In this paper, the approach is to represent acoustic properties of a phoneme inventory by means of explicit formulation of existing phonetic-acoustic knowledge. In this fashion the need for training can be considerably reduced if not eliminated.

The phonetic-acoustic knowledge can be formulated in different ways. In the bottom-up fashion, rules are written to classify a part of the speech signal if certain conditions are present. Expert systems for speech recognition are normally designed in this way. A problem with this method is that many rules are needed to cover all possible variation of the realisation of a phoneme due to phonetic context, reduction, etc.

Another way to express the knowledge is in the form of a speech production system. This is a top-down approach. A hypothesis is generated at a high level of a recognition system. Rules are then used to generate a synthetic version of the part of the utterance. The hypothesis is verified or rejected at the acoustic level. Each individual rule can describe a local effect. The results of different rules are combined to a single output. If the rules are independent, the number of rules can be lower than in the bottom-up case. The top-down approach is chosen in this report.

A question is if the current knowledge is enough to yield sufficiently high recognition accuracy. The quality of speech generated by synthesis-by-rule systems has reached a point where it is well understood even by naive listeners.

There is still a lot of improvements to be made though, before synthetic speech comes close to the quality of natural speech. It is therefore not expected that the recognition results of recognition-by-synthesis systems will be better than those which use natural speech for reference generation.

RECOGNITION SYSTEM OVERVIEW

In previous reports we have used a synthesis-by-rule system to build word templates for recognition by a dynamic-time-warping based isolated-speech recognition system (2). In this paper, we generate a library of synthetic phoneme spectra instead of whole word templates. The recognition algorithm is also changed. The vocabulary and syntax are compiled into a finite-state network with each node represented by a phoneme and its frequency spectrum. The recognition process uses a dynamic programming technique to find the phoneme sequence in the network that gives the lowest spectral distance to the input utterance. One difference to the work in (2) is that optional pronunciation of each word is allowed, which should improve recognition accuracy. Another difference is that, in this report, the phonemes are treated mainly as stationary segments, disregarding coarticulation effects. This should lower the accuracy. In future work, we will integrate the two features and thus hope to improve the recognition rate.

Since also the time position is computed of each phoneme in the selected phoneme string, it is possible to use the program for labelling of large speech data bases. In that case the phonetic network consists of the correct phoneme string with optional pronunciation alternatives.

ACOUSTIC-PHONETIC REPRESENTATION

The phonemes in the reference library are described in terms of synthesis control parameters. For vowels these are the frequencies and bandwidths of the lowest six formants and the voice source parameters. The frequencies of the lowest four formants were taken from male speaker data in (3) for vowels in stressed position in sentences. Data for Swedish consonants could not be found, so it was computed using an analysis-by-synthesis technique for one male speaker. The same speaker was later used in the recognition experiment described in this report. Nasalized vowels and nasal consonants sounds are specified not only by six formants, but also by two zeroes. Unvoiced fricatives and plosive bursts have two poles and one zero.

There is an important argument for storing synthesis control parameter in the phoneme reference library while doing the matching in the spectral domain. Application of coarticulation and reduction rules is easier on the control parameters than at the acoustic level. Although these rules mainly concern shifts of formant and antiformant frequency values, we don't need to perform an error prone tracking of spectral poles and zeroes on the incoming speech wave. Instead, the modified control parameters are transformed to the spectral domain, where the matching is made.

The main argument against using the control parameters is that the production model used for speech synthesis is a simplification and is not capable of making a perfect copy of natural speech. New and better production models will develop, however. They will be easy to incorporate into the system.

It will also be difficult to make an automatic adaptation of the phoneme library to a new speaker, since it requires a transformation of spectral data to control parameter values. If the identity of the adaptation material is known, though, the search space for the parameters can be considerably reduced, which will minimise the risk of tracking errors.

SPECTRAL ANALYSIS

The speech signal is recorded using a sampling frequency of 16 kHz. The incoming speech is transformed to the spectral domain by an FFT procedure using an analysis window of 25 ms and a frame rate of 10 ms. The frequency range from 200 to 5000 Hz is divided into 16 channels, linearly separated on a Bark scale.

The synthetic spectra are generated by computing a transfer function from the pole-zero specification of the phonemes. For voiced sounds, a voice source is superposed, using the model of Ananthapadmanabha (4). For unvoiced sounds, a white noise source is assumed. The synthetic spectra are transformed to the same representation as that of the incoming speech.

RECOGNITION EXPERIMENT

At this stage, a preliminary recognition experiment has been performed. The task was to recognise utterances consisting of 3 connected Swedish digits. One male speaker recorded 100 tokens of 3-digit strings in a sound-proof booth. The word and string recognition rates were 88 and 63 % respectively. A parametric display of a recognised utterance is shown in figure 1.

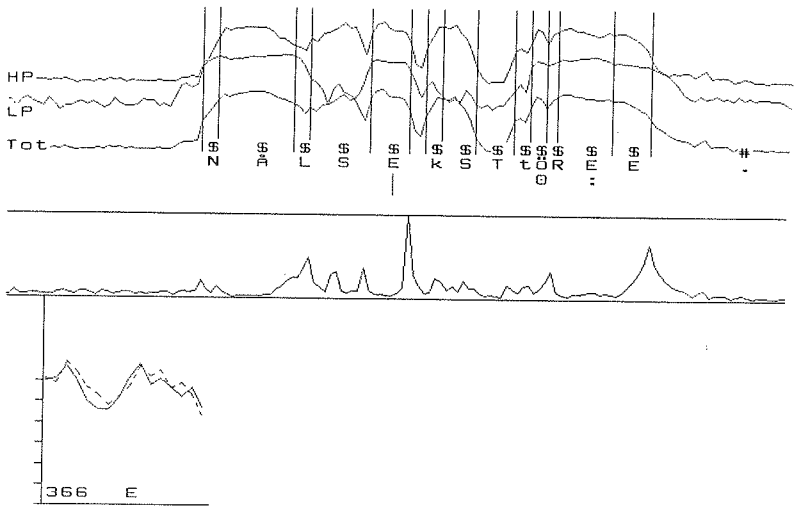


Figure 1. Display of a recognised utterance. The identity is 063 with pseudophonetic transcription /nålsekstre:/. The three intensity parameters are from bottom to top overall intensity, low pass 400 Hz and high pass 500 Hz. The marker below the phonetic symbol /E/ shows the time point of the spectrum section, drawn in solid line. The dashed line is the reference spectrum for /E/. The curve below the intensity parameters is a spectral distance curve, showing the acoustic distance between the prototype spectra of the labeled phonemes and the input spectrum at each time point.

A confusion matrix is given in table 1. A large proportion of the errors were confusion of the digits 3 ('tre') and 5 ('fem') with 7 ('sju').

The error probability seems to be dependent on the position of the digit in the utterance. The error rate for the last digit in the string was twice as high as for the first or second positions. This might be explained by the subglottal pressure drop at the end of the utterance, which causes an amplitude decrease. Also, the glottis opens at the end of the utterance and the voice source spectrum gets a low frequency bias. Furthermore, the duration increase of the final syllable might be responsible for some errors, although the durational constraints for the phonemes are quite wide.

	Recognised digit									
	0	1	2	3	4	5	6	7	8	9
Correct										
0	29	.	.	1	.	.	.	.	.	.
1	.	27	.	.	1	.	1	.	1	.
2	.	.	30	.	.	.	.	.	.	.
3	.	.	3	19	.	.	.	7	.	1
4	3	.	2	.	22	1	.	1	.	1
5	.	.	.	.	.	22	.	8	.	.
6	.	.	.	.	.	1	29	.	.	.
7	.	.	.	.	.	.	.	30	.	.
8	1	.	1	.	.	.	.	.	28	.
9	.	.	.	.	.	.	.	.	.	30

Table 1. Confusion matrix between individual digits.

DISCUSSION

The results presented are encouraging. Many errors could be interpreted in a phonetic manner and it is then possible to correct them by including the appropriate knowledge into the system. Especially, better treatment of voice source dynamics and transitionary parts at phoneme boundaries are believed to increase the accuracy substantially. Future work will also take care of coarticulation effects and speaker adaptation.

REFERENCES

(1) M. Blomberg and K. Elenius: "Time Alignment of Speech to a Phonetic Transcription," STL-QPSR 2/85, KTH,Stockholm 1985.

(2) M. Blomberg, Rolf Carlson, Kjell Elenius, B. Granström: "Speech Recognition Based on a Text-to-Speech Synthesis System", Proc. of the European Conference on Speech Technology, Edinburgh 1987.

(3) U. Stålhammar, I. Karlsson, G. Fant: "Contextual effects on Vowel Nuclei", STL-QPSR 4/1973, KTH, Stockholm, 1973.

(4) T.V. Ananthapadmanabha,: "Acoustic Analysis of Voice Source Dynamics",STL-QPSR 2-3/1984, KTH, Stockholm, 1984.

SPEECH RECOGNIZER FOR VOICE CONTROL OF MOBILE TELEPHONE

M. Blomberg, K. Elenius*, B. Lundström[x], L. Neovius**

* *Department of Speech Communication and Music Acoustics, KTH, Box 70014, S-100 44 Stockholm, Sweden*

[x] *Ericsson Radio Systems AB, S-163 80 Stockholm, Sweden*

ABSTRACT

Infovox is marketing a speaker-dependent, pattern-matching word recognition system, developed at KTH. The algorithms in the system have been modified for noise immunity, and performance has been evaluated in moving cars. The main problems were word detection and noise compensation. After simulations we decided to use a close-talking microphone and a "noise addition" method, where we added the measured noise in the moving car to the reference patterns recorded in a parked car. Using this method, the recognition rate was improved from 69% to 97% on a ten-word vocabulary using the best microphone. A more extensive test was performed on the modified recognition system using two cars and twelve speakers, seven male and five female. Most of them were naive speakers. The twenty-word vocabulary contained some confusable words and was trained in a parked car. During 98 sessions, 1,960 words were read under different conditions with an average recognition rate of 86%. With closed windows at 90 km/h the mean was 91%. An open window at the same speed decreased the result to 82%.

INTRODUCTION

In spite of the reported 99+ per cent accuracy of word recognizers together with simplicity of use and attractive pricing, the commercial breakthrough is still pending. One reason could be shortcomings revealed when the systems leave laboratory environments and are exposed to the practical case of, e.g., the voice control of cellular telephones with noisy background, changing microphone distance and different manners of speaking in noise and quiet. Voice control of a vehicular telephone should be a self-evident application for a speech recognition system. Eyes and hands are occupied by driving the car. Simultaneous manipulation of a keypad together with monitoring a dashboard display is a clearly dangerous task for the driver as well as others. There is already legislation in some countries against making telephone calls during driving and other countries will probably follow.

Besides the above-mentioned difficulties, most specification requirements for this application fall well within the performance of a standard isolated-word speaker-dependent recognizer:

- The vocabulary could be restricted to less than 40 words.
- The recognition time is not very critical due to the progress time of the call which is system-dependent and can be tenth of seconds.
- The stringent requirements of simplicity and considerably lower cost than the telephone itself can be met by using standard components like codecs, signal processors and 8-bit CPUs.
- Speaker-dependence and training of the system is acceptable if this can be made in a stationary car.
- If error feedback control is used, recognition accuracy above 90 per cent should be sufficient.

BACKGROUND

Reduction of the noise problem can be done at different stages in the speech recognition process. At the microphone level, high-directivity and close-talk features may be used to improve the signal-to-noise ratio in existing recognition systems. Research with microphone arrays for dynamic focusing to the speaker's position will enable a user to move more freely than with headset microphones or fixed direction microphones (ref 1).

Once the noise signal has been picked up, it will affect the recognition accuracy. To diminish the effects of the noise, one could try to separate it from the speech. This can be done based on statistical knowledge of one of the signals. Methods for doing this has been developed for purposes of speech enhancement (refs 2, 3). Another technique is to adapt the recognition system to the measured environmental noise. The noise can be measured just before and/or after the sampled word. Reference templates trained in a silent environment will have noise added during recognition, to simulate that the training and recognition have occurred under the same environmental conditions. This method has been reported by Klatt and others (refs 4, 5, 6). Still another possibility is to make the recognition algorithm less sensitive to noise. This could be accomplished by giving more weight to high energy regions of the input signal (ref 7).

The word boundary detection problem can be approached by techniques for recognition of continuous speech or by including some samples before and after the detected word endpoints into the sampled words and thus allowing for some uncertainty in the endpoint detection. This method has been used in the Infovox RA-201 and is also later reported by Haltsonen (ref 8). The former method is obviously better, but at moderate noise levels, the latter technique may prove quite adequate and it was also used in this study.

A problem that cannot be solved by the techniques mentioned above is the change of the speaker's voice in high noise conditions. Experiments have shown that this effect can have the same influence on the recognition rate as the noise itself (ref 9).

We decided to test a system using the above techniques of noise adaptation of the reference templates and allowing for some uncertainty in the endpoint detection.

DESCRIPTION OF THE RECOGNITION SYSTEM

The Infovox RA-201 is a speaker-adaptive word recognition system using pattern recognition and dynamic programming. The speech signal is preemphasized by +6dB per octave, sampled by 10 kHz and fed into a NEC 7720 signal processor, which is programmed as a 16 channel filter bank. The filters are spaced according to the critical band scale and the output from the filters is rectified and integrated over 25 ms giving 40 filter sections per second. Seven cepstral coefficients C_1 - C_7 are calculated from the 16 amplitude values. The first coefficient, C_0, that contains the overall energy, is not used, so the cepstral coefficients only describe the shape of the spectrum and are not at all sensitive to the signal amplitude. Every coefficient is represented in one byte. After endpoint detection, the words are linearly normalized to 32 samples giving them a nominal length of 32 * 25 = 800 ms. Each word now occupies 7 * 32 = 224 bytes.

In order to test the proposed noise-adaptive recognition algorithms, we carried out some preliminary experiments on a simulated system using a Data General Eclipse minicomputer. We used recordings of a ten-word vocabulary (a total of 100 test words) by one male and one female
speaker. The recordings were made with three different microphones to compare their performances. We got the best results using a position-adjustable close-talking microphone placed about 10 cm from the mouth of the speaker. In the original Infovox system the endpoint detection was based mainly on the energies in the frequency range from 200 to 500 Hz. The recordings showed that the environmental noise in the car

had a maximum in the same range - especially with all windows closed - making the sampling of the words very hazardous. It turned out that using the summed energy of all filters resulted in a safer endpoint detection.

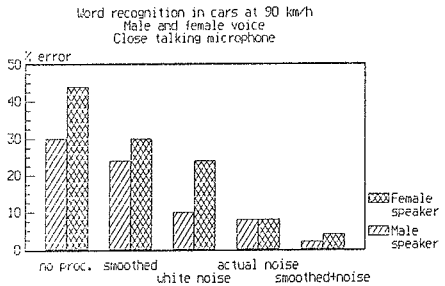


Figure 1. Different noise processing schemes.

The results of some different schemes for noise adaptation are shown in Figure 1 for both speakers. First we see the percentage of errors using the adjustable microphone, the endpoint detection described above and no adaptation to the noise. The amplitude variation between successive filters could be very large. Reducing the difference between filters to 10 dB gave the results shown. Adding white noise instead of the measured noise to the reference templates gave some improvement but not as much as adding the actual noise measured before the sampled word. Finally we see the results using the actual noise plus smoothing. The last results show a substantial improvement of the recognition compared to the first cases.

RESULTS USING TWELVE SPEAKERS

To more extensively test the best method above - noise adaptation plus spectral smoothing - mobile test equipment was developed. The hardware consisted of a standard, transportable PC-compatible computer and an Infovox RA 201/PC speech recognition board. The software was modified to implement the new algorithms. Special testing software was also developed, containing functions for training and recognition with randomized lists, logging and displaying of the results.

The test was run with seven male and five female speakers. It was a rather difficult 20-word vocabulary, with phonetically close pairs like "nio"-"tio", "hem"-"fem" and "Erik"-"Hiby" (nine-ten, home-five, Erik-Hiby), but also some longer words like "Soeharjo" (a name) and "kontoret" (the office). Training the system was made in a parked car repeating each word five times. For safety reasons the recordings were made by the person sitting next to the driver. The speeds varied from 50 to 110 km/h. The conditions also varied with different combinations of window opened/closed, rain or no rain, radio on or off. During 98 sessions, 1,960 words were read under different conditions. The tests were performed without using any sort of rejection of bad matchings and the total number of substitutions were 285, giving a recognition rate of 86%. One-third of the substitutions came from the three phonetically similar word pairs. The mean result was 91% at 90 km/h with closed windows, varying between 78% and 100% per speaker. An open window at the same speed decreased the mean to 82%, the individual results varied between 72% and 97%. We also tested a threshold on the matching distances to reject bad matchings among the 1,960 words. As an example, one of the tested thresholds resulted in 4% substitutions and 22% rejections.

DISCUSSION

We consider the results obtained quite satisfactory, especially when considering the vocabulary. But there are, of course, still some unresolved questions, e.g., weak phonemes at the endpoints can obscure the noise measurements. It is not obvious whether it is better to measure the noise before and/or after a word. Another problem is how to set the energy threshold for endpoint detection - whether it should be fixed or modified by the noise and/or speaking level. The method of rejection could be a fixed threshold rejecting words that have too large a matching distance to the best reference, or a relative threshold rejecting words that have too small a distance to the second best reference. A combination is also possible.

A complication using noise adaptation of the references is that the calculation of the cepstral coefficient of the templates has to be done according to the actual noise of each input word. This means that the reference patterns have to be stored as filter amplitudes, requiring 32 * 16 = 512 bytes per reference compared to 224 bytes when stored as cepstral coefficients. To get a noise-adapted reference, the 16 amplitudes of the noise spectrum have to be compared to each of the 32 spectral sections of the reference template, keeping the maximum of the two amplitudes in each point. To calculate the cepstral coefficients needs a total of 7 * 16 multiplications and additions per template. By using the NEC7720 signal processor for this, the time needed is 12 ms per template which is quite acceptable for a vocabulary of less than fifty words. The time to match references using dynamic programming in the same processor varies between 10 ms and 15 ms.

REFERENCES

1. J. L. Flanagan, "Bandwidth Design for Speech-Seeking Arrays," Proceedings of ICASSP, 1985, Boston.
2. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Transactions on ASSP, Vol. ASSP-27, No. 2, April 1979.
3. J. E. Porter, S. F. Boll, "Optimal Estimators for Spectral Restoration of Noisy Speech," Proceedings of ICASSP, 1984, San Diego.
4. D. H. Klatt, "A Digital Filter Bank for Spectral Matching," Proceedings of ICASSP, Philadelphia, 1976, pp 573-576.
5. J. N. Holmes, N. C. Sedgwick, "Noise Compensation for Speech Recognition Using Probabilistic Models," Proceedings of ICASSP, 1986, Tokyo.
6. B. P. Landell, R. E. Wohlford, L. G. Bahler, "Improved Speech Recognition in Noise," Proceedings of ICASSP, 1986, Tokyo.
7. H. Matsumoto, H. Imai, "Comparative Study of Various Spectrum Matching Measures on Noise Robustness," Proceedings of ICASSP, 1986, Tokyo.
8. S. Haltsonen, "Improved Dynamic Time Warping Methods for Discrete Utterance Recognition," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol ASSP-33, No. 2, April 1985.
9. P. K. Rajasekaran, G. R. Doddington, J. W. Picone, "Recognition of Speech Under Stress and in Noise," Proceedings of ICASSP, 1986, Tokyo.

SOME REMARKS ON TONAL ASPECTS OF MOTHER-CHILD COMMUNICATION.

Anne-Christine Bredvad-Jensen
Child Language Research Institute, Department of Linguistics and Phonetics,
Lund University

In mediating messages the verbal signal is naturally considered as the most important means of communication, even if both somatic and vocal means are used at the same time. For the little child the situation is reversed as the child during his first year has to rely on vocal and on somatic means before he begins to develop verbal language during his second year. Even after his second year non-verbal means will continue to play a very important part in his communication. In his communicative efforts the child is supported by the adult partner who generally is an attentive, watchful and cooperative person. It can be argued (from an acquisitional point of view) that one of the main aims of the adult-child dialogue is to encourage and promote the child's endeavours to take and keep the initiative in a dialogue. What strategies does the child use to signal his intentions to the adult partner? What strategies does the adult use in order to make it possible for the child to fulfill his dialogue intentions? Are there specific tonal strategies which can be used, either by the adult or by the child?

If the child wants to call the adult's attention to something, he may use more or less extensively either of the three communication channels (or a combination of two or three), the verbal, the vocal and the somatic. Verbally through supplying or asking for information: "Look, it's like Granny's.", "What's this?" Somatically through grasping or pointing to something. Vocally through using a special voice quality or tonal configuration. For example a child may mumble loudly but inarticulately. Now is there a definite set of tonal configurations for the "attention-getters"? Is there any difference between those "attention-getters" which are mainly verbal (and/or somatic) and those which are mainly vocally manifested (or at least without a verbally formulated message)? Is there a difference in effectiveness among different "attention-getters"? (Effective = resulting in an adequate answer from the adult.) Well, these are the questions I am interested in, but here I will only give a few illustrative examples.

The dialogue episodes have been chosen from a child oriented point of view*. A dialogue episode is considered to begin when the child initiates a "new paragraph" in the ongoing dialogue, by introducing a new topic/subtopic. It seems relevant to set the boundaries in this way because then the interactionally most interesting situations will be chosen for analysis.Will the adult respond to the initiative in an adequate way and so join the dialogue on the child's conditions so that the child may carry out his intentions? (Does the child receive an adequate answer or does the adult e.g. proceed with giving more redundant information which

*The material chosen for analysis originates from two different projects.1) Three-year-old children from Umeå in a project headed by Jan Winberg and Peter de Chateau, see 1982 and references there. This project investigates the long term effects of different birth routines. One part of the project deals with linguistic matters and it is headed by Ragnhild Söderbergh, see 1987 for the latest version of her dialogue model based on this material. 2) A longitudinal project, headed by Ragnhild Söderbergh, investigating the syntactic development of five preschool children from Stockholm. See Söderbergh 1973 and Lange & Larsson 1977. In project 2 the dialogue material has been recorded on audiotapes supplemented with a simultaneous commentary tape recording whereas in project 1 both videotape and audiotape recordings have been made. In both projects children are involved in free play situations/bookreading.

is not asked for.)

In the dialogue model of Söderbergh (latest version in Söderbergh & Bredvad-Jensen 1987) there is a verbal component consisting of two model codes, one of which is the dialogue flow which is relevant for this discussion. The dialogue flow shows "how topics are introduced and how the two partners jointly contribute to the development of these topics and to dialogue coherence" (op. cit. p. 371). Among symbols used in the dialogue flow are the following:

==> introduction of topic

--> continuation of topic implying that new aspects are introduced

<--tying on to the partner's utterance

In the first dialogue below called "The girl-doll", Embla (2 years, 6 weeks) and her mother are looking in a children's book together. Embla introduces a new topic by asking the name of a big doll in the book (Ch1) at the same time as she points to it. Embla then calls the doll "That girl." Mother does not answer the question but questions whether it really is a girl. Embla consents and reformulates her question (Ch3), asking the name of the "doll she", (a possible interpretation according to the transcription, which is the result of at least four independent listeners). Mother says she does not know and turns the question back to Embla, who gives no answer (Ch5). Mother argues that it is a doll. Apparently she has not interpreted Ch3 as a consent. Embla ascertains that it is a doll (Ch7) not a girl (Ch9). Mother confirms. Embla introduces a new aspect of the topic by asking whether the doll can shut its eyes (Ch11) and mother supposes that this is the case.

THE GIRL-DOLL

Ch1: /he:/ DEN flickan?*(What's the name of THAT girl?)*

M2: Är de DÄR en flicka. - TYCKER du det? *(Is THAT a girl. - Do you THINK so?)*

Ch3: Nä. - Heter /då ho/ - Hm. *(No. - What's the name /doll she/ - Mm.)*

M4: Jag VET inte. Vad TROR du den heter? *(I don't KNOW. What's her name, do you THINK?)*

Ch5: Hä. *(Ha.)*

M6: Jag tror det är en DOCKA ser du. *(I think it's a DOLL, you see.)*

Ch7: Hm. - Docka. *(Mm - Doll.)*

M8: Mm. *(Yeah.)*

Ch9 INTE flicka. *(NOT girl.)*

M10: Nää. - EN FLICKDOCKA. *(No. - A GIRL-DOLL.)*

Ch11: Kan den BLUNDA den? *(Can she SHUT her eyes, the doll?)*

M12: Ja DE/det kanske den kan. *(Yes THAT/that maybe it can.)*

Comparing Embla's questions in this dialogue, Ch1 will be coded as ==> , as a new topic is introduced, whereas Ch3 and Ch11 will be coded as --> instead, as these questions add new aspects to the old topic. Looking at the tonal configurations of these questions they seem somewhat contra-intuitive, as the fo-range is considerably wider for Ch3 (ca 80 Hz) and Ch11 (ca 150 Hz) than for Ch1 (ca 45 Hz) and as the fo-maximum is considerably higher for Ch3 and Ch11, see table 1. One could hypothesize that calling a person's attention by introducing a new topic should result in a wider range and a higher maximum in comparison with continuing the development of the topic, which is a joint enterprise, where the partner's attention has already been attracted, but here the opposite is the case. Is this due to the simultaneous

pointing gesture in Ch1? It might be argued that the attention-calling function of the verbal/ vocal aspects of the message might be diminished if an attention-calling gesture is added, as in this case. And/ or perhaps it is too complicated for the little child to expand on all three aspects at the same time. Among the material presented in Söderbergh & Bredvad-Jensen 1987 there is a child who stops playing with an object (with her hands) while formulating new sentences (p. 376). Naturally it is impossible to draw any conclusions whatsoever from this sample alone, but Schaffer, 1984, has shown that in non face-to-face dialogues between adults there was a strong correlation between high beginning Fo-range and topic continuation as opposed to topic change. This also holds for face-to-face communication, but only as a clear tendency. If we compare Ch7 and Ch9 in "The girl-doll", both of which can be classified as topic continuation without adding new aspects, that is <-- , with Ch1 it is evident that the same relation holds for them as for Ch 3 and Ch11 in comparison with Ch1. This is maybe more startling as Ch7 and Ch9 are statements (although with contrastive function) whereas Ch3, Ch11 and Ch1 (!) are Yes/No - questions.

TABLE 1

	fo-range	fo-max	fo-min
Ch1	43	312	269
Ch3	78	370	292
Ch7	73	416	343
Ch9	116	400	284
Ch11	151	400	249

In the next dialogue, "The towel", the child tries to take away a towel which is fixed onto a wall in a doll house asking mother how to do it (Ch1). Mother explains why it should not be taken away. After a pause the child argues that it might be used in this case because 1) it is intended for drying oneself (Ch3) and 2) Daddy is wet, because he washed his hands (Ch 5). Mother consents (M6).

THE TOWEL

Ch1: Men hur går de å få bort de HÄR? *(But how is THIS to be taken away?)*

M2: Den tror jag inte man SKA ta bort. Jag tror faktiskt att den e FASTKLISTRAD här. Den ska liksom bara HÄNGA där,så att man - ser att de brukar hänga handdukar där. *(I don't think you SHOULD take that away. I think it has been PASTED on here. It should only sort of HANG there, for you to - see that they usually hang towels there.)*

(pause)

Ch3: Mh det här kan ju tvätt/ öh TORKA sig. *(But this you may wash/ eh DRY yourself.)*

M4: Mm. *(Yeah.)*

Ch5: Han e så våt/ han öh därför han öh TVÄTTA händerna - <PAPPAN>*. *(He is so wet/ he eh WASHED his hands - the <DADDY)>*.)*

M6a: <Mm>*. *(<Yes>*.)*

M6b: Mm. *(Yes.)*

*uttered simultaneously

An analysis of this episode is presented in Söderbergh & Bredvad-Jensen 1987. It is evident from that analysis that this mother is a clear representative of the *supportive* style (as Wells *et al.*, 1979, put it). Here we can see that she gives an adequate answer, which stimulates the child to verbal creativity and that she waits with a long pause for the child to formulate his verbal message and then only stimulates him by uttering the backchannel items in M4, M6a and M6b. The tonal configurations as such of these backchannel items also exhibit supportiveness because they are in the mother's very upper range; a range which more naturally belongs to the fo-range of the child. In two cases (M4 and M6a) the mother even begins with the same fo-value as that at the end of the child's preceding utterance. Looking superficially at the curves it looks like the child continues to speak instead of the mother. These utterances exhibiting tonal coherence on the part of the mother are quite different from the "normal" utterances in M2 all of which are uttered in a lower range.

This is probably an example of a specific tonal strategy used by an adult to promote the child's linguistic performance but to draw any conclusions from one single episode is impossible, as is also the case with the girl-doll episode.

BIBLIOGRAPHY

Lange, S. & Larsson, K., 1977. Studier i det tidiga barnspråkets grammatik (Studies in early child grammar. With an English summary). Stockholms universitet: Institutionen för nordiska språk, Projektet barnspråkssyntax, Stockholm.

Schaffer, D., 1984. The role of intonation as a cue to topic management in conversation. Journal of Phonetics, 12, pp. 327-344.

Söderbergh, R, 1973. Project child language syntax and project early reading. Stockholms universitet: Institutionen för nordiska språk, Projektet barnspråkssyntax, Stockholm.

Söderbergh, R. & Bredvad-Jensen, A.-C., 1987. A dialogue model and its application in a free play situation. In Lilius, P. & Saari, M. (eds.) The Nordic Languages and Modern Linguistics 6, Proceedings of the Sixth International Conference of Nordic and General Linguistics in Helsinki, August 18-22,1986, pp. 369-395, Helsinki University Press, Helsinki.

Wells, G., Montgomery, M. & MacLure, M., 1979. Adult-child discourse: Outline of a model of analysis. Journal of Pragmatics, 3, pp. 337-380.

Wiberg, B. & de Chateau, P., 1982. Long term effect on mother-infant behavior of extra contact during the first hour post partum. IV. Study design and methods. In Anthony, E.J. (ed.) The Child and his Family, pp. 105-128, Wiley and Son, New York.

Winberg, J. & de Chateau, P., 1982. Early social development: Studies of infant-mother interaction and relationships. In Hartrup, W.W. (ed.) Review of Child Developmental Research, 6, pp. 1-44, The University of Chicago Press, Chicago.

DIALOGUE PROSODY

Gösta Bruce, Ursula Willstedt, Paul Touati and Antonis Botinis
Department of Linguistics and Phonetics, Lund University, Sweden

The present paper is the first report from a research project, which started recently in January 1988 and which is planned to run for three years. The project is called Contrastive Interactive Prosody - with the acronym KIPROS based on the project title in Swedish - and has got financial support from the Bank of Sweden Tercentenary Foundation. This paper contains mainly a summary of the research plan for the project and also reports on the ongoing research work.

There are three important starting points for the KIPROS project. The first starting point is the research on prosody that has been conducted in phonetics in Lund through the years and that covers many aspects of prosody. Research on Swedish prosody led to the development of the so-called Lund model of prosody (cf. Bruce 1977, Bruce & Gårding 1978, Gårding & Bruce 1981, Gårding 1982, Bruce 1985). The second starting point is the contrastive study of prosody conducted in Lund and directed towards the development of a general model of prosody particularly by Gårding. Examples from this research are Gårding (1981), Gårding, Botinis & Touati (1982), Gårding, Lindau, Norlin & Svantesson (1986) and most recently Touati (1987). A third starting point is the great interest in recent years taken in the study of interaction and dialogue analysis within linguistics and child language research in Lund (cf. Sigurd 1986, Söderbergh & Bredvad-Jensen 1987).

Our current, accumulated knowledge of prosody is based mainly on the study of prosody in fairly simple, well controlled experimental situations, i.e. so-called laboratory speech. The kind of methodology that we have been advocating then has typically been the simulation of a very simple dialogue - a question-answer-paradigm - where our informant has been playing the roles of both interlocutors (cf. Bruce 1977). Against the background of this research we now consider it possible and important to take the step from studying prosody in these simple situations towards investigating the role of prosody in communicatively more natural and relevant, but also more complex situations such as in interactive speech and dialogue.

A nice demonstration of the role of prosody in interactive speech is the comparison of two different tape recorded versions of the same dialogue, where one is the authentic version of a part of a spontaneous dialogue and the other is the artificially spliced together version consisting of the corresponding read, isolated utterances occurring in the same order as in the

original conversation. When listening to the artificially spliced together version of such a dialogue, one is struck by the absence of interactive prosody, and the important role of prosody in human spoken interaction becomes apparent, which may otherwise pass unnoticed in the natural, genuine version of the same dialogue.

KIPROS is about how prosody is being used in human spoken interaction. The purpose of the project is to investigate dialogue prosody in a contrastive perspective involving a few European languages that display interesting structural differences from a prosodic point of view: French, Greek and two varieties of Swedish (South and Standard Swedish). The ultimate goal of our research will be the development of a model for French, Greek and Swedish prosody in an interactive perspective. In such a model we will have to relate prosodic properties to suitable interactive categories and on the basis of the analysis propose rules for the generation of prosody in dialogues. The study of interactive prosody will also permit the calibration of our earlier research on prosody.

In our analysis of interactive prosody we will focus our attention particularly on how prosody contributes to the construction of a dialogue and how speakers make it develop through the use of prosodic markers such as particular tonal and temporal patterns. Our interest in the study of dialogue prosody will be both how the speaker uses prosody - accentuation, phrasing, boundary signalling - to make the internal organization of his contributions to the dialogue, and how this is related to the other speaker's contributions, for example how prosody is used in cooperation between speakers, as well as how prosody signals the actual regulation of turns in a dialogue. By analyzing prosody in French, Greek and Swedish dialogues we hope to elucidate the specific character of dialogue in these different cultures but also and most importantly to increase our understanding of the structure and function of prosody in these languages.

The research described here will be primarily basic research, but there are apparent possible applications of this research in language teaching, automatic speech synthesis and recognition as well as in speech therapy.

In the study of prosody in a dialogue perspective we are as phoneticians faced with a number of new problems. The first major concern of the KIPROS project is the choice of suitable tape recorded material for phonetic and acoustic analysis. We are presently searching along the dimension constructed - authentic dialogue; i.e. on the one hand experimentally arranged and more controlled situations, where the topic and the development of the dialogue is predictable to some extent, and on the other hand situations where this is more or less not the case, for example dialogues recorded from radio and TV

programs. The actual choice of dialogue has also to be related to our interest in those dialogues where prosody is expected to play an especially important role for the development of the dialogue.

In the present first period of the project we have been recording and studying different types of dialogue for each of the three languages. We have been doing this in order to explore the field and get an overview and to be able to choose the most interesting types of dialogue for special study. For Greek we started recording a constructed type of dialogue in the laboratory, for Swedish we have been recording a more authentic kind of dialogue from a well known radio program, and for French we have begun studying the most authentic type of dialogue: interviews about more existential topics. Although we start at different points for the three languages, we intend to cover the same dialogue types for French, Greek and Swedish.

The second major concern of the project is the analysis of the recorded material. This phase consists of three different parts: a fairly rough analysis of the structure of the dialogue - we are presently testing different models for dialogue analysis -, a selective prosodic transcription covering those prosodic categories that appear to be relevant for our purpose, and most importantly the qualitative and quantitative study of mainly tonal and temporal patterns from acoustic recordings of F0 and durations. We expect to find a fair degree of variation and optionality among the possible, prosodic markers in interactive speech.

A third important concern of the project will be the task of formulating rules for the use of prosody in spoken interaction. We will test these rules perceptually through editing of the relevant parameters of recorded speech and speech synthesis. By isolating stereotypical patterns of interactive prosody through the use of speech synthesis, we intend to be able to evaluate the optionality and suitability of different prosodic markers for interactive categories. One step in this analysis by synthesis may be the use of stylization of F0 contours for the establishment of the relevant pitch patterns, where the equivalence of F0 contours will be evaluated perceptually. This method has been successfully exploited for the analysis of intonation by the Dutch school (cf. for example 't Hart & Collier 1975).

The three phases of recording, analysis and synthesis occur logically one after the other. In the project work they will recur cyclically. We will actually try to come the third phase - rules and synthesis - fairly soon after having made a careful analysis of a relatively small speech sample. In this way we hope to get an idea rather soon about the degree of success in our analysis of dialogue prosody.

REFERENCES

Bruce G. 1977. Swedish word accents in sentence perspective. Gleerup, Lund

Bruce G. 1985. Structure and functions of prosody. In Guerin & Carré (eds.) Proceedings of the French Swedish Seminar on Speech, 549 - 559. Grenoble

Bruce G, & E. Gårding. 1978. A prosodic typology for Swedish dialects. In Gårding et al. (eds.) Nordic Prosody, 219 - 228. Department of Linguistics, Lund University

Gårding E. 1981. Contrastive prosody: A model and its applications. Studia Linguistica 35: 146 - 166

Gårding E. 1982. Swedish Prosody. Phonetica 39, 288 - 301

Gårding E, & G. Bruce. 1981. A presentation of the Lund model for Swedish intonation. In Fretheim (ed.) Nordic Prosody II, 33 - 39. Tapir, Trondheim

Gårding E, A. Botinis & P. Touati. 1982. A comparative study of Swedish, Greek and French intonation. Working Papers 22, 137 - 152. Department of Linguistics, Lund University

Gårding E, M. Lindau, K. Norlin & J.-O. Svantesson. 1986. Final report: Phonetic analyses of some non-European languages (LUCLA)

't Hart J, & R. Collier. 1975. Integrating different levels of intonation analysis. Journal of Phonetics 3, 235 - 255

Sigurd B. 1986. Computer simulation of dialogue and communication. In Karlsson (ed.) Proceedings of the 3rd Conference of Scandinavian Computational Linguistics. Department of Linguistics, Helsinki

Söderbergh R, & A.-C. Bredvad-Jensen. 1987. A dialogue model and its application in a free play situation. In Lilius & Saari (eds.) The Nordic Languages and Modern Linguistics 6. Helsinki University Press

Touati P. 1987. Structures prosodiques du suédois et du français. Lund University Press

MODELLING DIFFERENT WAYS OF SPEAKING IN A TEXT-TO-SPEECH SYSTEM.

Rolf Carlson och Björn Granström
Department of speech communication and music acoustics
KTH, Stockholm
Phone 468 - 790 7568

Introduction

Currently available text-to-speech systems are not characterized by a great amount of flexibility, especially not when it comes to varying of voice or speaking style. On the contrary the emphasis has been on a neutral way of reading, modelled after reading of non-related sentences. Most of the testing of these systems have been carried out with isolated words, frequently monosyllabic nonsense words. Varying the voice of the synthesis, if at all possible, has frequently been modelled as a simple transformation in the synthesis parameter domain. Likewise basic variation of speaking rate has been thought of as a linear change in the time domain. Vowels, consonants and pauses are in this case affected by the same factor, irrespective of stress, essentially by changing the time interval between synthesis parameter updates. We have started to look into some of these shortcomings in the context of our text-to-speech system.

The demand for different speaking styles

There is a very practical need for different speaking styles in text-to-speech systems. Such systems are now used in a variety of applications and many more are projected as the quality is developed. The range of applications ask for a variation close to the one found in human speakers. General use in reading stock quotations, weather reports, electronic mail or warning messages are examples where humans would choose rather different ways of reading. The most common application today is in aids for the handicapped.

Visually impaired persons have very different needs (Carlson & Granström, 1986). On one extreme end of the style continuum they want hypercorrect speech that gives them maximally exact information how the text is written. Spelling is in this case a possibility, but it is too slow and gives hardly any understanding of the text content. The other extreme is very fast speech for information scanning. Using the device as a speech prosthesis implies the general need of variation that any human feel. There is, however, a problem to make the variation available to the handicapped person in this case, since communication speed is the primary concern. Much of the variation found in natural speech is not controlled deliberately, at least not at a very conscious level. The possibility has been discussed to connect the degree of emphasis to some myoelectric signal rather than controling it with special symbols in the text . To our knowledge this has not yet been successfully tried.

Apart from these practical needs in text-to-speech systems there is the scientific interest to formulate our understanding of human speech variability in explicit models

Current possibilities

In our current system there are several possibilities to vary the speech. On the global level there are possibilities to vary speech tempo, level and voice parameters like amount of aspiration, mean pitch and pitch dynamics, vocal tract length, general speed of articulation etc. Some of these parameters have been combined to define different

voices similar also to female or child voices. The simulations of such voices are, however, still not very convincing. One aspect that we currently are working on is to include a more sophisticated voice source in the model. This new model will make it easier to simulate different ways of speaking.

The text-to-speech system consists of a combination of rules and lexica. It is always possible to enter a phonetic transcription, thereby getting full control of allophone selection. In the lexica, information on basic pronunciation is given but also on the dichotomy: function word/content word. So called strong and weak forms of words can both be given and selected by later rule governed processes. Lexical items can also be marked with extra information such as parts of speech or tags for phonological reduction processes.

In the rule components all sorts of effects could be modelled i.e. as long as they are rule governed. The conditioning factors for these rules could either be given as analysis by the system, such as syntax or some measure of predictability of words, as commands to the system or it could be given as extra information in the input text. One example of the latter is the emphasis control in the present system. By adding a number before each word in the phonetic string we can over-rule such things as default sentence stress assignment and function word reduction and also force different degrees of emphasis.

Experiments with speaking styles for British English

Although the concept of a multi-lingual text-to-speech system is a familiar one, rather little attention has been given to the question of the variety of each language that is synthesized. Language variation is currently being incorporated into our system both as concerns dialect and style (Bladon et al.,1987).

Our British text-to-speech implementation has been extended to provide a "style variable", a user-set range of ten values. This device can be used, for example, to propagate more affrication with a "lower" style number. The area of the system in which we first explored this style variable was in fact that of the forty or so function words ('can, have, for, them' etc.) of British English whose pronunciation, though not their spelling, varies considerably with sentence context and style. As an example, the word 'can' in a phrase 'I can go' may have a large number of realizations, some of which may be just acoustically specifiable subtleties, but some at least of which can be rendered transcriptionally: [kæn , kən , kəŋ , kŋ ,ˌŋ]
It is probably reasonable to rank these forms from left to right as graded from most formal to casual. They can therefore be synthesized with style variable values of say 9, 7, 5, 3, and 1, respectively. There are doubts whether style is possible to model on a one dimensional variable, especially for an entire utterance. Style may also change within the utterance and there may well be style variants that are lexically, and individually, determind. An illustration of the simple approach can be seen in figure 1. The three utterances are controlled by the same phonetic string, and interestingly there is no rule specifically concerned with speech rate. The faster speech of the "casual" style is due primarily to the greater reductions allowed.

To undertake this style ranking more widely through English phonetics is, in the present state of knowledge, rather an uncertain exercise. The normative data have hardly been collected at all. At the same time, there are two particularly good motives for pressing ahead. One is that, at present, the text-to-speech developer is faced with some uncomfortable decisions of simplification when specifying such a highly variable word as "can". Another reason is a research issue.

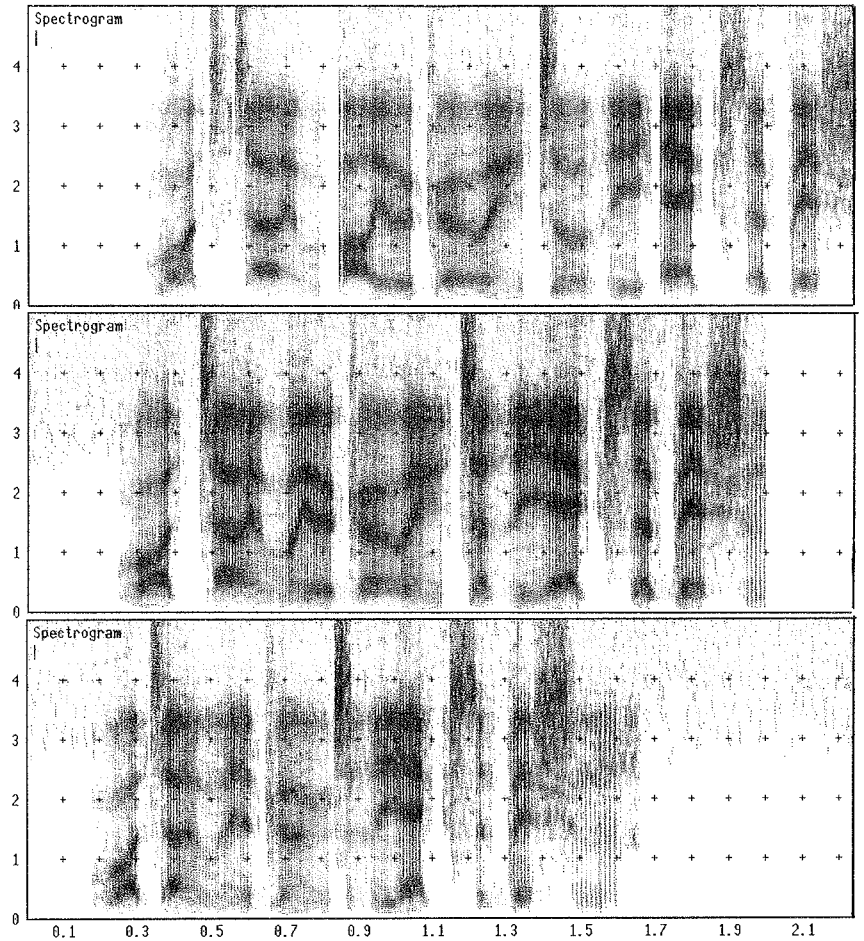


Figure 1.The synthezised sentence "What time are you going to the exhibition" spoken in "formal"(top), "normal"(middle) and "casual"(bottom) speaking style.

Fast speech for the visually impaired

Even if there are means of varying the speaking rate in the normal text-to-speech system they are not appropriate for the the extremely high speaking rates demanded by the visually impaired. Normal speaking rate is often estimated to be around 150 words per minute (wpm). This measure will be language and text dependent and it will also increase if pauses are not included. The demand from the blind is to obtain speaking rates of around 500 wpm to approximate fast silent reading by sighted persons. One way to obtain these "super human" rates would be to ignore progressively more of the low information content of the text. This would mean some kind of keyword reading without any language stucture. After preliminary tests we abandoned this idea. One problem with this solution is how to predict where the keywords are. It also seemed to us that the lack of linguistic structure was very confusing. If at all possible it is our conviction that we should model the speech on human performance. At these high

speeds, however, there is no good human templet. We don't want to extrapolate from the point where human speech production breaks down.

The rationale in the present attempt to increase speed further is that the phonetic component could be changed for a reduced and faster component at runtime appropriate for very high speaking rates. Especially prosodic rules are simplified or taken away. The differences between stress and unstressed syllables are still marked by duration and fundamental frequency, a new set of inherent durations are established and mean pitch also needed to be increased. To most listeners the speech is close to unintelligible at this speed though it is claimed to be useful by experienced blind listeners. In figure 2 two reading speeds are shown.

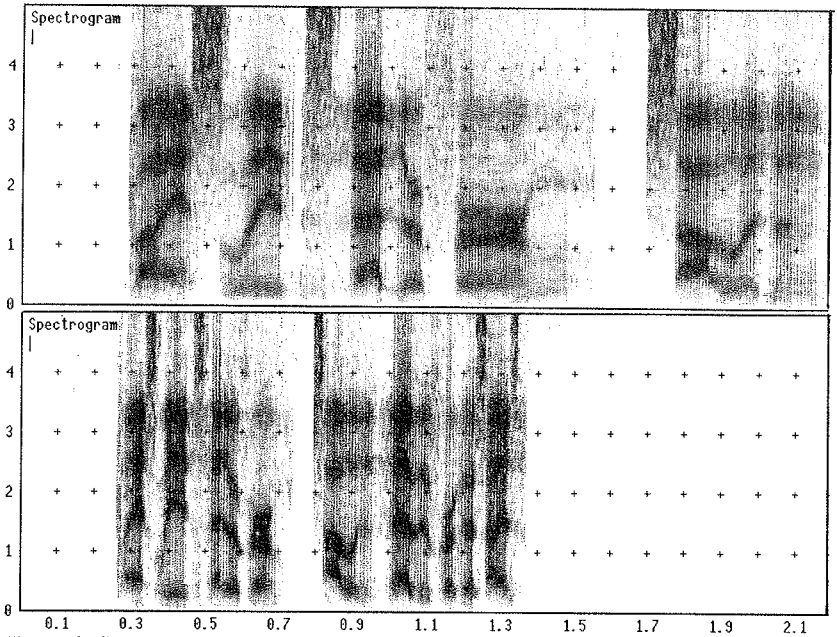


Figure 2. Spectrogram of the synthesized sentence "Buy swedes that are firm, solid and heavy for their size" at normal speaking rate(150 wpm, truncated) and at very high rate (500 wpm)

References

Bladon A., Carlson R., Granström B, Hunnicutt S. & Karlsson I.(1987): "Text-to-speech system for British English, and issues of dialect and style",European Conference on Speech Technology, vol. 1, Edinburgh, Scotland.

Carlson R. & Granström B. (1986): "Applications of a multi-lingual text-to-speech system for the visually impaired", pp. 87-96 in (P.L. Emiliani, Ed.): Development of Electronic Aids for the Visually Impaired, Martinus Nijhoff/Dr. W. Junk Publ., Dordrecht.

WHICH FOREIGN ACCENTS CAN SWEDES IDENTIFY?

Una Cunningham-Andersson and Olle Engstrand
Institute of Linguistics
University of Stockholm

INTRODUCTION

One of the major questions addressed by the current project "Attitudes to Immigrant Swedish" at Stockholm University is to what extent are attitudes to immigrants related to attitudes to the way in which immigrants speak Swedish?

There are a number of ways to test this relationship, but one of the most intuitively attractive is to let a group of listeners hear a number of readings of a text in the belief that all the speakers are different individuals. The listeners are given information about the speakers' nationalities. Unknown to the listeners, they hear several speakers twice, but believe that in each case they are hearing two different individuals with different mother tongues. If then the listeners' attitudes to the speakers (as regards parameters such as intelligence, education, friendliness and honesty) are elicited, any differences between attitudes to both versions of the speakers heard twice must be attributed to the difference in information fed to the listeners regarding the speakers' origins.

It seems probable that some kinds of false information about accent origin are easier to believe than others, simply because some accents are more easily identified than others. It is probably difficult to persuade a native speaker of Swedish that a certain accent is, for example, Finnish if it lacks most of the features used by Swedes to identify Finnish accents. Conversely, a typical Finnish accent is unlikely to be accepted as an example of, for example, a Greek accent.

A series of experiments was designed to test the capabilities of two groups of native speakers of Swedish: 18-year old students of non-humanities subjects, and 6 teachers of Swedish as a second language. Accented Swedish material was obtained from the IRIS database[1]. A large number of recordings of immigrants from different linguistic backgrounds reading the well-known passage "The North Wind and the Sun" in Swedish are contained in the database. 35 versions of the first half of the text (about 25 seconds) were obtained from the database and elsewhere (4 native Swedish versions were included).

EXPERIMENTS

Free response method

Two methods were used to elicit accent identifications. For half the students (17), an answer sheet was developed where the same three questions (compare Ryan 1983[2]) were asked about each of the 35 voices the listeners heard: (a) Is the speaker a native speaker of Swedish?; (b) If the answer to (a) was "no", which

---

[1] Engstrand, O. (1987) The IRIS speech data base - a status report. RUUL 17, 121-126. Inst. of Linguistics, Uppsala University.

[2] Ryan, E.B. (1983) Social psychological mechanisms underlying native speaker evaluations of non-native speech. Studies in Second Language Acquisition 5(2) 148-159.

part of the world does the speaker come from?; (c) What is the speaker's mother tongue? To help them to answer (b), the students were given a map of the world with numbers representing areas such as "North America", "Far East", Eastern Europe" etc. The teachers of Swedish as a second language were asked the same questions, except that for them, (b) was changed to: If the answer to (a) was "no", which language group does the speakers native language belong to, or which part of the world does the language belong to? No map was provided for the teachers.

The results obtained from the students were disappointing in the sense that they seldom answered question (c), and did not always answer question (b). This was also true of the teachers, though to a lesser extent. It is however possible to compare the degree of accuracy with which the part of the world where the speaker's native language is spoken was identified. For the purposes of this experiment, the world was divided into eight zones:

A	Scandinavia	B	N.America, N.Europe, Australia
C	C. & S.America, S.Europe	D	N.Africa, Middle East
E	E.Europe, Soviet	F	Far East, S.E.Asia
G	Africa	H	Indian sub-continent

The following table shows how many percent of each listener group identified each speaker correctly in terms of the above eight zones:

ZONE	ACCENT	STUDENTS	TEACHERS
A	Finnish	94	100
A	Lappish	59	83
A	Swedish 1 (biling Sw-Tur)	94	100
A	Swedish 2 (Northern)	100	100
A	Swedish 3 (Scanian)	94	100
A	Swedish 4 (Finnish-Swedish)	100	100
A	Swedish 5 (Uppland)	100	100
A	Swedish 6 (Scanian-Uppland)	94	100
A	Norwegian	94	100
B	British English	94	100
B	American English	18	33
B	German	53	83
C	Spanish 1 (Spain)	12	50
C	Spanish 2 (Columbia)	24	50
C	Spanish 3 (Uruguay)	29	100
C	French 1	18	50
C	French 2	71	100
C	Greek 1	18	67
C	Greek 2	29	17
D	Turkish	12	50
D	Kurdish	12	83
D	Persian	0	0
D	Arabic	47	67
E	Polish 1	6	83
E	Polish 2	6	17
E	Serbo-Croat	0	33
E	Russian	47	83
E	Czech	6	83
F	Japanese	6	17
F	Korean	0	0
G	Swahili	0	67
G	Kinyarwanda	0	0
G	Tigrinya	6	33
G	Yoruba	17	67
H	Bengali	0	0

As can be seen, the teachers were considerably better than the pupils at identifying the geographical origin of the speakers. At least 40% of both the teachers and the pupils identified the Finnish, Lappish, native Swedish, Norwegian, British English, German, French, Arabic and Russian speakers' zone of origin, while the teachers could also place Spanish, Greek, Turkish, Kurdish, Polish, Czech, Swahili and Yoruba accents in the appropriate part of the world. Moreover, the teachers were more often able to correctly specify the speakers' native languages than were the pupils.

Guided response method

The figure on the next page shows the results for the remaining group of 19 pupils. They were tested using the same tape with 35 native Swedish and accented readings, and a form similar to the matrix shown in the figure, with possible native languages for the speakers listed down one side and the number of the speaker across the top. The pupils were instructed to place a cross in the box corresponding to the native language they believed each speaker to have and they were encouraged to place crosses in more than one box if they were unsure. The boxes with solid frames are those which were crossed by at least 40% of the pupils. The correct answers are written in bold typeface. The numbers in the boxes represent the number of pupils who crossed each possible native language for each speaker. The numbers at the top and bottom of the figure represent the order in which the speakers were presented to the pupils, and the numbers at the right of the figure represent the sum of responses for each possible native language.

The results were very similar to those obtained using the free response method. At least 40% of the pupils could identify French, native Swedish, Norwegian, Lappish and Finnish accents. This method permits an analysis of the wrong answers obtained. As can be seen from the figure, the following incorrect identifications were made by at least 40% of the pupils: Finland-Swedish was mistaken for a Finnish accent; Bengali, British and Swahili accents were heard as American (in the case of Bengali and Swahili, this can be attributed to the speakers having learned English prior to their arrival in Sweden); Yoruba was slightly more often marked as Kinyarwanda, but notice that Swahili also got quite a few votes – it seems likely that the speaker was simply identified as "African"; one of the Polish speakers was mistaken for Serbocroat (again, a correct identification of the speaker's native language group); the other Polish speaker and the Arabic speaker were identified as having Turkish as their native language; and the American English speaker (who spoke Swedish with only a slight American accent) was believed to be a Norwegian speaker.

In conclusion, these results show which accents are easily recognised by native Swedish listeners, and which false statements about accent origin listeners would be likely to believe, thus providing a basis for the kind of attitude experiment mentioned in the introduction.

FIGURE. (PUPILS WERE GIVEN ORAL INFORMATION ABOUT WHERE THE LESS FAMILIAR LANGUAGES ARE SPOKEN.)

SPEAKERS

Speaker no:	21	32	5	20	14	12	3	13	17	23	7	15	4	18	22	24	28	31	25	11	26	35	27	2	10	29	9	33	8	30	16	19	34	1	6	Summa
Persian	–		3			4	1			2													1	1		3	1		1		3	2			2	24
German		**5**			1			2			4		1		1			1		2	3			2	4	1						1				28
Polish		4	**1**	**3**	2	1		3	1	3										1		1	2	4	4		1	1	2	5						39
American	3				–	1															14			10				5	8				2			43
Kurdish	1		4	3		**7**	3	2	1	2	1							1	1	1					2	5	2		1	1	2	3		5	48
Spanish		1		2			–	**3**	**3**		1								2	1			3		1		2	1		2	2		1			25
Arabic	1		3			2	2	4	1	**3**	1	1					1	1					2		1	4	2	1	3			2			4	39
French		1								1	**2**	**17**											6					1			1					29
Swedish					2	1					1		**19**	**19**	**19**	–	**18**	**10**																		89
Serbocroat	2		8	1		4	2	1	2	3					2				**3**				1		1	3	3	1	2	3	4	2	1		1	50
Lappish				1	2	1	1	1			2					5		6		**9**							1	1				1	2	6		39
Bengali	2			1	1	2													1		–					1	1	3				2	2		1	17
Norwegian		1			8				1		3		5	1	1		4	4				**19**														47
Kinyarwanda	1		1	1		2	4	1	2	1									3		2		–		1	4	1	8			1	1	1		2	37
British	4				1		1									1					4		1	**3**				1	7			1			1	25
Korean	4	2				2		1			2	1			1	1			7	1	2			2	–		5		1	1					1	34
Tigrinya			2	1		2	2		3	1									2	1	1				1	**4**		3		3		2	1		1	31
Turkish		3	9	4		3	6	1	2	8	1								1						4	4	**3**		2	3	2	3		3	62
Yoruba			1	1	2	3	4		2		1								3		3					4		**7**		1		2	7		2	43
Swahili	2		2		2	2	4		2										3		3					5		7	–	1		2	7		2	44
Russian			2	4	1				2	1								2		1			1		3		2	6	1	**6**	2	1	2			37
Greek		1	4	2	1	2	4			6	1								1	2			1		3	3	4		1		**3**	**4**	2		7	52
Czech		2	2	4		3		2	3	2	2									4			2		4		2	2	2	2	2	1	**6**			47
Finnish																18		1		1	1												1	18		40
Japanese	2	2	1	1		2		1		1	2					1			4		2						4		1	2		3	1		**1**	31
Speaker no	21	32	5	20	14	12	3	13	17	23	7	15	4	18	22	24	28	31	25	11	26	35	27	2	10	29	9	33	8	30	16	19	34	1	6	

DISCONTINUOUS VARIATION IN SPONTANEOUS SPEECH[1]

Olle Engstrand & Diana Krull
Institute of Linguistics, Stockholm

1 Introduction

In spontaneous speech, words and phrases are frequently heard to be strongly reduced in relation to their alleged phonological base forms. In a subset of the vocabulary[2], we observe an apparent quantal alternation between word and phrase representations in the sense that the amplitude of a phonetic gesture corresponding to an underlying segment displays no continuous transition between levels of reduction or elaboration. As a rule, moreover, it seems as if speakers tend to adhere to a very restricted number of such discontinuous realizations. Fig.1 provides an example. The figure displays spectrogram tracings of the first and second formants pertaining to all occurrences of the grammatical conjunction så att /so att/ (so that, such that) as pronounced in a spontaneous speech situation by a male Stockholmer (see Sec. 2.1 for details). There are two basic, quite distinct sets of curves pertaining to one reduced (left three columns) and one elaborated (rightmost column) level of pronunciation. The reduced version sounds very much like the word satt /satt/ (sat).

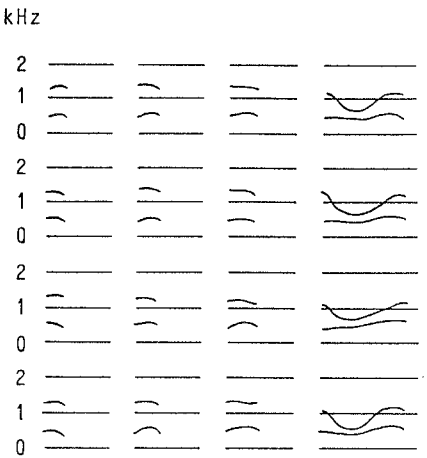


Fig. 1. Tracings from broad band specrograms illustrating two basic varieties of the pronunciation of the grammatical conjunction så att. The left three columns represent a reduced level of pronunciation and the rightmost column represents an elaborated level. Subj. JS. (See section 2 for details.)

Do strongly reduced forms such as these instances of så att as a rule contain any kind of acoustic traces of the segments constituting the corresponding phonological base forms? Let us refer to the affirmative answer to this question as the weak hypothesis of segmental invariance[3].

Mere listening and transcribing is probably an insufficient method of investigating this hypothesis. For example, even when a clear segment realization is heard in the natural speech situation, the effect frequently seems illusory when listened for in an excized portion of the speech flow. To test the hypothesis, therefore, careful measurements are necessary. The remainder of this paper discusses a set of such measurements and their subsequent evaluation in terms of listener reactions.

2 Experiments

2.1 Recording and formant measurements

We used for this experiment approximately half an hour of recorded speech produced by a male native speaker of the Stockholm dialect of Swedish (subj. JS). The greater part of the recording is a lively monologue supported by brief questions and comments from one of the experimenters[4].

Out of the material, we selected and spectrographed (1) all instances of the grammatical conjunction så att (so that), where å stands for the phonologically tense back rounded vowel /o/, and a stands for the phonologically lax open fronted vowel /a/; (2) all sequences saC, where /C/ is any non-nasal consonant; and (3) all preconsonantal instances of the word så (so).

Onset and nuclear (mid vowel) frequencies of the first four formants (F_1–F_4) were identified and measured for all occurrences of these utterances. F_3 and F_4 were practically identical and will not be shown here. Measurement data for F_1 and F_2 are given in Table 1.

Table 1. Onset and nuclear frequencies for the vocalic segments in så att, saC and så. Subj. JS.

NO.	PHRASE	ON/NUC	FORMANT	N	MEAN	S.D.
1	så att	on	F1	12	470	49
2	så att	on	F2	12	1332	57
3	så att	nuc	F1	12	553	42
4	så att	nuc	F2	12	1319	62
5	saC	on	F1	16	516	78
6	saC	on	F2	16	1348	70
7	saC	nuc	F1	13	596	85
8	saC	nuc	F2	13	1325	83
9	så	on	F1	10	420	32
10	så	on	F2	10	1395	102
11	så	nuc	F1	9	408	40
12	så	nuc	F2	9	1373	101

The table shows slight differences between the three cases. In particular, a t-test reveals statistical significance at the 1% level for mean F_1 onset values both for så vs. så att and for så att vs. saC.

We next turned to the question whether these small differences are at all able to influence listeners' perception in the direction of the respective utterances så att and satt with these portions of the speech flow excized out of their contexts.

2.2 A listening test

A series of stimuli containing 12 instances of så att and 12 instances of satt was presented to nine native Swedish speakers of roughly the same dialect as subj. JS. Since the sequence satt occurred only four times in our recorded material, copies of those tokens were introduced randomly to give an equal number of both stimulus types. The resulting set of stimuli was subsequently duplicated and a new random order was created for the duplicate. This procedures gave a total of 432 responses from the nine listeners whose task it was to decide for each stimulus whether it represented så att or satt.

The results are shown in Table 2 in terms of the outcome of a multiple regression analysis using 1-5 predictor variables. The data refer to correctly identified instances of så att. The values of R for the correlation between number of correct answers and each individual variable is given in the middle column, and the cumulative effect is shown in the right column. Calculated across all predictor variables, this effect is R=0.89.

Table 2. Result of multiple regression analysis: prediction of number of correctly identified instances of the phrase så att using 1-5 predictor variables. 9 subjects.

PREDICTOR	MULTIPLE R	
	Individual variables	Cumulative
F1on	-0.60	-0.60
F2on	-0.60	0.77
F1nuc	-0.35	0.77
F2nuc	0.11	0.85
Duration	0.51	0.89

We thus obtain a high cumulative correlation with contributions from several variables (including nuclear duration which was slightly greater for satt than for så att).

3 Comments

The above relationships, of course, are purely statistical, and do not tell us much about the perceptual contribution of each variable. However, F_1 onset is the acoustically strongest effect and may therefore play a decisive role. The contribution of a possible spectral difference in the preceding /s/ should also be evaluated. As demonstrated at the Conference, however, the measured effects are likely to be very marginal as a perceptual criterion in natural speech. It is, however, interesting in certain other respects.

Firstly, the example suggests that this speaker in fact tends to preserve slight traces of underlying segment structure even in very strongly reduced forms.

Secondly, since the speaker is very close to completely eliminating a segment, for what reason does he nevertheless preserve it to some extent?

These points may be discussed in terms of various constraints on phonetic speaker-listener interaction. Thus, on the one hand it can be assumed that speakers, in accordance with Lindblom's view of adaptive variability (Lindblom 1987), tend to shape the speech wave so as to achieve a good balance between articulatory effort and perceptual clarity. If this balance were optimal, however, the vowel corresponding to /o/ in så att would be reduced completely since it is highly predictable in context and therefore semantically insignificant to the listener.

Reasonably, however, an optimal adaptation would in itself require a fair amount of data processing. Thus, rather than designing a particular utterance plan to cope with predictable segments, it would seem to be a more convenient speaker strategy to manipulate the basic motor scheme by means of a somewhat more coarse-grained adaptation of gestural amplitude. In particular, the speaker would simplify his task by aiming for either a low (hypo) or a high (hyper) level variety with phonetic traces of even perceptually negligible underlying segments preserved at both levels.

As a third possibility, it may be hypothesized that complete segment reduction is marked stylistically, so that traces of underlying segments will remain and be perceived even though their strictly semantic value is insignificant. The above experiment demonstrates that listeners are, in fact, sensitive to this kind of variation. The difficulty of dealing with it from a semantic-perceptual point of view suggests the need for adding the phonostylistic dimension to the issue of segmental invariance in spontaneous speech.

Footnotes

1) Supported by the Swedish National Board for Technical Development (STUF) and The Bank of Sweden Tercentenary Foundation.

2) Most typically in grammatically functional morphemes, words and phrases.

3) The strong version would involve the claim that some acoustic aspect of a segment appears unaltered across a variety of contextual transformations.

4) This recording session was conducted by Rolf Lindgren of the Institute of Linguistics.

Reference

Lindblom, B. 1987. Adaptive variability and absolute constancy in speech signals: two themes in the quest for phonetic invariance. Proc. XIth ICPhS, Tallinn, Estonia, Aug. 1987, Vol. 3, pp. 9-18.

CONTRIBUTIONS TO TEMPORAL ANALYSIS OF READ SWEDISH

Gunnar Fant & Anita Kruckenberg
Dept. of Speech Communication and Music Acoustics
Royal Institute of Technology, Stockholm

Introduction

This is a summary of some of the findings from the last few years' work on a project centered around the data bank now developing within our group at the Royal Institute of Technology. Earlier reports dealing with both segmental and prosodic aspects have appeared in Fant, Nord, & Kruckenberg (1986; 1987). A more comprehensive report on these studies will be given in a forthcoming issue of the STL-QPSR; see also Fant & Kruckenberg (1988). The prosody project to be reported here aims at studies of: (1) General relations between objective measures - durations and pitch excursions - and subjective estimates of syllabic stress. (2) Inter-stress intervals and their interaction with syntactic boundaries and pauses. (3) Relations between objective boundary region measures and subjective estimates of degree of juncture-boundary marking. (4) Distribution and individual variations of pauses.

Text material and processing

For the analysis performed in the present study, we have selected a paragraph of nine sentences from a novel by Kerstin Ekman, in all 133 words of about 50 seconds' reading time. This is but a small part of the data bank text material. A major part of the analysis was devoted to the reading of a reference subject, ÅJ, who has a clear and engaging reading style without mannerisms. We also analyzed one of the nine sentences read by 15 other subjects, including five females.

Computer-generated spectrograms with synchronous FO, oscillogram, and intensity trace were made of this limited speech material. A segmentation into successive speech sounds corresponding to a broad phonetic transcript was undertaken. The occasional difficulties and ambiguities involved have been discussed previously (Fant, Nord, & Kruckenberg, 1986).

Objective and subjective measures of syllabic stress

Swedish is a stress-timed language with sequences of unstressed syllables alternating with stress syllables. A stressed syllable carries a nucleus of a long vowel followed by one or two short consonants, or no consonant, or the vowel is short and followed by one consonant, or a consonant cluster. A stressed syllable also carries one of two contrasting tones, accent 1 or accent 2.

Duration appears to be the main correlate of stress in Swedish, at least it is more readily quantifiable than associated FO measures. The zone of durational increase with increasing stress is the entire syllable but a larger part is confined to the vowel and the following consonant. This VC nucleus will serve as our major objective measure but we have also been studying durational patterns of entire syllables and vowel-to-vowel units. For each of these different objects, we have constructed normalized measures, syllable duration indexes, to account for variations with the number of phonemes within the unit.

We define the syllable duration index by

$$S_i = 1 + (T-T_u)/(T_s - T_u),$$

where T is the measured duration of a syllable or a vowel-to-vowel unit or a VC unit and T_u is the typical duration of unstressed units and T_s of stressed units determined separately for the specific number of phonemes in a unit. The duration index S_i is thus attained by interpolation/extrapolation with respect to mean values of 1 for unstressed and 2 for stressed syllable units. Most of our studies are based on VC and syllable final or single V units. For our reference speaker ÅJ, we noted T_u=53 ms for V, T_s=185 for V:, and T_u=109 ms for unstressed VC. For stressed VC, we made use of separate references of V:C=236 ms and VC:=212 ms.

A similar process was carried out for S_i calculated on the basis of syllables and V-V units. These two latter alternatives gave rather similar results. In other words, with proper normalization, the durational correlates of stress are very much the same independent of the unit of observation.

SYLLABIC STRESS. OBJECTIVE AND SUBJECTIVE.

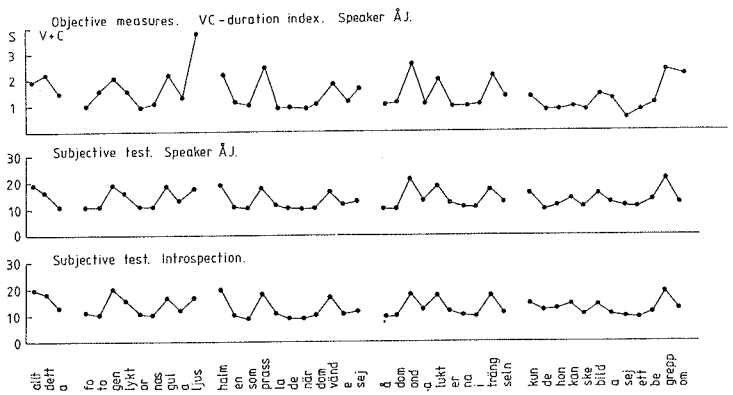


Fig. 1. Syllabic stress. Objective and subjective measures.

In Fig. 1, the VC duration index has been compared to subjective evaluations. Fifteen subjects were asked to make a direct estimate of perceived prominence of each of the syllables in sentence 7. They first made an introspective evaluation from silent reading of the text and then listened to subject ÅJ reading the same passage. The consistency was quite good, standard deviations for a single estimate were of the order of three units within the given frame of 10 for unstressed and 20 for typical stressed syllables. Apart from minor deviations in unstressed syllables, the overall profiles display an apparent similarity. Deviations between objective and subjective scalings of ÅJ's reading may in part be explained by the occasional influence of more extreme inherent durations, such as for /s/ and /a/. The sentence contains four phrases. Those that end with a stressed syllable receive a final lengthening which apparently is ignored in the subjective estimate.

A conspicuous trait is the great similarity between the introspective performance and the listening to subject ÅJ. This could imply that subjects might rely more on their inner "top down" expectancy than on what they hear. However, control experiments involving listening to

several subjects' readings of one and the same sentence showed that listeners estimates did follow the individual variations of produced stress patterns, as evidenced from objective measures and expert listening. One conclusion is that subject ÅJ's interpretation of the text is quite similar to the average of the subjects in the listening test and thus not extreme.

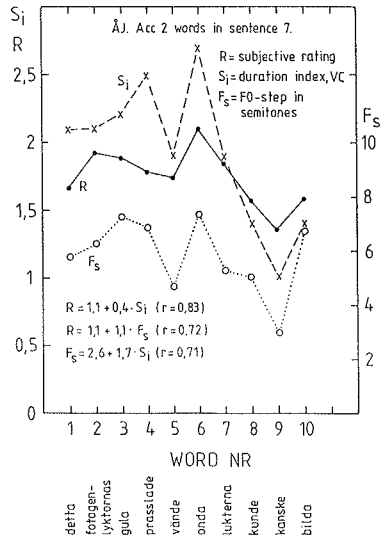


Fig. 2.

In Fig. 2 we have included local F0 excursions as a second objective measure. For this purpose, we have selected the ten stressed syllables of sentence 7 carrying accent 2. The F0 fall within the voiced part of a syllable has been plotted in semitones together with S_i values corrected for inherent durations and the subjective response. the overall correlation is apparent and of the order of 0.7 to 0.8. Larger durations of the voiced part allow larger total F0 excursions, as expected. We have also been able to demonstrate similar covariations with the relative F0 rise in accent 1 syllables. Male and female subjects produce similar F0 contours when measured on a semitone scale.

Connective F0 traits

A few comments will be made on the F0-contour of the reference sentence, Fig. 3. It contains five intonational phrases supporting the syntactic structure. Following Eva Gårding (1984), we have sketched an intonation grid. In the accent domains the grid has a width of the order of half an octave. The overall declination within a phrase is also of this order, i.e., six semitones. The final rise at the end of the sentence could be described as indicating a focal domain. Standard descriptions of Swedish intonation, e.g., Bruce (1977), treat the secondary stress of grave accent words as a discrete unit. The secondary hump is thus considered to be present only in compounds or under the influence of a sentence accent. We often find a weak secondary F0-peak appearing in the second syllable after the main stress, e.g., "vände sig", "kunde hon", "bilda sig", i.e., a pattern of alternation. We feel that in connected non-laboratory speech there is place for a more continuous aspect of stress, the measurable correlates including a gradual appearance of a secondary F0 accent 2 peak. It is an open question whether the high F0 in the second syllable of the adjective "gula" belongs to the accent 2 domain of the word or whether it is a high F0 reference point for the following accent 1 domain of the word "ljus". Perhaps we have an additive effect here.

Other aspects of stress correlates that need to be quantified are voice source characteristics and reduction phenomena including consonant-vowel contrast.

The speech code

A special attention has been laid on the study of rhythmical qualities in text reading. We have fresh evidence supporting a suggestion of Lea (1980) that inter-stress intervals, defined by distances from a stressed vowel to the next stressed vowel of the sequence, act as synchronizing impulses for an internal clock which guides the time we take in making pauses between phrases and sentences.

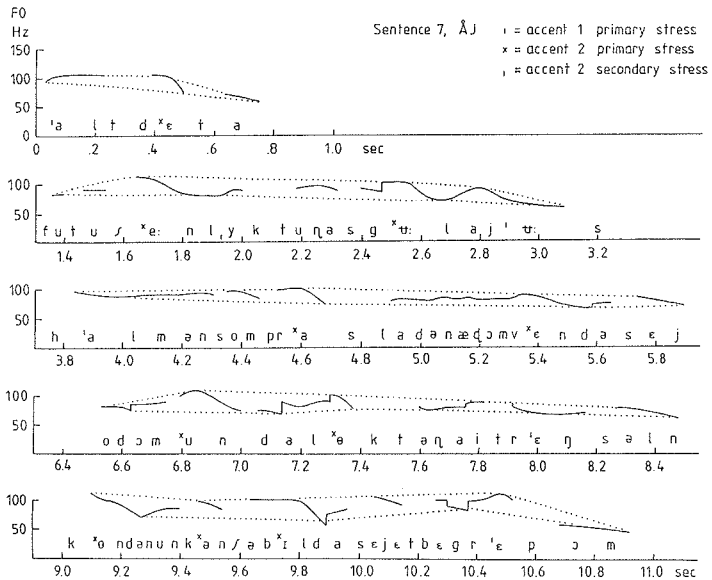


Fig. 3.

We have found that the average duration of the inter-stress intervals not spanning a syntactic boundary sets the basic temporal module of the internal clock. In rhythmical reading, the duration of an inter-stress interval spanning a phrase boundary with a pause tends towards the sum of an inter-stress interval predicted from the number of phonemes contained plus one modular unit of the inner clock, i.e., an average inter-stress interval is added. At sentence boundaries with longer pauses, there is a tendency of one or two additional clock units being added. In other words, noting that we always have a terminal lengthening before and to some extent after pauses, we may restate this finding as an expectancy that the sum of the physical pause and terminal lengthening equals an integer multiple of the rhythmical time constant. This is illustrated in Fig. 4.

Durations of inter-stress intervals are thus largely imposed by language and do not appear to be adjusted o retain isochrony while the rhythmical demand becomes apparent in the planning of pauses. Even without such perfect synchrony, e.g., when a phrase boundary is realized without a pause, the final lengthening alone is capable of signalling the appearance of the boundary.

The constants of the regression line relating inter-stress intervals to number of phonemes appear to be a key to the analysis of components of individual reading and talking styles. These constants should also be of interest in contrastive language studies.

The subjective markedness of phrase boundaries as determined by listening tests was found to correlate well with the duration of the boundary spanning inter-stress interval and with F0-dip measures and appearance of voice creak. In spite of large inter-subject variations of pause durations, the group means showed a clear tendency towards what was observed as a rhythmical norm for a single speaker.

Relative pause time, i.e., the ratio of pause time to reading time, was found to be a consistent speaker-dependent characteristic. The general tendency in speech that the variance of a larger unit is greater than the sum of the variances of its parts as in, e.g., V:C and C:V units apparently also holds for pause durations and terminal lengthening as parts of an inter-stress interval. There is also a weak tendency in this direction relating the total reading time to its parts, effective speech time, and pause time. The larger unit is more stable than the parts.

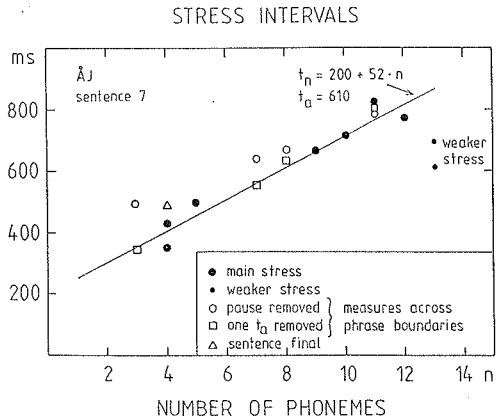


Fig. 4.

References

Bruce, G. (1988): Swedish Word Accents in Sentence Perspective, CWK Gleerup, Lund.

Fant, G. & Kruckenberg, A. (1988): "Temporal structures in Swedish text reading", STL-QPSR 2-3/1988.

Fant, G., Nord, L., & Kruckenberg (1986): "Individual variations in text reading. A data-bank pilot study", STL-QPSR 4/1986, pp. 1-17.

Fant, G., Nord, L., & Kruckenberg, A. (1987): "Segmental and prosodic variabilities in connected speech. An applied data-bank study", pp. 102-105 in Proc. XIth ICPhS, Tallinn, USSR, Vol. 6, Estonian Academy of Sciences.

Gårding, E. (1984): "Comparing intonation", Working Papers No. 27, Linguistics Dept., University of Lund, pp. 75-99.

Lea, W.A. (1980): Trends in Speech Recognition, Prentice Hall Int.

A PHONETIC INVESTIGATION OF THE STØD IN STANDARD DANISH

Eli Fischer-Jørgensen
Copenhagen

The Danish stød is a prosodic phenomenon, whose distribution corresponds, roughly, to accent 1 in Swedish and Norwegian. However, in contradistinction to the tonal accents, the stød requires a long vowel or a short vowel plus a voiced consonant for its manifestation.

The present investigation started as a teamwork at the Institute of Phonetics in Copenhagen, but I have later enlarged the material considerably. It comprises 13 subjects in all, but all did not participate in the whole investigation. The acoustic analysis is based on around 1400 word pairs, the physiological investigations on somewhat smaller numbers. The words were read in frame sentences. The results have been published in ARIPUC 21, 1987, p. 55-265, and, in an abbreviated form, they will be published in two articles in Phonetica.

In older descriptions, including Jespersen, the stød was described as a glottal closure, i.e. acoustically as a pause. This has, however, not been confirmed by instrumental investigations, and in the present material a closure was found in only one, strongly emphatic ,example. Irregular vibrations were, however, found in 70% of the corpus (see Fig. 1), but the degree of irregularity is very variable.

The most consistent characteristic of the stød is a decrease in intensity , beginning about 6 cs after vowel start and comprising 12 dB, on the average (see Fig. 2b) This decrease was found for all 13 subjects. A restricted spectrographic analysis showed that particularly the lower part of the spectrum is weakened.

Moreover, in words with stød all subjects have a higher fundamental in the first part of the syllable than in words without stød, the difference being about 21 Hz. In the second half of the syllable (about 10 cs after vowel start) the fundamental often starts falling, sometimes rather abruptly (see Fig. 2a).This fall was found in 55% of the examples, but in most of the other cases the measurement of F_0 was made impossible by irregularities. This means that

a weak stød is characterised by a frequency drop, a strong stød by irregular vibrations covering the frequency drop. Even or rising F_0 in the stød syllable is very rare. Because of the high start the second syllable of disyllabic words may be lower than the first syllable in contradistinction to the normal rising contour in the Copenhagen Standard. This happens particularly for speakers with a Jutlandish background (see Fig. 2a).

Inverse filtering of a few words read by two subjects showed decrease of negative spikes, and the integrated curve showed lower flow in the second half of syllables with stød.

It is of interest to look for the physiological production mechanism behind these acoustic facts.

In his thesis (1944) Svend Smith advances the theory that the decisive factor is a strong ballistic contraction and relaxation of the expiratory muscles (demonstrated by means of surface electrodes). This will produce a corresponding rise and fall of the subglottal pressure. The pressure rise may cause a proprioceptive reflex activity in the vocal folds, and if their tension is not adjusted to the quickly falling pressure, irregularities may arise. The falling pressure also causes a fall in intensity.

Measurements of airflow (6 subjects) showed consistently lower airflow in the second part of the syllable.

It was not possible to record the expiratory muscles, but subglottal pressure was recorded for one subject. It was normally higher in the beginning of a word with stød than in words without stød (see Fig. 2e), but the difference was only 1-2 cm H_2O, and the fall was very slow. This cannot explain the fall in intensity nor the frequency contour.

The constriction of the glottis was examined by means of fiberoptics for 6 subjects (50 frames per second). Vowels with stød show constriction of the glottis in the latter part of the syllable, and there is often a contraction of the false vocal folds, but with large intersubjective variation (see Fig. 3).

Electromyography of the muscles of the larynx did not show any particular activity of posticus or transversus; but 5 out of 7 subjects had strong activity in the vocalis

muscle in words with stød with a peak at the start of the frequency drop (see Fig. 2d). There was also often a (slightly earlier) peak in cricothyreodeus, which explains the high start of the syllable with stød. But there was generally no valley corresponding to the fall in frequency (except sometimes for one subject Fig. 2e). Lateralis (3 subjects) had a peak for the stød like vocalis.

The extra rise in subglottal pressure often precedes the extra vocalis activity, and in these cases the latter might be a reflex innervation, but the rise in subglottal pressure may also start later, and there is much overlapping between the subglottal pressures in words with and without stød, whereas there is hardly ever overlapping in the vocalis recordings, which means that there are many cases of strong vocalis activity without any higher subglottal pressure.

Probably the stød involves an independent contraction of several muscles (expiratory muscles, vocalis, lateralis, cricothyreodeus, and possibly more). The stød seems to be a type of dynamic accent; and the irregularities are hardly due to maladjustment, as Smith assumes, but rather part of an intended special phonation type.

However, two subjects did not show any particular vocalis activity in words with stød. In one of the cases the electrodes may not have been in the right position. But in the other case this is not probable, since this subject was recorded twice with the same result. This subject also had very strong contraction of the false vocal folds, and he often had rising fundamental in the second half of the syllable in stead of the normal fall.

It is possible that he uses other muscles for the stød, e.g. lateralis and the constrictor muscles. It was, unfortunately, not possible to try these muscles (It is well known that EMG of the larynx muscles is not very pleasant for the subjects, and only few want to repeat the experience).

The present investigation should be supplemented by a perceptual analysis of the relative importance of the different acoustic cues.

Fig. 1

Stød with irregular vibrations

Fig. 2

Average curves
of lo recordings
(see the text)
Subject BF

Fig. 3

Fiberoptics of the
vocal folds in the
first (A) and the
second (B) part of
the vowel with stød
1 subject HU
2 subject JR

FOCAL DEFORMATIONS OF STOCKHOLM ACCENT PATTERNS AND THEIR FUNCTIONS

Eva Gårding

For Bengt Loman

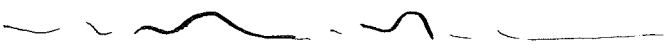
In this paper I will show how focus signals not only the most important part of the message but that it can also signal modality and have other functions.

My introductory example is a passage from a recording of a spontaneous conversation which was moderated by Bengt Loman at the Studio of the Malmö branch of Radio Sweden in the early sixties. I was asked to analyse it prosodically for the project Svenskt Talspråk (see Gårding 1964,1967 and Loman 1967). The text of this passage and the fundamental frequency curves are presented below.

ä harninåra MINNEN i från ä

SKOLÅLDERN attä LÄRARE har KRITISERAT

ert uttal eller försökt å gjort ä



VÄRDERINGAR om vilka LJUD man ska använda å va man

ska UNDVIKA åsåvidare

The curves are copied from mingograms. At the time I called the contours marked by thick lines central stresses because they marked points that were central to the message without producing the impression of contrast.

My auditive and acoustic analysis of the spontaneous Stockholm material showed that there were two types of manifestations of central stresses (focal accents), both rising but with different

ranges. One of them covered half of the speaker's range and the other the whole range.

To understand this phenomenon, a special recording was made in which the main informant was asked to make a series of imitations of his own focal accents in one and the same carrier phrase. He did very well with the less marked focus manifestation but had a lot of resistance towards the strong one (Gårding 1967). I shall return to these difficulties later.

It is only recently that I have understood that such accentuations can express modality. To support this statement I have other examples from the same material, many observations that I have made listening to the radio, in particular to a program with many interviews called *Kanalen* 'the channel', a renewed analysis of my old material (Fig.1) and an analysis of a new material (Fig.2).

To begin with I will comment on the term *deformation* which I have used in the title of my talk. I regard the focal accents, like the expressions of other intonational features at the phrase and sentence level as deformations of certain basic contours of the accents.

What are these basic contours? I went through my own Stockholm material and the material that Gösta Bruce had presented in his dissertation (1977) to find common characteristics of the accents in different contexts. It became apparent that the distinctive features could only be defined in a negative way, i.e. A1 does not fall and A2 does not rise in the accented syllable of polysyllabic words.

To make the notion of deformation meaningful (which I think it must be particularly for the understanding of how accentual and tonal patterns are recognized in connected speech) I decided to regard the citation form as basic and the influence of phrase and sentence intonation as deformations of this form.

Figure 1 shows the citation forms of the two accents and below each of them examples of the two different deformations which have arisen from the two different manifestations of focal accents.

Let us look at the superposed curves of A2, *värden* 'values' and A1, *världen* 'the world', singled out from a carrier phrase *det gäller -- här* ' 'we are for -- here' in statement intonation with a falling terminal contour. The curves represent average values of five repetitions. In the curves local maxima and minima have been enclosed by grids according to a rule of thumb , described earlier (Gårding 1984).

It is easy to see for A2 that the highs and lows have roughly retained their positions relative to the segments although all the absolute values have changed. These changes can be captured in an economical way by means of the grid as an expression of the intonation of the focussed part.

The grid lines also show that the stronger degree of focus manifestation is accompanied by a stronger inclination of the grid. Note that the speaker, in order to achieve this effect has lowered the low. Another possibility would have been to raise the last high. That is a strategy used by two of my new informants. A fourth informant retains the same degree of inclination but uses a wider grid i.e. a larger amplitude for the pitch movements of the more marked focus.

The corresponding focus manifestations of A1 are more influenced by the global contour. The strong focus lies in a

rising part of the contour with its highest point in the second syllable. In contrast, the highest point of the weak focus occurs in the first syllable. Correspondingly, the first part of the strong focus is curved upwards and that of the weak focus is curved downwards. In Bruce's investigation only strong focus was taken into account (1977).

After this description of two different deformations in focus let me now pass on to their communicative functions. The spontaneous material has many instances of the stronger focus manifestation appearing in questions and in different discourse situations that for the sake of simplicity we may call emphasis or contrast.

To test this view I conducted a new experiment with different informants. The new speakers, all from Stockholm, were given a list of questions modeled on questions I had heard in the radio interviews with a sequence of focussed parts and answers to these questions with a comparable set of foci. There were also requests among the test sentences and reassuring answers to these requests (not treated here).

Figure 3 shows some examples from this investigation which started recently. I have used the beginning of the focussed words as line-up points for the pitch contours of the focussed parts. (Tempo variations and pitch differences outside of focus have been disregarded.)

The figure supports the contention that we have traces of the modal function all through the focus domain. A steep topline inclination is used in a question rather than in a statement. For this informant the ironic question has the steepest inclination, but for my other informants the lowest one.

I would like to sum up my observations about the communicative functions of focus and the focus domain in the following way. It expresses weight and brings out the lexical-distinctive function. It has a grouping and demarcative function and as we have seen in this paper it can express modality and emotion. To a certain extent this can also be said about the accent domains but it is only in focus that all the functions cooccur. To paraphrase a famous definition in segmental phonology: Focus is a bundle of intonational functions.

Among the functions of intonation we usually also count the extra linguistic ones, e.g. the power of intonation to reflect dialect. For this function I give priority to the focus domain. What happens to pitch in focus is perhaps the most powerful tonal dialect marker, the true *ostinato* of the dialect and the first feature to be copied in conscious imitation. Scanians living in Stockholm seem to take after the local ostinato unconsciously. Conversely the Stockholm ostinato tends to be suppressed for Stockholm speakers living in Skåne for a lengthy period of time.

The difficulties that my first informant had to produce marked focus in controlled speech have come into new light thanks to comments from my new speakers. A sequence with repeated strongly marked foci does not belong to the Stockholm prestige dialect. Rather, it seems to be typical of a dialect (Ekenssnack) spoken in some southern areas of Stockholm and something of a prosodic shibboleth.

References
Bruce G. 1977. Swedish word accents in sentence perspective. Travaux de l'institut de linguistique de Lund XII. Gleerup. Lund
Gårding E. 1964. Några observationer rörande tryckförhållanden i stockholmskt talspråk. Svenskans beskrivning 1, ed. C.-C. Elert, Stockholm
Gårding E. 1967. Prosodiska drag i spontant och uppläst tal. In Svenskt talspråk, ed. G. Holm. Almqvist & Wiksell, Stockholm: 40-85
Gårding E. 1984. Comparing intonation. WP 27. Ling. Lund University: 75-99
Loman B. 1967. Prosodi och syntax. In Svenskt talspråk, ed. G. Holm. Almqvist och Wiksell, Stockholm: 86-159

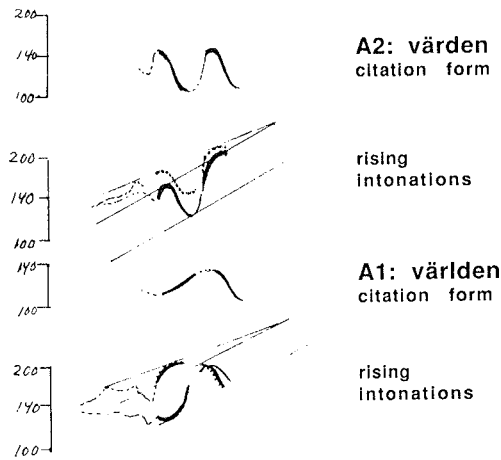


Fig.1 Strong _____ and weak ..... focus.

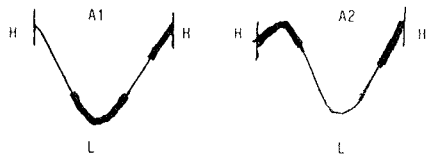
Fig. 2 Focal deformations in different modes
(Interviewer to politician): _____
Har du nån ANING om vad man ska GÖRA för att få ORDNING på det här?
(Question with suspicion): _ . _ . _ .
Har du nån ANING om vad man ska GÖRA för att få ORDNING på det här?
(Politician, assuring): . . ..
Jag har en ANING om vad jag ska GÖRA för att få ORDNING på det här.

EASTERN NORWEGIAN AND WESTERN SWEDISH INTONATON IN A COMMON DESCRIPTIVE FRAMEWORK

Eva Gårding and Per Lindblad

The purpose of this recently started project is to apply a method of intonation analysis to two neighboring dialects of two different languages which permits a systematic comparison. Our material consists of similar sentences, pronounced as statements and echo questions with focus in different places.

For this presentation we have selected the statements *Manne lever bedre nå* 'Manne lives better now' (Norwegian) and *Manne lever bättre nu* (Swedish). Figure 1 presents fundamental frequency tracings for typical utterances. We shall first comment on the Oslo speaker. The accented words are located in more or less V-shaped pitch patterns.In focus the two accents can be schematized as follows:



For the analysis we shall make use of the V-shape. Accent 1 (A1) and Accent 2 (A2) are placed in their respective V's according to different rules which give the words their characteristic tonal patterns. The low point of A1 is in the middle of the accented syllable. For A2 the corresponding low comes later, more precisely at the beginning of the postaccented syllable. A difference in the timing of the accent contour for the two accents is well documented in the literature for all Swedish and Norwegian dialects. Here we note in particular that the timing difference is present even at the first high point, which for A1 is the final high of an earlier accent and for A2 is in the middle of the accented vowel. The last High on the other hand, is similarly timed for both accents. Here A2 has caught up with A1 (Gårding and Lindblad 1973).

We shall call this V the focus domain and give it the representation HLH. Outside of focus we also find V-forms in connection with the accented syllables, which we call accent domains. Very often the focus domain contains not only the focussed word but also a group of following unaccented or deaccented syllables. The term *deaccented* means that the word has lost its lexical accent in the phrase. The internal weight of the syllables is maintained by duration.

We have several examples of deaccented words subordinated to a V. In spite of their deformation by larger- scale falls and rises, their inherent pitch shapes are still recognizable. This has interesting psycho-acoustic implications suggesting a constant representation at a higher level. However, in synthesis these portions of the curve can be generated by straight interpolation.

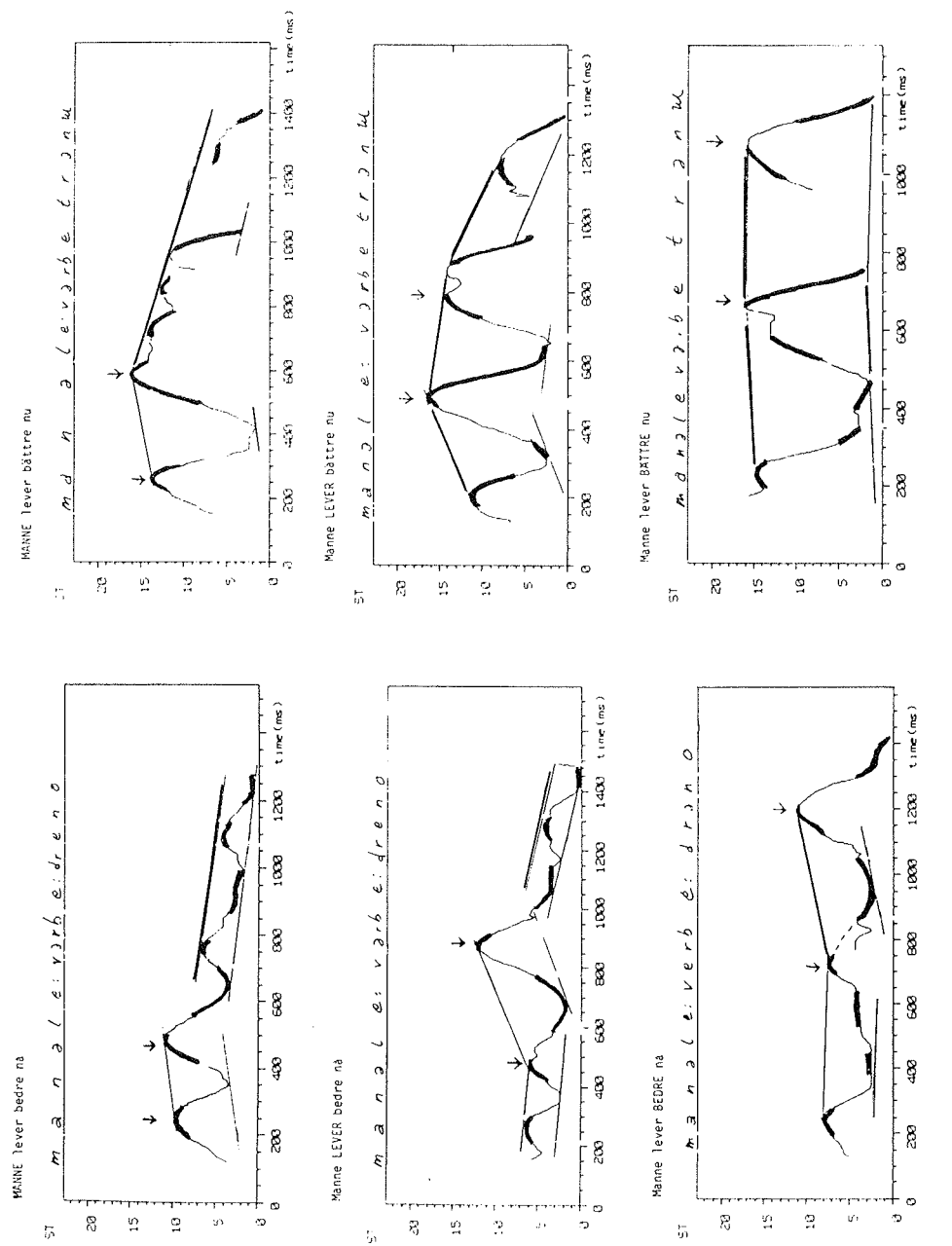
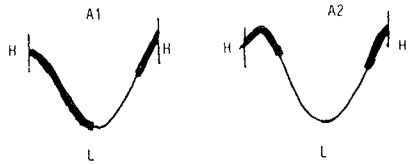


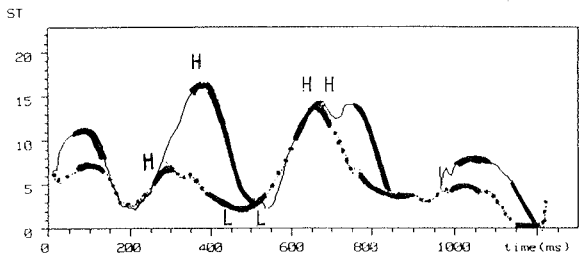
Fig. 1. Fundamental frequency tracings for typical Oslo and Göteborg utterances in semitone scale. Arrows show pivotal points. See text.

Let us now turn to sentence intonation. The focus domain, which has a larger range than the rest of the contour, is set off by two pivotal points (the arrows of Fig. 1).These pivotal points divide the intonation into smaller parts of which the last part expresses modality, here statement. To illustrate this, we can enclose the postfocal accents in a grid. The width is about one third of the total range. It is obvious that the declination of the grid is determined by the length of the phrase. In the focus domain the intonation is rising, acoustically manifested by a rising topline. Before the first pivotal point the intonation is falling or level. The size of the obtrusions from the topline reflects the degree of accentuation which is largest in focus.
In the Göteborg dialect the accents are also V-formed but the contours of the accents and positions in the V are different (see below and Fig. 1).



It is A1 which accounts for the most important difference, visual as well as auditive, from Norwegian. As for Norwegian, the last focus high is a pivotal point for the intonation. The final fall expresses statement and the rate of declination is determined by the length of the phrase. When *Manne* is in focus, our speakers have used two different strategies. Speaker LL puts *Manne* into a V of its own and deaccentuates *lever* on a high plateau. Two other informants, include *lever* in the right branch of the V.

The comparison between the dialects is summarized by the figure below. Dots denote the Oslo curve.



The figure shows the sentences in which *lever* has been in focus. The two utterances which were similar in overall duration have been brought to a common length by a linear change of the time scale. The superposition of the pitch contours demonstrates that the most conspicuous difference between the dialects can be tied to the manifestation of A1. HLH marks the focus domain. The accented syllable in Göteborg is in the left branch of the V (HL) whereas in Oslo it is at the bottom of the V-shape (LH). Another consistent feature is that the focus domain has a rising intonation manifested by a

rising topline in Oslo. In our Göteborg material the situation is less clear.

What is this V, that has been given such an important place in our description? We regard it as the phonetic domain of focus which makes it possible to give precise boundaries to the focal contour and predict this part of the intonation curve from information about focus and accented syllables. For the phonological domain of focus it would be reasonable to let it start with the focussed word.

In phonetics it is customary to differentiate phonetic and phonological entities. For segmental units and even prosodic ones like accents or tones this is regarded as a matter of course. In agreement with this we would like to make a plea for the importance of differentiating phonetic and phonological domains of prosodic units larger than a syllable. The distinction made here between the phonetic and phonological domain of focus is a case in point.

Let us end this comparison with some clarifications concerning our earlier work and also compare with the analyses of other researchers. In the analyses that were presented in the project Swedish prosody (Bruce and Gårding 1978), sentence accent was used as a term and concept instead of focus. No representation for focus domain or phrase was needed since phrase and sentence coalesced in the material designed to analyse intonation in different prosodic categories of Swedish. The postfocal high of the Göteborg dialect comparable to the last high of the focus-V in the present material was then interpreted as a sentence accent, SA, although the authors were well aware that the accent did not have the characteristic acoustic qualities of a sentence accent. In the present terminology it would be called a high pivotal point.

Thorstein Fretheim (1988) calls everything that happens from the accented syllable to the following accent a foot or an accent group (tonelagsgrupp). A similar notion is stress-group used by Nina Thorsen (1978) and Gösta Bruce (1987). The rhythmic counterpart is denoted speech tact by Eva Strangert. Fretheim calls the larger group that also includes proclitic words an intonation phrase which seems to be a phonological unit.

With the focus domain as a parameter it is possible, in this material at least, to show that the topline of the focus domain has a different inclination in questions than in statements. This is corroborated by material presented by Bredvad Jensen (1984) and Gårding (1979).

Deaccentuation has been found to be much more frequent in the Stockholm than in the Skåne dialect. Some dialect specific tendencies have been presented (Gårding 1964) and compared to tendencies in Danish (Gårding et al 1974). One result of the comparison of the dialects of Göteborg and Oslo is that deaccentuation of lexical accents is similar i.e. speaker dependent and situation dependent rather than dialect dependent.

We are also led to surmise that what in analyses of different prosodic systems appears under many different names, accent group, stress group, spreading tones, floating tones and Brückenakzente, may be the same phenomenon, i.e. phonologically deaccentuation or in a tone language neutralisation of tone,

which phonetically corresponds to an interpolation between the targets that the speaker chooses to accentuate.

References

Bruce G. 1987. How floating is focal accent? Nordic Prosody IV, ed. Gregersen K., Basböll H: 41-49

Bruce G. and Gårding E. 1978. A prosodic typology for Swedish dialects. Nordic Prosody. Travaux XIII, ed. Gårding E., Bruce B., Bannert R. Lund

Bredvad-Jensen A.-C. 1984. Tonal geography. Nordic Prosody III, ed. Elert C.-C., Johansson I., Strangert E.: 31-41

Fretheim T. and Nilsen R. A. 1988. Alternativspörsmål: opp som en löve, ned som en skinnfell. Manuskript

Gårding E.1964. Några observationer rörande tryckförhållanden i stockholmskt talspråk. Svenskans beskrivning 1, ed. C.-C. Elert, Stockholm

Gårding E. 1979. Sentence intonation in Swedish. Phonetica 36: 207-215

Gårding E. and Lindblad P. 1973. Constancy and variation in Swedish word accent patterns. WP 7. Ling. Lund University: 53-65

Gårding E., Bannert R., Bredvad-Jensen A-C, Bruce G., Nauclér K. 1974. Talar skåningarna svenska? Svenskans beskrivning 8, ed. Platzack C.:107-117

Strangert E. 1985. Swedish speech rhythm in a cross-language perspective. Umeå

THANKS to Ann-Christine Bredvad-Jensen for lending us her Oslo material!

A DANISH TEXT-TO-SPEECH SYSTEM USING A TEXT NORMALIZER BASED ON MORPH ANALYSIS.

Björn Granström
Department of Speech Communication and Music Acoustics,
Royal Institute of Technology, S-100 44 Stockholm, Sweden.
tel. (468)-7877568
Peter Molbaek Hansen and Nina Grønnum Thorsen
Institute of Phonetics, University of Copenhagen,
Njalsgade 96, Dk-2300 Copenhagen,Denmark.
tel. (451)-542211

Abstract

A Nordic cooperative project has been started to develop a text-to-speech device for the Nordic languages. The development is based on the system originally created in Stockholm. Language specific features have necessitated modifications of the original structure. For Danish, this primarily involves the inclusion of a morph based "text normalizing component". This paper presents the construction and function of the system and also discusses some preliminary use of the device.

Introduction

Speech synthesis has been a major line of research in our two departments for several decades. In Sweden, this effort has resulted in a multi-lingual text-to-speech system (Carlson & Granström, 1986), commercially available through Infovox AB. A joint effort within the project "A Nordic text-to-speech system", financed by The Nordic Committee on Disability, is aimed at making this device available to the handicapped in the Nordic countries.

Although the Nordic languages are mutually intelligible, Danish poses some special problems for a text-to-speech system because the relation between the standard orthography and pronunciation is rather complicated. To tackle this, we have included a unique component in the system that transforms words into an idealized normalized orthography. This is accomplished through a morphological analysis based on a set of moderately large morph lexica. With a limited set of rules, the result is transformed to a phonetic transcription, including stress.

In a phonetic rules component, special care has been taken to realize the prosodic structure of Danish which differs considerably from standard Swedish or Norwegian. There are also many other differences in structure such as the ample use of "stød", a kind of creaky voice unknown in the other Nordic languages but which corresponds roughly to the tonal word accent I in Swedish and Norwegian.

Morphological analysis and normalization

The inclusion of a normalization component (NC) is a deviation from the general philosophy of the KTH system which is rule based. The work done by

the NC corresponds to the work performed by certain early rules in the rule components for the other languages. There was a practical and a theoretical motivation behind the establishment at IPUC of a lexicon based NC for Danish. The practical reason was that Danish orthography corresponds very badly with pronunciation and that the number of rules in a rule based conversion system, and also the degree of arbitrariness of most of such rules, would have been prohibitive. For a more detailed description of the peculiarities of Danish orthography, see Molbaek Hansen (1983). The theoretical reason was that the use of a morph lexicon should reflect more closely the human process of reading, since we generally know the morphemes of our native language and only feel the necessity of relying on some sort of mental rule system when occasionally we come across a hitherto unknown word in a text.

The NC does two things: it supplies the correct morphological boundaries, and it normalizes the spelling of individual morphemes. A few examples illustrate this: A word like 'dal' ('valley') is pronounced with a long vowel whereas 'tal' ('number') is pronounced with a short vowel. The NC will identify these words correctly and output them as DAL and TAL, respectively, thus assigning to them a notation which is consistent - but not identical - with a phonetic transcription. The symbol represents an abstract consonant phoneme which will prevent the vowel from being lengthened and the consonant L from receiving the stød. A word like 'kvindeemancipationen' ('the emancipation of women') represents a more complex case: It is analyzed by the NC as consisting of the morpheme sequence 'kvind' (a native root), + 'e' (a native suffix) + 'e' (a latin prefix) + 'man' + 'cip' (latin roots) + 'at' + 'ion' (latin suffixes) + 'en' (a native ending), and it is output as: KVIND£mEO£iE£fMANCIPATION£pE0N. The symbol sequences £m, £i, £f, and £p represent various morphological boundaries with different phonological effects. No boundaries are inserted between 'man' and 'cip' nor between 'at' and 'ion' because such boundaries would not supply any information relevant to pronunciation. The symbol sequence E0 represents the vowel shwa. A word like 'hund' ('dog') is output as HU6ND. The sequence U6 represents a particular, abstract phoneme which resists an otherwise general phonological rule of Danish which lowers high vowels before homosyllabic nasals. Thus, the inventory of distinct symbols which may be output from the NC is considerably larger than both the number of letters in the alphabet and the number of phonemes needed to represent Danish speech. The problems of integrating the NC as such into the system have been few and small, since its output, i.e., the input to the rule system, is of the same type as the orthographic input, namely a sequence of ASCII characters.

The phonological rule component

The rule language developed at IPUC is of the same SPE type as the one used in the KTH system, and the rule component of the IPUC system could, therefore, be translated into the notation of the present system. However, owing to certain technical differences between the input scanning routines and feature interpreting procedures of the two systems, this translation could not be done in a simple rule-by-rule fashion. One difficulty arises from the fact that in the IPUC system a segment is identifiable exclusively by its feature composition whereas in the KTH system a segment is identifiable by its symbolic representation. Another main difference lies in the way a string of segments is scanned by a rule. In the

KTH system, the context is matched left to right, whereas the IPUC system starts the match at the structure to be changed and then matches the left and the right contexts. This has necessitated the reformulation of certain IPUC rules.

The phonetics of standard Danish

The vowel system is rich, with ten vowel phonemes /i e ɛ a y øœu oɔ/ which may be either short or long (the difference is phonological).Generally, the long and short vowels have identical phonetic quality, except for /a(:), o(:), (:)/, and due to a language specific variation in some of the vowels with the phonetic context, a total of 18 distinct phonetic vowel qualities must be distinguished: [i e ɛ æa a˗ɑy ø œ ɞ ɶu oɔɒʌ] plus [ə].

The consonant system is correspondingly restricted: /p t k b d g f s h v ð j m n ŋ l r/. It differs from the consonantal systems of the closest Germanic neighbours in the realization of, particularly, the stop series. Firstly, an opposition between /p t k/ versus /b d g/ is found only initially in syllables containing a full vowel (i.e., not /ə/). Secondly, the manifestation of the contrast is one of aspiration only. I.e., /p t k/ are unvoiced aspirated, and /b d g/ are unvoiced unaspirated. Both series are lenes, rather than fortes.

Stress is free and phonemic on the surface, although stress placement can to a large extent be predicted from the syllabic structure and the morphology, e.g., 'billigst, bilist' ['bilisd, bi'lisd] (cheapest, motorist). Phonetically, stress is signalled mainly through fundamental frequency variation, but (full) vowel quality and (longer) duration also contribute to the identification of stressed syllables.

A complication in the phonology and phonetics is the Danish "stød", a kind of creaky voice whose occurrence, like stress, is to a large extent predictable, but on the surface, stød versus non-stød distinguish words, like 'Møller, møller' ['møl^{l}ʌ, 'mølʌ] (a proper name, Miller).

Standard Danish intonation can be decomposed into (and synthesized from) the following components: a text contour, an utterance or sentence contour, which may be decomposed into a succession of two or more phrase contours, a stress group pattern, and a stod movement. These are all speaker controlled. Involuntary variations arise due to the intrinsic properties of the segments. These components are hierarchically organized so that components of smaller temporal scope are superposed upon and subordinate to components of larger temporal scope. Sentence intonation is signalled by the overall global course of the intonation contour, rather than by a special local (final) movement. Standard Danish lacks an obligatory sentence accent, which makes it prosodically rather simpler than, e.g., Swedish and English, see further Thorsen (1983). The synthesis incorporates the utterance, stress group, and stød components, and further adds certain segment conditioned fundamental frequency variations.

Illustrated in fig. 1 are the parameters and the acoustic output of the utterance 'Hvem har en soster der hedder Kamma?' ['vɛm' 'hɑ:' en 'søsdʌ dɑ heðʌ 'kɑma] (Whose sister is called Kamma). The corresponding transcription in the text-to-speech system looks like this:
[V'ÄMQH'A3:Q EN S'ÖSDÅ2 DA3 HEDHÅ2 K'A3MA1].

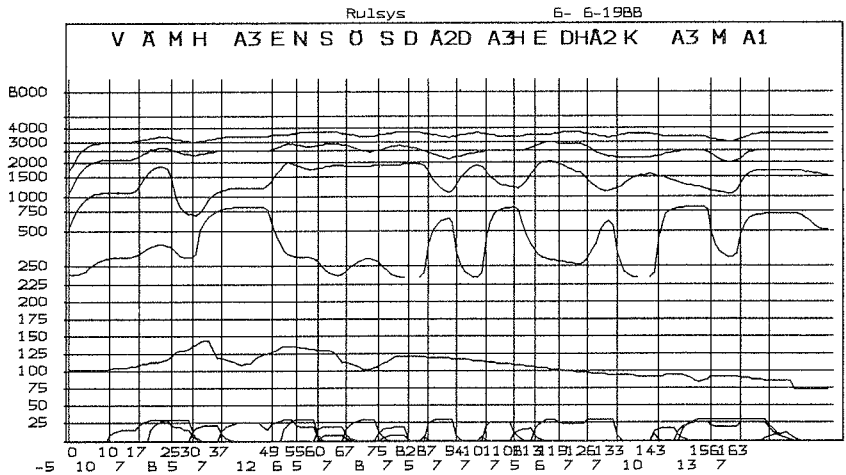


Fig. 1. The Danish utterance 'Hvem har en soster der hedder Kamma?' as produced by the text-to-speech system. The parameters are, from the bottom up: Diverse amplitude parameters, fundamental frequency, and the four lowest formants.

Concluding remarks

At this stage in the project, we have made no formal evaluation of the system. Preliminary versions of the program have been used in a project aiming at a Danish work station for visually impaired persons. Several imperfections still exist in both the text analyzing part and the phonetic realization part of the system. However, the output is unmistakably Danish and judged useful in a variety of applications. One improvement that is still needed is a somewhat faster and more reliable NC. It is a problem with the current implementation of that component that it makes quite a few wrong choices in cases of ambiguous input. We are at present developing a new version with a better performance.

References

Carlson, R. & Granström, B. (1986): "Linguistic processing in the KTH multilingual text-to-speech system", pp. 2403-2406 in IEEE-ICASSP, Vol. 4, Tokyo.

Molbaek Hansen, P. (1983): "An orthography normalizing program for Danish", Ann.Rep. of the Institute of Phonetics, University of Copenhagen 17, pp. 87-109.

Thorsen, N. (1983): "Standard Danish sentence intonation - phonetic data and their representation", Folia Ling. 17, pp. 187-220.

The CTH - Speech Database

An integrated multilevel approach

Per Hedelin, Dieter Huber, Per Lindblad *

Chalmers University of Technology, Department of Information Theory
* *University of Gothenburg, Department of Linguistics*
Göteborg, Sweden

INTRODUCTION

The scientific purpose of speech databases is essentially two-fold: (1) to provide the raw-material for investigative speech research, and (2) to provide the reference-material for simulative speech research.

Ideally, the same speech database serves both purposes simultaneously. In addition to that, it should also permit systematic comparison and exchange of speech data and analysis parameters between different databanks both nationally and internationally. To acheive these purposes, three basic requirements have to be fulfilled:
Standardization at all levels of data collection, registration, sampling and quantization, analysis, transcription, statistical and linguistic evaluation, database management, etc.
Integration of signal processing, analysis, and synthesis routines on the one hand, and between different levels of acoustical, statistical and linguistic analysis and evaluation on the other.
Adaptability, i.e. the speech material once collected and analysed to study one aspect of speech communication (e.g. the phonetic characteristics of certain speech sounds) should also be accessible for research on different aspects (e.g. long-time spectral properties), at different levels (e.g. prosodic variations), over different domains (e.g. complete texts), and from different points of view (i.e. auditory evaluation).

There is wide agreement in the speech research community on these points, as expressed lately during the databank meetings at the European Conference on Speech Technology (held in Edinburgh, September 1987), the Swedish Phonetics Conference (held in Uppsala, October 1987) and the Meeting of the Association for Computers and Humanities on Text Encoding Practices (held in New York, November 1987). Efforts in the direction of standardization, integration and adaptability have lately been reported among others in Itahashi [1], Ladefoged [2], and Pérennou [3].

APPLICATIONS

There are several applications of speech research that can share and use a large speech material. For speech coding it is important to have access to a variety of speakers and speaking conditions in order to evaluate the performance of a speech coder. In addition, modern speech coding is often based on large code-books, for instance for LPC-vectors. These code-books require very large training sets involving many speakers for adequate training. Essentially, the same code-books can be used in speech recognition for a raw classification of speech segments. In this case, the code-book represents the knowledge built into the recognizer regarding the mapping from phonetics to real signals. The purpose of the code-book is to split the signal space into subspaces that can be phonetically labelled. The validity of such a partitioning of the signal space, obviously, depends on the amount, and the quality, of the speech data available during training. Future development in both these areas is expected to strongly rely on the development of well trained and well balanced code-books. For speech recognition and synthesis it is vital to have a data-base for collecting information regarding, for instance, diphone data or intonation patterns. The speech material must, for these purposes, be carefully analysed and labelled from a linguistic, phonetic and prosodic point of view.

The CTH-SPEECH DATABASE

Our paper describes the approach taken at Chalmers University of Technology (Department of Information Theory) in building up an integrated multilevel speech database for the purpose of speech research (analysis and synthesis) and the development of speech coding techniques.

Given the research goals at our department, our material comprises today isolated speech sounds (phones and diphones) as well as short unrelated sentences and coherent texts. Data collection is, to start with, restricted to Swedish material and read speech. Registration of the speech samples was carried out under optimal conditions (sound insulated, unechoic studio) at the department of applied acoustics (CTH), using digital recording equipment (SONY PCM-1). Speakers were chosen in an attempt to minimize possible dialectal differences (Rikssvenska) and to represent a high standard of professionality in oral reading skills (radio journalists, experienced public speakers). The SAP signal analysis routines [4] are used for LPC, pitch estimation, inverse filtering, etc. Transcription of the entire material is performed interactively in 16 msec frames and comprises narrow phonetic transcription at the acoustical level (i.e. marking not only allophonic

variation, but even different phases in the production of various speech sounds) together with prosodic (accentuation, intonation units), sentence linguistic (constituent structure, parts-of-speech, frequency ratings) and textual analysis (text attribution, pause, breathing patterns, boundary phenomena).

THE PHONETIC LABELLING

All phonetic labelling is performed manually. Several automatic and semi-automatic methods have been considered and evaluated. With the aim of high reliability in the transcriptions in mind there were no doubts after the evaluation that manual classification is preferable.

Labelling is made in a ***sound segment*** domain. A sound segment is defined as an event in the acoustic signal. A continuant allophone consists of one sound segment. A glide has two sound segments. Stops have 4-5 sound segments. The sound segments have been chosen to ensure a close correspondence to the allophonic domain. The labelling distinguishes 52 allophones, silence included. The vowel labels are, following Elert's notation for Swedish: [ɑː], [a], [eː], [e], [ə], [iː], [ɪ], [ʊː], [ʊ], [ɯː], [ɵ], [yː], [ʏ], [oː], [ɔ], [ɛː], [ɛ], [æː], [æ], [øː], [ø], [œː] and [œ]. The consonants considered are [b], [d], [ɖ], [f], [g], [h], [j], [k], [kʰ], [l], [ɭ], [m], [n], [ɳ], [pʰ], [p], [r], [ʐ], [s], [ʂ], [ʃ], [tʰ], [t], [ʈ], [ç], [v]. Finally there is silence [ɾ], as a technical member of the allophone family.

Each of the 120 sound segment labels used is a part of a careful pronunciation of one of the 52 allophones. Labels are written in technical notation to faciliate computer handling of labels. The long vowel [iː] has the corresponding sound segment labels I11 and I12. I11 is continuant, whereas I12 corresponds to the final part in the [ij] pronunciation of /i:/. The stop [pʰ] has four corresponding sound segments, with labels P11, P12, P13 and P14. Here, P12 denotes oclusion, P13 the burst and P14 the aspiration.

The principle of using sound segments for the manual transcription allows considerable freedom. The method as such does not assume or imply that all realizations of a particular allophone use the same segments. In the actual labelling, the voice-bar of a [b], i.e. B12 can be omitted whenever needed to take one example.

SIGNAL PROCESSING

The standard processing done on all material is LPC-analysis and pitch extraction. Whenever needed, several other interactive analysis tools are available, such as spectral analysis of single frames or spectrograms of several seconds of speech. For close examination of the glottal excitation, there is an accurate and automatic inverse filtering routine.

LPC-analysis is performed using 48 ms Hamming windowed segments at an update rate of 16 ms. Adaptive pre-emphasis is employed. The auto-correlation approach to LPC is used. This is a fast and robust method that has almost no problems attached. The result of the LPC-analysis is a 10^{th}-order LPC-filter. The coefficients of the filter can be converted to the formant domain if formant estimates are required for the manual labelling.

Considerable effort has been spent on selecting an accurate and robust pitch extractor for the pitch analysis. Several different schemes have been evaluated including time domain methods such as the Gold-Rabiner algorithm, auto-correlation domain methods such as the SIFT-algorithm as well as cepstral methods. All methods tested have proven to generate errors, typically 95-99 % of the frames are correctly handled. Among the prominent problems are incorrect voicing decisions and pitch doubling. For female voices some methods occasionally give pitch estimates that are one octave below the true values. Some methods, such as the Gold-Rabiner algorithm, are more noisy in the sense that pitch estimates have low precision whereas auto-correlation methods can measure pitch with an accuracy much better than 0.5 Hz.

The final choice for the pitch extractor is an extended and improved version of the SIFT-algorithm. The original SIFT approach decimates the speech signal in order to decrease the computational burden. As an artefact, accuracy is lost in the pitch values. The basic SIFT procedure was augmented with a final analysis on the non-decimated speech signal in a vinicity of the preliminary pitch value. As a result, high precision pitch estimates were obtained with only a slight increase in computational complexity. The simple and straightforward voicing discrimination of the SIFT-algorithm was completely revised. A statistical hypothesis test was introduced. This algorithm uses speech level, spectral tilt, the normalized peak of the auto-correlation function, zero-crossing rate, rate of pitch change and formant information in order to form a composite voicing decision. In the evaluation, the pitch extractor thus designed outperformed all other methods included in the tests, in particular as regards voicing errors.

The final result of the pitch extraction contains voicing errors with typical error rates in the interval 1-2 %. Or, in other words, each sentence typically contains several incorrect voicing decisions. Manual corrections were made on a routine basis during the manual labelling step.

STATUS

The CTH-Speech Database comprises today about 2.5 hours of sampled speech data (ca 550.000 16-msec-frames). Signal processing and sentence linguistic/prosodic/textual analysis has been performed for nearly two thirds of this material (ca 320.000 frames = 1.4 hours). Narrow phonetic transcription has so far been completed for ca 42.000 frames (approx. 11 min). Further signal analysis and linguistic-prosodic-phonetic transcription is ongoing.

REFERENCES

[1] S. Itahashi (1986) A Japanese Language Speech Database, ICASSP 86, Tokyo, pp.321-324

[2] P. Ladefoged (1987) Revising the International Phonetic Alphabet, XIth ICPhS, Tallinn, Se.64.5.1

[3] G. Pérennou (1986) B.D.L.E.X.: A Data and Cognition Base of Spoken French, ICASSP 86, Tokyo, pp.325-328

[4] P. Hedelin (1986) Manual for SAP-tasks, CTH Technical Report No 5

[5] C-C. Elert (1966) Allmän och svensk fonetik, Gleerup. No 5

RECOGNITION OF PROSODIC CATEGORIES IN SWEDISH: RULE IMPLEMENTATION

David House, Gösta Bruce, and Lars Eriksson
Lund University, Department of Linguistics and Phonetics
Francisco Lacerda
Stockholm University, Department of Linguistics and Phonetics

INTRODUCTION

This paper represents a status report from an ongoing joint research project shared by the Phonetics Departments at the Universities of Lund and Stockholm. The project, "Prosodic Parsing for Swedish Speech Recognition", is sponsored by the National Swedish Board for Technical Development and is part of the National Swedish Speech Recognition Effort in Speech Technology. The primary goal of the project is to develop a method for extracting relevant prosodic information from a speech signal. We hope to devise a system which from a speech signal input will provide us with a transcription showing syllabification of the utterance, categorization of the syllables into STRESSED and UNSTRESSED, categorization of the stressed syllables into WORD ACCENTS (ACUTE and GRAVE) and categorization of the word accents into FOCAL and NON-FOCAL accents. We also hope to be able to identify JUNCTURE (connective and boundary signals for phrases). We are currently working with a restricted material of 20 prosodically varied sentences spoken by two speakers of Stockholm Swedish.

The type and structure of the information to be presented to the recognizer has been based on a series of mingogram reading experiments (see House, et al. 1987a, 1987b). Descriptive rules were then formulated and tested using two non-expert mingogram readers.

Our scheme for automatic prosodic recognition can be broken down into three main steps. First, intensity and fundamental frequency are extracted from the digitized signal. Second, intensity relationships and fundamental frequency information are used to automatically segment the utterance into "tonal segments" which ideally correspond to syllabic units. The prosody recognition rules are then applied to these tonal segments giving us prosodic categories as the output of the system.

AUTOMATIC SEGMENTATION

The automatic segmentation component of the recognition scheme has been designed using intensity measurements in much the same way as that described by

Mertens 1987. Similar algorithms have been described by Mermelstein 1975, Lea 1980, and Blomberg and Elenius 1985. In short, the algorithm uses relationships between maximum and minimum values of both filtered and unfiltered intensity curves to accomplish a broad segmentation. A -3dB threshold prior to the intensity maximum of each segment is applied to locate the onset of the vowel for each syllabic nucleus. The end of the tonal segment is marked at the point where voicing ends prior to the next vowel onset, or if voicing continues, the end of the tonal segment will coincide with the next vowel onset. These tonal segments comprise the basic syllabic units for prosodic recognition.

RULE IMPLEMENTATION

Our preliminary strategy has been to reduce the information available to the recognizer in an attempt to attain the best results with the least possible amount of information. In this way we hope to isolate the most salient cues and build upon them to improve our results. It is clear from our descriptive rule testing that fundamental frequency information is crucial to the recognition of prosodic categories, especially word and focal accents. Evidence from our rule testing indicated that an important area of Fo information is the average Fo level during the first 30-50 ms after vowel onset. This also corresponds to results from speech perception experiments (House 1987). Another important area of information in the rules is the syllable final Fo-level. We therefore decided to assign two Fo values to each tonal segment, average Fo during the first 30 ms (B) and average Fo during the last 30 ms of each tonal segment (E). This amounted to a linear stylization of the tonal contour. In order to test this stylization and see how much prosodic information is lost, we synthesized both speakers' productions of ten sentences using LPC synthesis with the stylized tonal contour as the pitch parameter. In several informal listening tests, the majority of the stylized sentences could not be distinguished from their original counterparts on the basis of intonation alone. These results give further strength to our preliminary method of reducing Fo information.

To incorporate Fo relationships between tonal segments, each segment is assigned two additional Fo values representing the high (H) and low (L) from the preceding (stylized) segment. Finally, two more values are assigned to each segment representing amount of (stylized) Fo change (C) during the segment and total duration (T) of the tonal segment.

In a first implementation of the rules using these six values, conditions for three word-accent categories (grave, acute+focal and acute+non-focal) were formulated based on the descriptive rules and on actual measurements of these

values from the categories in question in ten test sentences. The conditions are listed in table 1.

Table 1. Rule conditions for three word-accent categories.

Grave	**Acute+focal**	**Acute+non-focal**
C≤-20 Hz	C>5 Hz	-30 Hz<C<0 Hz
T>150 ms	T>100 ms	T>80 ms
B≥H-5 Hz	E≥H	B<H
E<L-5 Hz	B>L-5 Hz	E<L
	(B+E)/2>(H+L)/2	B<(H+L)/2

Where B=Fo beginning, E=Fo end, C=Fo change, T=duration of tonal segment, H=Fo high in preceding tonal segment, L= Fo low in preceding tonal segment.

A recognition routine checks each condition against the six values for each tonal segment. For each true condition, the segment receives one point for the category containing the condition. When all conditions are checked, the category having the most points is assigned to the segment. If two or more categories receive the same score, the following rule hierarchy applies: grave, acute+focal, acute+non-focal.

Finally a relative score threshold can be set where if the highest relative score does not reach the threshold, the syllable is assigned the category UNSTRESSED. If the score reaches the threshold, the category STRESSED is assigned by implication.

RESULTS

The automatic segmentation algorithm successfully detected 168 of 178 syllabic nuclei in ten test sentences. Five extra segments were added by the algorithm rendering a detection score of 92%. Four of the five extra segments were caused by a dental nasal [n] following the vowel. The vowel onset was not as successfully detected in all cases, especially when the vowel was preceded by a nasal or a liquid. In these instances the -3dB level often occurred in the middle of the consonant. The rule conditions for the three prosodic categories, with the relative score threshold set at 0.75, gave the following results: GRAVE 12 recognized of 13 occurrences, ACUTE+FOCAL 11 of 13, ACUTE+NON-FOCAL 7 of 10 and STRESSED 34 of 37. The category UNSTRESSED, however, was recognized in only 39 cases of 82 occurrences. In most cases, the missed unstressed syllables were categorized as ACUTE+NON-FOCAL.

DISCUSSION

Our preliminary results from the segmentation algorithm are promising as is the success of the rule implementation in separating the three accent categories tested. The major problem is of course that half the unstressed syllables are still categorized as stressed. To a certain extent, this reflects the results of the expert reader who identified 100% of the stressed syllables but only 73% of the unstressed. We hope to improve the results by using a seventh value representing the vowel duration of each tonal segment. It might also prove useful to replace the value for tonal-segment duration with a value representing duration from vowel onset to vowel onset. These new values will be more useful if we can improve detection of vowel onset locations. We are currently investigating the use of intensity curves from different filter bands as an aid to vowel onset identification.

Other problems such as identifying juncture cues and separating these cues from word-accent cues may necessitate the use of additional values for each tonal segment. For example, maximum and minimum Fo values could be added. Our recognition scheme will enable us to test these changes as well as further additions to the rules.

REFERENCES

Blomberg, M. and Elenius, K. 1985. Automatic time alignment of speech with a phonetic transcription. In Guerin and Carré (eds.), 357-366. Proceedings of the French Swedish Seminar on Speech, Grenoble.

House, D. 1987. Perception of tonal patterns in speech: implications for models of speech perception. Proc. of the Eleventh International Congress of Phonetic Sciences. 1:76-79. Academy of Sciences of the Estonian S.S.R. Tallinn.

House, D., Bruce, G., Lacerda, F., & Lindblom, B. 1987a. Automatic Prosodic Analysis for Swedish Speech Recognition. Proc. European Conference on Speech Technology, Edinburgh 1987: 215-218.

House, D., Bruce, G., Lacerda, F., & Lindblom, B. 1987b. Automatic Prosodic Analysis for Swedish Speech Recognition. Working Papers 31, Department of Linguistics, Lund University: 87-101.

Lea, W. 1980. Prosodic aids to speech recognition. In Lea (ed.) *Trends in Speech Recognition* , 166-205. Prentice-Hall, N.J.

Mertens, P. 1987. Automatic segmentation of speech into syllables. Proc. European Conference on Speech Technology, Edinburgh 1987: 9-12.

Mermelstein, P. 1975. Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am.* 58 (4) 880-883.

Laryngealization as a Boundary Cue in Read Speech

Dieter Huber

Chalmers University of Technology, Department of Information Theory, Göteborg
Lund University, Department of Linguistics

The function of laryngealization as a potential boundary cue at sentence, paragraph, and turn junctures has been noted by several researchers [1],[2],[4].[5]. In these studies, laryngealization is mostly treated as an either-or-phenomenon that occurs as a complementary feature together with other, more prominent boundary cues (pause, pre-boundary lengthening, phonological blocking, non-level intonation) to mark an extra degree of horizontal break. No differentiation between kinds of laryngealization has been suggested in the literature, and related terms such as creak, creaky voice, glottal fry, and glottalization are usually employed indiscriminately and in an synonymous way.

In a study of text intonation in Swedish [3] four different patterns of laryngealization were observed to occur consistently at different kinds of junctures. These patterns are examplified in figure 1, and will be denoted in the following as *creaky voice, creak, diplophonic phonation,* and *glottalization.*

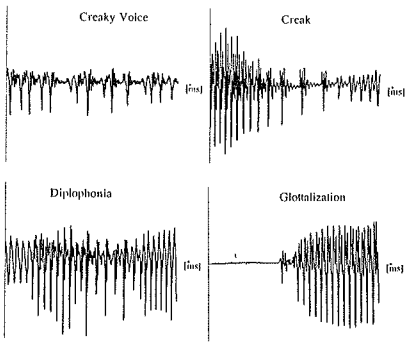


Figure 1 Patterns of laryngealization in the reading of coherent texts by four Swedish subjects.

Creaky voice and creak occurred predominantly at pre-boundary (utterance final), and glottalization exclusively at post-boundary (utterance initial) locations. Diplophonic phonation, i.e. alternations between strong and weak glottal excitations, were observed both at utterance internal and utterance final junctures.

Although each of the four speakers participating in this study (2 female, 2 male; 2 radio journalists, 2 experienced public speakers) used each of the four modes of laryngealization at some time or another during the text reading sessions (3 texts: 1 narrative, 1 descriptive, 1 argumentative; comprising a total of 2610 running words), there are clear indications in our data for some kind of preference patterning both with respect to speaker variability and kind of boundary.

Inter-speaker variability

It appears from these data (1) that both female speakers make more frequent use of laryngealization as a boundary marker than their male counterparts, and (2) that creak and diplophonic phonation are the two modes that were preferred by our female subjects, while the two men (in as much as they use laryngealization at all) apparently prefer creaky voice. The use of glottalization at post-boundary positions does not appear to differ significantly between our male and female speakers.

Boundary locations

The correlations of different patterns of laryngealization with different kinds (and/or degrees) of textual, sentential and prosodic junctures (paragraph, sentence, clause, phrase, constituent, intonation unit, declination resetting, speech inhalation pause) indicate the following general tendencies:

Creak appears to be closely associated with intonation unit boundaries that are not marked by a low terminal F_0 contour.

Creaky voice was found predominantly at intonation unit offsets with a low terminal F_0 contour.

Diplophonic phonation occurred either at intonation unit offsets with a low terminal F_0 contour (female), or in utterance internal positions at word junctures between adjacent voiced (mostly vowel) sounds. At intonation offset locations, diplophonic phonation occurred together with other vocal features (e.g. breathiness) or in the transition before a period of aperiodicity and/or devoicing.

Glottalization was most often used to mark sentence internal clause boundaries, irrespective of whether they coincided with the onset of a new intonation unit or not.

It thus appears that creaky voice, creak and glottalization serve as genuine boundary cues, the use of which varies between different speakers. Diplophonic phonation, on the other hand, occurs predominantly in our material as a transition phenomenon between adjacent

coarticulated voiced speech sounds and might thus be a consequence of source-tract interaction rather than an independently controlled laryngeal gesture used as a boundary cue.

REFERENCES

[1]G.FANT, L.NORD and A.KRUCKENBERG (1987) Segmental and prosodic variabilities in connected speech. An applied data-bank study, XIth ICPhS, Tallinn, vol 6, pp 102-105.

[2]E.GÅRDING (1967) Internal juncture in Swedish, Gleerup. Lund.

[3]D.HUBER (1988) On the communicative function of voice in text-to-speech, forthcoming.

[4]J.KREIMAN (1982) Perception of sentence and paragraph boundaries in natural conversation *Journal of Phonetics* **10**, pp. 163-175.

[5]I.LEHISTE (1979) Perception of sentence and paragraph boundaries, in: B: Lindblom & S. Öhman (eds.) Frontiers of Speech Communication Research.

CLINICAL USE OF NASAL AIRFLOW IN ASSESSMENT OF THE VELOPHARYNGEAL MECHANISM.

Birgit Hutters, Inst. of General and Applied Linguistics, Univ. of Copenhagen
Kirsten Brøndsted, Cleft Palate Departm.,Inst. for Speech Disorders,Hellerup

When inappropriate in degree or timing, coupling between the nasal cavity and the rest of the vocal tract may result in two related symptoms, namely nasality and nasal emission of air. Nasality refers to the altered resonance of voiced sounds caused by the inclusion of the nasal resonance system, while nasal emission of air refers to the escape of air via the nasal route during speech sounds produced with a velopharyngeal leakage. Nasality and nasal emission of air are central concerns in cases of velopharyngeal insufficiency. The clinical routine assessment is - in Denmark at least - normally based on: 1) the perceptual impression of speech, 2) oral mirror inspection, and 3) various traditional tests. However, important limitations are inherent in these methods: re 1) the relationship between the perceptual judgment of speech produced by speakers suffering from inadequate velopharyngeal function and its physiological causes is very complex, and re 2) and 3) the traditional examination procedure and tests can only be applied to non-natural speech - typically isolated sustained speech sounds - or to non-speech tasks, and there is general agreement that these conditions are poor indications of velopharyngeal behaviour in natural speech.

Thus, there is obviously a need for supplementary methods, which more directly inform about the behaviour of the velopharyngeal mechanism in (quasi-) natural speech. There are various possibilities, but most of them are not practicable as tools in the clinical routine assessment due to a complicated application procedure and/or because they are very expensive. One possibility, however, is registration of nasal airflow in order to detect and quantify nasal emission of air. Unfortunately, the relationship between the amount of nasal airflow and the degree of opening at the velopharyngeal port is not simple, since the amount of flow depends not only on the magnitude of the opening but on the behaviour of the whole speech apparatus - including the respiratory mechanism. This is a very important point when we deal with cleft-palate speakers, as they very often exhibit a deviant glottal, subglottal, and supraglottal behaviour, when they attempt to camouflage or compensate for their speech handicap. Therefore, in order to control these other factors, oral airflow and oral pressure should be recorded simultaneously with nasal airflow, if it were not for the complicated recording procedure. Thus, the question is whether nasal airflow - as a single measure - can be a useful tool in the assessment of the velopharyngeal mechanism in determining the choice of treatment and in the subsequent evaluation of this

treatment.

Nasal airflow may be monitored by various instruments based on different principles,but irrespective of principle, the requirements to an optimal flow meter are the following: linear registration, good frequency response, separation of expiratory and inspiratory airflow, non-sensitivity to local tubulence, independence of temperature, and a stable zero-level. The Cleft Palate Department at the Institute of Speech Disorders in Hellerup (abbr. CPD) possesses an Exeter Nasal Anemometer. However, this instrument fulfil none of these requirements, since it is based on the simplest version of the hot-wire principle. But it is a cheap and handy little box, which are important qualities seen from a clinical point of view.

In order to answer the question whether nasal airflow - with special reference to the Exeter Anemometer - can be a useful tool in the assessment of the velopharyngeal mechanism, wer have - at present - recorded and analysed nasal airflow from 1) 16 Danish speakers suffering from velopharyngeal insufficiency and from 2) a normal group comprising 10 speakers. The speech signal has also been recorded. All speakers have been examined by an ENT-doctor and by a speech therapist. For practical reasons, it has not been possible to match the two groups. For the patient's sake and due to the limited time available for the recordings, the test material is minimized to consist of 1) isolated, sustained speech sounds [i α u s m] and 2) the words 'pige' (*girl*), 'side'/'sure' (*side/acid*), 'lampe' (*lamp*) said in the frame sentence 'jeg siger ... i dag' (*I say ... to-day*). The isolated oral speech sounds were included for comparison with other investigations published in the literature - all dealing with non-Danish speakers - and with the normal routine examinations based on these same sounds. The Exeter Anemometer is delivered with a standard calibration curve, which seems to be only roughly reliable. Therefore, [m] was included as a relative reference in case the calibration should prove to be useless. The test words include a stop consonant, a sibilant, and a stop following a nasal consonant, as these sound categories are the most sensitive to velopharyngeal insufficiency. The test material was recorded several times for each speaker in order to observe the variation in airflow from token to token, which is supposed to be a powerful cue in the assessment of the velopharyngeal mechanism.

The recordings were made at the CPD by a speech therapist. As she has access only to an AM-tape recorder, two FM-tape adapters have been constructed by S.E. Lystlund and P. Dømler at the phonetics laboratory - one for the recording at the CPD, and one for the reproduction at the phonetics laboratory, where LP-filtered nasal airflow traces and standard curves for

segmentation are made. At the time being, the data processing is performed 'by hand'.

The figures below present nasal airflow traces obtained from two normal speakers (to the left) and from two patients (to the right) showing the inter-speaker variation within the two groups. As expected, the normal speakers in most cases show no nasal airflow during oral sequences as illustrated in fig.1. Only with low vowels - as it appears from [α] - more than a third of the normal speakers produce nasal airflow, and the amount of flow is of the same order of magnitude as seen in the patients. However, some normal speakers produce a small amount of nasal airflow during sequences containing only non-low oral vowels and obstruents, which is exemplified in fig.2.

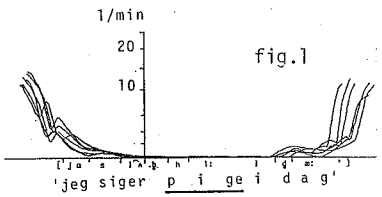
As for the patients, the airflow may be considerable and may show a pattern which tends to be similar to the pattern of intraoral pressure produced by normal speakers as shown in fig.4. In these patients the nasal airflow pattern reflects mainly the supraglottal articulation due to a very poor velopharyngeal function. With other patients the amount of airflow is smaller and does not show this 'pressure'pattern, which indicates a better velopharyngeal function - provided that it is not due to a compensatory or camouflaging articulation. The patient presented in fig.5 shows only a small amount of nasal airflow, which does not deviate very much from the normal speaker seen in fig.2. However, other parts of the material obtained from this patient show a considerable degree of airflow and an appreciable inconsistency from token to token, which reveals an inconsistent velopharyngeal mechanism. This is exemplified in fig.6, which should be compared to fig.3 showing the same normal speaker as in fig.2. Finally, patients may show airflow traces more normal in the isolated sounds than in the sentences - or vice versa - which also indicates an instable velopharyngeal function.

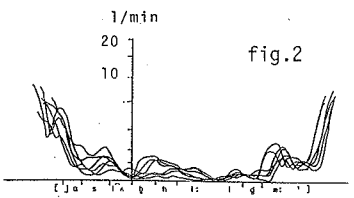
This kind of observations leads us to conclude that nasal airflow can distinguish speakers with velopharyngeal insufficiency from normal speakers, and it can serve to differentiate fair, poor, and very poor velopharyngeal mechanisms. Thus, nasal airflow - as a single measure - is a useful tool in the assessment of the velopharyngeal mechanism. But it should be emphasized that the patient's speech always has to be controlled in order to prevent misinterpretation of the airflow traces. Further, with the Exeter Anemometer one has to take into account especially its poor frequency responce and the non-separation of in- and expiratory airflow.

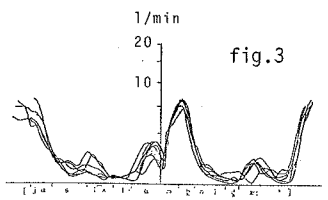
It was mentioned above that the Exeter Anemometer is a cheap and handy instrument. However, to produce airflow curves and curves of the speech signal involves quite a number of expensive accessory instruments, which are

normally not available to speech therapists, and thus, the Exeter Anemometer can no longer be considered a cheap and handy tool. In all fairness it should be mentioned that a more advanced - and a more expensive - version of the Exeter flowmeter is available. With this version the signals are recorded on a small cassette, which can be mailed to a processing centre in England, and from the centre the speech therapist receives a dual-trace chart containing speech and airflow signals (which, howver, do not seem very informative to us). We are convinced that establishing a Nordic centre for signal processing, comprising all kind of signals which may be relevant for the clinical assessment, would be received with open arms.

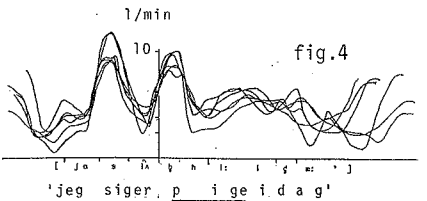
NORMAL SPEAKERS

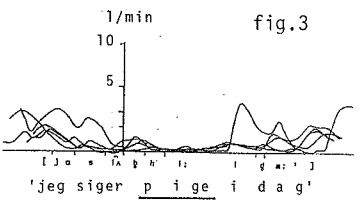


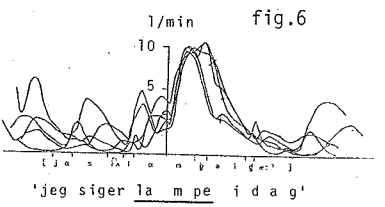




PATIENTS







A FORMANT-BASED MODEL FOR PREDICTING PERCEPTUAL DISTANCES BETWEEN SWEDISH STOPS

Diana Krull
Institute of Linguistics
University of Stockholm

This paper constitutes a part of a larger investigation (Krull, 1988) whose aim is to evaluate the capacity of different perception models to predict listeners' responses. Such models are needed especially in automatic speech recognition.

The investigation is based on the notion of a perceptual space where distances are inversely related to perceptual similarity, that is, the shorter the distance, the greater the similarity. The perceptual distances should be predictable, given acoustic data and a good perception model.

To obtain a quantitative measure of the perceptual (dis)similarity, listening tests were performed using fragments of nonsense words of the form $V_1C:V_2$ where V=[ɪ,ɛ,a,ɔ,ʊ] and C=[b,d,ɖ,g] The tapes with the stimuli were presented to 20 listeners whose task was to try to identify the consonant. The greatest number of confusions were obtained using stimulus fragments of 26ms beginning at the moment of consonant release. The results were accounted for in 25 confusion matrices, one for each vowel context. Three examples of the matrices are shown in Fig.1, for a complete set of matrices see Krull (1987, 1988). The confusions displayed a regular pattern which was clearly dependent on the vowel context, especially the following vowel. For example, /g/ before a front vowel was often confused with a dental or a retroflex but seldom with a labial, when followed by a back vowel, however, /g/ was often confused with a labial but almost never with a dental or retroflex. The perceptual distance is thus short between velar and dental in front vowel context and long between velar and labial. In back vowel context the reverse is true.

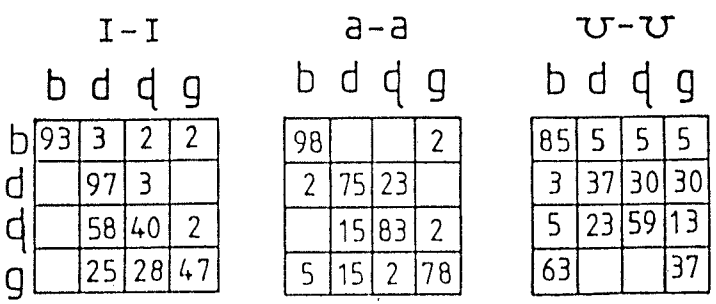


Fig.1 Three examples of confusion matrices

The listeners' confusions were now plotted as a function of the calculated distances. The resulting plot contained much noise in the form of over and underestimations of perceptual confusions. Part of these were due to asymmetries in the listeners' responses - the same calculated distance would thus be related to different percent confusions, as for example in the case of b→g and g→b in the /ugu/ stimulus (see Fig.1). An analysis of the asymmetries showed that they were dependent on the acute-grave dimension of the following vowel in a regular way and consequently predictable. The listeners' answers were therefore temporarily symmetrized by calculating the mean values of the confusions in the two directions. Secondly, there were differences in the formant-based distances depending on the following vowel: all distances were relatively short before front vowels, and longer before back vowels. Also, the calculated regression lines for percent confusions as a function of acoustic distance had a steeper slope with stimuli with front V_2 than with back V_2. Thirdly, in calculating distances, reference values with the same V_2 as the stimulus were used. Listeners may not have recognized the vowel and may have used references with another - for example neutral - vowel context.

Regression analyses were performed taking all these aspects into account. The calculations were performed separately for each of the three groups: stimuli with a front vowel, /a/ and back vowel as V_2. An additional listening test was performed, showing that V_2 could not always be recognized, especially, back vowels after a dental or retroflex stop were perceived as front or neutral vowels. Experimenting with different reference values showed that those with V_2=/a/ gave the best correlation to perceptual confusions except for labial and velar stimuli with back V_2 where the vowel could be clearly recognized. In these cases references with the same V_2 as the stimulus were used.

The results now showed that for stimuli with V_2=/i,e/ or /a/ the confusions diminished with distance up to about 2 Bark and remained unaffected by further increasing distance. With back V_2 too, large distances gave few confusions. However, there were several labial-velar pairs which were relatively seldom confused in spite of a short distance between them - listeners must have used some additional cue(s). A possible cue in this case was the length of the noise segment after consonant release: it is known to be long for velars stops and short for labials. The noise burst was therefore measured for all stimuli and the differences in its length between stimulus and reference were calculated. Thereafter, these differences were included in the distance measure using the equation

$$Dm_{i,j} = \sqrt[p]{(w_1 * Df_{i,j})^p + (w_2 * Db_{i,j})^p} \qquad \text{Eq.(3)}$$

where Dm is the modified distance, Df the formant based distance and Db the difference in burst length, i and j different stimuli, w_1 and w_2 different weighting factors and p a variable. In this case, w_1=1.0, w_2=.1 and p=2 gave the best results. (That means that 10ms was given about the same perceptual weight as 1 Bark.)

This new distance measure gave markedly better correlations between acoustic distance and percent confusions for stimuli with back V2, and a slight improvement for the other stimuli (Fig.3).

What acoustic properties correspond to these perceptual distances? I found a possible example of such properties in Fant (1973) where F_2 and F_3 at CV-boundaries were plotted against each other, C in this case was /b,d,g/ followed by nine different vowels, the syllables were read by a male Swedish speaker. The F_2-F_3 points for /g/, for example, were near those of dentals and retroflexes in front vowel context, and near labials in back vowel context. A corresponding plot with the material of the present investigation showed similar acoustic distances (Fig.2). However, the overlap between F_2-F_3 points for dental and retroflex consonants appeared to be too great in relation to the confusions between these two places of articulation, therefore F_4 was added - the frequency of this formant constitutes an important difference between the dental and the retroflex place of articulation. The formant frequencies had been measured in Hertz. In order to get distances better corresponding to what is received by the human ear, the frequencies were first converted to Bark, using the equation from Traunmüller(1983)

$$z=(26.81*f/(1960+f))-.53 \qquad \text{Eq.(1)}$$

where f is the frequency in Hz and z the value in Bark corresponding to that frequency. Thereafter, the Euclidean distances in the three-dimensional formant space were calculated according to the equation

$$D=\sqrt{(\Delta F_2)^2+(\Delta F_3)^2+(\Delta F_4)2+)^2} \qquad \text{Eq.(2)}$$

where F_n is the critical band rate in Bark of the n-th formant.

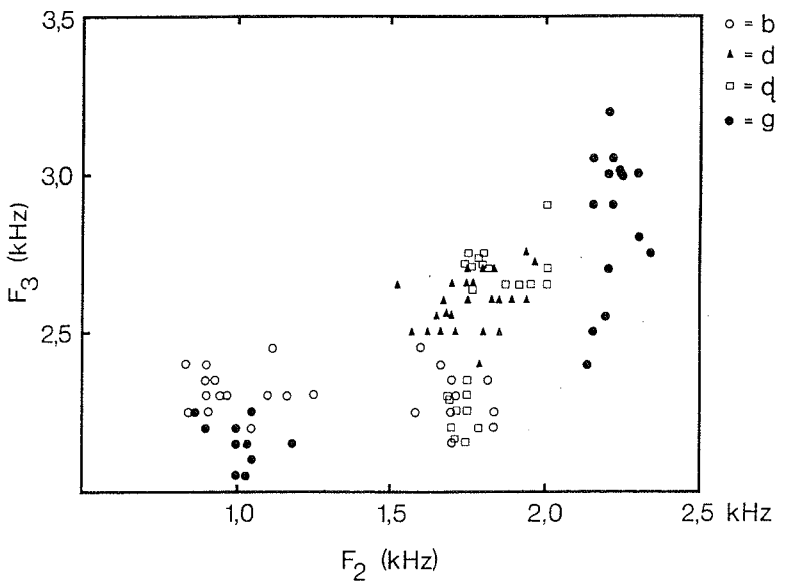


Fig.2 F_2 and F_3 at the CV-boundary.

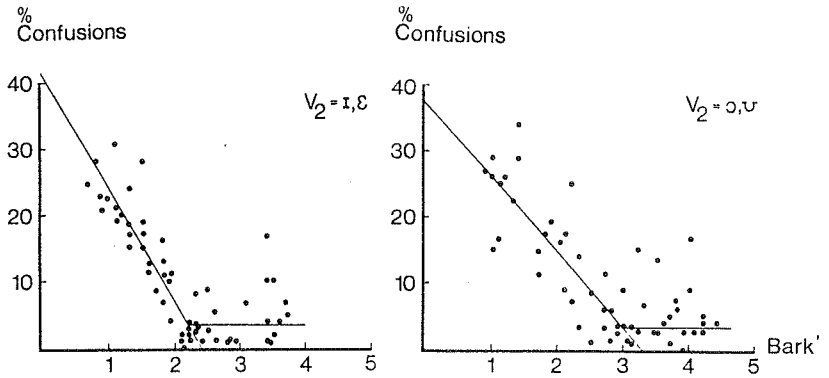


Fig.3 Percent confusions as a function of calculated distances based on formant frequencies in combination with differences in burst length (Bark').

Finally, the asymmetries in the listeners' confusions were calculated using formant values (Bark) at the CV-boundary and in the middle of V_2 as predictors (see Krull, 1988 p.108 for details), and percent confusions for each stimulus was predicted. The correlation coefficient between the predicted and observed percent confusions was r=.85. The predictions were calculated separately for the V2 contexts front, /a/ and back. Normalizing the V_2-dependent differences in the formant-based distances (Krull,1988, p.92f) made it possible to calculate percent confusions also for all vowel contexts together, which resulted in about the same correlation between the predicted and observed values, r=.86.

The formant-based model is an improvement compared to the spectrum-based one described in Krull (1987). Now percent confusions can be predicted not only for mean values for each V_2-context but also for single stimuli. In particular, the asymmetries in the listeners' answers can be predicted. It is also more satisfactory to be able to perform the calculations for stimuli with all vowel contexts together.

REFERENCES

Fant, G.(1973), "Stops in CV-syllables", in G. Fant: Speech Sounds and Features, MIT Press

Krull, D.(1987), "Spectrum and dynamics in the perception of stop consonants", Papers from the Swedish Phonetics Conference, RUUL 17, Department of Linguistics, Uppsala University

Krull, D. (1988), Acoustic Properties as predictors of perceptual responses, PERILUS VII, Institute of Linguistics, University of Stockholm.

Traunmüller, H. (1983), "Analytical expressions of a tonotopical sensory scale", (Part of a doctoral dissertation), Inst. of Linguistics, University of Stockholm

Presentation of a Method for Comparing Child and Adult Phonological Forms

Ingrid Landberg and Liselotte Roug
Department of Linguistics
University of Stockholm

In this international project we need criteria in common with the other groups to decide what shall be counted as a word. This is a presentation of a procedure for determination of words prepared by Dr. Marilyn Vihman at Stanford University and Dr. Lorraine McCune at Rutgers University. In the Swedish branch of the project we have used this procedure with certain modifications.

In this project we concentrate on early words based upon adult forms, i.e. we do not consider 'protowords'. In early child language there is on one hand phonetic forms mirroring an attempt to produce an adult form within the constrictions of child language reduction rules and phonological processes such as fronting, stopping, gliding, and consonant/vowel-harmony , on the other hand there is an intentional use of this phonetic form in due context, submitted to semantic rules of child language.

The adult wordform may be nominal as well as onomatopoetic, illustrating animal and vehicle sounds or eating. There are two basic criteria which should be met by the child form. One is that the utterance of the child must have a minimal phonological similarity to the adult form. By minimal phonological similarity we mean at least a two segment match. The words must have been used in a plausible context related to an event or object.

If these functional and phonological conditions are met, the utterance achieves the status of a 'word candidate' and will be submitted to a thorough examination before the final decision regarding word status is taken.

After an evaluation including a very strict phonological scoring procedure as well as observations and estimations of determinative context and other evidential factors such as frequency of use and stability of form and context the word status of the utterance is settled. This procedure is described by Table I.

Table I

BASIC CRITERIA

Plausible form? Plausible context?

W O R D C A N D I D A T E

FORM	CONTEXT
Exact match	Determinative context
Complex match	Identification by the mother
Prosodic match	

ADDITIONAL EVIDENCE
Frequent use in and/across sessions
Stability in context and form

W O R D

Concerning phonological matching there are forms that either show an exact match to the adult form or a complex match, i.e. more than the minimum two segments of the adult form. There are also forms that show a prosodic match in fundamental frequency or voice quality.

An explicit scoring-procedure is set for phonological matching, where bonus is given for agreement with the corresponding adult form and penalty for certain deviances, could they not be related to common processes of child language.In that case they will be accepted without penalty.

We find in this elaborate scoring system a sympathetic aim at crediting every reflection of the adult form on segmental, syllabic, or prosodic level. The fundamental principle is that each segmental agreement or each feature that is not expressed elsewhere in the child form should be credited with one bonus point. Demands of agreement, however, are heavier between consonants than between vowels. Vowels are allowed to deviate one featue along the parameters high/mid/low, front/central/back, rounded/unrounded, while consonants should have full agreement to get bonus, except for the feature voiced/voiceless and the feature place for sibilants. Additionally scores are given for deviant segments where you can trace a common phonological process as a cause for the deviation. Bonus is given also for language-specific traits of English such as off-glide and vowel length.

On the syllabic level agreement in phonotactic structure is given bonus. Since it is believed that post-tonal syllables in poly-syllabic words are difficult for the child to perceive, one bonus point is given to syllables that follow the post-tonal syllable, irrespective of numbers, provided there are post-tonal syllables in the adult form. On the prosodic level bonus vill be given for pitch and/or voice quality matching the adult form.

Penalty score will be given for added segments and for substitutions that cannot be understood as a product of common phonological processes in child language. The heavier demand on agreement of consonants than of vowels is partly due to the greater difficulty in reaching acceptable reliability inter transcribers when dealing with vowels. Added syllable including a 'true' consonant will be scored two penalty points, while added syllable without 'true' consonant will be scored one penalty point. Mismatch in unstressed vowel will render one penalty point. Segmental errors in pre-tonic syllables as well as in unstressed syllables are disregarded.

To apply a scoring system like this is somewhat problematic. It is hard to guarantee that you have accounted for all factors included. One major problem is to decide on what level, phonetical or phonological,the adult forms should be represented, whether on phonetic or phonologic level. One form representing a plausible articulation might show a great variation on account of reduction degree, speech rate, focus etc. By all means the transcripts of child forms are more narrow than the corresponding representation of the adult form.

The results from the other groups will show to what

extent the scoring system has been adapted to the different languages in the study. The Swedish group has proposed the following modifications for the Swedish data: Bonus scores should be rendered for vowel and consonant matching length, for matching accent II, and for correct number of syllables.

We do not consider vowel off-glide bonus relevant for Central Standard Swedish. Nor do we use the classification pretonal, tonal and post-tonal syllable as references for scoring since the reduction rules and stress rules of American English are significantly different from those of Swedish.

Phonetic reality of the mora in Eskimo

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

Introduction

Based on the results of a pilot study, this paper discusses to what extent such an abstract unit as "mora" is realized phonetically in West Greenlandic Eskimo (henceforth WGE). The mora is generally defined as a unit of timing that takes approximately the same length of time (e.g. Ladefoged 1975, p.224). Heavy syllables like CVV and CVC may be regarded as consisting of two moras, CV-V and CV-C, respectively.

Pitch patterns of WGE words

The relevance of such a unit as the mora for the description of WGE has been most convincingly demonstrated in relation to pitch patterns. According to Rischel (1974, p.79), "phrase final neutral intonation contour high-low-high" is placed on the last three vowel moras rather than on the last three syllables.

Examples

anāāna	ataāta	but	ataatāā
('mother')	('father')		('his father')

Examples of the pitch curves are shown in Fig. 1. Note that in all the words, pitch movements of the last three moras are very much alike when they are lined up at the end point. An important observation is that the pitch change takes place at mora boundaries and not at syllable boundaries. For example, in anaana, the most notable pitch change (rise) occurs in the last [na] starting from the onset of the consonant. Similar pitch movement is observable in the last [ta] for ataata though the earlier part is masked by the voiceless consonant. In the word ataataa, however, this pitch rise occurs only in the latter part of the last syllable [taa]. We assume that this portion that has the rise movement corresponds to the last mora [a]. Similar observations were made for the phrase internal intonation where the last mora, rather than syllable, has a pitch fall. The relevance of the point of pitch change for the perception of accent categories is discussed elsewhere in more general perspective (Nagano-Madsen and Eriksson forthcoming)

Durational patterns of WGE words

It is then reasonable to assume that the mora also plays a role in determining the durational pattern of WGE words. Since the mora is defined as a constant time unit regardless of its segmental composition, we expect that the duration of a word can be captured, at least as a first approximation, in terms of the number of moras it contains. The following short experiment was designed in order to test this hypothesis.

Four phonological words made up of a nominative stem 'ataata (=father)' plus different affixes were prepared. They have an increasing number of moras and also contain various types of mora (i.e. CV, V, or C). They are all real words. A female subject read the list of words twelve times with the same intonation at moderate speaking rate. The durations of individual segments as well as of the whole word were measured from mingograms. The result is shown below:

	word	number of mora	duration in ms (S.D.)	ratio
1.	ataata	4	675 (31.0)	4.0
2.	ataataa	5	833 (35.7)	5.0
3.	ataataata	6	997 (41.9)	6.0
4.	ataatatta	6	978 (37.0)	5.9

In no. 2 one mora [a] is added, in nos. 3 and 4, two moras (a-ta or t-ta) are added. The mean duration of the word ataata was 675 ms. When one mora was added, it increased by 158 ms. When two moras were added, it increased either by 322 ms (no. 3) or by 303 ms (no. 4). Two observations were made. First, each additional mora increased the total word duration by roughly the same amount (152-161 ms). This is in contrast to languages with lexical stress such as English and Swedish that tend to keep the duration of a word relatively constant by adjusting the duration of a stressed vowel systematically when more and more unstressed syllables are added (Lehiste 1972, Klatt 1973, Lindblom and Rapp 1974, Port 1981). As there is no lexical stress in WGE, the duration of each segment in ataata, except for the last [ta] that is affected by prepausal position, stays basically the same for all four words.

Second, whether the mora is a CV, a vowel V, or a consonant C does not affect the duration of mora much. However there is evidence from the data above that a consonant mora C is slightly shorter than the other types of mora (cf. no. 4). This is in agreement with the figures obtaind for the single-geminated ratios of this language. Geminated vowels have a higher ratio than consonants exceeding sometimes well over three times the length of a single one (Nagano-Madsen

forthcoming) while that of geminated consonants generally ranges between 2-3 that of a single one (Mase and Rischel 1971).

These points become clearer when the duration of each word is expressed in ratios with reference to the duration assigned to one mora. The total of the mean durations (3483 ms) was divided by the total number of moras contained in the material (21). We get 165 ms by this method and assume that it represents one mora at this speaking rate. The duration of each word was then simply divided by this figure. The ratio shows such a good correspondence to the number of moras that one can roughly predict the number of moras in a given duration, at least for the kind of material used here. Note that the regularity observed can only be captured in terms of constant temporal units (moras) and not in terms of segments or syllables. From a given duration, we can neither tell the number of segments nor the number of syllables it contains.

A syllable, defined in a traditional way, does not seem to play a significant role in determining the durational pattern for this language. Both nos. 1 and 2 consist of three syllables but their durations are very different. There is, however, an indication that a syllable is a relevant articulatory unit for this language as well since the duration of the last [t] in ataata was very similar in the first and second words when it occured in prepausal syllable but much shorter in the third and fourth words.

The kind of regular mora-timing observed above was found not to exist to the same extent in a connected discourse (Nagano-Madsen forthcoming) but this is not contradictory to the perceptual impression since WGE sounds more mora-timed when spoken slowly. Likewise a strong version of mora, e.g. that each mora in WGE tends to be constant regardless of its inherent segment duration, was refuted (Nagano-Madsen 1983). At this point then, I am inclined to think that the reality of mora in the temporal dimension comes most strongly from the manifestation of the so-called single-geminated dichotomy discussed above. Since WGE is a typical quantity language in which most segments have this dimension of contrast, they are carefully manifested. Further study is needed in order to confirm some of the points made in this paper.

References

Klatt, D. 1973. 'Interaction between two factors that influence vowel duration'. J. Acoust. Soc. Am. 54, 1102-04.

Ladefoged, P. 1975. A course in phonetics. New York: Harcourt Brace Jovanovich.

Lehiste, I. 1972. 'The timing of utterances and linguistic boundaries'. J. Acoust. Soc. Am. 51, 2018-2024.

Lindblom, B and K. Rapp. 1973. 'Some temporal regularities of spoken Swedish. Papers from the Institute of Linguistics'. Univ. of Stockholm, publication 21, 1-59.
Mase, H and J. Rischel. 1971. 'A study of consonant quantity in West Greenlandic'. Annual Report of the Inst. of Phonetics (Univ. of Copenhagen) 7, 1-98.
Nagano, Y. 1983. 'Syllable-timing in West Greenlandic Eskimo'. Paper presented at the 4th World Congress of Phoneticians, Kobe, Japan.
Nagano-Madsen, Y. Forthcoming. 'Temporal characteristics in mora-timed and syllable-timed languages: evidence from Eskimo and Yoruba'.
Nagano-Madsen, Y. and Eriksson L. Forthcoming. 'The location of the turning point in the identification of stress/tone categories'.
Port, R. 1981. 'Linguistic timing factors in combination'. J. Acoust. Soc. Am. 69, 262-274.
Rischel, J. 1974. Topics in West Greenlandic Phonology. Copenhagen: Akademisk Forlag.

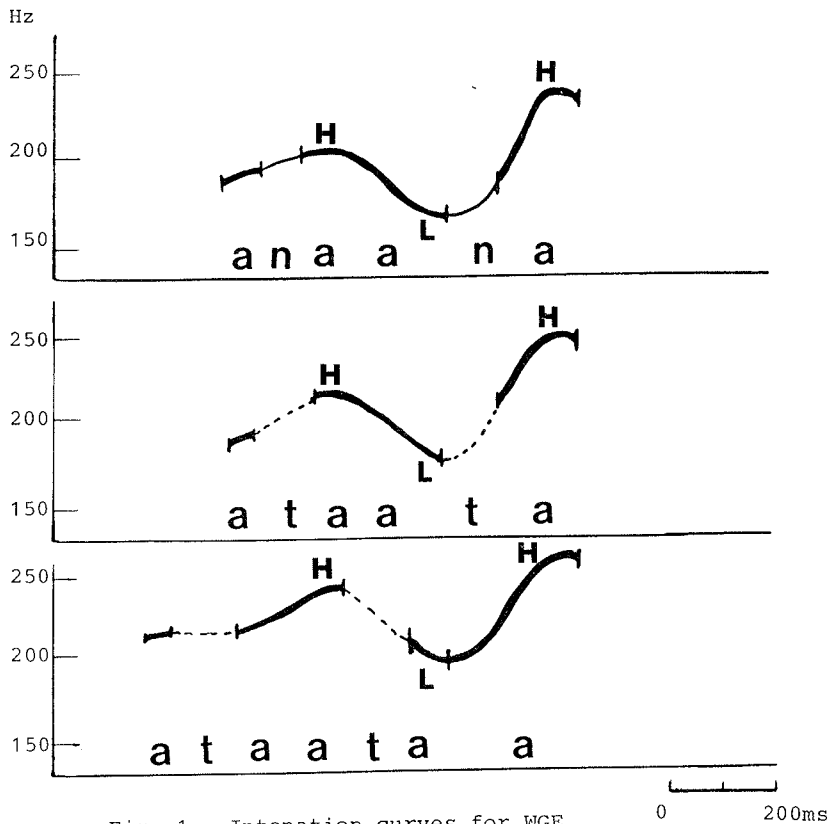


Fig. 1. Intonation curves for WGE.

PHONEME AWARENESS, SYLLABIC STRUCTURE, AND PHONETIC SUBSTANCE

Eva Magnusson and Kerstin Nauclér
Department of Linguistics and Phonetics
Lund University

The research presented in this paper is part of a longitudinal project where groups of language disordered and normally speaking children are studied with the aim of identifying the linguistic abilities that are most important, or even indispensable, for learning to read and write. The project started four years ago when the children were six years old, i.e. one year before they started school. Beside testings in the pre-school, testings were also done in the first and third grades and a follow-up is planned in grade four

In the discussion of reading and writing acquisition and subsequent problems, a lot of interest has been shown in recent years for the role of linguistic, or more precisely phonological, awareness and for the relation of phoneme awareness to reading and writing. Some researchers regard phonemic awareness as a prerequisite for learning to read and write, while others regard it as an effect of reading and writing acquisition. Those who hold phoneme awareness to be a prerequisite often base their opinion on work with children, e.g. on studies where the effect of phoneme awareness on reading and writing achievements, and the training of such awareness, are studied in beginning readers (e.g. Mann & Liberman 1982, Bradley & Bryant 1985). Among those who see phoneme awareness as an effect of literacy, we often find researchers who have worked with older readers or with illiterate adults who, in some studies, have been found to be unaware of phonemes (Morais et al. 1979). An intermediate position is taken by those who argue that phoneme awareness is both a prerequisite and an effect of reading acquisition (e.g. Ehri & Wilce 1979, Valtin 1984).

It is well known that reading ability is not indispensable in order to become aware of phonemes as evidenced by all the non literate pre-schoolers who enjoy rhyming and other playful activities that require an awareness of phonemes. On the other hand, from the ability to read does not automatically follow an awareness of phonemes as shown by the findings that readers of non-alphabetic writings systems are phonemically unaware (Mann 1986, Read et al. 1986).

Linguistic awareness is often discussed as if children were either linguistically aware or totally unaware. In our studies of non literate pre-school children (e.g. Magnusson & Nauclér 1987), we have noticed that the same subjects appear to be more or less aware depending on the type of task they are given: more children show linguistic awareness on a rhyme recognition task than on a phoneme identification task. Furthermore, it seems to be important which type of segments they are asked to manipulate and which sequences or structures these segments are part of. The present study was undertaken with the aim of looking into how syllabic structure and segment type (or phonetic substance) influence children who are in the process of developing phoneme awareness .

PROCEDURE

For the study we have used data from our longitudinal project. At the

time the children were six years old and still non literate pre-schoolers. From the various meta-linguistic tasks the children were given at the pre-school testing (Magnusson & Nauclér 1987), we have chosen two tasks: identification and segmentation of phonemes.

Identification of phonemes In the identification task the children were asked to identify two consonants, a fricative /s/ and a plosive /t/, and two vowels, a back vowel /u/ and a front vowel /i/ in a number of familiar words. The target sounds appeared in either initial, medial, or final position of the word - or not at all (see table 1). The children were asked if they could hear the target sounds in words said by the experimenter and their task was to indicate whether or not the sound - in their opinion - was part of the sound sturcture of the word.

Table 1. Words used in the identification task.

Ident. of /s/	Ident. of /t/	Ident. of /u/	Ident. of /i/
sol	nalle	bil	Ida
Lisa	katt	Ola	mage
apa	mus	mor	docka
myra	tak	kaka	bi
hus	äta	dörr	bok
båt	nål	ko	pil

Segmentation of phonemes In the segmentation task the children's task was to indicate the number of phonemes by selecting the correct number of markers of some kind. The words used in the task varied as to the number of phonemes as well as to syllabic structure, e.g. VC, CV, VCC, CVC, CCVC, CVCC. Both monosyllabic and bisyllabic words were included (see table 3 below).

RESULTS

Even if both the tasks were designed to measure the awareness of phonemes, the identification task turned out to be easier than the segmentation. More children understood the task and suggested a solution for each of the test items in the identification task than in the segmentation task (92 out of 114 as compared to 83). This was so although the identification task contained more test items (24) than the segmentation task (18).

Identification Both segment type and position of the segment in the word influenced the children's ability to identify phonemes. As regards segment type, Vs were easier to identify than Cs (see table 2). There was no difference as to how well the children identified the

Table 2. 92 subjects' correct identifications of phonemes. ( ) number of possible identifications.

Type of phoneme		Position		
C (1104)	V (1104)	Initial (368)	Medial (368)	Final(368)
834	916	306	263	262
/s/ 447	/i/ 463			
/t/ 387	/u/ 453			

two Vs, /i/ and /u/, but there was a difference between the Cs so that the fricative /s/ was more often correctly identified than the plosive /t/ and almost as often as one of the Vs, namely /u/. If position is considered irrespective of segment type, we find that more phonemes were correctly identified in initial than in either medial or final position.

Segmentation The length of the word seemed to influence children's ability to segment words into phonemes so that words with four phonemes were more difficult to segment than words with three phonemes (see table 3). However, such a difference was not found between words with two or three phonemes.

In words with two phonemes, syllabic structure was important so that CV syllables were easier to segment than VC syllables. Segment type, whether a fricative or a plosive, did not influence the children's ability to manage the task as in the identification task.

Table 3. 83 subjects' correct segmentations.

CV	VC		
se 62	ös 51		
gå 60	ek 55		
CVC	VCV	VCC	CCV
katt 49	apa 56	ost 59	stå 43
sol 52	åka 50	arm 43	bra 38
	CVCV	CVCC	CCVC
	bada 28	dans 22	glas 30
	titta 16	mask 24	spik 14

In words with three phonemes, the segmentation was influenced by whether or not there were consonant clusters, the position of the cluster, and to some extent by the type of cluster. Words with clusters (e.g. 'stå', 'bra') were more difficult to segment than words with only singletons (e.g. 'sol', 'apa') and words with initial clusters (e.g. 'stå') were more difficult than word with final clusters (e.g. 'ost').

To be able to segment the cluster /st/ in final position as in 'ost' did not ensure that the children were able to segment the same cluster in initial position as in e.g. 'stå'. Nor did the ability to segment /st/ in 'ost' and 'stå' guarantee the segmentation of sC-clusters in words like 'mask' and 'spik' whith one additional segment. These longer words make larger demands on short term memory, and may for some children exceed their short term memory capacity. When tested for short term memory it was found that some of the children had a memory span of only two or three items.

The type of cluster made the task more or less difficult. Clusters with /s/ were easier to segment than clusters with /r/ ('stå' was easier than 'bra'), although we cannot argue that clusters with /s/ were easier than all clusters containing liquids as e.g. 'glas' was segmented correctly by more children than 'spik'.

SUMMING UP

Segment type is important for the children's results on the identification task. This is consistent with the way the task was presented; the children were given a sound as a model and their task could be described as finding a perceptual match to the target sound. In order to do this, they have to perceive words not just as entities but have to have gained at least some insight into the possibility of segmenting words into smaller units, even if the demands on their knowledge about the segmental structure are not as heavy as in the segmentation task. Position in the word also influences the children's identifications so that segments in initial position are the easiest to identify.

In the segmentation task, on the other hand, segment type does not seem to be as important as in the identification task, while word length and syllabic structure play a more important role. Provided that the number of phonemes in the word does not exceed the short term memory limit (which we have reason to assume in a number of cases) syllabic structure is important, whether it is a CV or VC syllable, whether or not there are clusters, and whether the clusters are word initial or word final.

In conclusion it can be said that both type of segment (or phonetic substance) and syllabic structure are crucial for the achievements of children who are in the process of developing an awareness of phonemes. The importance of each factor is partly dependent on the type of task the children are given.

REFERENCES

Bradley, L. & Bryant P. 1985. Ryme and reason in reading and spelling. Ann Arbor: The University of Michigan Press

Ehri, L.E. & Wilce, L.S. 1979. The mnemonic value of orthography among beginning readers. Journal of Educational Psychology 71, 26-40

Magnusson, E. & Nauclér, K. 1987. Language disordered and normally speaking children's development of spoken and written language: preliminary results form a longitudinal study. In RUUL, Reports form Uppsala University Department of Linguistics, no 16, 35-63

Mann, V.A. 1986. Phonological awareness: The role of reading experience. Cognition 24, no 1.2, 65-92

Mann, V.A. & Liberman, I.Y. 1982. Phonological awareness and verbal short term memory: Can they predict early reading problems? Haskins Laboratories: Status Report on Speech Research SR-70

Morais, J., Cary, L,. Alegria, J., & Bertelson, P. 1979. Does awareness of speech as a sequence of phones arise spontaneously? Cognition 24, 45-64

Read, C., Zhang, Y., Nie, H., & Ding, B. 1986. The ability to manipulate speech sounds depends on knowing alphabetic writing. Cognition 24, no 1-2, 31-44

Valtin, R. 1984. Awareness of features and functions of language. In Downing & Valtin (eds), Language awareness and learning to read. Berlin: Springer Verlag

This research was supported by a grant from HSFR, Swedish Council for Research in the Humanities and Social Sicences.

MEASURING SPEECH COMPREHENSION

Robert McAllister & Mats Dufberg
University of Stockholm
Institute of linguistics

1. **BACKGROUND.** In many areas related to the study of language, both theoretical and applied, there is a need for measurements of speech comprehension. Two such areas dealt with in this report are the assessment of functional hearing loss and oral speech comprehension in second language aquisition. As we see it, understanding speech is dependent on two sources of information, signal dependent and signal independent information, respectively. (Cf. top-down and bottom-up, respectively.) The interaction between the two sources is shown in fig. 1. According to this model, the more signal dependent information that is available for the listener the less signal independent information is needed and vice versa. Common audiometrical measuring methods have a bias towards signal dependent information, and the correlation between tone and speech audiometry, on the one hand, and speech comprehension, on the other, is commonly questioned.

Here we will report on a pilot study which is part of a larger study we have recently started, which we will present in McAllister & Dufberg (forthcoming). The aim of this pilot study is to test two available methods and one new method that, at least partly, claim to test speech comprehension. If our model of interaction between signal dependent and signal independent information is correct and it is correct to assume that noise will mask the signal it is reasonable to assume, firstly, that native speakers will be able to stand more noise (that is, lower signal-to-noise ratio) than L2 speakers, secondly, that persons with normal hearing can stand more noise than hearing impaired persons.

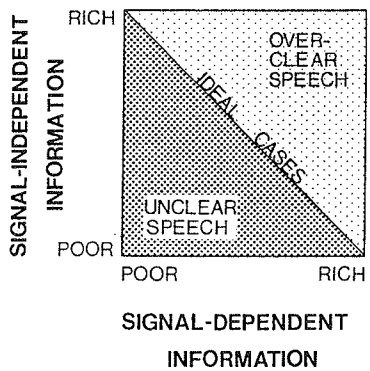


Fig 1. Mutality of speaker-listener interaction. (Reprinted from Lindblom, 1987.)

2. **SUBJECTS.** In this pilot study we used two groups of subjects; one group of foreigners with Swedish as second language who speak broken Swedish, henceforth referred to as L2 speakers, and one group of hearing impaired persons with Swedish as their first language.

The L2 speakers were four persons working at our lab,

three of which are trained phoneticians and the fourth belonging to the technical staff. All four of them speak Swedish well. The group of hearing impaired persons consisted of only two persons. They are both members of the staff of our department. Their hearing loss is fairly severe, 50 to 70 dB. As a control group we used six native speakers of Swedish with normal hearing, all of them members of the department staff.

Both experiment groups have perturbed speech comprehension; the L2 speakers on their ability to use signal independent information, and the hearing impaired persons on their ability to use signal dependent information. We regard measuring functional hearing loss and testing speech comprehension in language learning being special cases of measuring speech comprehension.

3. **METHODS.** In this study we tested three related methods. The first method is one designed by Walker & Byrne (1985). In this method the subject is listening to a text and asked to set a noise level so that s/he can only barely follow the text. The signal-to-noise ratio at this threshold level of comprehension is taken to be the measure. This method has been successfully used by A. Risberg and M. Dahlquist with relatively high level of reliability (personal communication). In the following we will refer to this method as the Threshold method.

The second method is similar to the first. Instead of letting the subject choose the noise level s/he pressed a button which gave an increasing noise level. When the subject released the button the noise level fell to a minimum. The instruction to the subject was to press the button and to release it when s/he could not follow the text. The method has also been tested by Risberg and Dahlquist who report that it is less reliable and has a greater learning effect than the Threshold method (personal communication). In the following we will refer to it as the Ramp method.

With our third method, which to our knowledge has not been tried before, we tried to create a less unnatural situation where comprehension really is tested, not the subjective impression of comprehension. We presented questions in noise to our subjects and asked them to answer the questions. When they gave a correct answer we increased the noise level one decibel and in the case of an incorrect or no answer we decreased the noise level one decibel. The questions presented were so called Helen questions (Ludvigsen 1975) which anybody speaking Swedish would know the answer to. ("What color is a lemon?" etc.) We took the signal-to-noise ratio at a correct answer after an incorrect one (that is when we had decrease the noise level one decibel) as the measure of this method. We will refer to it as the Helen method in the following.

The noise we used in this study had the same long time spectrum as male speech and was low frequency modulated. It is the noise described and used by Hagerman (1984). We assume that

It fairly well matched the male voice we used for our speech material.

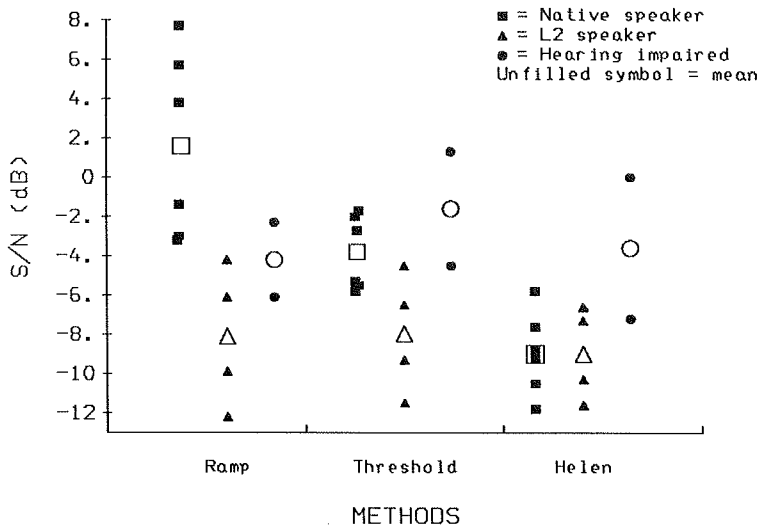


Fig 2. In the figure a filled symbol represents the mean of a few repetitions of one subject. Each subject is represented three times in the figure: once for each test method. An unfilled symbol represents the mean of the test group for each test method.

4. RESULTS. In figure 2 the results of our study are presented. If we can assume that the Helen method is more valid for testing speech comprehension and that our speech material for the different tests are comparable, we can conclude that native speakers seem to underestimate their speech comprehension in noise. (Compare squares for the three methods. Note that the decrease on the S/N scale indicates more noise.) The L2 speakers, on the other hand, seem to be more accurate in their estimation (triangles). The hearing impaired group (circles) is too small to let us draw any conclusions, but we can notice that one of them is not worse than the control group.

Let us finally compare the results for the three test groups for the Helen method. The means of the control group and the L2 speakers are identical. All L2 speakers are in fact within the control group. One of the hearing impaired is also within the control group, the other is well above.

5. DISCUSSION. For two of the methods, the Threshold and the Ramp methods, the relationship between test result and speech comprehension is at best indirect. Firstly, we cannot assume

that all subjects have the same criterion for deciding that they can only barely follow the text or cannot follow the text. Secondly, both tasks are very different from normal speech interaction. The Helen method, on the other hand, involves no subjectivity. What is measured is the ability to do the task, i.e. answering the questions. If you have not heard the question you cannot guess, and if you have heard it you will certainly be able to answer it. But there are certainly problems with the speech material we used in the Helen method, i.e. the Helen questions. Firstly, they are very predictable in form (just wh-questions), secondly, it is very limited what they question (color-of, opposite-of etc.). These two factors make the risk of learning quite high.

In figure 2 we can, firstly, see that the L2 speakers can stand more noise than the control group for both the Threshold and the Ramp methods. We suggest that L2 speakers are more used to low comprehension level and therefore have a lower requirement for what they consider to be their subjective threshold. Secondly, and more surprisingly, it seems like the noise has the same effect on speech comprehension for native speakers and L2 speakers, assuming that the Helen method really tests speech comprehension. Such an result is clearly counter intuitive. Both authors of this report clearly feel that they are more sensitive to noise in L2 than L1 situations, and that is a widespread view as far as we know. (One suggestion is that the noise type is very special. In our forthcoming paper we will report from studies with a different noise type.)

6. CONCLUSIONS. In this study we cannot see that second language speakers are more sensitive to noise masking than native speakers, which contradicts our expectations. We cannot as yet give any reasonable explanation for the result.

REFERENCES

Hagerman, B (1984): *Some aspects of methodology in speech audiometry* (dissertation). Karolinska Inst, Stockholm.

Lindblom, B. (1987): "Adaptive variability and absolute constancy in speech signals" in *Perilus V*, 2-20. Inst of linguistics, Univ of Stockholm.

Ludvigsen (1975): "Construction and evaluation of an audio-visual test (the Helen test)" in *Visual and audiovisual perception of speech* (Sixth Danavox symposium 1974). Stockholm: Almquist & Wiksell.

McAllister & Dufberg (forthcoming): paper to appear in *Perilus VIII*. Inst of linguistics, Univ of Stockholm.

Walker, G. & Byrne, D. (1985): "Reliability of speech intelligibility estimation for measuring speech reception threshold in quiet and in noise" in *Australian J of Audiology, 7:1*, 23-31.

DATA BANK ANALYSIS AND NASAL SYNTHESIS

Lennart Nord

Department of Speech Communication and Music Acoustics,
Royal Institute of Technology (KTH), Box 70014,
S-100 44 Stockholm, Sweden.

1. INTRODUCTION

This is a brief summary of recent work on segmental analysis and synthesis. Specifically, the aim is to improve the quality of nasal sounds in our text-to-speech synthesis system, developed at the Speech Department in Stockholm [1].
Our acoustic-phonetic work is presently making use of some new software facilities, developed at the Department. Another recent feature is the support from our speech data bank that is being structured. A number of earlier reports have dealt with different aspects of the data bank project [2-6].
For expanded versions of this report, see [7,8].
Theoretical and modelling work has given us a good insight into the acoustics of nasals and the effects of nasal coupling [9-14]]. Still, perceptual evaluations of synthesis systems, usually give high error rates for nasals compared to other speech sounds.

2. NASAL SYNTHESIS

2.1 Serial synthesizer

Synthesis work at the Speech Communication Department traditionally exploits a serial type of terminal analog formant synthesizer, the OVE III as a part of our text-to-speech system [15]. There is thus a fixed relation between formant frequencies and relative spectral levels, which has to be taken into account in the spectral matchings. As there are numerous extra resonances and zeroes to be found in nasal spectra, the matchings are difficult.

2.2 Evaluation of present text-to-speech system

Systematic listening tests have been performed during the years, using /VCV/ words with V-V = /a-a, I-I, U-U/. As a complement a separate test was made with /CV:Ca/ words with /V:/ = /a:, i:, u:/ and the consonant C = all combinations of /m,n,l/, giving words like *mila, lama, nola.* An analysis of the nasal errors show errors in manner as well as in place of articulation. Specifying the probable sources of error will give a representative repertoire of the nasal acoustic cues. In Table I some of the typical errors are listed, together with a brief comment of source of error. Error scores are not given here as they fluctuate considerably with sound quality and listener sophistication.

TABLE I

	Probable Main Source of Error
/ni/ -> /mi/	formant transition
/no/ -> /mo/	formant transition and poor nasal resonance
/mo/ -> /vo/	transition rate and poor nasal resonance
/ɲi/ ->/ji, gi/.	intensity level and poor nasal resonance

2.3 Continuing synthesis work

2.3.1 Software facility.
We are presently using a convenient software tool, RULSYS, developed at the Speech Department. The program package will allow us to select a natural speech sample,

make an FFT analysis of it and display a spectrogram on the monitor together with the rule synthesis output in the form of synthesis parameter tracks. The resulting synthesis can be played back and compared to the sound of the natural speech sample. At any point along the time axis a spectral section can be derived both of the synthesis output and the speech signal, a necessary component of the matching procedure.

2.3.2 Pattern matching and testing synthesis strategies.
The software facility is used in two ways. Firstly, to generate a set of detailed pattern matchings of vowel-nasal-vowel syllable, trying to achieve synthetic copies with as good quality as possible. Once we get perceptually and acoustically acceptable matches, we start degrading the parameter tracks to establish the relative importance of the acoustic cues. Formant transitions, transition rates, bandwidth values, nasal segments as opposed to nasalized vowels, etc are evaluated. The other line considers the testing of various synthesis strategies. As there are some built-in restrictions of the synthesizer, this will give us a quick feedback regarding some of the principles that should be used in synthesizing nasals. The nasal segment can thus consist of the single nasal resonant, the vowel formants with proper bandwidth modifications etc. For earlier work on nasal synthesis, see [16,17]

3. A SWEDISH SPEECH DATA BANK

3.1 Purpose
An important component in improving the text-to-speech system is the supply of real speech data and especially the possibility to analyse large materials. Our data bank is labelled at a phonemic level, and a systematic search can be made through the material by the use of phoneme strings. Our long-term goal is to include ten males, ten females and ten children, but so far only four males and one female have been recorded reading the entire material, and only two of the male readings have been partially labelled. Eventually, it will be very interesting to get data on suffixes from running text material, that frequently consists of vowel-nasal syllables.

3.2 Recording and labelling
The speech is digitized at a sampling frequency of 16 kHz and edited on a computer by means of a speech wave editing program. After editing, the material is stored on 80 Mbytes computer discs. First, a phoneme string is generated from the text-to-speech system, Carlson & Granström [1], using the orthographic strings as input. The phoneme strings, including prosodic markers, are synchronized to the speech wave by means of the speech editing program. To facilitate the placing of labels, part of an automatic speech recognition system, Blomberg & Elenius [18] can be used. By using computer generated spectrograms as a complement, the automatically produced segmentation is adjusted by moving the marks along the oscillographic tracing on the computer screen, while listening to the signal. This process is very time-consuming, and as we at the same time are trying to improve our segmentation criteria there is no way to make this phase of the work automatic. When finally an accepted labelling of the material is attained, the label files will be stored together with the speech files on the 80 Mbyte discs and the material is ready for usage.

3.3 Analysis
The analysis will also rely on the text-to-speech program that generated the phoneme strings, in an automatic search procedure. The search is made by specifying the phoneme or the class of phonemes that are of interest, together with prosodic markers. The result can either consist of automatically derived measures of segment durations, spectral envelopes or simply a subset of phones for visual inspection and subsequent later analysis. Depending on the detailed prosodic marking, contextual specifications may include details of stress levels and position within the word, thus, for example

enabling an analysis of unstressed verses stressed syllables. As an example of data bank usage this routine has been performed on another set of speech material by Carlson & Granström, who reported on duration measures of Swedish and American vowels and consonants, all of which were derived in this fashion [4].

3.4 The phonemic principle

Our label string is thus basically phonemic, and allophonic variations are not taken into account, unless they are clearly the result of some higher level rule (e g the dialectal pronunciation of proper names). This means that any allophonic variation will be seen in the distribution of analysis data. If we for example ask for a formant plot of all stressed /ö/ phonemes, we will probably get a bimodal distribution in the formant plane, in this particular case due to the fact that there are two contextually dependent /ö/ allophones in Swedish. There are two reasons for this 'phonemic' approach. Firstly, it will lessen the demands on part of the labeller, who otherwise could spend an infinite amount of time, trying to chose among IPA symbols and diacritics. Secondly, as we are trying to improve the quality of synthetic speech, we are usually in a situation where, starting from a phoneme string, the goal is to generate a satisfactory acoustic output. The statistical distribution of how certain phonemes strings will be modified by coarticulatory processes and realized is easily found by making a search in the material, using the string of interest.

3.5 Texts

Our choice of texts tries to satisfy a number of demands from a phonetic-acoustic point of view. The material consists of speech sounds in various contexts, ranging from isolated phonemes to text passages. Vowels, consonants, syllables and words pronounced in isolation with a lexical stress are included as a means to establish a basic acoustic mapping of the speakers, and also enabling studies of reduction properties when going from careful pronunciation to more relaxed. Different types of text: isolated sentences, newspaper articles of different complexity (politics, general news bulletin, speech research), and part of a novel by a Swedish novelist, complemented by one page that is read with some variation in style (normal, clear, weak, strong) and a short piece of dialogue from the same novel constitute the entire text material.

3.6 Choice of speakers

Speakers are chosen according to two principles. On the one hand we are interested in trained speakers, such as radio announcers, that can read well and come close to a natural sounding, almost spontaneous speaking style. On the other hand we are also interested in normal, non-professional speakers, and especially for recognition work it would be fatal not to include variation for the analysis. In the work on Swedish prosody by Fant, Nord and Kruckenberg [6] and Fant and Kruckenberg [19], one of the aims is to understand the personal variations in reading styles.

3.7 Principles of segmentation and labelling

Most of the labelling difficulties occur in the text passages. There is indeed a great difference between this kind of material and isolated sentence material. In the fluently read texts strong reductions often occur, and in contrast to the isolated sentences, large parts of the text, sometimes whole phrases lie out of focus.

5. REFERENCES

[1] R CARLSON & B GRANSTRÖM,"Linguistic Processing in the KTH multilingual text-to-speech system", Conference record, IEEE-ICASSP, Tokyo (1986)

[2] R CARLSON & B GRANSTRÖM,"Rule Controlled Data Base Search", STL-QPSR 4/1985, p 29 (1985)
[3] R CARLSON & B GRANSTRÖM, "Swedish Duration Rules Derived from a Sentence Data Base", STL-QPSR 2-3/1986, p 13 (1986)
[4] R CARLSON & B GRANSTRÖM, "A Search for Durational Rules in a Real-Speech Data Base", Phonetica 43, p 140 (1986)
[5] G FANT, L NORD & A KRUCKENBERG,"Individual Variations in Text Reading . A Data-Bank Pilot Study", STL-QPSR 4/1986, p 1 (1986)
[6] G FANT, L NORD & A KRUCKENBERG,"Segmental and Prosodic Variabilities in Connected Speech. An Applied Data-Bank Study", Proc XXIth ICPhS, Tallinn, USSR, Vol.6, p 102 (1987)
[7] L NORD, "Acoustic-Phonetic Studies in a Swedish Speech Data Bank", paper presented at the FASE symposium, Edinburgh (1988) and forthcoming STL-QPSR
[8] L NORD, "Synthesis of Nasal Sounds in a Text-to-Speech Framework", paper presented at the FASE symposium, Edinburgh (1988) and forthcoming STL-QPSR
[9] A HOUSE & K STEVENS, "Analog Studies of the Nasalization of Vowels", J Speech Hear Dis 21, p 218 (1956)
[10] S HATTORI, K YAMAMOTO & O FUJIMURA,"Nasalization of Vowels in Relation to Nasals", J Acoust Soc Am 30:4, p 267 (1958)
[11] O FUJIMURA,"Spectra of Nasalized Vowels", Q Prog Rep 58, Res Lab Electron M I T, p 214 (1960)
[12] O FUJIMURA & J LINDQVIST-GAUFFIN,"Sweep-Tone Measurements of Vocal-Tract Characteristics", J Acoust Soc Am 49:2, p 541 (1971)
[13] J LINDQVIST-GAUFFIN & J SUNDBERG,"Acoustic Properties of the Nasal Tract", Phonetica 33, p 161 (1976)
[14] S MAEDA,"The Role of the Sinus Cavities in the Production of Nasal Vowels", Proc of ICASSP 1982, Paris, France, p 911 (1982)
[15] R CARLSON, B GRANSTRÖM & S HUNNICUTT,"A Multi-Language Text-To-Speech Module", Proceedings IEEE-ICASSP, Paris (1982)
[16] L NORD,"Perceptual Experiments with Nasals", STL-QPSR 2-3/1976, p 5 (1976)
[17] L NORD,"Experiments with Nasal Synthesis", STL-QPSR 2-3/1976, p 14 (1976)
[18] M BLOMBERG & K ELENIUS,"Automatic Time Alignment of Speech with a Phonetic Transcription", STL-QPSR 1/1985, p 37 (1985)
[19] G FANT & A KRUCKENBERG,"Some Durational Correlates of Swedish Prosody", paper presented at the FASE symposium, Edinburgh (1988)

PHONETIC ASPECTS OF LARYNGECTOMEE SPEECH

Lennart Nord
Department of Speech Communication and Music Acoustics,
Royal Institute of Technology (KTH), Box 70014,
S-100 44 Stockholm, Sweden
and

Britta Hammarberg
Department of Logopedics and Phoniatrics,
Huddinge Univ Hospital,
S-141 86 Huddinge, Sweden

INTRODUCTION

The alternative speech methods that are available today for the laryngectomized have each one its advantages and disadvantages. and each method seems to have its proponents. It is important to investigate the features of the different techniques and also to get a deeper insight into what signifies an acceptable voice following laryngectomy. Moreover, the basic mechanisms of the different kinds of alaryngeal phonation are not very well known, and a comparison with the function of the normal glottis is therefore of interest.
We are presently investigating three different techniques: (i) esophageal speech, (ii) tracheo-esophageal speech (TE-speech) where part of the esophagus is used as voice source, and the driving pressure is generated by the lungs via a valve and finally (iii) electrolaryngeal speech, using an electromechanical device (Servox) that emits pulses through the pharynx wall. For a survey of the area, see [1,2].

STUDIES OF ALARYNGEAL SPEECH

The introduction of the tracheo-esophageal technique has in many ways implied a significant improvement for the establishment of a new voice after laryngectomy. A number of recent studies have compared the different types of alaryngeal speech. Speech intensity, intonation range, intelligibility and listener acceptability have been investigated [3-5], as well as prosodic aspects of the different speaking methods [6,7]. We will present some preliminary findings of acoustic-phonetic analysis of esophageal and tracheo-esophageal speech. Parts of this project have been presented earlier [8,9]. For more extensive reports, that will also cover some clinical aspects, see [10,11].

SPEECH MATERIAL

The speech material was designed with the aim of testing prosodic aspects of communication. Sentences with contrastive word emphasis and short questions and statements typical of a dialogue were thus included. A few vowels were embedded in carrier phrases and the subjects were asked to produce them with low, medium and high intensity. A reading of a short paragraph gave us the material to extract a long term average spectrum of the voice as well as establishing the pitch range of the subject. To investigate air pressure and air flow conditions and also to get an estimate of the voice source characteristics, a flow mask [12] was used in a separate reading of nonsense words.

SPEAKERS

Presently, we have recorded three male esophageal speakers and three male TE-

speakers. A few laryngeal speakers will also be recorded and used as reference. So far, only one male speaker of the same age group has been analysed. Our intention is to collect a representative amount of speakers, that are using the different techniques and also record some multi-technique speakers.

ANALYSIS

The acoustic analysis includes measurements of pitch, intensity, duration, speaking rate, long time average spectral shape and voice source characteristics, derived by inverse filtering. Detailed spectral analysis of articulation and source features are also performed at selected points.
As an important part of the study the acoustic measurements will be correlated to perceptual evaluations, using different groups of listeners.

Overall speech quality

The speech quality differs a lot between the speakers in our material. In Figure 1 spectrographic illustrations of two of the speakers are shown. As can be seen the esophageal voice looks quite normal (top), the deviant quality is here mainly perceived in the prosodic pattern and the low pitch. The TE-speaker has a very harsh voice (bottom) with irregular formant excitations and superimposed noise.

Pitch and timbre

Pitch extraction. Pitch extraction turned out to be difficult and uncertain for these deviant voices. The reasons for this seem to be that the pitch values are usually extremely low and sometimes very aperiodic. Moreover, in contrast to the case of a laryngeal low pitched voice, inverse filtering reveals that the instant of excitation in a tracheo-esophageal voice is not well defined. In Figure 2 a short interval of the vowel /a/ is shown together with the extracted voice source pulse, for a tracheoesophageal speaker, who is using a Panje-valve (note the difference in time scale).
In the frequency domain the fundamental (F0) is very weak relative to the formants, which means that a pitch extraction algorithm based on a spectral measure will work poorly.

Voice source spectrum

When changing from strong to weak voice a typical feature of laryngeal voices is that the fundamental will dominate the spectrum. This is not the case for these voices. Instead, the fundamental will maintain a low level irrespective of the total intensity. This lack of dynamics probably contributes to the unnatural quality of the voice. The perceptual dimension *hyperfunctional*, or tense voice also correlates well with the spectral dimension of relatively weak fundamental [13].

INTONATION

In the production of voiced sounds there are no intricate laryngeal-like mechanisms to stretch and manipulate the vibrating parts in the esophagus. Although a pseudo-glottis is developed, we know little about its function, apart from the fact, that for some speakers it will serve as a surprisingly good substitute.
The analysis of the sentences containing emphatic word stress was made manually, using a speech wave editing program, developed at the Speech Communication Department. Each subjectively judged periodic cycle was marked and the pitch was automatically calculated. Figure 3 shows a set of emphatic sentences *Gunnar målade dörrarna gröna* for one of the TE-speakers, where the word stress was moved one step to the right for each display a) to d). Two things are worth noticing. First, the mean F0 value is rather low for this speaker. Second, the signalling of emphasis seems to be

reasonably well managed by the speaker, in these plots an F0 hump is seen moving to the right for each new sentence (the approximate location of the emphasized word is underlined).

REFERENCES

[1] Shanks,J (Editor) (1986) "Current Strategies of Rehabilitation of the Laryngectomized Patient" Seminars in Speech and Language, Vol 7, No 1, February 1986

[2] Hammarberg, B.(1987) "Erfarenheter av rehabilitering med Panjeventil." Lary nr 2/3 1987, sid 16-20 (Tidskrift utgiven av Riksförbundet för laryngektomerade)

[3] Robbins,J, Fischer,H, Blom,E & Singer,M. (1984) "A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production" J of Speech and Hearing Disorders, Vol 49, pp 202-210, May 1984

[4] Robbins,J , Fischer,H, Blom,E & Singer,M. (1984) "Selected acoustic features of tracheoesophageal, esophageal, and laryngeal speech" Arch Otolaryngol, Vol 110:670-672, Oct 1984

[5] Tardy-Mitzell,S, Andrews.M & Bowman,S (1985)"Acceptability and intelligibility of tracheoesophageal speech" Arch Otolaryngol, Vol 111, pp 213-219, April 1985

[6] Gandour,J & Weinberg,B (1985) "Production of intonation and contrastive stress in esophageal and tracheoesophageal speech" Journal of Phonetics 13, pp 83-95

[7] Weinberg,B, Gandour,J (1986) "Prosody in alaryngeal speech" from 'Current Strategies of Rehabilitation of the Laryngectomized Patient' Seminars in Speech and Language, Vol 7, No 1, pp 95-106, February 1986, Shanks,J (Editor)

[8] Hammarberg,B. & Nord,L. (1987) "Matstrupstal kontra tal via talventil - akustisk analys av två sätt att tala efter laryngektomi." Paper presented at the symposium 'Tal, Ljud, Hörsel', Lund, October 1987

[9] Hammarberg, B. & Nord, L. (1987) "Traditional esophageal speech versus tracheo-esophageal speech - acoustic and perceptual aspects." Paper presented at the 'Seminar on Post Laryngectomy Voice Rehabilitation', Göteborg, November 1987

[10] Hammarberg,B & Nord,L (1988) Phoniatric and Logopedic Progress Report, Huddinge Univ Hospital (forthcoming)

[11] Nord,L & Hammarberg,B (1988) STL-QPSR report from the Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH) (forthcoming)

[12] Rothenberg, M (1973) "A new inverse-filtering technique for deriving the glottal airflow waveform during voicing", JASA 53, pp 1632-1645

[13] Hammarberg,B. (1986) "Perceptual and acoustic analysis of dysphonia." Studies in Logopedics and Phoniatrics No 1 (Doct Dissertation), Dept of Logopedics and Phoniatrics, Huddinge Univ Hospital, Huddinge

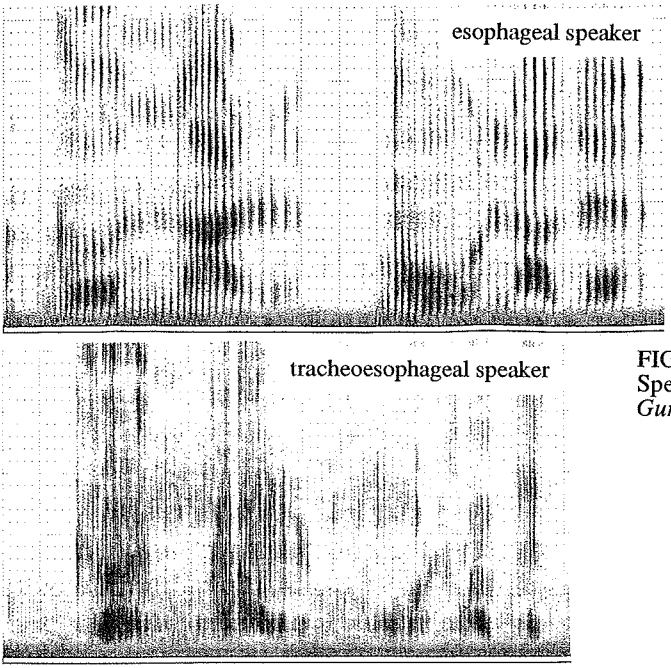


FIGURE 1
Spectrograms of
Gunnar målade...

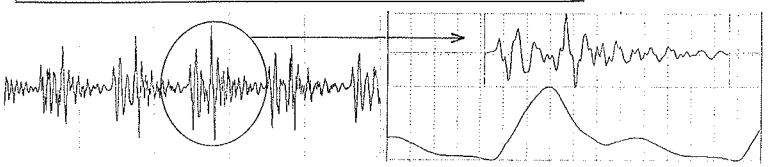


FIGURE 2 Vowel /a/ Extracted voice source pulse (TE-speech)

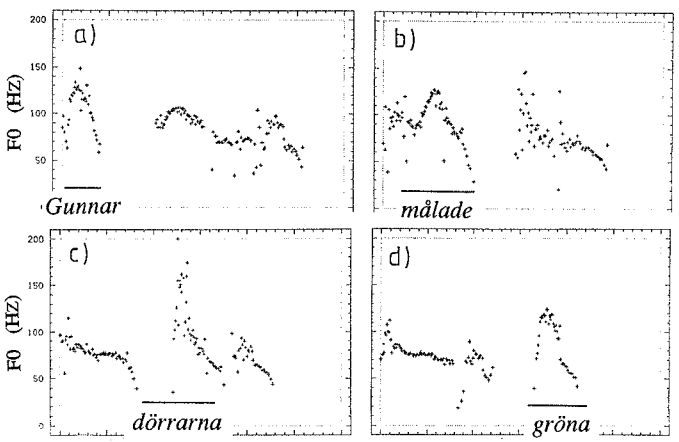


FIGURE 3 F0-tracing of sentences with emphatic stress (TE-speech)

THE ROLE OF INTRINSIC FUNDAMENTAL FREQUENCY IN THE PERCEPTION OF SINGING

Niels Reinholt Petersen
Institute of Phonetics
University of Copenhagen

Introduction

It is a well established fact that in speech high vowels have higher fundamental frequencies (Fo) than low vowels, everything else being equal. Although the mechanisms underlying these intrinsic Fo differences among vowels are still subject to discussion, it is generally aggreed that, in contradistinction to the Fo variation associated with higher level prosodic categories (such as stress, tone, intonation), they are not under the voluntary control of the speaker.

In singing where a high degree of precision of Fo control is crucial for the realization of musical categories intrinsic Fo differences between high and low vowels are also observed, although they are smaller than those found in speech. In a study by Greiffenberg and Reinholt Petersen (1982) on unaccompanied solo-singing, two trained singers were instructed to sing an interval of an ascending minor third (embedded in a short carrier tune) on the vowel sequences *a-a, a-u, u-a,* and *u-u.* As it turned out Fo in the vowel *a* was lower than in *u* by 40 cents on the average (range 21-50 cents) at the same intended pitch, and - as a consequence - the minor third was produced almost one semitone wider in the *a-u* vowel sequence than in the *u-a* sequence (see figure 1). The two singers also read the text of the song, and here the intrinsic Fo differences were far larger than in singing, averaging 320 cents in stressed and 75 cents in unstressed syllables.

Intrinsic Fo differences in singing have also been also reported by Ternström, Sundberg, and Colldén (1983 and 1987) who had singers sustain a tone at a given pitch and make a change of vowel quality at mid-tone. A change between high and low vowels produced a fundamental frequency change approaching one semitone in a great majority of cases, when the singers' auditory feed back was masked by noise. With normal auditory feed back the Fo changes were smaller, but equally frequent.

From the results outlined above it is evident that the effect of vowel height on Fo is reduced in singing as compared to speech, but it is not eliminated and it is still well above the threshold of detection. On this basis the question arises why the singers listening to their own voices did not detect their deviations from the intended musical intervals or their fundamental frequency changes in the sustained notes.

The experiment reported below was intended to try out the possible explanation that in unaccompanied solo singing (choir-singing and accompanied singing may be different) listeners, including singers listening to their own voices during singing, perceptually compensate for the intrinsic Fo differences among vowels, i.e. the perceptual system expects, as it were, the same note to be realised at slightly different fundamental frequencies depending on the vowels sung.

Method

The test stimuli were synthetically produced tunes consisting of four *dV* -syllables. There were four different vowel sequence conditions, namely *dadadada, dadadadu, dudududa,* and *dudududu.* The first three syllables all had a fundamental frequency of 110 Hz (A2). In the fourth syllable Fo was varied in 11 steps from 124 to 139 Hz, i.e. from 207 to 405 cents above the first three syllables. The stimuli are schematically displayed in figure 2. The steps were

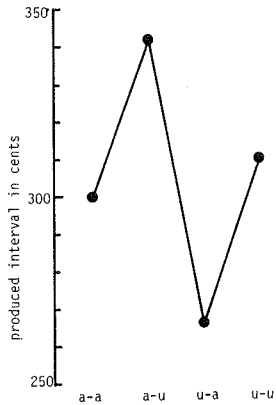


Figure 1: An intended minor third as produced under different vowel sequence conditions (adapted from Greiffenberg and Reinholt Petersen 1982).

intended to be 20 cents, but had to be rounded to the nearest integer number of Hertz, resulting in unequal step sizes. The eleven Fo steps were combined with each of the four vowel sequence conditions, making altogether 44 stimuli which were arranged in 10 different random orders in a test tape.

The test was taken individually by two musically trained female listeners, both of whom were university students of musicology and phonetics as well as experienced singers. The task of the listeners was to decide for each stimulus whether the musical interval between the third and the fourth syllable in the tune was a *major second,* a *minor third,* or a *major third.*

Results

The response patterns of the two listeners were very similar, and it was, therefore, thought jusitified to pool them. The identification functions for the three musical intervals are displayed in figure 2 (left graph), together with the 50% cross-over points (right graph).

It is clear that the boundaries are shifted systematically as a function of vowel height. The boundary between a major second and a minor third is about 20 cents higher in the *a-u* sequence than in the *u-a* sequence. The same pattern of variation is seen in the boundary between a minor and a major third, although the effect of vowel height is less pronounced. It is also seen from figure 2 that the median of the stimulus values having received minor-third-responses is influenced systematically by vowel height. An Extended Median Test showed the effect to be significant for the lower boundary ($p<0.001$) and for the median of the minor third responses ($p<0.05$), but not for the higher boundary ($p>0.05$).

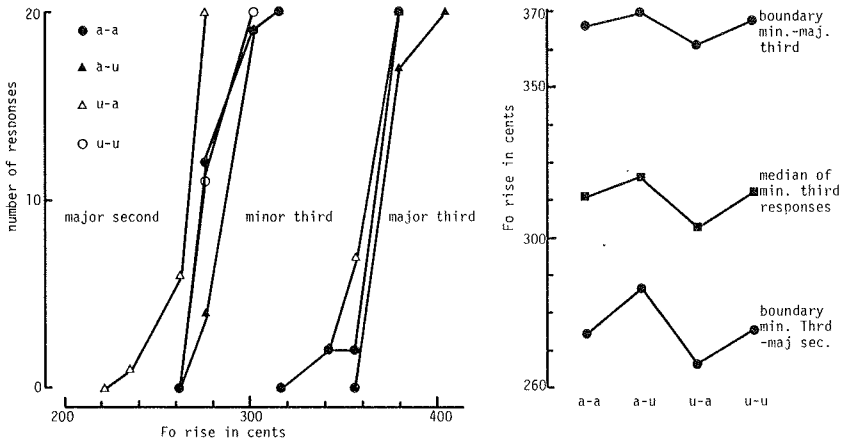


Figure 2: Identification functions (left), and 50% cross-over points and median values of minor-third-responses (right).

Discussion

The results of the experiment reported above has shown that the perception of musical intervals is systematically influenced by vowel height, and is thus giving support to the hypothesis that listeners to solo-singing perceptually compensate for vowel height. In addition to this main finding there are, however, two points which will also have to be considered in the discussion of the nature of the compensatory effect.

Firstly, the effect of vowel height revealed in the present experiment is far smaller than should be expected on the basis of the production data reported in Greiffenberg and Reinholt Petersen 1982 (cf. figure 1 above). The difference between extremes (i.e. a-u and u-a sequences) was about 80 cents in the production of an intended minor third, whereas the corresponding difference between medians of minor-third-responses in the listening test was only about 13 cents.

Secondly, as can be seen from figure 2 the effect is greater by a factor of more than two at the boundary between a major second and a minor third than at the boundary between a minor and a major third.

Research into the perception of pitch of vowels seems to give evidence for the existence of two types of vowel quality effect on perceived pitch. Although both are working in the direction to be expected under a hypothesis of perceptual compensation for intrinsic Fo, the two types of effect have - in my view - distinctly different origins (a more detailed treatment of this point may be found in Reinholt Petersen 1986).

One effect is a psychoacoustically conditioned *pitch bias effect* (as it has been aptly termed by Chuang and Wang 1978), and can be accounted for by reference to pitch perception processes common to the general class of complex periodic signals, of which vowels are, of course, a subset. The pitch bias effect is quite small, ranging from less than 1% to less than 3%.

The other type of vowel height effect on the perception pitch may be assumed to be speech specific, conditioned not by the spectral properties as such but by the phonetic characteristics of the vowels. It is demonstrated in experiments where listeners are to decide about prosodic categories of language (Rosenvold 1981, Silverman 1985), and - in contradistinction to the pitch bias effect - the magnitude of the phonetic compensation effect is very close to what should be expected from the intrinsic fundamental frequency differences found in speech production, i.e. about 10%.

In the perception of solo-singing it might very well be assumed that only the pitch bias effect and not the phonetic effect were involved. Actually, the relatively small boundary shifts observed in the present experiment point in that direction. This explanation does not, however, account for the fact that far larger deviations from an intended musical interval go unnoticed by singers listening to their own voices. Nor does it account for the shift being greater at the boundary between a major second and a minor third than at the boundary between a minor third and a major third.

On this basis it may be argued that the phonetic compensation effect is involved, after all, in the perception of solo-singing, but contributing to the perceived pitch to varying degrees depending on factors having to do with the individual listener, or with stimulus or listening conditions in general. Within this theoretical framework the discrepancy between the results of the Greiffenberg and Reinholt Petersen experiment and those of the present could be accounted for as follows: Listening to a human voice, and particularly, perhaps, to one's own voice during singing may be assumed to increase the degree of involvment of the phonetic effect in the perception of pitch, whereas listening to a synthetic voice will reduce it. Similarly it may be hypothesized that some musical intervals, and consequently their adjacent boundaries, are more resistant - so to speak - than others to the phonetic influence on perceived pitch, possibly depending (as was suggested to me by Robert McAllister, personal communication) on the functional importance they have in music. If this view is tenable, it could be the explanation for the the vowel height effect being less pronounced at the boundary between a minor and a major third that at the boundary between a major second and a minor third.

References

Chuang, C.-K. and Wang, W. S.-Y. 1978: Psychological pitch biases related to vowel quality, intensity difference, and sequential order, *J. acoust. Soc. Am. 64,* p. 1004-1014.

Greiffenberg, M. and Reinholt Petersen, N. 1982: The effect of high and low vowels on the fundamental frequency in singing: Some preliminary observations, *Ann. Rep. Inst. Phon., Univ. Copenhagen 16,* p. 101-111.

Reinholt Petersen, N. 1986: Perceptual compensation for for segmentally conditioned fundamental frequency perturbation, *Phonetica 43/1-3,* p. 31-42.

Rosenvold, E. 1981: The role of intrinsic Fo and duration in the perceptuion of stress, *Ann. Rep. Inst. Phon. Univ. Copenhagen 15,* p. 147-166.

Ternström, S., Sundberg, J., and Colldén, A. 1983: Articulatory perturbation of pitch in singers deprived of auditory feedback, *STL-QPSR 2-3,* p. 144-155.

Ternström, S., Sundberg, J., and Colldén, A. 1987: Articulatory perturbation of pitch in singers, *Rep. Uppsala Univ. Dept. of Linguistics 17,* p. 84-86.

Progress report:From Babbling to Speech

Liselotte Roug och Ingrid Landberg
Department of Linguistics
University of Stockholm

The project "From Babbling to Speech" constitutes the Swedish part of a larger international project. The study is persued in collaboration with the linguistic departments in Stanford, Paris and Padua. The aim of this work is to determine how and when the young child first aquires language specific phonetic and phonological characteristics from the ambient language. By observing children acquiring different mother tongues - American English, French, Japanese, Italian and Swedish - we hope to be able to define at what point and by what means this acquisition takes place.

Method

Six children, three girls and three boys are audio- and video-recorded on a bi-weekly basis, in their homes from the age of nine months until they spontaneously use 25 words in a recording session, the so called 25-word point. Each session lasts 30 minutes and consists of 20 minutes of mother-child interaction followed alternatively by 10 minutes of child playing alone with private toys or 10 minutes of child playing with a group of standardised toys brought by the assistant. One session per month is transcribed in order to determine the lexical stage of the child. At the 4- and 15-word point an extra recording is made to ensure sufficient data from these points. Six sessions, one each from the 0- and 25-word point and two each from the 4- and 15-word point are chosen for analysis. These word-points are chosen since they represent stable levels in the lexical development (Vihman and Ferguson 1986). It is expected that the child has at least twice the number of words in his/her total lexicon than that which occurs in a session. This means that the 4-word point is equivalent to a lexicon of at least 8 words and the 25-word point to at least 50. The word-point sessions are transcribed by two independent transcribers. For a transcription to be accepted as the basis for statistical analyses, agreement of at least 80 percent between the transcribers on place and manner for consonants and on number of syllables in the utterance is required. When the child is 13 and 16 months, an imitation test is made. This test is repeated if the child is older than 17 months when the 25-word point is reached. The aim of these tests and of the group of standardised toys mentioned earlier, is to allow for an evaluation of the cognitive development of the child. Once a month the assistant and the mother go through a questionnaire concerning the child's general communicative development. In connection with the imitation test mentioned above, they also go through a more detailed questionnaire of the same type. The mother is asked to keep a diary on the child's lexical and general communicative development.

What constitutes a "word" ?

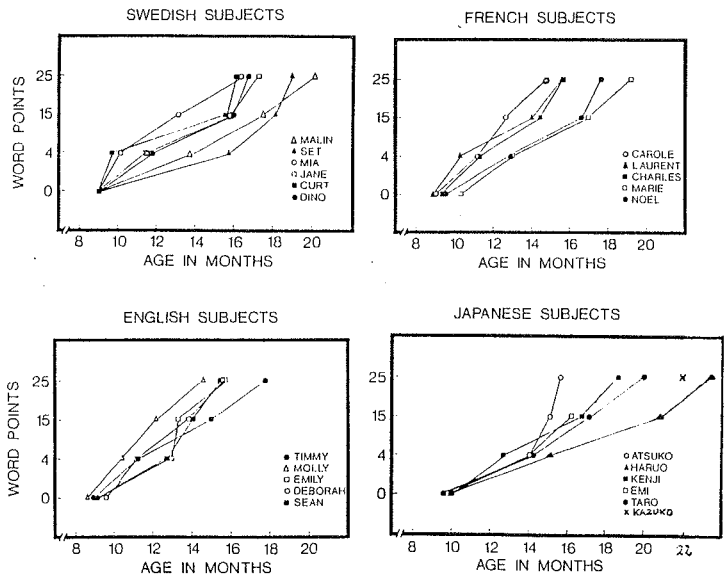
In order to determine the lexical stage of a child, it is necessary to identify the words that the child is using. A method of distinguishing early words from babbling has been developed at Stanford by Lorraine McCune and Marilyn Vihman (1988). My colleague Ingrid Landberg will describe this method in greater detail in the following presentation. Here it is sufficient to say that there are a number of criteria which a

vocalization suspected of being a word must meet in order to achieve "word-status". First of all, the vocalization must meet the criteria of plausible form and plausible context, whereby it achieves the status of "word-candidate" and is awarded two points, one for form and one for context. Each word-candidate must thereafter meet a number of additional criteria in order to be accepted as a word. Each criteria met is rewarded by a point. A word-candidate must achieve at least five points in order to be accepted as a word.

Analysis

The groups are analysed with respect to certain phonetic and phonological characteristics. These are tentatively divided into universal and language specific. The universal features are those that can be expected to occur in all children, irrespective of language background, these are: reduplicated babbling by ten months of age, preference for open syllables, single consonants and front vowels. Further, it is expected that the amount of "true" consonants (not glottals and glides) will increase during the period studied and that stops will constitute at least 50 percent of these. The universal features can be said to be a result of the child being exposed to a human language in general.

Figure I



These Figures show the age of the individual children in the four language groups at the four word-points.

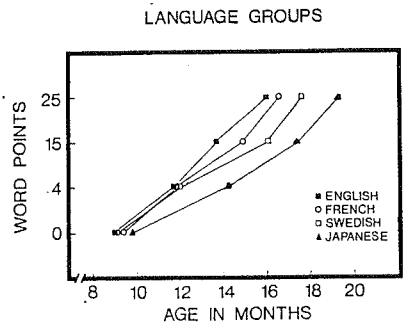
The language specific effects on the other hand, are those that can be said to result from the child being exposed to a specific language. These effects are divided into prosodic and segmental characteristics. The prosodic characterstics are: degree of reduction, length of vocalization, fundamental

frequency variation within words and vocalization closure. The segmental characteristics are: geminated (long) consonants, diphthongs, nasal vowels, front rounded vowels and liquids. The language groups are expected to differ in frequency of occurrence of these features. As mentioned earlier, six recording sessions will be analysed. The analysis consists of frequency counts of phonetic symbols and additional acoustic analysis. Further, certain parts of the mother's speech to the child will be analysed. This procedure is the same in the five research groups and constitutes the basis for comparison between them.

Preliminary results

The recordings have been completed in the English, French, Japanese and Swedish groups. In the Italian group the recordings have just begun. In Figure I we see the age of the individual children at the four word-points for the four language groups. From this Figure we conclude that the individual variation within each group is quite large. In the Swedish, French and English groups there appears to be a relatively stable amount of individual variation across time while in the Japanese group, it seems to increase. In Figure II we see the mean for each of the four language groups at the four word-points. We see that the Japanese children as a group, reach the 4-word point, about two months later than the others. The mean age in the English, French and Swedish groups lies at 12 months while the Japanese have an mean age of 14 months. At the 15-word point the four groups have spread over a three month period in the order from earliest to latest: English, French, Swedish and Japanese. This order remains at the 25-word point. Again, we want to stress that these are preliminary results that might be subject to later revision.

Figure II



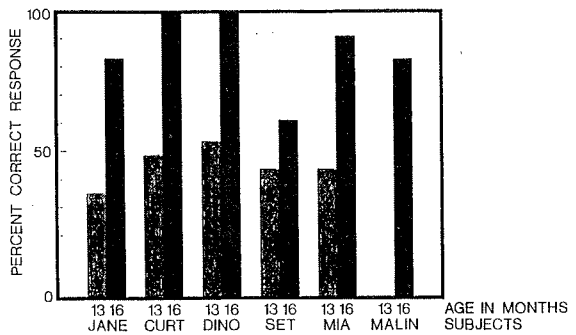
This Figure shows the mean in age at the four word-points.

We also want to point out that the individual variation within each group is so large that it seems difficult to draw any direct conclusions about language specific effects based on these data only. Even if it is too early to say anything definite about when and how the language groups differ, there are some characteristics which seem to surface differently in the groups. One such characteristic is the occurrence of longer babbling sequences, known as jargon. French and Japanese are traditionally analysed as having longer surface phonological

units than Swedish and English (Ferguson 1985). This is supposed to be reflected in a higher occurrence of polysyllabic vocalizations in the French and Japanese infants. This expected difference seems to be confirmed by the available data. Further, there appears to be a difference in the use of final consonants. The English infants have a higher use of final consonants than do the Japanese and French. There also appears to be a difference in preference of manner of articulation for consonants, so that the English group shows a higher proportion of stops while the French have a more even distribution of stops, fricatives, nasals and laterals.

In Figure III we present the results from the imitation test at 13 and 16 months. From the histogram we read that there is a dramatic increase in performance between 13 and 16 months. The tendency to imitate increases by 47 percent in three months (if the sixth child, Malin is excluded, the increase drops to 39 percent but is still impressive). Four of the children have reached their 4-word points by the time the 13-month test is conducted. The same four children have reached their 15-word point by 16 months and they all have over 80 percent correct responses at that test. The fifth child (Set) is just as successful as the other four at 13 months but fails to do as well at 16 months. The sixth child (Malin), does not respond correctly at all at 13 months but does as well as the first four at 16 months. Both these children (Set and Malin) were the last to reach the 4-, 15- and 25-word points. The purpose of this test is to determine whether the tendency to imitate can be correlated with the lexical development. A tentative conclusion is that there seems to be a correlation between these two factors so that the lexical "spurt" that occurs between the 4- and 15-word point co-occurs and to a certain extent might be said to depend on, an increased imitative ability.

Figure III



This Figure shows percent correct response of the four Swedish children at 13 and 16 months imitation test.

A final report on the Swedish findings of this project should be available by autumn 1989.

References

Ferguson, C.A. 1985. The Emergence of Ambient Language Effects on Children's Vocalizations. Proposal to the National Science Foundation. Mimeographed report.

Vihman, M.M. and C.A. Ferguson 1986. Phonological Development from Babbling to Speech: Common Tendencies and Individual Differences. Applied Psycholinguistics, 7 , 3-40.

Vihman, M.M. and L. McCune 1988. When is a Word a Word? Critique, methodology and analysis. Mimeographed report.

ON UTILIZING SIMILARITIES BETWEEN LANGUAGES: THE PERCEPTION OF DUTCH WORDS BY SWEDISH SPEAKERS*

Eva Strangert
Department of Phonetics, University of Umeå

Introduction

Swedish and Dutch are closely related languages with apparent similarities phonologically as well as phonetically. In consequence it is possible for those who have Swedish as their native language to understand Dutch to a high degree. People without any previous knowledge of Dutch may understand about 60 per cent of the words in newspaper texts, while spoken words in a textual context are somewhat more difficult (Hedquist, 1985). If only Dutch words related to Swedish are considered and if these words are presented in isolation, the figures for correctly understood written and spoken words are about 60 and 35 per cent respectively.

This paper focuses on some of the difficulties with spoken Dutch leading to misunderstanding by Swedish listeners. Only isolated words and words related to Swedish are accounted for. The data to be presented refer to 12 subjects lacking previous experience with Dutch. The subjects translated the words and noted their pronounciation by representing their component sounds with letters.

The specific difficulties to be discussed are:

- /x/, the unvoiced velar or uvular fricative which corresponds to /g/, /k/ or /f/ in Swedish. Compare Du. "goed", /xut/ and Sw. "god", /gu:d/; Du. "schat", /sxat/ and Sw. "skatt", /skat/; Du. "macht", /maxt/ and Sw. "makt", /makt/; Du. "kracht", /kraxt/ and Sw. "kraft", /kraft/.

- Consonant clusters with /r/ or /l/ and a following consonant into which there is often a vowel, described as /ə/, inserted between the consonants. Compare Du. "film", /filəm/ and Sw. "film", /film/.

Some erroneous translations of words with /x/, /sx/, and clusters with /r/ and /l/ are listed below:

	Dutch	Swedish	Translation
/x/	god	goed	rot
	elegant	elegant	elefant
	weg	väg, väck	vers
/sx/	schip	skepp	tripp,skräp
	schat	skatt	skratt
	schouwen	skåda	skruva
cons clust	merk	märk, märke	märklig
	volk	folk	följa
	film	film	fylla
	urn	urna	öron, ören

Regarding /x/ and /sx/ the translations contain an /r/ in the position corresponding to /x/, the only exception being the word "elegant". Also, the words with /r/- and /l/-clusters are translated into two-syllable words. To find out why such translations occur the following questions need to be answered:

- How are the sounds perceived? And, more generally:

- What is the relation between the translation and the perception of the sounds in a word?

The perception of /x/, /sx/, and clusters with /r/ and /l/

The perception of /x/ and /sx/ in different positions is shown in Table I.

Table I. The perception of /x/ in initial, medial, and final position and /sx/ in initial position for different groups of words as represented with letters by the subjects. Distribution of answers in per cent. N = the number of words in a specific category (the number of words in the group x 12 subjects).

	sch	ch	h	schr	chr	hr	sr	r	rch	*
/x/ init N=144	15	22	10	6	20	15		6		7
/x/ med N=60	5	55	2		3	10		13	5	7
/x/ fin N=96	3	76	1		2			6	10	1
/sx/ init N=144	10			62	1		14			13

* miscellaneous low-frequency answers

Transforming the letters into sounds the high frequency of /r/- or fricative + /r/-representations appear from the table. The figures also point to a strong tendency by the listeners to adhere to the principle of sonority (see e.g. Sigurd, 1965) in those cases where /x/ is represented as a combination of /r/ and a fricative: /r/ occurs after the fricative in initial position, before or after in medial position, and before in final position. Further, the tendency to perceive an /r/ is extra strong in the case of /sx/.

These representations thus explain why there are translations with /r/ in words with Dutch /x/. Table II (below) likewise explains the occurrence of two-syllable answers for words with /r/- and /l/-clusters.

Table II. The number of perceived vowels and their quality in clusters with /r/ and /l/ as represented with letters by the subjects.

	i	ü	e	ä	a	u*	o	å
nerf			2					
verf	1		4	1				
merk	6		3					
urn			10			1		
hoorn			11					
eekhoorn			12					
film	2		5			1		
kalm			7		1	2		
elk			6	1	4	1		
melk			4	1	5		2	
golf		1				2	7	2
volk						2	8	2
wolk			2			2	7	1

* includes ə

There is a general tendency to perceive a vowel between the consonants in the cluster, though this tendency is stronger in some cases and weaker in others. The quality of the vowel seems moreover to be contextually determined with /r/ and /l/ as well as the preceding vowel as conditioning factors. Thus the listeners perceive several different vowel qualities rather than one single schwa-like vowel.

It is reasonable to assume that the different perceived qualities reflect different formant patterns due to coarticulation with the surrounding sounds. Similar coarticulation effects on the formants of reduced (unstressed) vowels have been described by Nord (1986) for Swedish.

The relation between the perception of the sounds in a word and its translation

Apparently, knowledge of how the sounds of Dutch are perceived may be used to explain how the words are translated. However, this is not always the case. Often there is a clear discrepancy between a subject's representation of the sounds in a word and the translation. Thus, not everyone giving an /r/-representation for Dutch /x/ gives a translation with an "r" in it. And likewise, the translation "elefant" for Dutch "elegant" does not reflect the perception of /x/ in this word. None reports hearing an /f/ for /x/. The listener apparently is very flexible as regards the sound impression, thus indicating a complex relation between sound and translation.

The observations make it reasonable to assume that normally the analysis of the sounds in a word is very fragmentary. The listener interprets the meaning of a word on the basis of 'percept skeletons' which are derived from both input driven and knowledge driven cues (see Sajavaara, 1986, p 75; see also Bannert, 1987). The analysis is seldom detailed. It is the whole, the 'gestalt', rather than the parts that determines how a word will be translated. Also, it is reasonable to assume that disregarding the details and concentrating on the whole is a prerequisite for understanding a related language in which all words deviate to a greater or lesser degree from those in the native language.

References

Bannert, R. 1986. From prominent syllables to a skeleton of meaning: a model of prosodically guided speech recognition. Working Papers 29. Department of Linguistics, Lund University, 1-30.

Hedquist, R. 1985. Nederländares förståelse av danska och svenska. Skandinavisk-Nederländsk Språkförståelse, Report 3. Departments of Phonetics and Nordic Languages, Umeå University.

Nord, L. 1986. Acoustic studies of vowel reduction in Swedish. Speech Transmission Laboratory uarterly Progress and Status Report (STL-QPSR) 4. Department of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm, 19-36.

Sajavaara, K. 1986. Transfer and second language speech processing. In Kellerman, E. & Sharwood Smtih, M. (eds.) Crosslinguistic Influence in Second Language Acquisition. New York: Pergamon Institute of English, 66-79.

Sigurd, B. 1965. Phonotactic Structures in Swedish. Lund: Scandinavian University Books.

Strangert, E. Hedquist, R. 1988. Hur svenskar uppfattar och förstår nederländska ord. (Forthcoming).

* This paper elaborates on some of the results in a more comprehensive study of the perception of written and spoken Dutch conducted in co-operation with Rolf Hedquist, Department of Nordic Languages, University of Umeå, see Strangert & Hedquist (1988).

SPEECH DIRECTED TO CHILDREN

Ulla Sundberg

DEPARTMENT OF LINGUISTICS

UNIVERSITY OF STOCKHOLM

In speech communication different modifications or changes are made in order to adapt the speech to the listener and to different conditions. The speech style used when talking to young children has not been thoroughly investigated and yet it is an important and interesting field as a potential in speech acquisition.

In this article I will present a short background and give some examples from the literature. I will also make a short presentation of a pilot study regarding fundamental frequency, pitch range, stress, duration, and tempo.

Previous research

The earliest studies of child directed speech (CDS) were mostly discriptive and did not pay attention to what the child said or did. With these studies one wanted to refute the view that language acquisition was based on the child's innate capacity and almost independent of the language environment. Chomsky claimed at this time, i. e. the late sixties, that the utterances which the child hears are ill-formed, contain false starts and ungrammaticalities and that they break against rules. Later studies made quite opposite findings; the CDS is easily transcribed thanks to its clearness and well-formedness and that it is characterized by few hesitations, few disfluencies and that false starts only rarely occurred, to give some examples.

Bellinger points out the consistency and absence of variability in mother's speech to their children. He describes how children's ages very precisely could be predicted from the mothers' speech. The rate at which the mothers' speech changed was greatest when the children were between 1;8 (one year and eight months) and 2;3 and least when they were between 2;3 and 5;0.

According to Kaye (1980), CDS has some universal characteristics falling into five general classes:

1. Prosodic features - higher pitch, greater range of frequencies, more varied intonation.
2. Lexical features - special forms like pippi or vovve in Swedish.
3. Complexity features; shorter utterances, fewer embedded clauses, fewer verb auxiliaries etc.
4. Redundancy features; more immediate repetition and more repetition of the same words or phrases over a period of time.
5. Content features; restriction to topics in the child's world.

Some of these phenomena are verifications of findings earlier made by Ferguson (1984) in a cross-language study of Baby Talk.

He regards e. g. (1) repetition, (2) the exaggerated intonation-contours, and (3) reduplication of words and parts of words as probable universal features. He also claims that the most common phonological changes are simplifications of consonant clusters and replacement of /r/ by another consonant. Also, there seems to be a general tendency of using labial and apical stops and nasals in Baby Talk.

The modifications of speech directed to children must also be related to the age and sex of the addressee as well as the sex of the speaker him-/herself.

Concerning the age of the addressed child Bellinger's study was mentioned

above. His findings are to some extent in coherence with results obtained by Garnica (1977) in an investigation of prosodic and paralinguistic features of speech to two - and five year old child listeners and adult listeners. She noticed, for example, that the average fundamental frequency was significantly higher in the speech directed to the two year old than to the five year old children. The frequency range of the speaker's voice was greater in speech directed to the smallest children. In the speech directed to the two year old listeners instances occurred of rising sentence final pitch terminals with no grammatical motivation.

Regarding the relationship between speech directed to children, the sex of the addressee, and the speaker Johansson (1982) found that the sex identity was reflected in the fundamental frequency as follows; men used higher FO and greater frequency range in their speech to six year old girls than to boys while women more often used high vocal pitch and extended pitch range in their speech to boys in comparison to girls.

There is a variety of explanations and more or less speculative theories concerning the function of the modifications in speech style that are being used when addressing young children. The speech lesson theory claims that the speech adjustments would make it easier for the child to learn language structure. This theory is based on a "fine-tuning" hypothesis meaning that the speech directed to children is perfectly and optimally adjusted to it's capabilities in different aspects.

Fernald (1983) discussed an interesting theory on small children's perceptual and affective predispositions. In her opinion the expanded pitch range, the higher degree of tonal and temporal continuity contribute to make CDS suitable for the child from a perceptual point of view.

Garnica (1977) suggests the special features of CDS to have analytic and/or social functions. The former function is said to facilitate the child's analysis of linguistic material; for example the rising pitch terminals might help the child to locate sentence boundaries. The social function of certain features is to initiate and maintain communication between adult and child.

As we can see there is no commonly accepted explanation why the modifications in CDS occur. Also, similar changes of speech occur when people talk to flowers (Plant Talk) or to pets (Animal Talk) and even sometimes between adults in intimate situations.

Pilot study

From the recently closed Jollerprojektet at The Dept. of Linguistics SU, I chose typical utterances made by three mothers. They were addressing their 14 months old children. These utterances are called SV (spontaneous version). They were written down and presented to the mothers again. The women read, in a neutral way, their own utterances in an anechoic chamber. These utterances are called NV (neutral version). The recorded speech signal was fed into a computer programmed to analyze fundamental frequency and intensity in terms of means and standard deviations. The programme is from The department of Speech Communication and Music Acoustics at KTH (Royal Institute of Technology), Stockholm.

The syllables with primary stress have been underlined. Pause length is measured between main clause and "tag"-question.

The fundamental frequency was, as expected, higher in the SV: between 15 and 60 Hz. In one utterance, one of the mothers reached a top of about 400 Hz while in the NV the corresponding value was 280. Utterance K 6, however, is an exception. This phrase was pronounced in a slightly reproachful way.

Frequency range and standard deviations were greater in SV as compared to NV. The figures give examples of the exaggerated intonation contours, that many investigators have found in CDS. K 6 is an interesting example of fundamental frequency expansion in the lower parts of the frequency range.

Table 1. Results from computer analysis

UTTERANCES		MEAN FUNDAMENTAL FREQUENCIES Hz	PITCH RANGE Hz	STANDARD DEVIATION Hz	LENGTH OF UTTERANCES sec
J16	SV	260		(63)	3,3
	NV	220		(43)	2,5
J12	SV	270	250-440	(58)	0,9
	NV	207	150-280	(33)	0,7
K6	SV	140	100-200	22	4,5
	NV	185	160-240	32	3,9
K11	SV	210	170-280		3,1
	NV	190			3,9

UTTERANCES.

J 16: "Jätteduktig har du vart! Jaa!" ("You have been so good! Yes!")
J 12: "Tea time"
K 6: "Du får inte utveckla ett sånt där humör gumman, kan vi inte stå ut med, nä." ("You mustn't develop such a temper, girl, we can't put up with that, no.")
K 11: "Va är du ute på för äventyr? Ha?" ("What are you up to? Say?")

Figures in parenthesis are uncertain.
SV = spontaneous version, NV = neutral version

Table 2. Results from spectrogram/computer analysis

UTTERANCES		DURATION OF PRIMARY STRESSED SYLLABLE	PAUSE LENGTH	LENGTH OF "TAG"-QUESTION
K8	SV	45 csec	205 csec	40 csec
	NV	22	0	36
V4	SV	36	a) 80 b) 40	30
	NV	24	a) 32 b) 24	15
J16	SV	85	22	78
	NV	42	40	40
V10	SV	28		
	NV	39		

Utterances:

K 8: "Va håller du på med? Hörru!" ("What are you doing? Hey you!")
V 4: "Hur många magar har du egentligen va, säj? Ha?" ("Tell me, just how many tummies do you have? Hey!")
J 16:"Jätteduktig har du vart! Jaa!" ("You have been very good! Yes!")
V 10: "Inga fossingar i maten va!" ("No feet in the food, ha!")

Table 1 also shows that, in most cases, the speech was slower in SV than in NV. The primary stressed syllables had longer durations in SV as compared to NV; in two cases, K 8 and J 16, the difference was doubled. Utterance V 10 showed the opposite situation, which in my speculation might be due to this mother's elaborated attempt toward acting during the recording of the NV. These findings agree with those reported by Garnica, who agrues that a prolonged duration of stressed words might help the child

in identify the "key" words in the sentence. She even found many cases of two primary stresses within a sentence, but only in speech directed to the two year old children and not to five year old children or adults.

The length of pauses and length of "tag"-questions were greater in SV compared to NV. The function of tag- questions is discussed in a paper by Blount and Padgug (1977). They believe that tag-questions in adult-directed speech have several purposes: confirmation, query for information or it may serve as a device for interaction. In child-directed speech, however, there is only one function; to initiate and/or maintain interaction. The "tag"-questions are often characterized by a final fundamental frequency rise, and, according to these authors children react very early to this phenomenon.

The observations presented in this pilot study should be taken with some caution as they were based on a rather small number of measurements. Also, the comparison between SV and NV may reflect not only the difference between speech child-directed and neutral speech but also the difference between spontaneous speech and loud reading. The advantage in this type of comparison is that the linguistic material is the same.

In summary, the results of this investigation seem promising as they, by and large, agree with previous observations, and hence the method should preferably be applied to a larger material. For instance, the issue of speech reduction would be interesting to apply to child-directed speech in future investigations. Also, cross-language studies may reveal further interesting similarities between child directed speech in different speech communities.

REFERENCES:

Bellinger, D. (1980) Consistency in the pattern of change in mother's speech: some discriminant analyses. Journal of Child Language 7, pp 469-487.

Blount, B. and Padgug, E. (1977): Prosodic, paralinguistic and interactional features in Parent-Child speech: English and Spanish. Journal of Child Language, nr 1,vol.4, pp 67-86

Chomsky, N. (1967) The formal nature of language, in Lenneberg, E.H. (ed) Biological Foundations of Language, Appendix A, pp 397-442.

Ferguson, C.A. (1964): Baby Talk in six languages. American Anthropologist, 66, pp 103-114.

Fernald, A. (1983) The Perceptual and Affective Salience of Mother's speech to Infants. The Origin and Growth of Communication. Feagans et al.

Garnica, O. (1977) Some prosodic and paralinguistic features of speech to young children, in Snow, C.E. and Ferguson C.A. (eds) Talking to children: Language Input and Acquisition, Cambridge: Cambridge University Press.

Johansson, I. (1983) Sex-related and sex-directed features in questions of adults to sex-year-old children, in Dahlstedt, K-H, Hansson, Å, Hedquist, R. & Lindblom B. (eds) From Sound to Words. Essays in Honor of Claes-Christian Elert. Acta Universitatis Umensis 60, 1983, pp 169-178.

VOICELESS STOPS AND F_0 IN KAMMU

Jan-Olof Svantesson
Institute of East Asian Languages, Lund University

It is a well-documented fact that voiceless stops increase fundamental frequency in the first part of a following vowel, while voiced stops do not have this effect. This has been shown for several languages, including the tone languages Thai and Yoruba. It was found in these investigations that the F_0 perturbation due to a voiceless stop lasts for a relatively long time period, more than 100 ms in non-tonal languages such as English and Swedish, but only for 30-40 ms in the tone languages that have been investigated. See Hombert 1978 and Silverman 1987 for a general discussion and a survey of the literature.

In Kammu, an Austroasiatic language spoken in northern Laos, two tones (high level, denoted ´ and low level, `) have developed in connection with the merger of voiced and voiceless initial stops, *voiceless stops (or other consonants) giving rise to high and *voiced stops to low tone, as in the example: **kaa* > *káa* 'seal'; **gaa* > *kàa* 'to climb'. The original state with an opposition between voiced and voiceless stops is preserved in some Kammu dialects (see Svantesson 1983 for details).

Experimental procedure

In order to find out if the voiceless stops of Kammu have a raising effect on F_0, recordings of 24 Kammu words consisting of open syllables were investigated. These words were obtained by combining each of the four voiceless unaspirated stops in Kammu (*p, t, c* and *k*) with the three (out of ten) long vowels *aa, ii* and *uu,* and the two tones. There are also aspirated stops (p^h, t^h, c^h, k^h) and implosives (*ɓ, ɗ*), which can be followed only by high tone; they have not been investigated here.

A list consisting of these 24 words inserted into a sentence frame *Ò lʌ̀ʌ ___ pàar pàt.* 'I said ___ twice.' was read six times (each time in a different, random order) by a male Kammu speaker, Kàm Ràw, living in Sweden. The recording was made in the sound studio at the Department of Linguistics, Lund University. The recording was digitized at the sampling rate 20 KHz, and the length of the six first periods of the vowel, as well as the duration of the occlusion and aspiration phases of the stops were measured directly from computer oscillograms, using an interactive wave form editor developed by Lars Eriksson. The vowel period durations were converted into frequencies. The main results are shown in Table 1 and Figure 1.

F_0 raising

The F_0 raising effect of voiceless stops is present in Kammu, and its magnitude (the average is 17.7 Hz in low-tone and 25.8 Hz in high-tone words) is comparable to what has been found for other languages. On the other hand, its duration is small, comprising mainly the first period of the vowel (i.e. 6-9 ms), whose F_0 is in all cases except for *káa* and *kíi* significantly higher than that of the second period. In most cases the second period has higher F_0 than the third, but the difference is usually not significant, and F_0 does not change significantly in

periods 3-6. In Table 1, the average F_0 of periods 3-6 is given, representing a stable pitch value for the beginning of the syllable. At the end of the syllable, fundamental frequency usually falls slightly, but the distance between the two tones remains fairly constant.

As seen in Figure 1, the raising effect is, in all cases except for *caa*, smaller for the low than for the high tone. Because of the rather large F_0 variation in the first period (see Table 1), this difference is significant only for *paa* ($p<1\%$), *taa* ($p<5\%$) and *kuu* ($p<5\%$). According to Silverman 1987:4.10, F_0 raising does not increase with increasing F_0.

Stop duration

Since the voiced stops have been devoiced fairly recently in Kammu, one possible cause of the different size of F_0 raising for high and low tone may be a remaining lenis/fortis difference in the stops, although oscillograms show that the *voiced stops have become completely devoiced. In order to investigate this, the duration of the stops was measured. As is usually the case, the 'unaspirated' stops have a short aspiration phase, longer (averaging 26.6 ms) for *k* than for *t* (12.3 ms), and still shorter (and difficult to measure) for *p*. Although analysed as a stop, *c* has a short (37.1 ms) fricative/aspiration phase.

There is no systematic difference in the total stop duration between words with high tone and those with low tone. Before the vowels *aa* and *ii*, the total consonant duration is greater in high-tone than in low-tone words, but before *uu* it is shorter. (The mean differences were 10.6, 10.0 and -1.6 Hz) This difference is significant only for *paa, pii, tii* ($p<5\%$) and *kaa* ($p<0.1\%$), however.

Intrinsic pitch

My data also show that the effect of vowel intrinsic pitch is smaller for the low than for the high tone. The mean F_0 for periods 3-6 is on the average 20.5 and 19.4 Hz higher for *uu* and *ii* , respectively, than for *aa* in the high tone, but only 11.4 and 11.5 Hz higher than *aa* in the low tone. Although the intrinsic pitch effect increases with increasing F_0 (Silverman 1987:4.4), the difference found here is much larger than the expected proportional increase.

Conclusion

The F_0 raising effect of a voiceless stop has a short duration in Kammu, affecting only one or two periods of the vowel. Raising is smaller for the low than for the high tone, and this can probably not be attributed to any remaining difference in the stop itself. These facts, as well as the fact that the intrinsic pitch effect is smaller for the low than for the high tone, suggest that Kammu speakers control these features in order to avoid the fundamental frequency of the low tone becoming so high that discrimination of the two tones is endangered.

This active control contradicts the theory that F_0 raising after voiceless stops and intrinsic pitch are due to purely aerodynamic factors. The simultaneous suppression of these two effects in the low tone may suggest a common control mechanism for them, probably involving the crico-thyroid muscle, whose activity has been shown to be correlated with intrinsic pitch by Dyhr 1988 and with F_0 raising after voiceless stops by Baer et al. forthc.

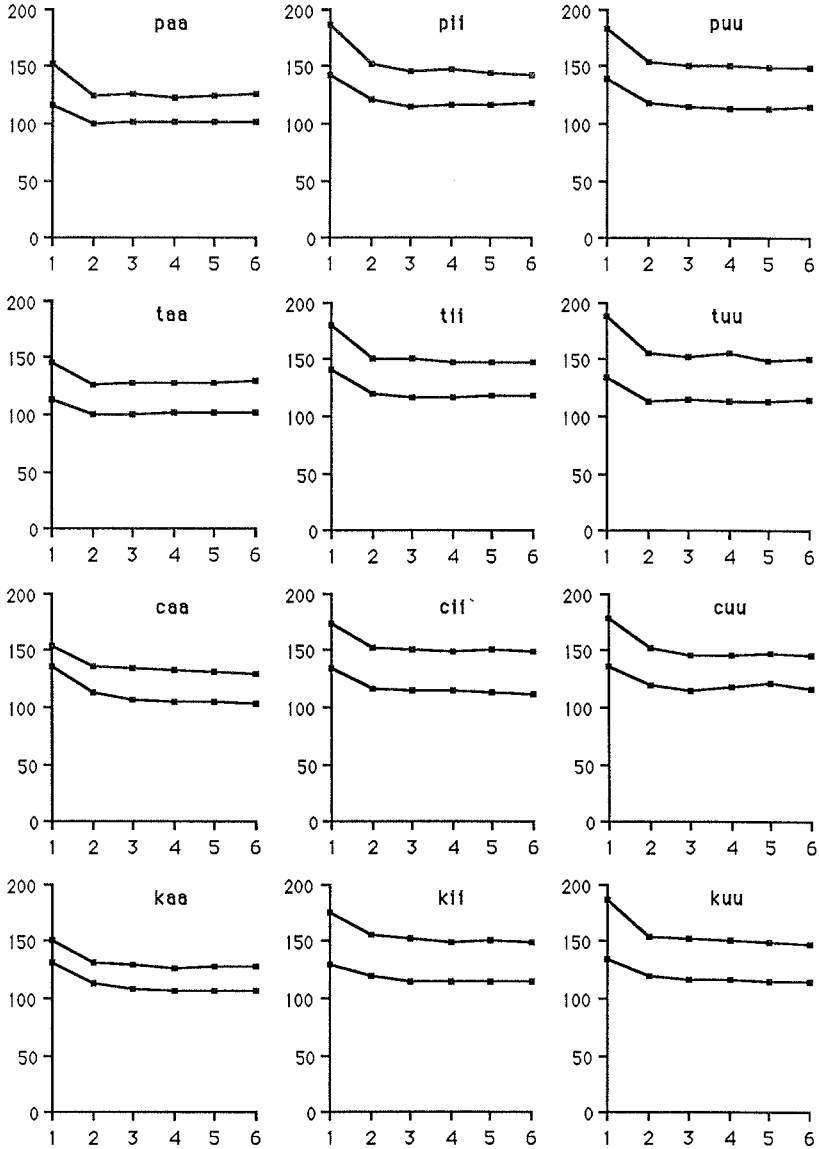


Figure 1. Fundamental frequency in the first six periods (Hz). Each point represents the average of six values.

Table 1. Mean values and standard deviations for the total duration of the stops, the duration of their aspiration or fricative phase, and for F_0 in the first two periods as well as the average F_0 in periods 3-6 (n=6 in all cases).

	Total duration (ms)		Asp./fric. (ms)		Period 1 (Hz)		Period 2 (Hz)		Period 3-6 (Hz)	
	x̄	s	x̄	s	x̄	s	x̄	s	x̄	s
pàa	122.5	6.57	-	-	116.6	5.01	100.5	3.65	102.1	2.66
páa	132.8	5.25	-	-	153.1	11.28	124.3	3.78	125.1	2.69
pìi	136.3	6.64	-	-	143.4	12.67	120.8	5.25	116.4	4.30
píi	147.6	8.96	-	-	186.1	8.58	151.7	8.70	145.0	3.48
pùu	142.8	3.80	-	-	140.1	3.47	118.0	3.50	113.8	3.37
púu	140.2	8.58	-	-	183.0	24.35	153.7	4.19	149.9	3.35
tàa	125.9	6.10	10.3	0.69	112.6	9.02	100.1	4.23	101.1	3.73
táa	129.3	4.43	8.9	0.68	146.1	4.72	126.9	3.80	128.6	3.28
tìi	130.9	5.83	17.0	1.90	140.2	9.18	120.4	4.63	117.2	2.39
tíi	140.6	4.92	13.3	2.51	179.5	24.77	150.7	5.47	148.1	4.23
tùu	145.7	4.49	13.0	1.72	133.8	8.77	113.7	4.09	114.4	4.65
túu	144.5	8.93	11.3	1.90	189.0	13.75	156.6	4.34	151.9	1.02
càa	118.0	6.90	27.1	3.41	135.7	10.80	112.7	5.15	105.1	4.89
cáa	123.7	5.55	28.2	1.41	153.5	19.27	135.6	5.81	134.7	5.48
cìi	137.4	10.65	56.4	7.29	133.8	17.91	116.1	5.91	113.6	3.33
cíi	144.4	15.35	39.7	5.85	173.4	14.27	152.0	9.19	150.0	8.35
cùu	145.9	11.07	38.8	6.84	135.5	14.33	119.3	5.84	117.5	2.57
cúu	145.1	8.13	32.4	6.04	179.1	23.92	152.4	4.33	146.4	4.61
kàa	105.1	7.32	24.2	4.55	131.1	19.59	112.7	6.25	107.0	1.34
káa	127.8	7.94	21.1	3.52	150.9	26.39	130.6	6.42	127.7	3.62
kìi	130.7	8.61	32.0	4.49	129.3	9.53	119.6	8.21	114.5	2.74
kíi	142.5	15.79	24.7	4.84	174.6	20.44	156.2	10.33	150.3	9.35
kùu	135.3	7.61	32.5	5.52	134.7	17.10	120.4	5.35	115.4	2.61
kúu	133.5	8.72	24.8	4.54	187.2	16.38	154.7	3.68	149.8	3.76

References

Baer, Thomas, Anders Löfqvist, Nancy S. McGarr and Robin Seider Story. forthc. 'The crico-thyroid muscle in voicing control'. ms.

Dyhr, Niels. 1988. *Intrinsik tonehøjde og aktiviteten i musculus cricothyroideus.* Specialeopgave i fonetik, Københavns universitet.

Hombert, Jean-Marie 1978. 'Consonant types, vowel quality, and tone'. *Tone: a linguistic survey,* ed. Victoria Fromkin, 77-111. New York: Academic Press.

Silverman, Kim. 1987. *The structure and processing of fundamental frequency contours.* Ph.D. thesis, University of Cambridge.

Svantesson, Jan-Olof. 1983. *Kammu phonology and morphology.* Lund: Gleerup.

DEFAULT SENTENCE ACCENTS AND FOCAL SENTENCE ACCENTS

Nina Grønnum Thorsen, Institute of General and Applied Linguistics,
University of Copenhagen

This paper summarizes part of the results of a comparative investigation of prosodic phenomena in Standard Copenhagen Danish, Regional Bornholm Danish, Skanian (Malmö) and Stockholm Swedish. For a complete treatment, documentation and references, see Thorsen (1988).

'Sentence accent' (SA) is to be understood thus in the following: one (or possibly more) of the stressed words in an utterance is (are) perceived as being more prominent than the other stressed words; this extra prominence is brought about with tonal means (larger and quicker pitch movements).

These are the points I wish to make:

(A) There are (at least) two different kinds of sentence accent
  (1) a prosodically or syntactically determined, final, *default* sentence accent - and
  (2) a contextually (semantically and/or pragmatically) determined *focal* sentence accent

(B) Neither default nor focal sentence accents are universal, i.e. there are (regional) languages without a compulsory default accent, and likewise focus does not necessarily get signalled prosodically.

(C) When focus does get signalled prosodically, there are (at least) two different ways to go about it
  (1) with a sentence accent
  (2) or merely by reduction of the prominence on the surrounding stressed words.

The material analysed consisted of three utterances, recorded in three different contexts:

Kamma stammer fra Svaneke (Næstved)./Kamma kommer från Svaneke.
Anders og Kamma skal til Sandvig (Fakse)./Anders och Kamma ska till Sandvik.
Torbens søster hedder Kamma./Torbens syster heter Kamma.

(Kamma comes from ... . Anders and Kamma are going to ... . Torben's sister's name is Kamma.)

These utterances occurred in isolation and as answers to two questions which were intended to provoke a focus signalling either ON 'Kamma' or on some other word in the utterance, as indicated by the underlining above.

The material was recorded by seven speakers from Bornholm, and two speakers from Copenhagen, Malmö and Stockholm, respectively. The tapes were processed by hardware instrumentation, whose output was registered on a mingograph. Fundamental frequency (Fo) was measured on a logarithmic scale to an accuracy of ± 0.25 semitones, the distance in time of each measuring point from the first one was likewise measured, with an accuracy of ± 0.5 centiseconds. The number and location of measuring points were determined in a manner which permits a faithful average reproduction of the six renderings of each utterance.

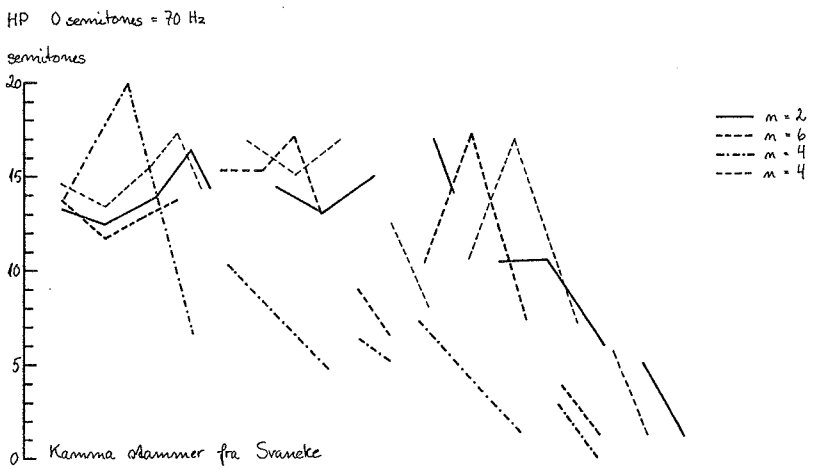
An auditory search for sentence accents (and focus indication by stress reduction) yielded the following results:

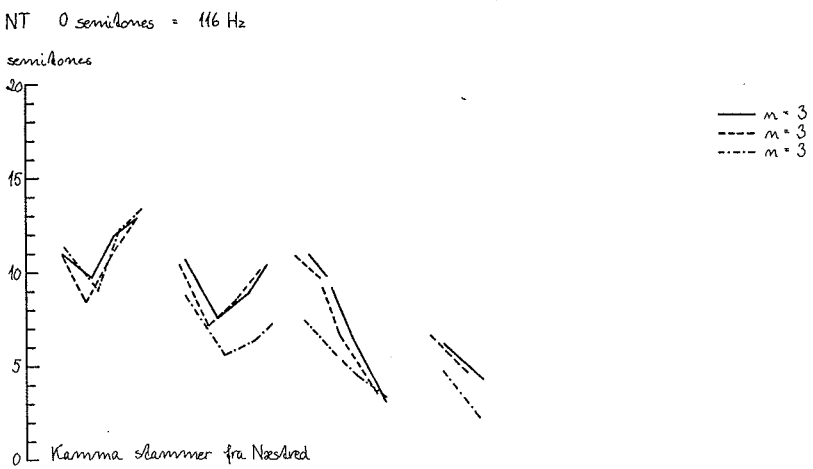
Stockholm: Isolated utterances receive a final, default, sentence accent always.
Contextually invited initial, medial and final focus gets signalled with a (focal) sentence accent always.

Bornholm: Isolated utterances receive no sentence accent in 64% of the items. They receive a final, default, sentence accent in 30% of the items. (These latter are due mainly to three of the seven speakers, the rest of whom did not produce default sentence accents at all.)
Invited initial focus gets signalled (with an SA) nearly 100%.
Invited medial focus gets signalled (with an SA) in 57% of the items.
Invited final focus gets signalled (with an SA) in 80% of the items.

Malmö: One speaker did not produce any sentence accents, be they default or focal, at all, nor any other focus indication.
The other speaker never produced any default accents in isolated utterances, and produced focus signals only reluctantly:
Invited initial focus got signalled in 41% of the items.
Invited medial focus got signalled in 17% of the items.
Invited final focus got signalled in 33% of the items.

Copenhagen: One speaker did not produce any sentence accents at all, nor any other focus signals.
The other speaker produced neither default nor focal sentence accents, but signalled focus (by stress reduction) in two positions:
Invited initial and medial foci got signalled every time, but no final foci got any prosodic marking.





The figures above are average tracings from a Bornholm speaker (HP) and a Copenhagen speaker (NT). HP illustrates the difference between an isolated utterance without default accent (full line), and one with a final default accent (thinner broken line). The default accent is phonetically distinct from the invited final focal sentence accent (thicker broken line): the focal accent involves a greater change in the preceding Fo course, i.e. the first stress group ('Kamma') is lower in the range and has a smaller movement, and the downwards twist of the second stress group ('stammer fra') is initiated sooner. A final focal accent may also be somewhat more

"boosted" than the default accent (not apparent in this example). Further, a final focal accent shortens the preceding part of the utterance more than does a default accent (and this is NOT due to the different conditions: isolated versus second utterance in a text, see further Thorsen, 1988, p. 64 f). The initial focal accent (dotted/broken line) is clearly "boosted" and nearly annihilates any autonomous Fo movements in the succeeding part of the utterance. Contrast this with NT, where the isolated utterance (full line) is indistinguishable from the utterance where a final focus signal was invited (broken line), and note how an initial focus (dotted/broken) line does nothing to affect the focussed item, but lowers and shrinks succeeding Fo patterns.

In brief: The default accent is compulsory in Stockholm, it is facultative in Bornholm (some never produce it, some do so, but not consistently), it is non-existent in Malmö and Copenhagen. Focus signalling is compulsory in Stockholm, it is facultative in Bornholm (where it is more frequent than the default accent, and there is no speaker completely without it), it is also optional in Malmö (but it is comparatively rare, and one speaker does not produce any prosodic focus signal at all). The situation in Copenhagen as regards focus signalling resembles Malmö, except that focus signals seem to be excluded finally.
The focus signal is a sentence accent in Stockholm and Bornholm whose manifestation is of the same kind as the default accent, only stronger. In Malmö and Copenhagen focus gets signalled, not by any particular boosting of the focussed item, but rather by a reduction of surrounding stress groups.

These are rather striking differences in the prosodic systems of languages that are otherwise obviously and closely related, and one is left to wonder if and how this may be tied up with differences in the spoken language syntax. This is obviously an immense field for future research.

Thorsen, Nina Grønnum, 1988: "Intonation on Bornholm - between Danish and Swedish", Ann. Rep. Inst. Phonetics, Univ. Copenhagen 22, 25-138.

PERCEPTION OF THE QUANTITY DISTINCTION IN SWEDISH /VC/-SEQUENCES.
Hartmut Traunmüller and Aina Bigestans

INTRODUCTION

In some germanic languages, including Swedish, there is a complementary relationship between the durations of [V] and [C] in [VC]-sequences. In positions carrying word-accent, Swedish phonotax allows only /V:/, /V:C/, and /VC:/ (Elert, 1964). In search for an invariant description of the durational properties of these sequences, invariant with respect to variations in speech rate and context, it has been proposed to consider the ratio of vowel duration to consonant duration (V/C), or the ratio of vowel to [VC]-sequence duration (V/(V+C)), (Bannert, 1976). Whithin the scope of the present investigation, the validity of these ratios as descriptors of the quantity distinction was tested, also considering the perception of the durations of any kind of speech segments in general.

METHOD

A female speaker of standard Swedish produced each word among the minimally distinctive pairs [ɕɛ:na] - [ɕɛn:a], [dɛ:sa] - [dɛs:a], and [mɛ:ta] - [mɛt:a] in a carrier phrase [de va ju ____ ja sɒ:], that was uttered at five different rates of speech, distributed over the whole range of variation ocurring in natural speech.

The utterances were recorded and a computer program allowing the elimination or repetition of single glottal periods, or sections of similar duration in unvoiced speech segments, was used to manipulate the durations of the [V]- and [C]-segments in such a way as to obtain a number of utterances in which these segments had durations interpolated between those occurring in the natural utterances with /V:C/ and /VC:/ sequences. The actually obtained durations of vowel, consonant, and utterance are listed in Table 1 for one example. The final vowel of the carrier phrase has been excluded from what we here refer to as "utterance duration". An extra fast speech rate was obtained by shortening the whole utterance spoken with the fastest speech rate our speaker could achieve by uniformly distributed removal of glottal periods.

Table 1: Segment durations (in ms) in one series of stimuli (1 to 5): Manipulated versions of [de va ju mɛ:ta ja sɒ:].

	Orig. [ɛ:t]	1	2	3	4	5	Orig. [ɛt:]
"Utterance"	1144	1136	1132	1128	1141	1129	1165
Vowel	220	185	155	129	107	92	101
Consonant	154	181	207	229	264	267	266

The utterances were presented via headphones in order of increasing speech rate, and in randomized order within each rate, to 25 adult subjects, all natives of Central Sweden, using a computerized system of response collection. The subjects had to identify the word containing the manipulated segments. The response alternatives were 1 (/CV:Ca/), 2 (/CVC:a/), and X (undecided).

RESULTS AND DISCUSSION

Figure 1 shows the results obtained for two series of stimuli based on original utterances opposed in quantity only. There were 18 such pairs of stimulus series, but the following presentation is limited to one crucial point in each of these: the mean position of the quantity

boundary, i. e., the mean of the two crossover points of the dashed and whole-drawn curves labeled 1 and 2 in Figure 1.

The durations of the vowel and the consonant for stimuli at the /V:C/-/VC:/-boundary are shown in Figure 2 as a function of total utterance duration. Figure 3 shows the ratio V/(V+C) as a function of utterance duration.

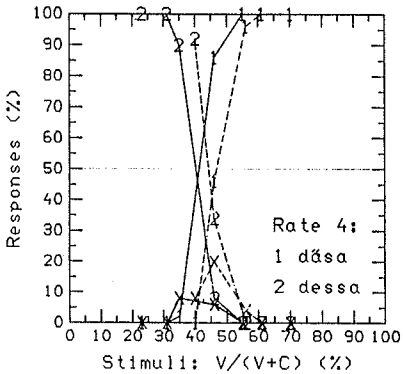


Figure 1: Identification results (see text).

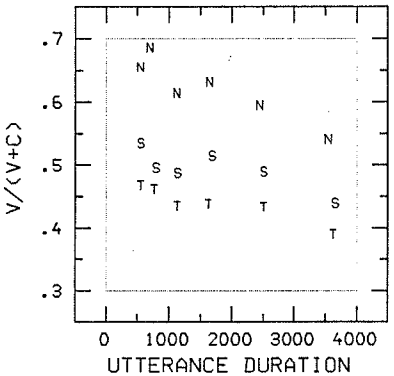
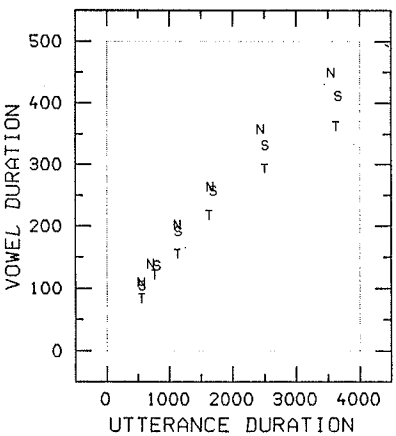


Figure 3: The durational ratio V/(V+C) at the /V:C/-/VC:/-boundary, as a function of utterance duration in ms.



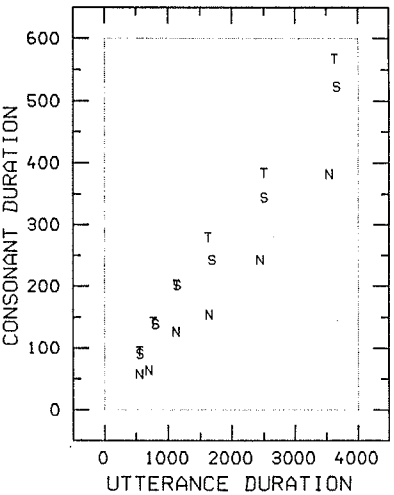


Figure 2: Vowel (left) and consonant durations (right) at the /V:C/-/VC:/-boundary, as a function of utterance duration. Durations in ms.

Footnote: There was an [h]-like segment (preaspiration) at the end of the vowel when followed by [t] or [s]. In Figure 1 this is included in the consonant duration, while it is included in the vowel duration in all the following figures.

If all segmental durations would vary proportionally with changing speech rate, implying invariant ratios V/C and V/(V+C), the data points in Figure 2 would fall on straight lines going through origo. In Figure 3 they would fall on lines in parallell with the x-axis. While this may be said to be the case in rough approximation, we observe regular deviations from such a course.

Non-proportional variations in segment durations have been observed and taken into consideration previously (Klatt, 1979; Lindblom et al, 1981). According to the approaches taken by these authors, our data points in Figure 2 should still fall on straight lines, but these need not go through origo. While an improved fit would be obtained in this way, inspection of our figures led us to the conjecture that a certain class of curved lines which do go through origo might describe our data still better and more adequately.

Figure 2 (right) shows substantial differences in duration between the consonants [n], [s], and [t]. In addition, the duration of the vowel is dependent on the following consonant, as can be seen in Figure 2 (left). Both observations agree qualitatively with what has been observed previously (Elert, 1964; Lindblom et al, 1981).

Figure 4 is equivalent to Figure 2, except for the scaling of the durations, which now is logarithmic. The linear regression lines shown in Figure 4 can be seen to fit the data very well. According to these lines, the relation between the durations Dy (y-axis) and Dx (x-axis) is described by $\log(Dy) = p \log(Dx) + c$, where p is the slope of the line and c its displacement. The slope of the line indicates the compliance, with respect to changes in speech rate, of the segment whose duration is plotted along the y-axis, as compared with the one plotted along the x-axis. If these segments vary proportionally in duration, then p = 1.0. Figure 4 shows the compliance of the vowel segment to be distinctly lower (0.76) than that of the whole utterance (1.00). This figure also shows the influence of the following consonant on the duration of the vowel to be roughly constant, expressed in logarithmic units, i. e., also in terms of a percentage.

All this means that a power-law with exponent p holds between Dy and Dx. We expect this kind of law to hold, in satisfactory approximation, for any kinds of segments, including also pauses, whose compliance p will be higher than that of any real speech segments. The power law also holds satisfactorily for the duration of the [VC] segment, as produced by our speaker, with p = 0.86, relative to the duration of the whole utterance.

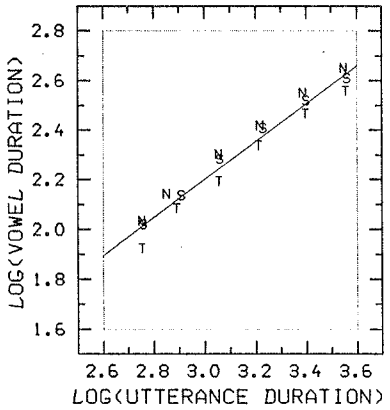
In this connection, it should also be noted that a power law with p = 2.0 has been found to hold between speech rate measured physically (syllables per minute) and psychophysically (Grosjean and Lane, 1981).

The power-law for segmental durations, in the following form, can be used, i. a., to vary the speech rate of synthetic speech:

$$D = k^p D_0 ,$$

where D is segment duration, D_0 its initial value, k the factor by which an utterance is to be streched in duration, and p the compliance of the segment relative to that of the whole utterance. However, before this law can be used practically, the segment specific values of p must be known, in addition to the initial durations.

Figure 5 is equivalent to Figure 3, but with utterance duration and the ratio V/(V+C) both scaled logarithmically. The slopes of the regression lines shown in this figure reflect the difference in compliance p of the vowel [ɛ] and the consonants [n], [s], and [t].



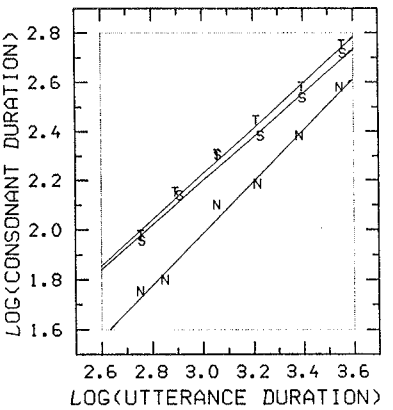


Figure 4: Vowel (left) and consonant durations (right) at the /V:C/-/VC:/-boundary, as a function of utterance duration. Durations logarithmically scaled. Same data as in Figure 2.

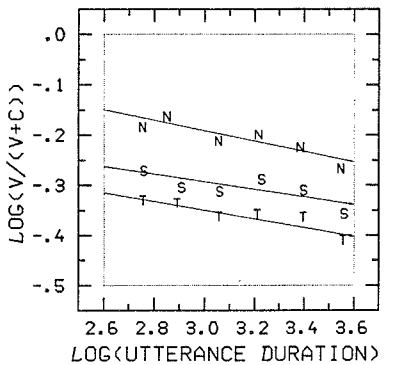


Figure 5: The durational ratio V/(V+C) at the /V:C/-/VC:/-boundary, as a function of utterance duration. Quantities logarithmically scaled. Same data as in Figure 3.

ACKNOWLEDGMENT

This research is supported by a grant from HSFR, the Council for Research in the Humanities and Social Sciences.

REFERENCES

R. Bannert (1976): Mittelbairische Phonologie auf akustischer und perzeptorischer Grundlage, Travaux de l' institut de phonetique de Lund, vol X, Gleerups, Lund.

C.C. Elert (1964): Phonologic Studies of Quantity in Swedish, Almqvist & Wiksell, Uppsala.

F. Grosjean, H. Lane (1981): Temporal variables in the perception and production of spoken and sign languages. In P.D. Eimas, J.L. Miller (eds.) Perspectives on the Study of Speech, Lawrence Erlbaum Ass., Hillsdale, New Jersey, pp. 297-237.

D.H. Klatt (1976): Linguistic uses of segmental duration in English: Acoustic and perceptual evidence, J. Acoust. Soc. Am., 59: 1208-1221.

B. Lindblom, B. Lyberg, K. Holmgren (1981): Durational Patterns of Swedish Phonology: Do they reflect short-term motor memory processes?, Indiana Univ. Linguistics Club.

F0 ANALYSIS PROBLEMS

Sidney Wood
Department of Linguistics and Phonetics
University of Lund

ABSTRACT
This paper reports an investigation of the sensitivity of speech analysis programmes to voice quality and signal quality, with a view to optimizing parameter settings for successful F_0 analysis. Results were more successful with data frame lengths of 10 ms than 5 ms and with better, less breathy, voices. In the speech samples used here, voice quality seemed to be more important than signal quality.

INTRODUCTION AND PROCEDURES

It is well known that F_0 analysis by ILS and similar programmes sometimes yields incongruous results with regard to phonetic properties like voicing and intonation. For example, they may fail to recognize the presence or absence of voice tone, on both vowels and consonants, and they may propose a high soaring tone on voiceless sibilants and plosive bursts. Anomalous analyses can always be cleaned up by editing the results, but this will obviously lead to circularity in some experimental contexts. It is particularly distressing if the analysis is required for a speech recognition device that is expected to identify segmental voicing/voicelessness or interpret an intonation pattern.

Obviously, phenomena like creak and vocal fry are readily identifiable discontinuities in the waveform that are perfectly audible disruptions to the otherwise melodic vibration of the voice tone. Cases like this are dealt with at length by Huber elsewhere in this volume. The puzzling cases are more subtle and examination of an expanded segment of the waveform does not always reveal just why the programme has overlooked the vocal vibrations during a vowel or has settled for harmonic vibration during the aperiodic hiss of a voiceless consonant. Sibilants and voiceless stops are particularly tricky.

Some voices seem to be more successfully analysed than others, suggesting that some personal factor is involved. One such

factor might be the degree of breathiness of the voice, which superimposes an aperiodic component and weakens the higher partials. Studio recordings from two speakers were therefore compared, both adult male native speakers of Bulgarian who read the same material, the one with a relaxed and non-breathy voice and the other very breathy.

One might also expect the quality of the recording to have some influence on the result, depending on the amount of distortion and on the signal/noise ratio. The presence of mains hum may also confuse F_0 analysis. A studio recording of originally good quality, of a trained professional speaker of Swedish, was degraded by being copied in several generations on primitive equipment in order to obtain a bad copy for comparison with the studio recordings of the two Bulgarian speakers.

As far as the programmes themselves are concerned, the speech samples were analysed by two methods for comparison: the ILS routines API (a modified cepstral processing technique) and SIF (the SIFT algorithm).

Finally, there are many programme parameters that can be set, and it need not be the case that the default settings are the optimum. The digitalized waveform is quantized into *data frames* of optional length, the longer the frame the poorer the temporal and spectral resolution but also the greater the chance that it will include at least one fundamental period. The default frame length, 64 data points, is optimized for input and output operations, not analysis. Frame lengths of 5 ms and 10 ms were used here. The LPC *analysis window* , of recommended length 15-35 ms, starts simultaneously with the data frame. It was set at 20 ms for this investigation. The analysis window is straddled by the *excitation buffer* that was kept at its default length of 32 ms; this buffer must be longer than the analysis window, around which it is shifted as the programme determines the periodicity of the current data frame. The periodicity decision depends on an excitation index, computed as a by-product of the cepstral analysis, exceeding a preset *excitation threshold*. This threshold (default 0) can be set by the user. Raising it sufficiently will purge the sequence of voiceless segments spuriously analysed as voice.

RESULTS AND DISCUSSION

With a *data frame* size of 50 ms, 11 out of 12 analysed sentences exhibited unexpected deviations in the F_0 curve. Lengthening the frame to 100 ms worsened one analysis, left 3

unchanged, somewhat improved 5 and produced an ideal result in 3. Lengthening the data frame thus improves F_0 analysis.

The choice of *analysis method* was inconclusive. Sometimes API gave a better result and sometimes SIF, apparently unrelated to the other factors investigated.

The results of the *poor signal quality* samples were only just slightly worse than those of the good studio recording of the relaxed non-breathy voice, and definitely better than those of the breathy voice. The speaker of the poor recording copy had a good, trained professional voice so that voice quality seems to be more important than signal quality for F_0 analysis.

The value of the calculated excitation index seemed to be slightly lower for longer data frames, which partly explains why lengthening the data frame improved the F_0 analysis. Raising the *excitiation threshold* certainly removed spurious voiced decisions, but often at the expense of losing some correctly voiced decisions. There may be an ideal setting for each individual voice, in which case the programme will need to be tuned to the speaker. Clearly, if the threshold has to be set differently for each sentence, we are back to an arbitrary and circular procedure again. One can certainly agree with the author of the ILS user notes for API, that analysis of the fundamental is an art and not science.

REFERENCES

Huber, D. (1988). Laryngealization as a boundary cue in read speech. P. 66 in this volume.

ILS. *The Interactive Laboratory System for Speech Analysis.*