Issues in the Perception of Prosody

Jacques Terken
Institute for Perception Research
P.O.Box 513, 5600 MB Eindhoven, The Netherlands

ABSTRACT

Three issues in the perception of prosody are discussed: the perception of pitch variation, of prominence and of phrase boundaries. Also, the relation of these issues to the transcription of prosody is treated, and a number of remaining questions are addressed.

INTRODUCTION

According to Crystal (1969) one of "the main problems that have yet to be faced" in prosody research is "the development of a more immediately meaningful system of notation than has been hitherto available". The quotation might equally well have been taken from a much more recent source, as is evident from the current concern with transcription systems for prosody (Bruce, 1989; Silverman et al., 1992). In this contribution I will argue that knowledge about the perception of prosody may provide guidelines as to what kinds of information should be transcribed and for the design of an appropriate coding scheme. Understanding the perception of prosody means that we understand how the listener interprets actual prosodic patterns in relation to his knowledge of phonological constraints as studied by prosodic phonology. Below, I will explore the perception of actual prosodic patterns in relation to the listener's abstract knowledge for three aspects of prosody.

THE PERCEPTION OF PITCH VARIATION

An important issue in the perception of pitch variation is how the listener extracts categorial tonal distinctions from the continuously varying pitch. There are two main approaches to the description of sentence melody: either by means of level tones or targets (High, Low), or by means of contour tones (Rising, Falling). Exponents of both approaches have claimed that their inventory of tones represents the perceptually and linguistically relevant units. This issue is not without importance because transcriptions are not easily translated from one approach to the other.

Instrumental analyses have helped little to clear up this issue. F0 contours are the result of many different contributions, both involuntary and voluntary ones (the latter due to the speaker's intention to produce certain melodic properties). Psycho-acoustic investigations of pitch perception have usually been restricted to very brief signals lacking the complexity of speech signals, so that their application to speech research is limited. Except for Thorsen (1979), who represents an early attempt to incorporate findings about pitch perception to the analysis of F0 contours in a principled way, most investigators have applied "trial-and-error" methods, based on an analysis-by-synthesis approach, to extract the information relevant to the perception of prosody (Bruce, 1977; Di Cristo, Espesser and Nishinuma, 1979; Fujisaki and Hirose, 1984; 't Hart, Collier and Cohen, 1990). For instance, by omitting increasingly greater variations from the F0 tracking (called "stylization"), making the result audible and comparing the synthetic contour to the original, 't Hart c.s. determined which variations were perceptually relevant (supposed to be intentional) and which were not. The "trial-and-error" methods give rise to stylized F0 trajectories, made up of sequences of (usually) connected straight lines or more sophisticated mathematical functions. The problem with such representations is that they *contain*, but need not *exactly represent* the perceptually relevant information. There may be alternative representations giving the same perceptual result.

Therefore, we need more insight into the perception of pitch in speech to differentiate between alternative representations. Recently, several studies on pitch perception in speech have been conducted in order to explore the perception of prosodic characteristics. House (1990) applies insights from psychoacoustic investigations to the properties of the speech signal. Stimuli which differ along several dimensions usually give higher sensitivity thresholds compared to stimuli which differ only along one dimension. The same applies to dynamic stimuli compared to static stimuli. E.g., the threshold for the perception of a pitch rise in speech-like stimuli is influenced by the rate of amplitude decrease (Van Der Horst, 1993). On the basis of such considerations, House argues that adequate models of pitch perception in speech must take into account the dynamic nature of speech: Speech exhibits continuous variation of spectral and amplitude properties, and areas showing fast spectral and amplitude changes such as CV transitions alternate with areas showing relatively little spectral and/or amplitude change such as the steady-state portions of vowels. In a series of ABX experiments he explored the influence of spectral variation on the perception of pitch. The results showed that the perception of pitch glide as such requires a spectrally stable portion of at least 100 ms, below which no pitch change will be perceived, but a succession of pitch levels. Although it is difficult to determine exactly which information listeners may have used as a basis for classification, his conclusions are confirmed by the outcome of an informal listening test which I conducted with stimuli constructed on the basis of his descriptions.

Further evidence for the perceptual reality of pitch levels in speech is provided by Hermes and Rump (this volume).

From the available data the following picture emerges. The perception of pitch change as such is impeded in portions with strong amplitude modulation and fast spectral changes such as consonants, and, most notably, consonant vowel transitions. Instead, pitch levels are perceived associated with the successive syllabic nuclei, or – at least – the most sonorant parts of the syllables. Actual pitch change will be perceived only if the F0 change starts in a spectrally stable portion of the speech signal, i.e., somewhere after the vowel onset and when the spectrally stable portion exceeds a certain minimum duration. It appears that in these cases the actual percept consists of a pitch level associated with the so-called *syllabic nucleus* followed by a pitch change. Categorial tonal distinctions appear to be dependent primarily on the pitch in successive syllabic nuclei and the presence or absence of an immediately following pitch change, at least for accented syllables. Near prosodic boundaries the situation seems different. An informal ABX test to assess the relative contributions of F0 in the vowel and in the post-vocalic consonant to perceived pitch suggested that in syllables preceding a silence the consonant F0 has a stronger influence on perceived pitch than the vowel F0.

There are a number of issues which remain to be cleared up. In the first place, the view proposed here implies that what counts is the pitch associated with the syllabic nucleus and the presence or absence of a pitch change following the nucleus, but that subtle differences in the realization of these pitch changes (e.g. having to do with the magnitude of a pitch change) will not play an important communicative role, because their perception is impeded by simultaneous spectral and amplitude variations. Secondly, the notion of "syllabic nucleus" remains to be defined. Generally, this is interpreted as a small region following the vowel onset, usually associated with the amplitude peak. It is clear, however, that High targets in pre-nuclear accented syllables may occur considerably later than this point (Silverman and Pierrehumbert, 1990), while is is evident that the details of their realization are communicatively relevant, e.g. for signalling relative prominence. In Hungarian, the height of pitch maxima following the syllabic nucleus is relevant for the distinction between declarative statements and yes/no questions (Gósy and Terken, 1993).

Thus, further investigations along the lines of those of House and Hermes and Rump are needed to determine how spectral and amplitude variations affect the perception of pitch variations before we can define the categorial distinctions in a principled way. Also, these insights will prove useful for the automatic extraction and labelling of melodic properties (see the contributions of d'Alessandro and Beaugendre, and Ten Bosch, this volume).

THE PERCEPTION OF PROMINENCE

Prominence has traditionally been related to three phonetic properties: loudness, perceived duration and pitch. It has been argued that pitch plays the most important role, because independent manipulation of the three phonetic properties showed that pitch cues could overrule the other cues. On the other hand, pitch variation doesn't appear to be essential to prominence perception: prominence can be perceived in the absence of pitch variation such as in whispered speech or in artificially monotonized speech (Cutler and Darwin, 1981). Nowadays, the interest has shifted towards investigating the combined effects of different properties on the perception of prominence, in particular in connection with the perception of degrees of prominence rather than the presence or absence of accent. It has been found, for instance, that perceived prominence varies as a function of the timing of pitch changes (Kohler and Gartenberg, 1991) and their phonetic realization (Repp, Rump and Terken, 1993). Also, the question has been addressed whether there is a trade-off between vowel duration and the timing of pitch changes (Rump, 1992; Fox, this volume). Finally, it has been suggested that conclusions with regard to the modest role of amplitude drawn from earlier research may be inadequate because the experiments involved inappropriate amplitude manipulations (Sluijter, this volume).

Still, it is evident that variations in the excursion size of pitch changes (i.e., variations in local pitch range) play an important role in the perception of degrees of prominence for focal accents, both for speakers and listeners. Speakers who are asked to speak a word with varying degrees of emphasis do so primarily by manipulating the local pitch range (Liberman and Pierrehumbert, 1984). In turn, listeners asked to judge the degree of prominence of an accented syllable do so very reliably as a function of variations in local pitch range (Gussenhoven and Rietveld, 1988). Therefore, models of prominence perception have in the first place aimed to capture the relation between pitch range variation and perceived prominence. One such model is proposed by Hermes and Rump (this volume), which was already briefly mentioned in the preceding section. It attempts to relate the paradigmatic aspect of prominence perception (i.e., having to do with intrinsic accent strength) to variations in local pitch range (excursion size).

Our understanding is still far from complete. In the first place, two accented syllables with the same excursion size measured on some appropriate dimension in the same phrase will not be perceived as equally prominent, due to the listeners' expectations about declination as the utterance proceeds (Pierrehumbert, 1979; Terken, 1991). The role of declination in the perception of relative prominence has been addressed by Terken (1991). Terken (1993) proposes a model for the perception of relative prominence, taking into account the observation that a given excursion size may result in varying degrees of prominence depending on how it is scaled in the overall pitch range of the speaker (see Figure 1). It does not yet incorporate the findings by Hermes and Rump (o.c.). Also, it is only tentative, and additional experiments are needed for evaluation. In fact, actual parameter values computed on the basis of earlier data did not give adequate predictions in two follow-up experiments (Repp et al., 1993), which means that either the parameter values may have to be adjusted or that the model is inadequate.

Secondly, it is not immediately obvious which phonological constraints are involved. Two influences besides declination need to be mentioned. Pitch maxima are varied in a gradient way to signal variations in prominence for paralinguistic reasons. In addition, the scaling of pitch maxima is also subject to downstep, a discrete phenomenon. However, the domain within which downstep operates, the conditions under which it may (if it is optional, i.e., contrastive) or must apply (if obligatory), and its relevance to everyday communication (as opposed to laboratory speech) are still unclear. As a consequence, investigating how the listener unravels the different influences is problematic.

Thirdly, the relation between paradigmatic and syntagmatic aspects of prominence perception (paradigmatic having to do with intrinsic accent strength and syntagmatic having to do with the strength of an accent relative to other accents) has not been cleared up yet. For instance, we don't know whether prominence gradations are perceived

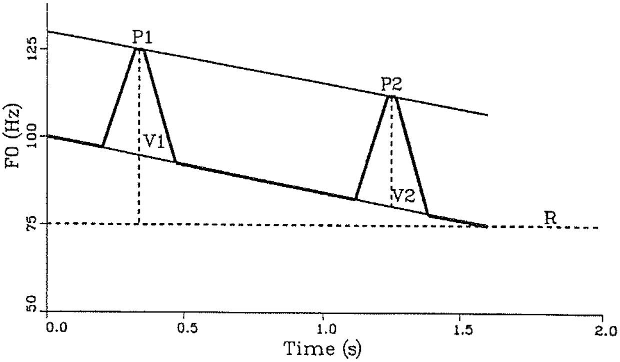

Figure 1: *Illustration of model relating judgments of relative prominence to pitch variation: accented syllables are perceived to be equally prominent if D2 = 0.9 (D1) + 0.23 (V1–R), with D: Distance between Peak P and Valley V, and R for a speaker-dependent fixed reference line. The values of 0.9 and 0.23 have been estimated on the basis of experimental data.*

primarily in a relative manner, due to the listener's ability to determine that one pitch change is perceptually more or less prominent than another, or whether listeners can also perceive gradations in perceptual prominence for individual pitch changes (and by implication may notice that one pitch change is not just more or less prominent but much more or much less prominent than another pitch change). In the latter case, the additional question is how such distinctions should be represented.

The frequently cited argument against a paradigmatic approach is that it implies categorial distinctions between different prominence levels primary, secondary etc.), and that such distinctions cannot be made very reliably which casts doubt on their categorial status. On the other hand, the experimental evidence from production and perception studies mentioned before favours the inclusion of paradigmatic aspects in the description. We can solve this problem by rejecting the assumption that prominence variations map onto a number of mutually exclusive categories such as 'primary' and 'secondary'. Instead, we may treat prominence as a scalar feature with fuzzy category boundaries which may be employed to signal linguistic or paralinguistic distinctions. It seems likely that listeners have some mental representation of the probability density function for excursion sizes of pitch changes in speech which they employ to interpret an actual excursion size, in the same way as has been proposed for durations (cf. the notion of *normalized duration* proposed by Campbell, 1992; Wightman et al., 1992). Both the theoretical implications and the way in which listeners unravel linguistic and paralinguistic contributions to local pitch range variation remain to be investigated.

THE PERCEPTION OF PHRASE BOUNDARIES

In transcribing prosody in speech data bases one has to decide as to how phrase boundaries are to be coded. The particular decisions taken reflect assumptions about the distinctions that can be perceived by listeners. Often, a tripartition is made between strong, weak and no boundary, but experienced transcribers will immediately agree that there is much more variation that listeners might employ. Therefore, finer distinctions have been proposed, either within a strictly hierarchical framework (Price et al., 1991) or in a more recursive one (Ladd, 1992). The latter proposal appears particularly attractive, because it includes both paradigmatic aspects (having to do with the classification of boundaries as belonging to a certain type) and syntagmatic aspects (having to do with the grouping of units at the same level). However, the more refined coding scheme has been applied mainly by experienced transcribers, and therefore it should be explored whether untrained listeners can also make finer distinctions in a reliable way. Once this has been established, it can be investigated how the distinctions relate to phonetic properties.

De Pijper and Sanderman (1992) conducted an investigation to find out which distinctions can be made reliably by non-expert listeners. They presented listeners with a number of isolated sentences and asked them to assign a digit between 1 and 10 to each word boundary to indicate the perceived degree of separation of the two words, the idea being that a higher score would express a stronger prosodic boundary. It was found that listeners could reliably distinguish more than three levels, and that this was not due to the lexico-syntactic information, as the results correlated very well with those of a test in which the same sentences were presented with the segmental information distorted so as to make them unintelligible. Furthermore, the levels appeared to be overlapping, suggesting that there are no clear category boundaries but rather a more gradient scale. These results may be taken to reflect contributions both of categorial factors and gradient factors. For instance, while the presence of a pause (in combination with a tonal boundary marker) would always result in the perception of a strong prosodic boundary (e.g. a boundary between intonation phrases), it appeared that the perceived boundary strength increased as a function of pause duration.

Although a number of methodological issues have to be cleared up before firm conclusions can be drawn, the findings are compatible with Ladd's (1992) analysis. Listeners can make categorial distinctions between different types of prosodic boundaries (Price et al., 1991). At the same time they can make judgments as to the relative strength of two prosodic boundaries of the same type in order to infer hierarchical relations. This implies that the most informative transcription would represent both paradigmatic properties (concerning types of boundaries) and syntagmatic properties (concerning the hierarchical relations between constituents of the same type).

A related issue concerns the employment of different kinds of phonetic information. Beach (1991) has shown that there is a trade-off between tonal information, usually supposed to evoke a categorial distinction between different tones, and duration information, which is supposed to be much more gradient. Furthermore, there is individual variation in the extent to which listeners employ different sorts of information (Bruce et al., 1992). Simple decision models may be designed to account for the combined use of different kinds of information and the individual differences.

CONCLUSIONS

Considerable progress has been made in our understanding of how prosody functions in human communication. A major problem seems to be conceptual: to incorporate the insights into prosodic theory. Prosody is often conceived of as part of phonology, since phonology is concerned with characterizing the sound forms which may transmit certain distinctions in languages (i.e., distinguish between different readings of a sentence or a text). This implies that the description of prosodic phenomena is subject to the restrictions imposed by phonological representation. I have argued that if one does so, one runs into problems when trying to account for the full range of prosodic phenomena at different levels. It is not difficult, for instance, to design a decision model accounting for the relation between temporal and pitch variation in signalling phrase boundaries, but it is difficult to incorporate such a model into a phonological description. Therefore, the direction should be inverted: in order to understand the perception of prosody and the way prosody functions in human communication, the phonological constraints against which actual prosodic patterns are interpreted should be investigated as an intermediate step.

REFERENCES

Beach, C. (1991), "The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations", *Journal of Memory and Language*, Vol. 30, pp. 644-663.

Bruce, G. (1977), *Swedish word accents in sentence perspective* (Travaux de l'Institut de Linguistique de Lund, Lund).

Bruce, G. (1989), "Report from the IPA working group on suprasegmental categories", *Working papers*, Vol. 35, Lund University, Dept. of Linguistics, pp. 25-40.

Bruce, G., Granström, B., Gustafson, K. and House, D. (1992), "Aspects of prosodic phrasing in Swedish", *Proceedings of International Conference on Spoken Language Processing 1992*, (eds. J. Ohala et al.), University of Alberta, Canada, pp. 109-112.

Campbell, W.N. (1992), *Multi-level timing in speech* (Ph.D. Thesis, University of Sussex).

Cristo, Di, A., Espesser, R. and Nishinuma, Y. (1979), "Presentation d'une methode de stylisation prosodique", *Travaux de l'institut de phonetique d'Aix*, Vol. 6, pp. 125-146.

Crystal, D. (1969), *Prosodic systems and intonation in English* (Cambridge University Press, London).

Cutler, A. and Darwin, C.J. (1981), "Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency", *Perception & Psychophysics*, Vol. 29, pp. 217-224.

Fujisaki, H. and Hirose, K. (1984), "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *J. Acoust. Soc. Japan*, Vol. 5, pp. 233-242.

Gósy, M. and Terken, J. (1993), *Phonological relevance of timing and excursion size of pitch change: Evidence from Hungarian* (Manuscript, Institute for Perception Research, Eindhoven).

Gussenhoven, C. and Rietveld, A.C.M. (1988), "Fundamental frequency declination in Dutch: testing three hypotheses," *J. Phonetics*, Vol. 16, pp. 355-369.

't Hart, J., Collier, R. and Cohen, A. (1990), *A perceptual study of intonation*, (Cambridge University Press, Cambridge).

Hermes, D.J. and Rump, H.H. (1993), *Prominence in speech intonation induced by rising and falling pitch movements*, (Manuscript, Institute for Perception Research, Eindhoven).

Horst, Van der, R. (1993), *On the sensitivity for pitch movements in speech during amplitude changes*, (Manuscript, Institute for Perception Research, Eindhoven).

House, D. (1990), *Tonal perception in speech* (Lund University Press, Lund).

Kohler, K. J. and Gartenberg, R. (1991), "The perception of accents: F0 peak height vs. F0 peak position", *Arbeitsberichte Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel*, Vol. 25, pp. 219-241.

Ladd, D.R. (1992), *Compound prosodic domains* (Manuscript, University of Edinburgh).

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

Pierrehumbert, J. (1979), "The perception of fundamental frequency declination," *J. Acoust. Soc. Am.*, Vol. 66, pp. 363-369.

Pijper, De, J.R., and Sanderman, A.A. (1993), "Prosodic cues to the perception of constituent boundaries", *Proceedings Eurospeech 93, Berlin, 21-23 September 1993*, to appear.

Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S. and Fong. C. (1991), "The use of prosody in syntactic disambiguation", *J. Acoust. Soc. Am.*, Vol. 90, pp. 2956-2970.

Repp, B.H., Rump, H.H. and Terken, J. (1993), *Relative perceptual prominence of fundamental frequency peaks in the presence of declination* (Manuscript, Institute for Perception Research, Eindhoven).

Rump, H.H. (1992), "Timing of pitch movements and perceived vowel duration", *Proceedings of International Conference on Spoken Language Processing 1992*, (eds. J. Ohala et al.), University of Alberta, Canada, pp. 1047-1050.

Silverman, K.E.A. and Pierrehumbert, J.B. (1990), "The timing of prenuclear high accents in English," in *Papers in Laboratory Phonology I: Between the grammar and the physics of speech*, ed. by J. Kingston and M. Beckman (Cambridge University Press, Cambridge), pp. 72-106.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992), "ToBI: A standard for labelling English prosody", *Proc. Internat. Conf. on Spoken Language Processing 1992, Banff, October 12–16 1992*, Vol.2, pp. 867-870.

Terken, J. (1991), "Fundamental frequency and perceived prominence of accented syllables", *J. Acoust. Soc. Am.*, Vol. 89, pp. 1768-1776.

Terken, J. (1993), "Baselines revisited. Reply to Ladd", *Language and Speech*, to appear.

Thorsen, N. (1979), "Interpreting raw fundamental frequency tracings of Danish", *Phonetica*, Vol. 36, pp. 57-78.

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