Generating Prosodic Structure for Synthesis of Swedish Intonation

Merle Horne
Department of Linguistics and Phonetics
Helgonabacken 12, S-223 62 Lund, Sweden
Merle.Horne@ling.lu.se

ABSTRACT

This article presents an outline of the prosodic constituent structure which will be incorporated in a linguistic preprocessor forming part of a text-to-speech system for generation of intonation in Swedish restricted texts.

INTRODUCTION

One of the goals of current research in text-to-speech systems is to improve the quality of intonation by developing algorithms for preprocessing texts in order to extract grammatical and discourse information necessary for the generation of appropriate prosodic patterns. In previous publications, we have reported on the work that we have done developing a preprocessor which tracks coreferential relations using lexical-semantic and morphological information to find referential identity between content words in restricted texts dealing with the stock-market (Horne & Johansson 1991, 1993, Horne et al. 1993a,b). This information is important in order to predict the location of the final focal accent in an utterance.

PROSODIC STRUCTURE AND PHRASING

Our current efforts are being directed towards the development of an algorithm which will allow further preprocessing of our restricted texts with the goal of using the information on coreferentiality obtained from the referent tracking algorithm together with further information on lexical category designation to group words together into a hierarchy of prosodic constituents such as those discussed in Bruce & Granström (1993). Information on prosodic structure is needed in order to better predict the location as well as the particular form of tone accents associated with utterance-internal prosodic boundaries.

Minimal Parsing

Following an approach similar to Bachenko & Fitzpatrick (1990), Quené & Kager (1993) and inspired by concepts within prosodic phonology (e.g. Nespor & Vogel 1986), we are attempting to determine how one, using a minimal amount of parsing, can obtain enough information to construct a hierarchical prosodic structure for each sentence in a text. Unlike other researchers, however, we are also using contextual information such as coreference in our approach to generating prosodic structure.

Prosodic Constituents

At least three levels of prosodic structure are required for Swedish in order to model all the prosodic information observed in our data. The smallest of these is the Prosodic Word which we will define as corresponding to a content word and any following function words up to the next content word within a given clause. At the beginning of a clause, the Prosodic Word can also begin with one or more function words. The Prosodic Word is characterized by a word accent and potentially a focal accent (Accent 1= HL*(H$^-$L$^-$), Accent 2 = H*L(H$^-$L$^-$) (We use H$^-$ and L$^-$ to represent respectively a focal high and the low tone accent following a focal high in order to distinguish them from the H and L associated with the word accents.). It is also marked by a boundary tone which is realized by a final rise in the case where the content word is not focussed (i.e. contextually given) (H#) or a fall when the content word is focussed (L#). This L# can be thought of as a potential low Prosodic Phrase boundary, i.e. given the proper contextual environment including sufficient duration, the L can be realized low enough to be interpreted as a L% boundary (cf. Bruce et al. 1993 who present experimental evidence to show that increasing the size of a Fo fall after a focal H can lead speakers to perceive a phrase boundary). The H# in its turn can be thought of as a potential H% boundary, e.g. a 'continuation rise' associated with nonfinality. Thus a Prosodic Phrase boundary always correlates with a Prosodic Word boundary but not vice versa. These boundary tones, we claim, play an important role in creating the transitions between consecutive Prosodic Words in a larger Prosodic Phrase. They are also points for potential pauses, e.g. before focussed content words (see Gårding 1967, Strangert 1993). The unit does not necessarily correspond to a syntactic constituent as the example in (1) illustrates ('–' represents the boundary between Prosodic Words). This type of 'nonsyntactic' grouping is perhaps more characteristic of well-planned read texts or spontaneous speech than of non well-planned texts read e.g. by a non-expert/non-professional. It can be characterized as more rhythmically-based than a grouping adhering strictly to syntactic phrase boundaries since it begins with a lexical word which has predominantly left-edge stress. We realize that this definition of the Prosodic Word is not the only possible one. However, it corresponds to the most common type of grouping for the speaker whose speech we are modelling and we have therefore decided to use it as a working definition for purposes of algorithm development.

(1) Kurserna på – Stockholmsbörsen – fortsätter att – falla.
Rates(det) on – Stockholm Stock Exchange(det) – continue to – fall
'Rates on Stockholm's Stock Exchange continue to fall'

Figure 1 illustrates the prosodic structure of (1) produced by the female speaker whose prosody we are modelling. She is an 'expert' speaker, i.e. she has detailed knowledge of the domain she is talking about (stock-market) and the well-planned impression her speech gives probably results both from this fact and from her long experience as the principal reader of stock-market reports on Radio Sweden (she retired in 1992).

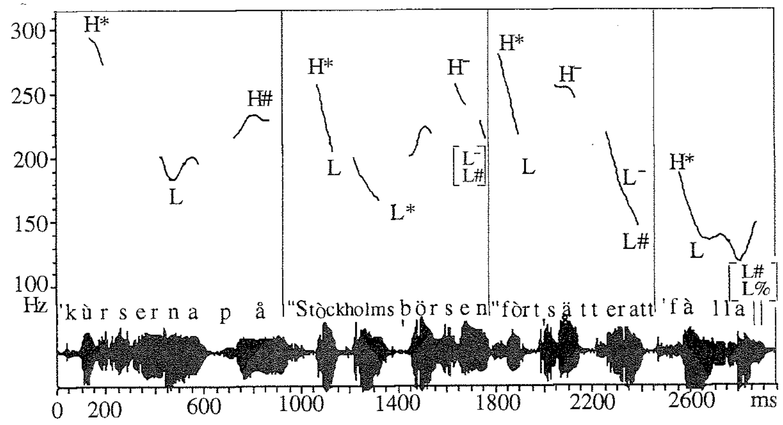

Figure 1. *Fo contour corresponding to the sentence in (1). Vertical lines correspond to Prosodic Word boundaries represented by L#; L% represents a Prosodic Phrase boundary.*

One or more Prosodic Words make up a Prosodic Phrase which is marked by a final L% or H% boundary tone accent. Factors which determine the location of Prosodic Phrase boundaries include the following: a) sentence boundary: A sentence boundary corresponds to the end of a Prosodic Phrase, b) new/given distinction: A Prosodic Phrase must contain at least one focussed Prosodic Word, c) length: A Prosodic Phrase will not exceed *x* syllables at a given rate of speech *y*. Finally, one or more Prosodic Phrases make up a Prosodic Utterance, which is bounded by pauses. It is further generally assumed that each prosodic constituent is characterized by a certain amount of preboundary lengthening (Gussenhoven & Rietveld 1992), and although we have not as yet made any detailed investigations of the phenomenon in our data which would allow us to quantify a lengthening index, we are assuming that, all other things being equal, the higher up in the hierarchy a prosodic constituent is placed, the greater the relative duration associated with its final syllable(s) will be (see Fant et al. 1992 who find that in "prepause" position, lengthening is on the order of 110 ms in stressed syllables and 70 ms in unstressed syllables).

Figure 2 presents in schematic form the prosodic constituents assumed for Swedish and their phonetic correlates. The tone accents (H and L) are assumed to be associated with syllables (S) according to principles outlined in Bruce (1977). It is also assumed that the realization of the tone accents is dependent to some extent on the number of syllables present in a particular word, i.e. the number of syllables in a given word dictates to a great extent how many tones will be realized phonetically.

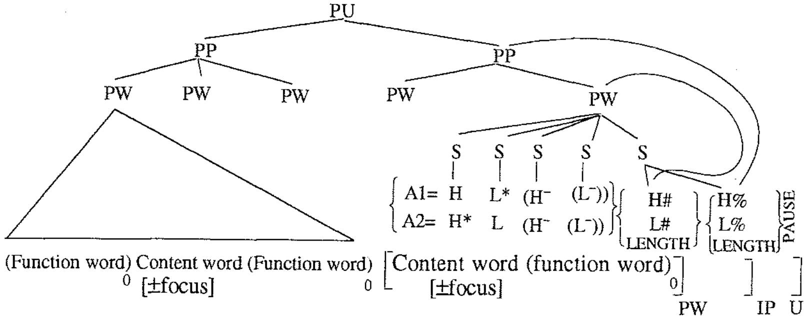

Figure 2. *Schematic presentation of the prosodic hierarchy assumed for Swedish and the associated phonetic correlates. Accent 1 is represented as $HL^*(H^-L^-)$ and Accent 2 as $H^*L(H^-L^-)$, where (H^-L^-) represents the focal High (H^-) and potential Low (L^-) associated with the focal accent. H# and L# represent the Prosodic Word boundaries and H% and L% designate the Prosodic Phrase boundaries. PW stands for Prosodic Word, PP for Prosodic Phrase and PU for Prosodic Utterance. (Function word)$_0$ stands for zero or more function words.*

CLAUSE AND WORD-CLASS RECOGNITION

In order to construct these prosodic constituents automatically, it is necessary to be able to unambiguously recognize internal clause boundaries and sentence boundaries as well as distinguish between content words and function words. Sentence boundaries correspond to a full stop (.) and internal clause boundaries are cued by e.g. commas (,) as well as conjunctions (*och* 'and', *men* 'but', the subordinate conjunctions *som* 'that/who', *att* 'that', *om* 'if', etc. As mentioned above, sentence boundaries correspond with Prosodic Phrase boundaries. Internal clause boundaries are also often associated with a Prosodic Phrase boundary, but not always. Considerations of other factors such as length must also be taken into consideration when determining the location of internal boundaries. The assignment of words to particular classes is not always straightforward either but one can say that in general, content words include the traditional categories of nouns, verbs, adjectives, adverbs, numerals, whereas function words consist of e.g. prepositions, pronouns, determiners, auxiliary verbs, interrogative/relative adverbs, etc.

In tackling the problem of word class recognition/disambiguation, we are currently considering two possible ways of proceeding. One involves using a tag set (e.g. Ejerhed et al. 1992) to annotate a large domain-specific text corpus which will then provide the basis for an automatic parser based on lexical and sequential occurrence probabilities (Eeg-Olofsson 1991). Another approach which we intend to explore is a Hidden Markov model available from Xerox (Cutting et al. 1992) used in combination with the computerized lexicon (Hedelin et al. 1987). This involves tuning the model on a non-tagged corpus by specifying a certain number of transition biases (for example, an infinitive marker is likely to be followed by a verb, determiners are not likely to be followed by verbs or prepositions).

ACKNOWLEDGEMENTS

This research has been supported by a grant from the HSFR/NUTEK Language Technology Programme.

REFERENCES

Bachenko, J. & E. Fitzpatrick. 1990. 'A computational grammar of discourse-neutral prosodic phrasing in English'. *Computational Linguistics* 16, 155-170.

Bruce, G. 1977. *Swedish accents in sentence perspective*. Lund:Gleerups.

Bruce, G. & B. Granström. 1993. 'Prosodic modelling in Swedish speech synthesis'. *Speech Communication* 13, 63-73.

Bruce, G., B. Granström, K. Gustafson, & D. House. 1993. 'Interaction of Fo and duration in the perception of prosodic phrasing in Swedish'. In Björn Granström & Lennart Nord (eds.), *Nordic Prosody VI*, 7-22. Stockholm: Almqvist &Wiksell.

Cutting, D., J. Kupiec, J. Pedersen, & P. Sibun. 1992. 'A practical part-of-speech tagger'. *Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy, April 1992.* Also available as Xerox PARC technical report SSL-92-01.

Eeg-Olofsson, M. 1991. *Word-class tagging. Some computational tools.* Univ. of Göteborg: Dept. of Linguistics.

Ejerhed, E., G. Källgren, O. Wennstedt, & M. Åström. 1992. *The linguistic annotation system of the Stockholm-Umeå corpus project*. Umeå: Dept. of Linguistics Report No. 33.

Fant, G., A. Kruckenberg, & L. Nord. 1992. 'Prediction of syllable duration, speech rate and tempo'. *Proceedings ICSLP 92*, 667-670.

Gussenhoven, C. & A.C.M. Rietveld. 1992. 'Intonation contours, prosodic structure and preboundary lengthening'. *Journal of Phonetics* 20, 283-303.

Gårding, E. 1967. 'Prosodiska drag i spontant och uppläst tal'. In G. Holm (ed.) *Svenskt talspråk*, 40-85. Stockholm: Almqvist & Wiksell.

Hedelin, P., A. Jonsson, & P. Lindblad. 1987. *Svenskt uttalslexikon: 3 ed.* Tech. Report, Chalmer's Univ. of Technology.

Horne, M. & C. Johansson. 1991. 'Lexical structure and accenting in English and Swedish restricted texts'. *Working Papers* (Dept. of Ling., U. of Lund) 38, 97-114.

Horne, M. & C. Johansson. 1993. 'Computational tracking of 'new' vs 'given' information: implications for synthesis of intonation'. In Björn Granström & Lennart Nord (eds.), *Nordic Prosody VI*, 85-97. Stockholm: Almqvist & Wiksell.

Horne, M., M. Filipsson, M. Ljungqvist, & A. Lindström. 1993a. 'Referent tracking in restricted texts using a lemmatized lexicon: implications for generation of prosody'. *Proceedings Eurospeech '93* (Berlin) Vol. 3, 2011-2014.

Horne, M., M. Filipsson, C. Johansson, M. Ljungqvist, & A. Lindström. 1993b. 'Improving the prosody in TTS systems: Morphological and lexical-semantic methods for tracking 'new' vs. 'given' information'. Proceedings ESCA Workshop on prosody, *Working Papers (Dept. of Linguistics, Univ. of Lund)* 41, 208-211.

Nespor, M. & I. Vogel. 1986. *Prosodic phonology*. Dordrecht: Foris.

Quené, H. & R. Kager. 1993. 'Prosodic sentence analysis without parsing'. In Vincent van Heuven & Louis Pols (eds.), *Analysis and synthesis of speech*, 115-130. Berlin: Mouton de Gruyter.

Strangert, E. 1993. 'Speaking style and pausing'. *Phonum* 2, 121-137.