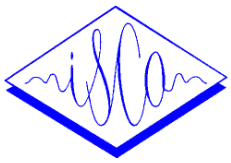

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

GENERATING PROSODIC STRUCTURE FOR SWEDISH TEXT-TO-SPEECH

Merle Horne and Marcus Filipsson

Department of Linguistics and Phonetics, Lund University
Helgonabacken 12, S-223 62 Lund, Sweden

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. It further discusses the structure of a number of algorithms (including a word-class tagger, a complex-word identifier and a prosodic parser) which enable the generation of the assumed prosodic structure.

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 [1-4]. This information is important in order to predict the location of the final focal accent in an prosodic phrase.

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 [5]. 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 [6], Quené & Kager [7] and inspired by concepts within prosodic phonology [8], 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 [9]. 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. [10] 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 [11-12]). 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

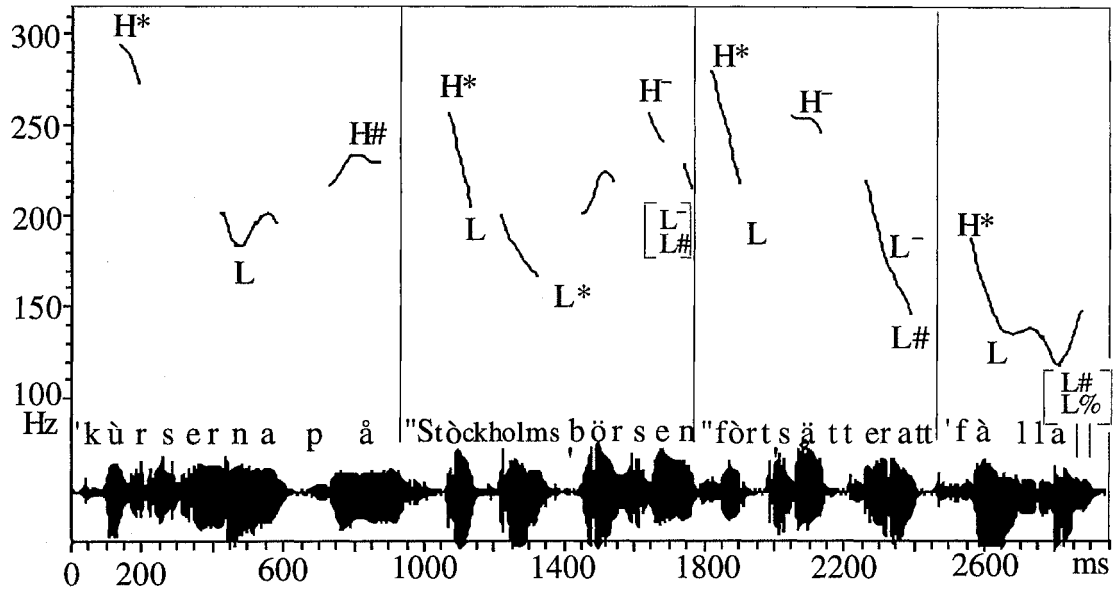

Figure 1. *Fo contour corresponding to the sentence in (1). Vertical lines correspond to Prosodic Word boundaries represented by L#; L% represents a Prosodic Phrase boundary.*

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).

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 [13-14], 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. [15] 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 [16]. 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.

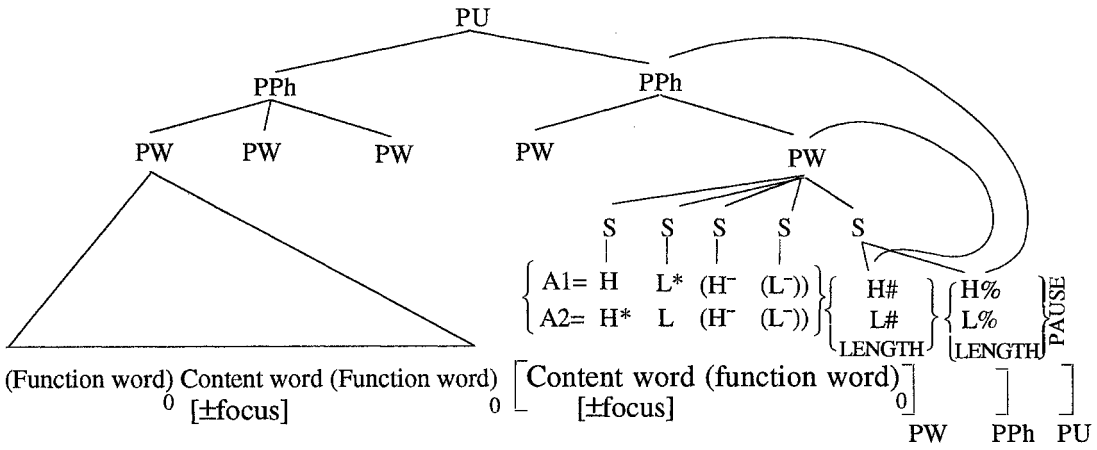

Figure 2. *Schematic presentation of the prosodic hierarchy assumed for Swedish and the associated phonetic correlates. Accent 1 is represented as HL*(H⁻L⁻) and Accent 2 as H*L(H⁻L⁻), where (H⁻L⁻) represents the focal High (H⁻) and potential Low (L⁻) associated with the focal accent. H# and L# represent the Prosodic Word boundaries and H% and L% designate the Prosodic Phrase boundaries. PW stands for Prosodic Word, PPh for Prosodic Phrase and PU for Prosodic Utterance. (Function word)0 stands for zero or more function words.*

COMPUTATIONAL DESIGN OF THE PROSODIC STRUCTURE COMPONENT

In order to construct these prosodic constituents automatically, a number of different analyses are required. The present system is based on a strictly modular approach, with each module having well-defined input/output formats. This allows one to easily replace a module with a new one should a more efficient algorithm be developed at some later stage. Figure 3 presents all the modules in the system and the output from each module.

The first task is to tokenize the text into a list of words (see Fig. 3 'explode'). At the same time, punctuation marks and paragraph boundaries are recognized. The next step is to look up the words in our domain-specific lexicon, which is an expanded subset of a larger computerized lexicon [17] (see Fig. 3 'labeler'). This process will generate multiple tags for some words. For example, the word *fast* 'firm', 'although' is tagged both as an adjective and as a subordinate conjunction. The next step is therefore the disambiguation of these tags (Fig. 3 'tagger'). In this endeavour, we are currently testing the performance of a stochastic parser based on lexical and sequential occurrence probabilities as well as overall tag probability [18]. The algorithm implements a first-order Markov chain and uses dynamic programming to estimate the best hypothesis for the whole sentence. A set of approximately 30 lexico-syntactic tags based on Ejerhed et al.'s tag set [19] have been chosen to train the system. These have been further assigned to the tagged words' lemma representations in the computerized lexicon, thus allowing recognition of all morphologically derived forms of a given head-word. Preliminary results indicate that the algorithm works quite well, but we intend to compare it with other approaches. One involves a Hidden Markov model such as in the Xerox Part-of-Speech Tagger [20]. Another approach is a rule-based one. Finally, we are considering combinations of these, e.g. using a rule-based system as the default method, and a probabilistic algorithm for cases where the rules fail.

After word classes are determined, the next stage is to recognize complex words, i.e. strings of content words that function as a single prosodic unit (Fig. 3 'grouper'). In the stock-market domain, these correspond to proper names (i.e. company/bank names and stock designations, e.g. *'Avesta Sheffield'*, *'S-E Banken'*, *'Hennes & Mauritz'*, *'Hasselförs Förvaltnings AB'* (*AB* 'CO.'). These strings are assigned a specific tag ('CX'-complex word) which, although it is not a lexical tag, is a member of the class of content word tags together with those associated with nouns, adjectives, verbs, adverbs, etc.

The next step is to recognize clause boundaries since the clause is the domain over which Prosodic Words are defined (see Fig. 3 'clausefinder'). Clause boundaries occur at certain punctuation marks, e.g. full stop, colon, semicolon, some commas (those not occurring in lists of words having the same word class), as well as before coordinate and subordinate conjunctions (*och* 'and', *men* 'but', *fast* 'although', *att* 'that'), and relative pronouns (e.g. *som* 'that', 'who').

The following stage involves classifying each word as either a content word ('CW') or a function word ('FW') (Fig. 3 'classifier'). The assignment of words to one of these classes is not always straightforward, but one can say that in general, content words include the traditional categories of nouns, verbs, adjectives, adverbs, numerals, whereas function words consist of prepositions, pronouns, determiners, auxiliary verbs, interrogative/relative adverbs, deictic adverbs, quantifiers, etc. Domain-specific considerations lead to the introduction of a number of unconventional tags, for example 'specifer' nouns and adjectives that occur after the head noun in complex proper names like *B fria* in the name *Electrolux B fria* 'Electrolux B free (shares)'.

The final stage of the system is the actual prosodic parser, which parses the list of words into a hierarchical structure with three levels: Prosodic Word, Prosodic Phrase and Prosodic Utterance (Fig. 3 'parser'). First, content words and function words are grouped together to form Prosodic Words (see Fig. 1). Second, clause boundaries currently generate Prosodic Phrase boundaries, although other factors such as length must also be taken into consideration when determining the location of these boundaries. These are currently being incorporated into the parser. Finally, a Prosodic Utterance boundary is generated at each sentence boundary in the present algorithm.

ACKNOWLEDGEMENTS

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

REFERENCES

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

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

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

[4] M. Horne, M. Filipsson, C. Johansson, M. Ljungqvist, & A. Lindström. "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, pp. 208-211, 1993.

[5] G. Bruce, & B. Granström. "Prosodic modelling in Swedish speech synthesis". Speech Communication 13, pp. 63-73, 1993.

[6] J. Bachenko, & E. Fitzpatrick. "A computational grammar of discourse-neutral prosodic phrasing in English". Computational Linguistics 16, pp. 155-170, 1990.

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

[8] M. Nespor, & I. Vogel. Prosodic phonology. Dordrecht: Foris, 1986.

[9] M. Horne, "Generating prosodic structure for synthesis of Swedish intonation." Working Papers (Dept. Ling., Univ. of Lund) 43, pp. 72-75, 1994.

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

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

[12] E. Strangert, "Speaking style and pausing". Phonum 2, pp. 121-137, 1993.
[13] C. Gussenhoven & A.C.M. Rietveld. "Intonation contours, prosodic structure and preboundary lengthening". Journal of Phonetics 20, pp. 283-303, 1992.
[14] C.W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf & P. Price. "Segmental durations in the vicinity of prosodic phrase boundaries". J. Acoust. Soc. Am. 91, pp. 1707-1717, 1992.
[15] G. Fant, A. Kruckenberg, & L. Nord. "Prediction of syllable duration, speech rate and tempo". Proceedings ICSLP 92, pp. 667-670, 1992.
[16] G. Bruce, *Swedish accents in sentence perspective.* Lund:Gleerups, 1977.
[17] P. Hedelin, A. Jonsson, & P. Lindblad, Svenskt uttalslexikon: 3 ed. Tech. Report, Chalmer's Univ. of Technology, 1987.
[18] M. Eeg-Olofsson, Word-class tagging. Some computational tools. Univ. of Göteborg: Dept. of Linguistics, 1991.
[19] E. Ejerhed, G. Källgren, O. Wennstedt, & M. Åström, The linguistic annotation system of the Stockholm-Umeå corpus project. Umeå: Dept. of Linguistics Report No. 33, 1992.
[20] D. Cutting, J. Kupiec, J. Pedersen, & P. Sibun, "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, 1992.

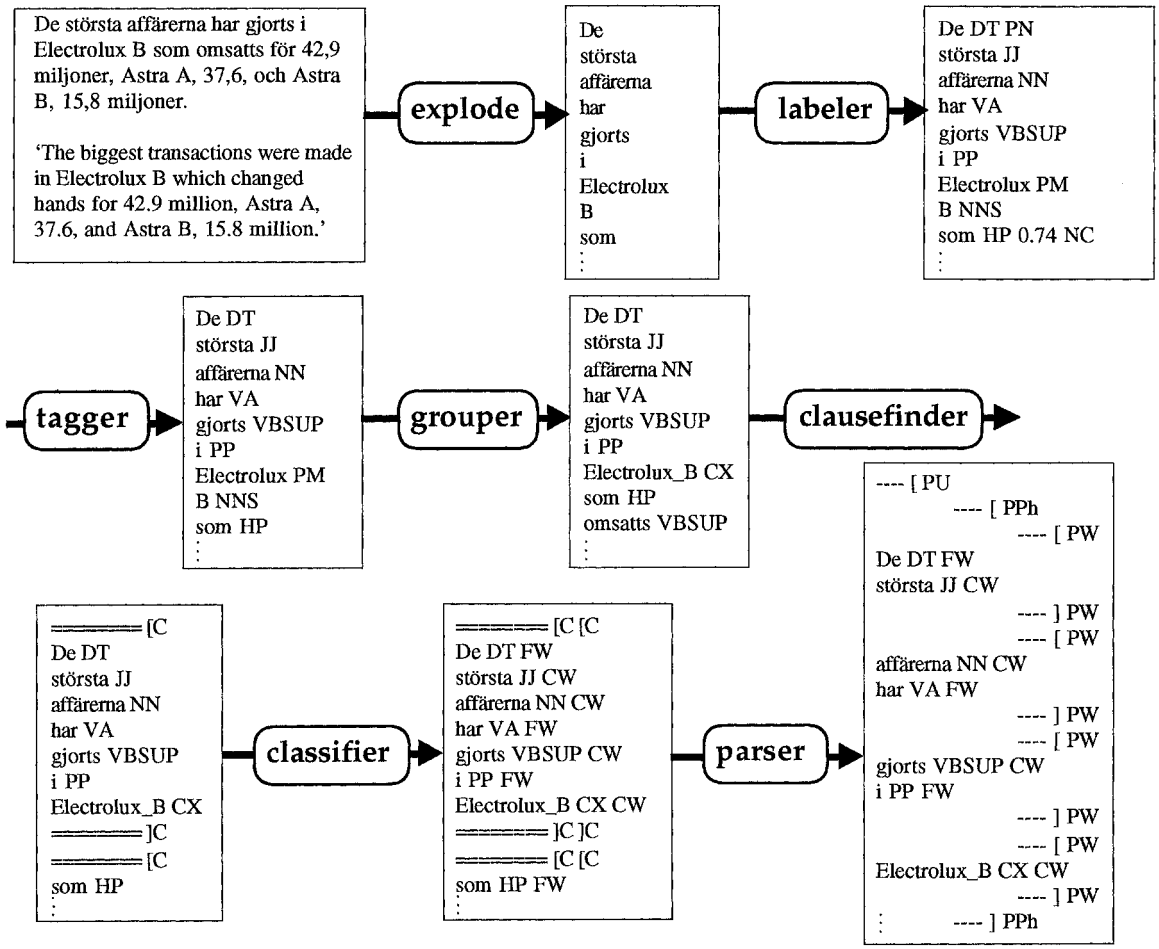

Figure 3. *Schematic presentation of the present computer system for prosodic parsing. The modules in the system are represented by a rounded corner rectangle. An excerpt of the output of each module (and consequently the input to the next module) is shown between each pair of modules. The first input is the stock market report newspaper text; the final output is the prosodically parsed text. (DT=Determiner, PN=Pronoun, JJ=Adjective, NN=Noun, VA= Auxiliary Verb, VBSUP=Supine form of Verb, PP=Preposition, PM=Proper Noun, NNS=Noun specifier, HP=Relative Pronoun, NC='non-clausal' conjunction)*