



## AUTOMATIC PROSODIC ANALYSIS FOR SWEDISH SPEECH RECOGNITION

D.House\*, G.Bruce\*, F.Lacerda+, B.Lindblom+.

### ABSTRACT

A mingogram reading experiment was carried out in which an expert in Swedish prosody was presented with computer simulated mingograms of unknown Swedish sentences and asked to identify the following categories: stressed and unstressed syllables, grave and acute word accents, focal accent, and terminal juncture. Out of a total of 178 occurrences of the different categories, 151 were correctly identified (85%). The categories were identified by using the Fo contour, energy envelope and a duplex oscillogram. On the basis of this experiment, a set of preliminary, hierarchically ordered automatic analysis rules have been formulated using Fo movement patterns synchronized with energy envelope peaks to define the prosodic categories. These rules have been tested by using two non-expert mingogram readers and are being implemented on an automatic prosodic analysis system.

### INTRODUCTION AND BACKGROUND

Prosodic information contained in the speech signal can be used by a speech recognition system to limit lexical access and provide information concerning semantic and syntactic structure. In Swedish the prosodic categories of stress, word accent, focal accent, and initial and terminal juncture are ready candidates for automatic recognition rule formulation.

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 Swedish Board for Technical Development and is part of the National Swedish Speech Recognition Effort in Speech Technology. Research in prosodic recognition is a natural continuation of a long tradition of research in prosody in Lund and Stockholm (ref 1-7).

Prosodic recognition can constitute one component of a larger, phonetically structured recognition system. The intention is to build on the existing knowledge of Swedish prosody thereby facilitating and enhancing Swedish Speech Recognition. This combination of prosody and recognition is a relatively new area of research. See Lea (ref 8) and Vaissière (ref 9).

\*Dept. of Linguistics and Phonetics, Lund University,  
S-223 62 Lund, Sweden.

+Institute of Linguistics, Stockholm University, S-106 91  
Stockholm, Sweden.

The primary goal of the project is to develop a method for extracting relevant prosodic information from a speech signal. Some issues relating to this goal are 1) What criteria can we use to recognize prosodic categories, 2) What kind of acoustic invariance relates to prosodic categories, and 3) What degree of success can we achieve in recognizing prosodic categories. Furthermore, by using a recognition approach to prosody, we hope to reach a better understanding of the mechanisms involved in human perception of prosody. The prosodic categories used in the project are STRESS (stressed and unstressed syllables) WORD ACCENTS (acute and grave), FOCUS (focal and non-focal accents), and JUNCTURE (connective and boundary signals for phrases).

In Swedish, the basic dichotomy between stressed and unstressed syllables exists as in English. This division provides the basic rhythmical structure of spoken Swedish, but gives no information about word boundaries or the number of words in an utterance.

In both Swedish and Norwegian, the primary stressed syllable is characterized by having one of two tonal accents: Acute (Accent I) or Grave (Accent II). Identification of GRAVE accent provides us with morphological information which can facilitate lexical access.

The identification of focal accents is important for a recognition system since a focal accent tells us that a word is emphasized and bears important information. Focal accents have a predictive value for both semantics and syntax.

The correct identification of juncture will assist a recognition system in isolating phrases. These phrases, or information chunks, are somewhat related to syntax. An interesting and challenging aspect of juncture is that the Fo representation of initial juncture bears a strong resemblance to that of the acute focal word accent while the representation of final juncture resembles that of a grave word accent. Moreover, juncture representations can interact and interfere with other prosodic categories.

#### MINGOGRAM READING (EXPERT READER)

Promising results from previous mingogram reading experiments (ref 10) led us to use this method as a basis for choosing acoustic criteria for prosody recognition and as a background for recognition rule formulation. By interviewing an expert reader, we should be able to isolate the most salient cues for use by the recognizer.

Ten sentences, unknown to the reader, were recorded in a Stockholm dialect of Swedish under laboratory conditions. The sentences were carefully designed so that both the placement of the prosodic categories and the syntax of the sentences were varied. Each sentence contained 10 to 15 syllables with 2 to 5 stressed syllables in each sentence. Computer simulated mingograms were then made from the

recordings displaying a duplex oscillogram, the Fo contour and a bandpass-filtered (1500-3500 Hz) intensity curve. The reader was given the task of identifying the above-mentioned prosodic categories on the basis of the mingogram registration.

Of a total of 178 occurrences of the different categories, 151 were correctly identified (85%). These results break down into categories as follows: GRAVE ACCENTS (both focal and non-focal) 13 of 13 (100%), GRAVE FOCAL ACCENTS 7 of 8 (88%), ACUTE ACCENTS (both focal and non-focal) 20 of 23 (87%), ACUTE FOCAL ACCENTS 12 of 13 (92%), ACUTE FOCAL FINAL ACCENTS 2 of 2 (100%), STRESSED SYLLABLES 37 of 37 (100%), UNSTRESSED SYLLABLES 60 of 82 (73%), and TERMINAL JUNCTURE 2 of 2 (100%). Clearly there is considerable prosodic information available in the acoustic signal alone.

#### RECOGNITION RULE FORMULATION

On the basis of the mingogram reading test results, descriptive rules were formulated and ordered so that the categories that were easiest to identify, i.e. those showing the greatest degree of signal invariance, should be identified first leaving the categories with more variable patterns to be identified last. Fo movement patterns were deemed to be the most salient information, and the rules, therefore, are based on these movements. However, a falling Fo contour, for example, can signal quite a number of prosodic categories. A vowel containing an Fo fall can be a stressed vowel with a grave accent, a grave focal accent, an acute focal final accent, or it can even be an unstressed vowel in final position. Invariance, therefore, lies not in the Fo movements alone, but in the synchronization of these movements with vowel onset along with the range and steepness of the Fo movement.

The first rule category is a falling Fo of a certain steepness and range where the beginning of the fall is synchronized with the vowel onset. This signals a stressed vowel with a grave accent. The reader first attempts to find all the grave accents in the sentences. Then the reader is instructed to look at the syllable following the identified grave accent. If there is a high or rising Fo in the following syllable, then the grave accent is also a focal accent. A rising Fo synchronized with a vowel onset signals an acute focal accent and a down-stepping of Fo from one vowel to the next signals an acute non-focal accent in the vowel receiving the down-stepping. Finally, a steep falling Fo signals terminal juncture. Each rule is elaborated with secondary rules concerning relative Fo highs and lows, range and steepness of movement, and restrictions such as the fact that one grave accent fall cannot be directly followed by another. Identification of word accents gives identification of primary stress since a stressed vowel will generally have one of the two word accents. Thus, the identification of stressed and unstressed vowels is mainly arrived at indirectly.

## RULE TESTING (NON-EXPERT READERS AND COMPUTER PROGRAM)

Two non-expert mingogram readers were given the same task as the expert reader. Of the 178 occurrences of the different prosodic categories, the first reader identified 138 (78%). The second reader, given a new set of ten prosodically comparable sentences, identified 139 of 202 category occurrences (69%). The major difficulties were found, surprisingly enough, in the identification of grave and grave focal accents. The readers spent nearly an hour working on each sentence.

The rules are currently being implemented on an automatic prosodic analysis system using a curve fitting program which is presented with the same data as the mingogram readers with the addition of segmentation information. A preliminary running of the program on the second set of sentences produced a promising 81% correct for grave accents. Out of the 202 category occurrences, the program identified 132 (65%). The major difficulty was in identifying acute and acute focal accents, and final juncture. Further results of this analysis system will be reported on at the Conference.

## REFERENCES

1. G Bruce, Swedish word accents in sentence perspective. Gleerup, Lund (1977)
2. G Bruce and E Gårding, A prosodic typology for Swedish dialects. In Gårding et al. (eds.) Nordic Prosody. Department of Linguistics, Lund University, 219-228 (1978)
3. E Gårding and G Bruce, A presentation of the Lund model for Swedish intonation. In Fretheim (ed.) Nordic Prosody II. Tapir, Trondheim, 33-39 (1981)
4. E Gårding, Swedish Prosody. *Phonetica* 39, 288-301 (1982)
5. B Lindblom, B Lyberg and K Holmgren, Durational patterns of Swedish phonology. Do they reflect short-term memory processes? Indiana University Linguistics Club, Bloomington (1981)
6. B Lyberg, Temporal properties of spoken Swedish. Monographs from the Institute of Linguistics, University of Stockholm no. 6 (1981)
7. S-G Svensson, Prosody and grammar in speech perception. Monographs from the Institute of Linguistics, University of Stockholm no. 2 (1974)
8. W Lea, Prosodic aids to speech recognition. In Lea (ed.) Trends in Speech Recognition, (Prentice-Hall, N.J., 1980) 166-205.
9. J Vaissière, A suprasegmental component in a French speech recognition system: reducing the number of lexical hypotheses and detecting the main boundary. *Recherches acoustiques CNET Lannion*, 7 82/83 (1983)
10. C W Welin and B Lindblom, The identification of prosodic information from acoustic records by phoneticians. *Journal of the Acoustical Society of America*, 99th ASA Meeting, S 65 (1980)