AUTOMATIC PROSODIC ANALYSIS FOR SWEDISH SPEECH RECOGNITION[1]

David House*, Gösta Bruce*, Francisco Lacerda+ and Björn 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 phrase boundaries 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.

[1] A shorter version of this paper was presented at the European Conference on Speech Technology, Edinburgh, 2-4 September, 1987

*Dept. of Linguistics and Phonetics, Lund University, S-223 62 Lund, Sweden.
+Institute of Linguistics, Stockholm University, S-106 91 Stockholm, Sweden.

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.

Research in prosodic recognition is a natural continuation of a long tradition of research in prosody in Lund and Stockholm. In Lund, this research has been aimed at investigating prosodic phenomena such as rhythm, accentuation and intonation and their relationship both to linguistic structure (phonology, morphology, syntax, semantics, pragmatics and text linguistics) and to the physical realization in the form of tonal and temporal patterns (Bruce, 1977; Bruce and Gårding, 1978; Gårding and Bruce, 1981; Gårding, 1982). Prosody research in Stockholm has primarily dealt with the temporal organization of speech (Lindblom, et al., 1981; Lyberg, 1981) and the relationships between prosody, grammar and speech perception (Svensson, 1974).

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 (1980) and Vaissière (1983).

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. Futhermore, by using a recognition approach to prosody, we hope to reach a better understanding of the mechanisms involved in human perception of prosody.

Among the many functions of prosody in spoken language communication, there are two that can be considered fundamental: the weighting function and the grouping function (Bruce, 1985). Speakers use stress and accentuation, for example, to give weight to syllables, words and phrases (related to semantics). Speakers use juncture to signal phrase boundaries and coherence signals within a phrase (related to syntax). 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 ACUTE accent can restrict and thereby facilitate the lexical search. Identification of GRAVE accent provides us with morphological information which can also facilitate lexical access. For example we know that the syllable following the stressed syllable of a grave accent word belongs to the same word.

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 (Welin and Lindblom, 1980) 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 test sentences, unknown to the reader, 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. (See Appendix 1 for a list of the test sentences.) The speech material was recorded in a Stockholm dialect of Swedish in an anechoic chamber, digitized and stored in disk files (one per sentence). The speech signal was sampled at 20kHz (16-bit/sample). Fundamental frequency contours were obtained by digital processing of the speech signal using a classical instantaneous Fo measuring technique for the Fo extraction. The Fo signal was smoothed and plotted in synchrony with the input speech signal and an intensity curve. The plots were stored, with low sampling frequency, in a 3-channel file for subsequent analysis by the automatic prosodic analysis system. Simulated mingograms of the speech signal displaying a duplex oscillogram, the Fo contour and a bandpass-filtered (1500-3500 Hz) intensity curve were presented to the expert reader (see Figure 1 for a sample mingogram). 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%). A confusion matrix display of these results is presented in Figure 2. Clearly there is considerable prosodic information available in the acoustic signal alone. It is even possible that these results could be improved by adjusting the reader's tendency to identify unstressed syllables as stressed, that is to say, by moving his stressed-unstressed boundary more toward the stressed syllables.

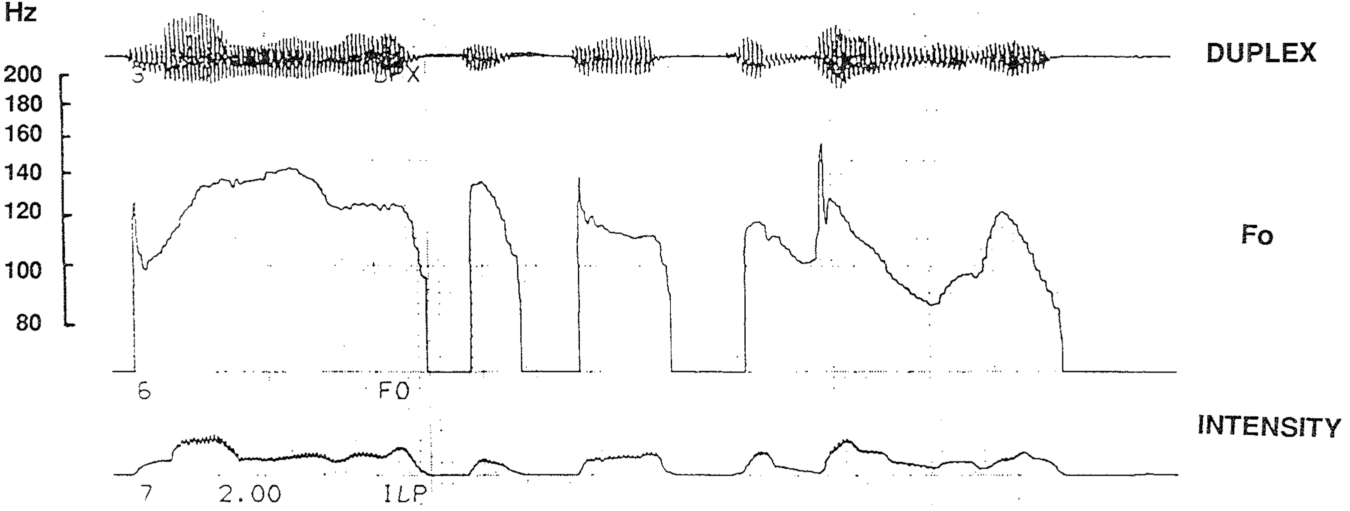

Figure 1. Sample mingogram used for the mingogram reading experiment.

Confusion matrix for expert reader: sentence set 1

IDENTIFIED AS (in percent):

CATEGORY

	G	GF	A	AF	AFT	TJ	SS	US	Ø	TOTAL
G	100									13/13 100%
GF	12	88								7/8 88%
A	4		87				9			20/23 87%
AF				92			8			12/13 92%
AFT					100					2/2 100%
TJ										
SS							100			37/37 100%
US							18	73	8	60/82 73%
Ø								(2)		

All categories = 151/178
85%

G	=	Grave
GF	=	Grave-focal
A	=	Acute
AF	=	Acute-focal
AFT	=	Acute-focal-terminal
TJ	=	Terminal juncture
SS	=	Stressed syllable
US	=	Unstressed syllable
Ø	=	Not a syllable or missed syllable
TOTAL	=	Total correct identification of the category in number and percent

Figure 2. Results of the mingogram reading experiment: Expert reader.

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. A complete English version of the rules is presented in Appendix 2.

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 a grave-accent fall cannot be directly followed by another grave-accent fall. 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, although the reader must also make use of duration and amplitude cues as formulated in rules 6 and 7.

RULE TESTING (NON-EXPERT READERS)

Two non-expert mingogram readers were given the same task as the expert reader. These readers spent nearly an hour working on each sentence. Of the 178 occurrences of the different prosodic categories, the first reader identified 138 (78%). A confusion matrix is presented in Figure 3. The second reader, given a new set of ten prosodically comparable sentences, identified 139 of 202 category occurrences (69%). See Figure 4 for the complete results.

The major difficulties for both non-expert readers were found, surprisingly enough, in the identification of grave and grave focal accents. The reason for this could lie in the ordering of the rules. The Fo fall for grave accent was deemed to be the easiest to identify and to be the most invariant category marker. Therefore this rule was placed first. The readers, however, showed a tendency to first reject Fo patterns that did not exactly fit the grave rule description. Once the grave category was rejected, the readers continued on to the following rules and may have felt forced to assign an acute or acute-focal-final category to the Fo fall even though the grave category would have been a better fit, simply because they had already rejected the grave category.

RULE TESTING (AUTOMATIC ANALYSIS)

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. At this stage of the project, the system uses a strategy that can be described in the following main steps:

1) Parametric description of the Fo contours of manually marked vowel segments

2) Selection of the "best" global parametric description for each vowel contour

3) Classification of the parameterized contours on the basis of a set of rules for identification of stress categories.

Confusion matrix for non-expert reader 1: sentence set 1

IDENTIFIED AS (in percent):

		G	GF	A	AF	AFT	TJ	SS	US	Ø	TOTAL
CATEGORY	G	54		8		8			31		7/13 54%
	GF		38	12		12			38		3/8 38%
	A			78					22		18/23 78%
	AF				85				15		11/13 85%
	AFT					100					2/2 100%
	TJ										
	SS							73	27		27/37 73%
	US							11	85	4	70/82 85%
	Ø				(1)				(3)		

All categories = 138/178
78%

G	=	Grave
GF	=	Grave-focal
A	=	Acute
AF	=	Acute-focal
AFT	=	Acute-focal-terminal
TJ	=	Terminal juncture
SS	=	Stressed syllable
US	=	Unstressed syllable
Ø	=	Not a syllable or missed syllable
TOTAL	=	Total correct identification of the category in number and percent

Figure 3. Results of the mingogram reading experiment: Non-expert reader 1.

Confusion matrix for non-expert reader 2: sentence set 2

IDENTIFIED AS (in percent):

CATEGORY

	G	GF	A	AF	AFT	TJ	SS	US	Ø	TOTAL
G	56		6	6	25		6			9/16 56%
GF	18	36			36		9			4/11 36%
A			71				5	24		15/21 71%
AF			14	71				14		10/14 71%
AFT					100					4/4 100%
TJ	18	9				54		18		6/11 54%
SS							90	10		35/39 90%
US							31	65	3	56/86 65%
Ø										

All categories = 139/202
69%

G	=	Grave
GF	=	Grave-focal
A	=	Acute
AF	=	Acute-focal
AFT	=	Acute-focal-terminal
TJ	=	Terminal juncture
SS	=	Stressed syllable
US	=	Unstressed syllable
Ø	=	Not a syllable or missed syllable
TOTAL	=	Total correct identification of the category in number and percent

Figure 4. Results of the mingogram reading experiment: Non-expert reader 2.

Confusion matrix for program testing 1.0: automatic parsing sentence set 2

IDENTIFIED AS (in percent):

CATEGORY	G	GF	A	AF	AFT	TJ	SS	US	Ø	TOTAL
G	81					6		12		13/16 81%
GF	27	54				9		9		6/11 54%
A	5	10	24					57	5	5/21 24%
AF	7	7		28				50	7	4/14 28%
AFT	25	25						50		0/4 0%
TJ	18	18	9			45		9		5/11 45%
SS							59	38	2	23/39 59%
US							10	88	1	76/86 88%
Ø										

All categories = 132/202
65%

G	=	Grave
GF	=	Grave-focal
A	=	Acute
AF	=	Acute-focal
AFT	=	Acute-focal-terminal
TJ	=	Terminal juncture
SS	=	Stressed syllable
US	=	Unstressed syllable
Ø	=	Not a syllable or missed syllable
TOTAL	=	Total correct identification of the category in number and percent

Figure 5. Results of rule testing: Automatic prosodic analysis system.

This simple curve-fitting procedure is being used as a rough test of the main components of the rules. 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%). See Figure 5 for the complete results. The major difficulty was in identifying acute and acute focal accents, and final juncture. These preliminary results differ considerably from the results of the mingogram reading experiment. Performance may be improved if the parametrization of the contours, at least in its present form, is abandonned. Some of the errors result not from the rules themselves, but rather from the discrepancies between calculated and actual Fo values.

REFERENCES

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

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

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

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

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

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

Lindblom, B., B. Lyberg and K. Holmgren. 1981. Durational patterns of Swedish phonology. Do they reflect short-term memory processes? Indiana University Linguistics Club, Bloomington.

Lyberg, B. 1981. Temporal properties of spoken Swedish. Monographs from the Institute of Linguistics, University of Stockholm no. 6.

Svensson, S-G. 1974. Prosody and grammar in speech perception. Monographs from the Institute of Linguistics, University of Stockholm no. 2.

Vaissière, J. 1983. 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.

Welin, C.W. and B. Lindblom. 1980. The identification of prosodic information from acoustic records by phoneticians. Journal of the Acoustical Society of America, 99th ASA Meeting, S 65.

APPENDIX 1: TEST SENTENCES

SENTENCE SET 1

1. Hon skriver ständigt om sina resor i sin dagbok.
2. Hon bekymrar sig över sitt skrikande barn.
3. Betoningen kommer sist i ordet.
4. Den intressantaste tavlan var en målning av molnen.
5. Mannen reser snart till Bollerup.
6. Vi flyger när vädret är perfekt.
7. Mannen målar, det vill säga han är yrkesmålare.
8. I gryningen ska vi vandra till flodens mynning.
9. Dom läser poesi hela natten.
10. Demonstranterna betraktades av polisen.

SENTENCE SET 2

1. Han skriver en klagodikt över sitt öde.
2. Hon lämnar mig många moderna målningar.
3. Flickan, hon som var på posten, skickade ett paket.
4. Det var den skämtande mannen som var starkast.
5. Hon talar försiktigt när hon vet att mikrofonen är på.
6. Vi får ofta arbeta på natten.
7. Dom har haft kontakt med honom via ett telefonsamtal.
8. Böcker och tidskrifter finns i huset bredvid.
9. Den vackra himlen inspirerade honom.
10. Mannen som gick, han med en blå jacka, betalade kontant.

APPENDIX 2: RULES USED BY MINGOGRAM READERS TO EXTRACT PROSODIC CATEGORIES

1. **Fo-FALL synchronized with a longer vowel having greater amplitude gives <u>GRAVE</u>** (risk of confusion with final juncture) (h)
   a) The beginning of the Fo fall is synchronized with the beginning (often abrupt onset) of a stressed vowel = longer duration + greater amplitude (6).
   b) Fo at the beginning of the fall is higher than or as high as the preceding Fo top.
   c) Fo at the end of the fall is markedly lower then the preceding Fo low.
   d) Range of the fall is greater than most other Fo movements.
   e) Steepness of the fall is relatively constant (45%).
   f) Duration of the fall runs throughout the entire vowel segment.
   g) **WARNING:** An Fo fall for grave accent cannot be directly followed by another grave-accent fall. A syllable containing a grave-accent fall is always followed by another syllable which belongs to the same word. This second syllable is either unstressed or carries secondary stress.
   h) **WARNING:** Risk of confusion with no. 5 below (acute+focus+final juncture)

2. **HIGH or RISING Fo in the vowel following an identified grave accent gives <u>GRAVE+FOCUS</u>**
   a) An Fo rise or a high Fo in the vowel following a grave-accent fall signals grave+focus.
   b) The Fo top in the vowel is almost as high, as high or higher than the previous Fo top.
   c) **NB!** If Fo is low during the vowel, the syllable is part of a non-focal grave accent.

3. **Fo-RISE synchronized with a longer vowel having greater amplitude gives <u>ACUTE+FOCUS</u>** (risk of confusion with initial juncture)
   a) The beginning of the Fo rise is synchronized with the beginning (often abrupt onset) of a stressed vowel = longer duration + greater amplitude (6).
   b) Fo at the beginning of the rise can be lower or the same as Fo in the preceding syllable.
   c) Fo at the end of the rise is higher than the preceding Fo top.
   d) Range of the rise is somewhat less than that of a grave accent fall and usually less than that of the rise for grave+focus (g).
   e) Steepness of the rise is relatively constant (30° - 60°) but varies more than the fall for grave accent.

f) The end of the rise (Fo top) comes near the boundary to the post-tonic syllable or in the post-tonic syllable.

g) **WARNING:** If the syllable comes at the beginning of an utterance, the range of the rise must be greater (= grave accent fall), otherwise the probable category is unstressed+initial juncture.

h) **WARNING:** If an identified grave accent fall is followed by one or more unstressed vowels separating the fall from an Fo rise in a stressed vowel, the sequence can be interpreted as either a compound word (grave+secondary stress) or as grave (non-focal) + acute focal. An Fo hump between the fall and rise signals grave + acute focal. If there is no hump the probable category is grave + secondary stress (compound word).

4. **DOWN-STEPPING of Fo from the previous vowel to a longer vowel having greater amplitude gives <u>ACUTE</u> (non-focal)**

a) Fo is lower in a stressed vowel (longer duration + greater amplitude) than in the preceding vowel.

b) Fo can even be falling in the stressed vowel. The important point is that Fo is higher in the pretonic syllable.

c) The range of the downstepping or fall is less than that of the grave-accent fall.

d) The range is less before focus, it can be greater after focus.

5. **Steep Fo-FALL preceded by a slight Fo-rise in a longer vowel having greater amplitude gives <u>ACUTE+FOCUS+FINAL JUNCTURE</u>**

a) An Fo fall in the beginning of a stressed vowel. This fall can be preceded by a slight Fo rise in the same vowel.

b) The end of the fall is often very low and is normally followed by a pause.

c) The range of the fall is large and is very similar to the grave-accent fall.

d) Steepness is constant and although similar to that of the grave-accent fall this fall is often steeper (70°).

6. **LONGER VOWEL + GREATER AMPLITUDE gives <u>STRESSED VOWEL</u>**

7. **SHORTER VOWEL + LESS AMPLITUDE gives <u>UNSTRESSED VOWEL</u>**