REFERENCES

Bolinger, Dwight. 1981. *Two kinds of vowels, two kinds of rhythm.* Bloomington: IULC.

Bruce, Gösta. 1981. 'Tonal and temporal interplay'. *Working Papers* 21, 49-60. Lund: Dept. of Linguistics.

Cooper, William and Stephen Eady. 1986. 'Metrical phonology in speech production'. *Journal of Memory and Language* 25, 369-84.

Gussenhoven, Carlos. 1986. Review of Selkirk 1984. *Journal of Linguistics* 22, 455-74.

Gussenhoven, Carlos. 1988. 'Lexical accent rules in English'. Unpublished manuscript, Instituut Engels-Amerikaans, Nijmegen University.

Horne, Merle. 1986. 'Focal prominence and the 'phonological phrase' within some recent theories'. *Studia Linguistica* 40, 101-21. (Also published in *Towards a discourse-based model of English sentence intonation, Working Papers* 32, 1987. Lund: Dept. of Linguistics.)

Liberman, Mark and Alan Prince. 1977. 'On stress and linguistic rhythm'. *Linguistic Inquiry* 8, 249-336.

Nespor, Marina and Irene Vogel. 1982. 'Prosodic domains of external sandhi rules'. *The structure of phonological representations,* ed. Harry van der Hulst and Norval Smith, 225-55. Dordrecht: Foris.

Selkirk, Elisabeth. 1980. *On prosodic structure and its relation to syntactic structure.* Bloomington: Indiana University Linguistics Club.

Selkirk, Elisabeth. 1984. *Phonology and syntax: the relation between sound and structure.* Cambridge, Mass.: MIT Press.

Sigurd, Bengt. 1981. 'Commentator. A computer system simulating verbal behaviour'. *Working Papers* 20, 67-89. Lund: Dept. of Linguistics.

Sigurd, Bengt. 1982. 'Text representation in a text production model'. *Text processing. Proceedings of the Nobel Symposium*, ed. Sture Allén, 135-52. Stockholm: Almqvist & Wiksell.

Sigurd, Bengt. 1983. 'How to make a text production system work'. *Working Papers* 25, 179-94. Lund: Dept. of Linguistics.

Sigurd, Bengt. 1984. 'Computer simulation of spontaneous speech production'. *Proceedings of Coling 84*, 79-83. Association for Computational Linguistics.

Strangert, Eva. 1985. *Swedish speech rhythm in a cross-language perspective.* Stockholm: Almqvist & Wiksell.

Lund University, Dept. of Linguistics
Working Papers 33 (1988), 153-161

Recognition of Prosodic Categories in Swedish: Rule Implementation

David House, Gösta Bruce, Lars Eriksson and Francisco Lacerda*

Abstract
Descriptive rules for recognition of prosodic categories in Swedish are currently being implemented in an automatic prosody recognition scheme. An algorithm is described in which the speech signal is segmented into syllables (tonal segments) using intensity measurements and fundamental frequency. Each syllable is then given six values related to fundamental frequency and duration. The values for each syllable are tested against conditions which describe the prosodic categories. The category attaining the highest score is assigned to the syllable. Preliminary results for two sets of rule conditions for ten test sentences are presented.

INTRODUCTION

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. The primary goal of the project is to develop a method for extracting relevant prosodic information from a speech signal. We hope to devise a system which from a speech signal input will provide us with a transcription showing syllabification of the utterance, categorization of the syllables into STRESSED and UNSTRESSED, categorization of the stressed syllables into WORD ACCENTS (ACUTE and GRAVE) and categorization of the word accents into FOCAL and NON-FOCAL accents. We also hope to be able to identify JUNCTURE (connective and boundary signals for phrases). We are currently working with 20 prosodically varied sentences spoken by two speakers of Stockholm Swedish.

The type and structure of the information to be presented to the recognizer has been based on a series of mingogram reading experiments (see House et al. 1987a, 1987b). In the first experiment an expert in Swedish prosody (Gösta Bruce) was presented with mingogram representations of ten unknown sentences showing a duplex oscillogram, fundamental frequency contour and intensity curve. On the basis of this information, he was able to identify 85% of all

* At Stockholm University, Department of Linguistics and Phonetics

occurrences of the prosodic categories referred to above. Descriptive rules were then formulated and tested using two non-expert mingogram readers. Their scores were 78% and 69%.

Our scheme for automatic prosodic recognition can be broken down into three main steps (see Figure 1). First, intensity and fundamental frequency are extracted from the digitized signal. Second, intensity relationships and fundamental frequency information are used to automatically segment the utterance into "tonal segments" which ideally correspond to syllabic units. The prosody recognition rules are then applied to these tonal segments giving us prosodic categories as the output of the system.

The system is being developed for use on an IBM-AT. Current testing of the segmentation algorithm, however, has been carried out using the ILS signal-processing package on a VAX 11/730.

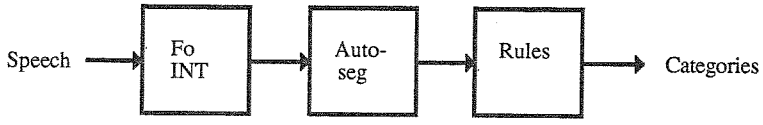

Figure 1. The main components of the prosody recognition scheme.

AUTOMATIC SEGMENTATION

The automatic segmentation component of the recognition scheme has been designed using intensity measurements in much the same way as that described by Mertens 1987. Similar algorithms have been described by Mermelstein 1975, Lea 1980, and Blomberg and Elenius 1985.

The speech signal is first low-pass filtered at 4 kHz (anti-aliasing) and sampled at 10 kHz. An intensity curve is obtained from this signal using the RMS intensity parameter in the ILS program package. This curve is referred to as the *unfiltered* intensity curve. Fundamental frequency is also extracted using a modified cepstral processing technique included in the ILS package. An additional intensity curve is obtained from a digital band-pass filtered version of the sampled signal (0.5-4 kHz, 72 dB/oct). This curve is referred to as the *filtered* intensity curve. Both intensity curves are smoothed (moving average). Figure 2 presents a graphic overview of the segmentation process where steps 1 and 2 represent the above described filtering, analysis and smoothing.

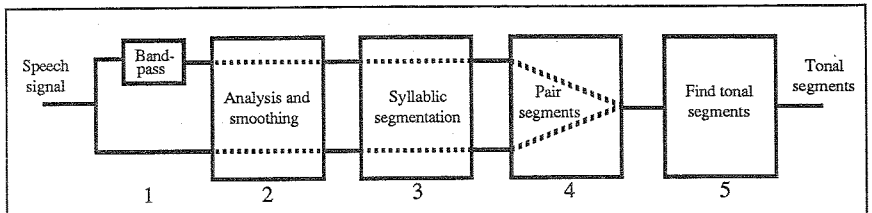

Figure 2. The five main steps of the automatic segmentation component. The parallel lines represent the two different intensity curves.

The next step in the segmentation procedure is a syllabic segmentation algorithm which is applied to both intensity curves (step 3 in Figure 2). This algorithm is also illustrated by the flow chart in Figure 3. Local maximum and minimum values are first marked for each curve. Then a broad segmentation is accomplished where local intensity minima are used as syllable boundaries. A syllable boundary is determined in the following way. Taking the first minimum as the first boundary, the program searches for the next maximum which exceeds 3 dB over the intensity level of the preceding boundary. From this maximum the next minimum which meets the following two conditions is taken as the next syllable boundary: 1) The intensity difference between the minimum and the highest preceding maximum in the syllable must be larger than 3 dB, and 2) The duration from the previous boundary to the minimum in question must be greater than 64 ms. This routine is applied to both intensity curves. The two curves are then compared and the syllable boundaries which are closer together than 64 ms are collapsed into one boundary which is placed halfway between the two original boundaries (step 4 in Figure 2).

The next step in the segmentation procedure is to more finely determine the beginning of each tonal segment, ideally corresponding to the onset of the vowel for each syllabic nucleus (step 5 in Figure 2). This is accomplished by finding the unfiltered intensity maximum in each syllable and defining the beginning boundary of each tonal segment as the point before the maximum where the unfiltered intensity is 3 dB weaker. If there is no voicing at the beginning of the tonal segment then the beginning boundary is adjusted to the right (towards the vowel) to the point where voicing begins. The end of the tonal segment is defined as the intensity minimum already marked by the algorithm (for example segment 2 in Figure 4) *or* the point in the segment where voicing ends (for example segment 5 in Figure 4). In other words, if voicing ends prior to the original boundary, the end boundary is adjusted to the left (towards the vowel) to the point where voicing ends.

To reduce the effects of jitter and wide variations of pitch values occurring at the onset and offset of voicing, an intensity value of 50% below the absolute

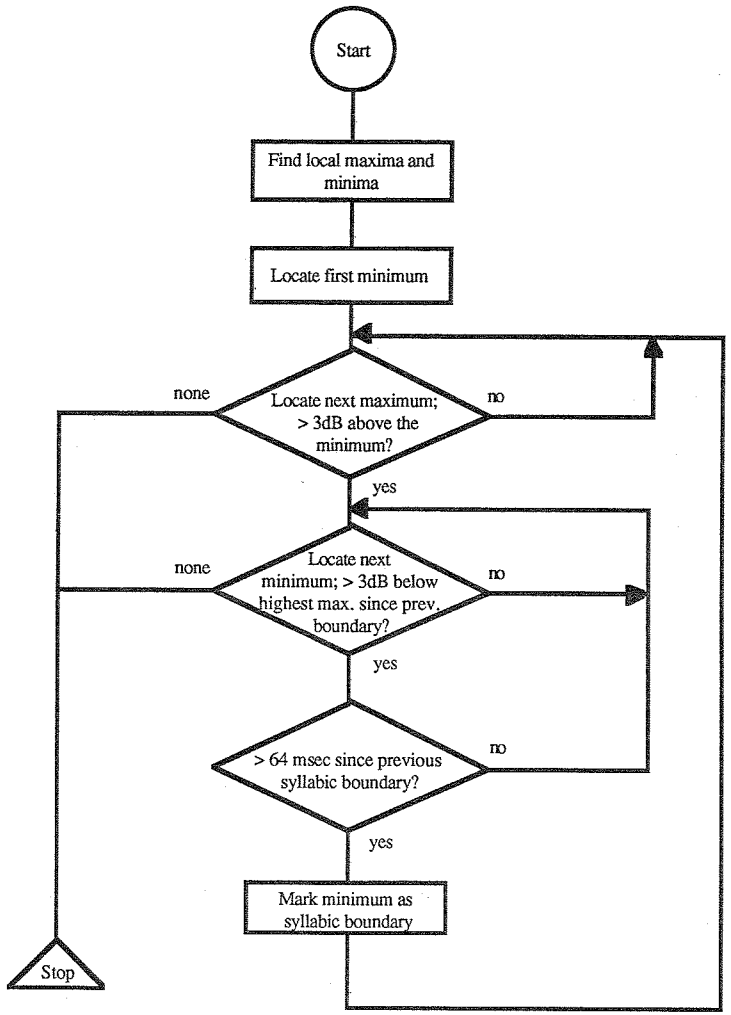

Figure 3. The syllabic segmentation algorithm (step 3 in Figure 2). This algorithm is applied to both the filtered and the unfiltered intensity parameters.

intensity maximum of the utterance was set as a threshold under which any Fo values are rejected, that portion being interpreted as voiceless (see Lea 1980).

A tonal segment, then, is defined as a portion of the speech signal stretching from vowel onset to the end of voicing prior to the next vowel onset. These tonal segments comprise the basic syllabic units for prosodic recognition.

The segmentation program allows free variation of all the above parameter values. As of yet no optimization of these values has been carried out.

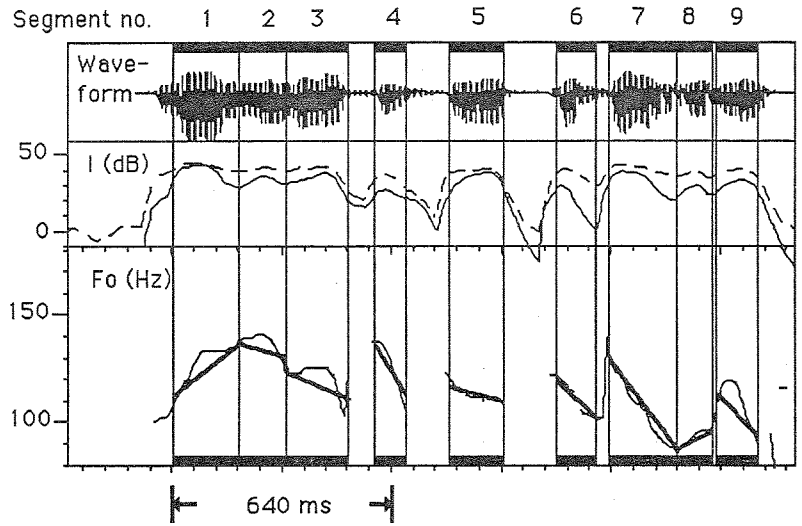

Figure 4. Example of the test utterance, *Mannen reser snart till Bollerup* ('The man will soon travel to Bollerup'), segmented into tonal segments. The unfiltered intensity is represented by the dashed line, the filtered intensity by the solid line. Analyzed Fo is represented by the thin line. The thick line represents the stylized Fo (see text, Rule Implementation).

RULE IMPLEMENTATION

Our preliminary strategy has been to reduce the information available to the recognizer in an attempt to attain the best results with the least possible amount of information. In this way we hope to isolate the most salient cues and build upon them to improve our results. It is clear from our descriptive rule testing that fundamental frequency information is crucial to the recognition of prosodic categories, especially word and focal accents. In our rule system Fo information is mainly expressed as relationships in Fo between successive syllables as this reflects the domain of accentuation.

Our task, then, is to reduce the analyzed Fo contour to a few values while maintaining critical information for recognition of prosodic categories. Evidence from our rule testing indicated that an important area of Fo information is the average Fo level during the first 30-50 ms after vowel onset. This also corresponds to results from speech perception experiments (House 1987). Another important area of information in the rules is the syllable final Fo level. We therefore decided to assign two Fo values to each tonal segment, average Fo during the first 30 ms (B) and average Fo during the last 30 ms of each tonal segment (E). This amounted to a linear stylization of the tonal contour (see Figure 4). In order to test this stylization and see how much prosodic information is lost, we synthesized both speakers' productions of ten sentences using LPC synthesis with the stylized tonal contour as the pitch parameter. In several informal listening tests, the majority of the stylized sentences could not be distinguished from their original counterparts on the basis of intonation alone. Although the reductions did give rise to a few cases of clearly audible tonal deviations, the overall results give further strength to our preliminary method of reducing Fo information.

To incorporate Fo relationships between tonal segments, each segment is assigned two additional Fo values representing the high (H) and low (L) from the preceding (stylized) segment. Finally, two more values are assigned to each segment representing amount of (stylized) Fo change (C) during the segment and total duration (T) of the tonal segment.

In a first implementation of the rules using these six values, conditions for three word-accent categories (grave, acute+focal and acute+non-focal) were formulated based on the descriptive rules and on actual measurements of these values from the categories in question in ten test sentences. The conditions are listed in Table 1.

A recognition routine checks each condition against the six values for each tonal segment. For each true condition, the segment receives one point for the category containing the condition. When all conditions are checked, the category having the most points is assigned to the segment. If two or more categories receive the same score, the following rule hierarchy applies: grave, acute+focal, acute+non-focal.

Finally a relative score threshold can be set where if the highest relative score does not reach the threshold, the syllable is assigned the category UNSTRESSED. If the score reaches the threshold, the category STRESSED is assigned by implication. For example with the threshold set at 0.75 (the value we are currently using) if grave receives two points, acute+focal three and acute+non-focal three, the segment will be assigned unstressed.

Table 1. Rule conditions for three word-accent categories.

Grave	Acute+focal	Acute+non-focal
$C \leq -20$ Hz	$C > 5$ Hz	-30 Hz $< C < 0$ Hz
$T > 150$ ms	$T > 100$ ms	$T > 80$ ms
$B \geq H-5$ Hz	$E \geq H$	$B < H$
$E < L-5$ Hz	$B > L-5$ Hz	$E < L$
	$(B+E)/2 > (H+L)/2$	$(B+E)/2 < (H+L)/2$

Where B = Fo beginning, E = Fo end, C = Fo change, T = duration of tonal segment, H = Fo high in preceding tonal segment, L = Fo low in preceding tonal segment.

RESULTS

The automatic segmentation algorithm successfully detected 168 of 178 syllabic nuclei in ten test sentences. Five extra segments were added by the algorithm rendering a detection score of 92%. Four of the five extra segments were caused by a dental nasal [n] following the vowel. The vowel onset was not as successfully detected in all cases, especially when the vowel was preceded by a nasal or a liquid. In these instances the -3 dB level often occurred in the middle of the consonant. The rule conditions for the three prosodic categories gave the following results: GRAVE 10 recognized of 13 occurrences, ACUTE+FOCAL 11 of 13 and ACUTE+NON-FOCAL 9 of 10 and STRESSED 36 of 37. The category UNSTRESSED, however, was only recognized in 31 cases of 82 occurrences. In most cases, the missed unstressed syllables were categorized as ACUTE+NON-FOCAL.

One of the interim goals of the project is to be able to quickly test and change the rule conditions. In an attempt to improve recognition of UNSTRESSED syllables the final condition for the ACUTE+NON-FOCAL category was changed from $(B+E)/2 < (H+L)/2$ to $B < (H+L)/2$, i.e. from "The Fo average of the actual segment must be lower than the Fo average of the previous one" to "The Fo beginning of the actual segment must be lower than the Fo average for the previous segment". The results for the two different condition sets can be seen in Table 2. A gain of ten category occurrences was achieved at the price of four occurrences giving a net gain of six.

Table 2. Recognition results for two rule condition sets, ten sentences.

Category	1st rule set	2nd rule set	change
Grave	10/13	12/13	+2
Acute+focal	11/13	11/13	±0
Acute+non-focal	9/10	7/10	-2
Stressed	36/37	34/37	-2
Unstressed	31/82	39/82	+8

DISCUSSION

Our preliminary results from the segmentation algorithm are promising as is the success of the rule implementation in separating the three accent categories tested. The major problem is of course that half the unstressed syllables are still categorized as stressed. To a certain extent, this reflects the results of the expert reader who identified 100% of the stressed syllables but only 73% of the unstressed. We hope to improve the results by using a seventh value representing the vowel duration of each tonal segment. It might also prove useful to replace the value for tonal-segment duration with a value representing duration from vowel onset to vowel onset. These new values will be more useful if we can improve detection of vowel onset locations. We are currently investigating the use of intensity curves from different filter bands as an aid to vowel onset identification.

During an additional mingogram reading session using material from the second speaker, our expert reader made greater use of intensity and duration information to differentiate between stressed and unstressed vowels than is currently present in our rules. Furthermore, more variation was found in the category ACUTE+FOCUS than is allowed for in our rules. The use of new values for duration and intensity will allow us to incorporate these findings in the rule implementation scheme.

Finally we anticipate that other problems such as identifying juncture cues and separating these cues from word-accent cues may necessitate the use of additional parameter values for each tonal segment. For example maximum and minimum Fo values could be added. Our recognition scheme will enable us to test these changes as well as further additions to the rules.

REFERENCES

Blomberg, Mats and Kjell Elenius. 1985. 'Automatic time alignment of speech with a phonetic transcription'. *Proceedings of the French Swedish Seminar on Speech,* eds. Bernard Guerin and René Carré, 357-366. Grenoble.

House, David. 1987. 'Perception of tonal patterns in speech: implications for models of speech perception'. *Proceedings of the Eleventh International Congress of Phonetic Sciences,* ed. Ülle Viks, 1, 76-79. Tallinn: Academy of Sciences of the Estonian S.S.R.

House, David, Gösta Bruce, Francisco Lacerda and Björn Lindblom. 1987a. 'Automatic prosodic analysis for Swedish speech recognition'. *Proceedings of the European Conference on Speech Technology,* eds. John Laver and Mervyn A. Jack, 1, 215-218. Edinburgh.

House, David, Gösta Bruce, Francisco Lacerda and Björn Lindblom. 1987b. 'Automatic Prosodic Analysis for Swedish Speech Recognition'. *Working Papers* 31, 87-101. Lund: Dept. of Linguistics.

Lea, Wayne. 1980. 'Prosodic aids to speech recognition'. *Trends in Speech Recognition,* ed. Wayne Lea, 166-205. Englewood Cliffs, NJ: Prentice-Hall.

Mermelstein, Paul. 1975. 'Automatic segmentation of speech into syllabic units'. *Journal of the Acoustical Society of America* 58, 880-883.

Mertens, Piet. 1987. 'Automatic segmentation of speech into syllables'. *Proc. European Conference on Speech Technology,* eds. John Laver and Mervyn A. Jack, 2, 9-12. Edinburgh.