RECOGNITION OF PROSODIC CATEGORIES IN SWEDISH: RULE IMPLEMENTATION

David House, Gösta Bruce, and Lars Eriksson
Lund University, Department of Linguistics and Phonetics
Francisco Lacerda
Stockholm University, Department of Linguistics and Phonetics

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 a restricted material of 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). Descriptive rules were then formulated and tested using two non-expert mingogram readers.

Our scheme for automatic prosodic recognition can be broken down into three main steps. 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.

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. In short, the algorithm uses relationships between maximum and minimum values of both filtered and unfiltered intensity curves to accomplish a broad segmentation. A -3dB threshold prior to the intensity maximum of each segment is applied to locate the onset of the vowel for each syllabic nucleus. The end of the tonal segment is marked at the point where voicing ends prior to the next vowel onset, or if voicing continues, the end of the tonal segment will coincide with the next vowel onset. These tonal segments comprise the basic syllabic units for prosodic recognition.

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. 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. 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. These 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.

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

Grave	**Acute+focal**	**Acute+non-focal**
C≤-20 Hz	C>5 Hz	-30 Hz<C<0 Hz
T>150 ms	T>100 ms	T>80 ms
B≥H-5 Hz	E≥H	B<H
E<L-5 Hz	B>L-5 Hz	E<L
	(B+E)/2>(H+L)/2	B<(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.

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.

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 -3dB level often occurred in the middle of the consonant. The rule conditions for the three prosodic categories, with the relative score threshold set at 0.75, gave the following results: GRAVE 12 recognized of 13 occurrences, ACUTE+FOCAL 11 of 13, ACUTE+NON-FOCAL 7 of 10 and STRESSED 34 of 37. The category UNSTRESSED, however, was recognized in only 39 cases of 82 occurrences. In most cases, the missed unstressed syllables were categorized as ACUTE+NON-FOCAL.

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.

Other problems such as identifying juncture cues and separating these cues from word-accent cues may necessitate the use of additional 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, M. and Elenius, K. 1985. Automatic time alignment of speech with a phonetic transcription. In Guerin and Carré (eds.), 357-366. Proceedings of the French Swedish Seminar on Speech, Grenoble.

House, D. 1987. Perception of tonal patterns in speech: implications for models of speech perception. Proc. of the Eleventh International Congress of Phonetic Sciences. 1:76-79. Academy of Sciences of the Estonian S.S.R. Tallinn.

House, D., Bruce, G., Lacerda, F., & Lindblom, B. 1987a. Automatic Prosodic Analysis for Swedish Speech Recognition. Proc. European Conference on Speech Technology, Edinburgh 1987: 215-218.

House, D., Bruce, G., Lacerda, F., & Lindblom, B. 1987b. Automatic Prosodic Analysis for Swedish Speech Recognition. Working Papers 31, Department of Linguistics, Lund University: 87-101.

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

Mertens, P. 1987. Automatic segmentation of speech into syllables. Proc. European Conference on Speech Technology, Edinburgh 1987: 9-12.

Mermelstein, P. 1975. Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am.* 58 (4) 880-883.