Prosodic parsing for Swedish speech recognition

House, D. and Bruce, G. and Eriksson, L. and Lacerda, F.

journal: STL-QPSR
volume: 30
number: 1
year: 1989
pages: 137-140

PROSODIC PARSING FOR SWEDISH SPEECH RECOGNITION

David House, Gösta Bruce*, Lars Eriksson*, and Francisco Lacerda***
**Dept. of Linguistics and Phonetics, Lund University*
***Institute of Linguistics, Stockholm University*

Abstract

A prosodic parsing system is described which uses tonal, intensity and durational information to recognize prosodic categories. An automatic segmentation algorithm first divides the utterance into "tonal segments" which roughly correspond to syllabic units. A vowel intensity threshold routine then sorts out probable unstressed syllables. Recognition rules are applied to the remaining syllables. The system has been fairly successful in identifying Swedish word accents, stressed syllables and phonetic focus when run on a small set of sentences spoken by two speakers of Stockholm Swedish.

INTRODUCTION

A system which can automatically recognize prosodic information in speech can be beneficial to a larger phonetically based speech recognition system. Information concerning stress, accent, phonetic focus and boundary signals can reduce lexical access time and provide information concerning phrase boundaries and syntactic structure (Lea, 1980; Vaissière, 1983; Gibbon & Braun, 1988). This paper represents a 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.

The primary goal of the project is to develop a method for extracting relevant prosodic information from a speech signal. Our objective is 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.

Our system for automatic prosodic recognition is comprised of four main steps. First, intensity and fundamental frequency are extracted from the digitized speech signal. Second, intensity relationships and fundamental frequency information are used to automatically segment the utterance into "tonal segments" which ideally correspond to syllabic units. Third, an integrated vowel intensity-duration threshold is applied to each syllabic unit. This threshold is set to sort out the most probable unstressed vowels. Finally, prosody recognition rules are applied to the remaining tonal segments giving us prosodic categories as the output of the system.

AUTOMATIC SEGMENTATION

A correct segmentation of the speech signal into syllabic units is of primary importance to the recognition system since the prosodic categories we are using are based on the syllable as the fundamental unit. It is also important that the system marks vowel onsets since vowel onsets make up the crucial synchronization points for identifying the prosodic categories using fundamental frequency movement (House, Bruce, Lacerda, & Eriksson, 1987).

The segmentation component 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 & Elenius (1985). For a complete description of our algorithm, see House, Bruce, Lacerda, & Eriksson (1988) and Lacerda, Bruce, House, & Eriksson (1988).

RULE IMPLEMENTATION

The recognition rules implemented in the system have been based on results from a series of mingogram reading experiments (House & al., 1987) where an expert in Swedish prosody (Gösta Bruce) was presented with mingogram representations of 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 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%.

The rules make use of fundamental frequency information and duration. The fundamental frequency contours are stylized using a linear interpolation method developed within the project. The stylized fundamental frequency information is then expressed as frequency movement within each syllable and relationships in frequency between successive syllables. The rules also take into account differences in duration between syllables. Based on this information, each syllable is given six values which are tested against conditions describing each prosodic category. The category attaining the highest score is assigned to the syllable. For a complete description of the rules, see House, Bruce, Lacerda, & Eriksson (1988).

Our preliminary strategy was to reduce the information available to the recognizer in an attempt to attain the best results with the least possible amount of information. This strategy was fairly successful when used to identify the following word-accent categories: GRAVE, ACUTE+FOCAL and ACUTE+NON-FOCAL. 83% of the word accents were correctly identified for speaker one. Our hypothesis was that we would be able to separate stressed from unstressed syllables as a by-product of correct accent identification. In other words, a syllable identified as having a word accent would automatically be given the category STRESSED while unstressed syllables would not be identified as having a word accent and therefore be given the category UNSTRESSED. The rules, however, seemed to show an oversensitivity to Fo movement causing over half of the unstressed syllables to be given word accents and therefore to be categorized as STRESSED. This lowered our total recognition score for speaker one to only 66% (103 recognized of 155 occurrences of the different categories).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. This problem can also be seen as an indication of the absence of a well-defined category boundary between the categories STRESSED and UNSTRESSED particularly in read speech.

INTEGRATED VOWEL INTENSITY

In an attempt to improve identification of the unstressed syllables, an integrated vowel intensity is measured for each tonal segment. The vowel is defined as the portion of the tonal segment where the intensity exceeds the maximum value for the segment minus 3 dB. The intensity is integrated within this vowel segment in relation to the average intensity for the whole utterance. An intensity value above the average intensity thus gives a positive contribution to the integrated vowel intensity and an intensity value below the average intensity gives a negative contribution.

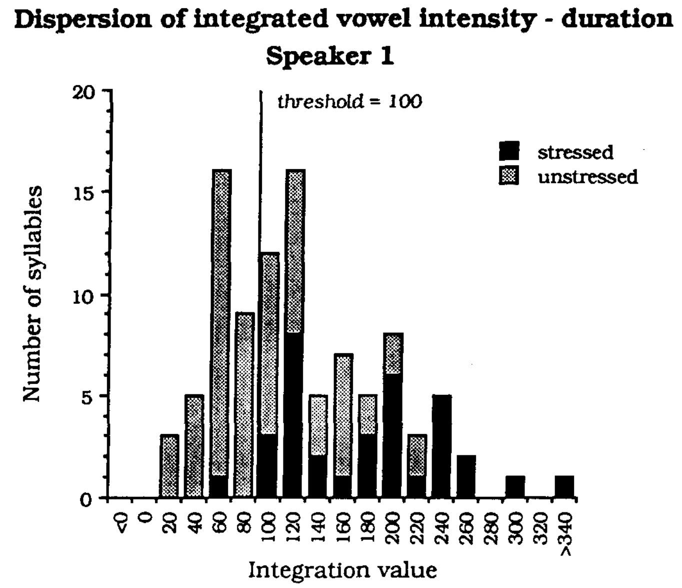

Fig. 1. *Graph showing the unstressed-stressed vowel continuum in terms of integrated vowel intensity and duration for speaker one. The vertical line indicates a threshold value of 100 [dB · cs].*

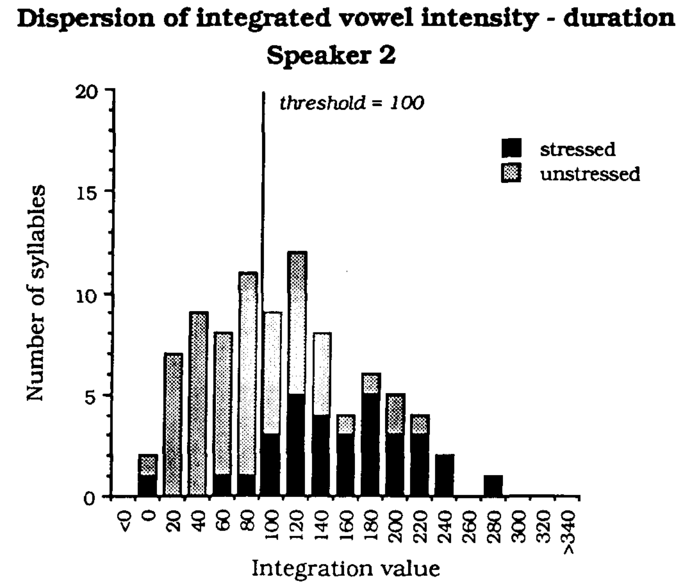

Fig. 2. *Graph showing the unstressed-stressed vowel continuum in terms of integrated vowel intensity and duration for speaker two. The vertical line indicates a threshold value of 100 [dB · cs].*

A threshold is then set to separate the stressed from the unstressed vowels. At present, this threshold is set toward the unstressed end of the stressed-unstressed continuum, i.e., the threshold should exclude only unstressed vowels while letting some unstressed and all stressed vowels through. A threshold of 100 [dB · cs] has been found useful for the test utterances from our two speakers (see Figs. 1 and 2). This threshold is applied to the tonal segments before the rule conditions are applied. Using ten test sentences for speaker one, recognition of the UNSTRESSED category improved from 39 of 82 to 54 of 82 with only one STRESSED category being changed to UNSTRESSED by the intensity threshold. This improved the overall results from 66% to 77%, a score which almost equals our best non-expert reader. Similar results were obtained for speaker two.

Although the addition of other categories such as juncture and the problems involved in separating these cues from those of word accent may necessitate the use of additional parameter values for each tonal segment, our strategy of reduced information and stylization of the tonal contour coupled with the vowel intensity threshold seems to be a promising means of achieving prosodic recognition. A further sharpening of the rules and testing on a larger set of speech material should lead to improved results and increase our understanding of prosody in a speech recognition setting.

References

Blomberg, M. & Elenius, K. (1985): "Automatic time alignment of speech with a phonetic transcription," pp. 357-366 in (B. Guerin & R. Carré, eds.) *Proc. of the French Swedish Seminar on Speech, Grenoble.*

Gibbon, D. & Braun, G. (1988): "The PSI/PHI architecture for prosodic parsing," pp. 202-204 in (D. Vargha, ed.) *Proc. of the 12th Int. Conf. on Computational Linguistics, Budapest.*

House, D., Bruce, G., Lacerda, F., & Eriksson, L. (1987): "Automatic prosodic analysis for Swedish speech recognition," pp. 215-218 in. (J. Laver & M.A. Jack, eds.) *Proc. European Conf. on Speech Technology*, CEP Consultants, Edinburgh.

House, D., Bruce, G., Lacerda, F., & Eriksson, L. (1988): "Recognition of prosodic categories in Swedish: Rule implementation," *Working Papers 33*, (Dept. of Linguistics and Phonetics, Lund University, pp. 161-169.

Lacerda, F., Bruce, G., House, D., & Eriksson, L. (1988): "Prosodic parsing of Swedish, status report: Segmentation strategy," pp. 1187-1195 in (W.A. Ainsworth and J.N. Holmes, eds.) *Proc. Seventh FASE Symposium*, Rayross Printers, Liverpool.

Lea, W. (1980): "Prosodic aids to speech recognition," pp. 66-205 in (W. Lea, ed.) *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, New Jersey.

Mertens, P. (1987): "Automatic segmentation of speech into syllables," pp. 9-12 in (J. Laver & M.A. Jack, eds.) *Proc. European Conf. on Speech Technology*, Edinburgh.

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

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, pp. 111-112.