

Modeling lexical stress in read and spontaneous speech. Joseph H. Polifroni and Alexander I. Rudnicky (School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213).

Although prosodic information has long been thought important for speech recognition, few demonstrations exist of its effective use in actual recognition systems. Lexical stress information has been shown to improve recognition performance by allowing the differentiation of otherwise confusable words (e.g., Rudnicky and Li, *DARPA Workshop on Speech Recogn.*, June 1988). In this study, we examine explicit lexical stress modeling for a spreadsheet system whose language that contains a significant number of confusable words (for example, EIGHTY and EIGHTEEN). Since performance characteristics obtained for read speech may not generalize to other types of speech, we evaluated our models on both read speech and on spontaneous goal-directed speech. A database of over 4000 spreadsheet and numeric utterances was available for training, sufficient data to produce robust models for both stressed and unstressed vowels. This database was used to train models for an (HMM-based) speaker-independent continuous-speech system with a vocabulary of 273 words and language perplexity of about 51. The testing data used in this study was based on data generated in a separate study examining the use of a spoken-language spreadsheet. Two testing sets were formed: *a*) a “spontaneous” set, composed of all parsable utterances from a spreadsheet task, from 5 different talkers. *b*) a “read” set, consisting of the above spontaneous sentences read by their original speakers. The use of lexical stress models was found to reduce the error rate for spontaneous speech by [foo%], demonstrating that training with read speech provides sufficient power for lexical stress modeling. A comparison with the read data provides an insight into the nature of the improvement.

Technical Committee: Speech Communication

(PACS) Subject Classification number: 43.72.Ne

Method of Presentation: Prefer lecture but willing to give as poster.

Telephone number: 412/268-2622 (A. Rudnicky)

Modeling lexical stress in read and spontaneous speech

Joseph H. Polifroni
Alexander I. Rudnicky

*Carnegie Mellon University
Pittsburgh, Pennsylvania
USA 15213*

*Paper FF 9
118th Meeting of the Acoustical Society of America
30 November 1989*

Abstract

Although prosodic information has long been thought important for speech recognition, few demonstrations exist of its effective use in actual recognition systems. Lexical stress information has been shown to improve recognition performance by allowing the differentiation of otherwise confusable words (e.g., Rudnicky and Li, *DARPA Workshop on Speech Recognition*, June 1988). In this study, we examine explicit lexical stress modeling for a spreadsheet system whose language that contains a significant number of confusable words (for example, EIGHTY and EIGHTEEN). Since performance characteristics obtained for read speech may not generalize to other types of speech, we evaluated our models on both read speech and on spontaneous goal-directed speech. A database of over 4000 spreadsheet and numeric utterances was available for training, sufficient data to produce robust models for both stressed and unstressed vowels. This database was used to train models for an (HMM-based) speaker-independent continuous-speech system with a vocabulary of 271 words and language perplexity of about 52. The testing data used in this study was based on data generated in a separate study examining the use of a spoken-language spreadsheet. Two testing sets were formed: *a*) a “spontaneous” set, composed of all parsable utterances from a spreadsheet task, from 5 different talkers. *b*) a “read” set, consisting of the above spontaneous sentences read by their original speakers. [Work supported by DARPA.]

The Problem

Hidden Markov Models provide an effective technique for modeling speech. However, recognition accuracy depends on the nature of the speech representation being used.

Other things being equal, the richer the representation the better the performance that one can expect.

Lexical stress has been shown to influence the acoustic realization of phones, and there are demonstrations that it can be used to improve lexical access (e.g., Waibel, 1986, for isolated words; Rudnicky and Li, 1988, for connected speech).

The utility of stress information for HMM-based recognition systems, however, has not been clearly established.

Questions

- Can stress information be used to improve the performance of a HMM-based recognition system?
- Is the contribution of stress information different for the recognition of read speech *versus* the recognition of spontaneous speech?

Experiment

A 4012 utterance training database for a spreadsheet vocabulary of 271 words was available for training.

We performed the following types of training:

- Trained monophone models, using stress-marked and non-marked lexicons.
- Trained triphone models, using stress-marked and non-marked lexicons.

Two sets of test data were used:

- The spontaneous speech testing data consisted of 353 utterances (1346 words) from users performing a spreadsheet task.
- Talkers were asked to return and record read versions of their spontaneous utterances. Only parsable utterances were recorded. There were 263 such utterances (982 words).

Results

The models trained with the different configurations were tested on the spontaneous and read versions of our test utterances.

Monophone Models (word error rate)

Training style	Spontaneous	
Simple Models	7.8%	
Stress Models	10.4%	

Triphone Models (word error rate)

Training style	Spontaneous	Read
Simple Models	3.5%	6.3%
Stress Models	3.3%	6.7%

Summary

- For monophone modeling, differentiating vowel models by stress *degrades* recognition performance.
- For triphone modeling, differentiating vowel models by stress appears to have no effect (though there might be some improvement for spontaneous speech).

Conclusions

- The monophone result is anomalous. It cannot be attributed to the reduction in training data for each model, since a comparable reduction was not observed in the triphone condition.
- The triphone result suggests that stress information is contextual in nature and may need to be defined in terms of acoustic relationships within a larger context rather than in terms of simple vowel stress value.

Since our data indicate minimal effect for triphone contexts, it suggests that the appropriate context for lexical stress information may be the *word*.