LEXICAL STRESS DETECTION ON STRESS-MINIMAL WORD PAIRS

Goangshiuan S. Ying, Leah H. Jamieson, Ruxin Chen, Carl D. Michell

School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285
URL: http://purcell.ecn.purdue.edu/~speechg

Hsin Liu[1]
Computer Science, Cornell University, Ithaca, NY 14853

ABSTRACT

We present a study on the use of lexical stress classification to aid in the recognition of phonetically similar words. In this study, we use a simple pattern recognition approach to determine which syllable is lexically stressed for phonetically similar word pairs (e.g., PERfect, perFECT) extracted from continuously spoken sentences. We use a combination of two features from the acoustic correlates of lexical stress, and assume multivariate Gaussian distributions to form a Bayesian classifier. The features used are normalized energy and duration of the vowel for each syllable of the word. We evaluate several normalization methods. Two sets of sentences were designed for this study. For the pilot experiment, the classification accuracy on words from the natural sentence set was 89.9% and on words from the control sentence set was 100%. To improve the performance, three-feature classifiers, which included two normalized energy features and one normalized duration feature, were developed. The classification accuracy on words from the natural sentence set was 97.23%.

1. INTRODUCTION

Stress is an important prosodic feature in American English. Changes in a combination of *intensity*, *fundamental frequency* (F_0), and *duration* signal different stress levels. A stressed vowel tends to have higher intensity, longer duration, and higher F_0. Along with phonemic structure, lexical stress patterns help to identify a word. In English, patterns of stress on syllables can be used to distinguish the different syntactic roles of a word. For example, the word "COMbine" (first syllable stressed) is a noun, but "comBINE" (second syllable stressed) is a verb. This study centers on lexical stress detection for words that have very similar phonetic structure but differ in their location of stress. Following [3], we refer to each pair of phonetically similar words as "stress-minimal" word pairs. We report on results from a single speaker pilot study, then use these results as the basis for experiments on an expanded database of five speakers.

[1]Supported by the 1994 NSF/CRA Distributed Mentor Program.

2. LEXICAL STRESS DETECTION

Related Work Several lexical stress determination algorithms have been proposed [1, 3, 6, 8]. In [1], Aull and Zue used 350 multisyllabic words spoken in isolation to train a template-based stress detector. In [8], Waibel used a Bayesian classifier to determine whether each syllable is stressed or unstressed. In [3], Freij et al. used two HMMs trained with frame-based features to recognize stress-minimal bisyllabic word pairs extracted from continuous speech.

2.1. Bayesian Classifier

Following [8], pattern recognition with Bayesian classifiers is chosen as the detection algorithm for this experiment. Since the experiment is done on words which have two possible locations of syllable stress, two Bayesian classifiers are used, one classifier for each syllable. Each classifier has only two classes, S (for stressed vowel) and U (for unstressed vowel). We assume that the energy and duration features can be jointly modeled by an N-dimensional normal distribution for each class,

$$\begin{aligned} p(x|x \in S) &= (2\pi)^{-N/2}|\Sigma_S|^{-1/2} \\ &\quad \exp\{-\frac{1}{2}(x-\mu_S)^T\Sigma_S^{-1}(x-\mu_S)\} \\ p(x|x \in U) &= (2\pi)^{-N/2}|\Sigma_U|^{-1/2} \\ &\quad \exp\{-\frac{1}{2}(x-\mu_U)^T\Sigma_U^{-1}(x-\mu_U)\} \end{aligned}$$

where N is the number of features. Since the number of stressed and unstressed syllables are equal in our task, the prior probabilities provide no discriminatory information. (i.e., $P(x \in S) = P(x \in U) = 1/2$). Hence, the Bayesian classifier reduces to a maximum likelihood classifier which can be written as follows:

$$\begin{aligned} &\exp\{(x-\mu_S)^T\Sigma_S^{-1}(x-\mu_S)\} + \log|\Sigma_S| < \\ &\quad \exp\{(x-\mu_U)^T\Sigma_U^{-1}(x-\mu_U)\} + \log|\Sigma_U| \end{aligned}$$

We have found that $\log|\Sigma_S| \approx \log|\Sigma_U|$, and that omitting these terms has a negligible effect on classification accuracy.

2.2. Feature Selection

The most successful features for stress detection have been energy, duration, and F_0 [1, 3, 6, 8]. Each of these acoustic correlates can be calculated in a number of ways. For instance, one could use the duration of a syllable, the duration of the vowel in the syllable, or the duration of the vocalic portion of the syllable. Researchers have suggested that obstruents (i.e., fricatives, affricates, and stops) are far less affected by stress than are sonorants[4]. Similarly, measurements taken on the vocalic portion of a syllable would be highly dependent on the vowel context. To simplify normalization, feature measurements are done on the vowel of each syllable.

Vowel boundaries and duration values were obtained using an explicit-duration HMM phone recognizer [2, 7]. The energy is calculated with a root mean square energy ($RMSE$) algorithm. Several different normalization methods on both energy and duration are applied in this study. The first energy normalization, NEng1, adjusts only for relative energy across sentences so that the recording level will not affect the results of the experiment. The first duration normalization, NDur1, adjusts only for speaking rate.

- $NEng1$: average vowel energy in each syllable normalized by the overall energy in the word.
- $NDur1$: vowel duration in each syllable normalized by total word duration.

The second normalization, NEng2 and NDur2, uses the total vowel energy and duration in the word (e.g., the sum of the two vowel energies or durations in a disyllableic word).

- $NEng2$: average vowel energy in each syllable normalized by the overall vowel energy in the same word.
- $NDur2$: vowel duration in each syllable normalized by total vowel duration in the same word.

The third normalization, NEng3 and NDur3, attempts to remove vowel-specific effects, e.g., the phenomenon that some vowels are typically longer than others, independent of lexical stress or context. The average energy and duration for each vowel were computed over the TIMIT database.

- $NEng3$: $NEng2$ normalized by the average energy for that vowel.
- $NDur3$: $NDur2$ normalized by the average duration for that vowel.

Most previous studies have included F_0 information in the feature vector. However, in continuous speech, phrase and sentence level intonation have strong effects on the word level pitch contour. Pitch accent, i.e., a stressed syllable produced by higher pitch value than other syllables in the word, can be de-accented by the influence of the intonation of the larger unit. A detailed study of how the pitch contour affects the word level stress value in continuous speech will be investigated in future work. F_0 related information is excluded in the feature selection of this experiment.

3. SPEECH CORPUS

Eleven word pairs were chosen for this experiment, where each word pair consists of two words that have similar phonetic structure but different stress patterns. Each word in this database is called a *target word*. The words are: *attribute*, *combine*, *compact*, *conduct*, *convert*, *convict*, *object*, *perfect*, *present*, *project*, and *suspect*. All word pairs but *attribute* are disyllabic words. For the words "ATtribute" and "atTRIbute", we include only the first two syllables in this experiment, since only the first two syllables can be stressed.

The initial databases contain speech from a male speaker. The first database was created by embedding the target words into carrier sentences. 112 sentences were created to cover the 11 target word pairs, with each word appearing ten times. The sentences were designed so that each target word appears in different positions in the sentences to avoid modeling the effects caused by the word position, such as lengthening the final syllable of the word at the end of a sentence. Each target word also appears in different surrounding phonetic contexts (e.g., "The town gossips suspect that ..."; "I suspect she will object ..."). This database is referred to the *natural sentence set*.

The second database is called the *control sentence set*. After each natural sentence was read, the speaker was prompted to repeat each of the target words from that sentence, but in a fixed context: "Please say _____ again." For example:

A wooden object does not conduct electricity.

Please say "OBject" again.
Please say "conDUCT" again.

The speech was sampled at 16KHz using a 12-bit A/D converter. The target words were manually extracted from the utterances using *xwaves* [5]. The phone transcription for each word in the database was labeled manually. In the pilot study, an HMM phone recognizer, which was trained on TIMIT database, automatically segmented each word into phones [7].

4. PILOT STUDY

The pilot study was designed to examine how the choice of features affected classification accuracy. A Bayesian classifier was used to train and test on the database which contained the speech data from a male speaker.

Two-feature Classifier: Two evaluation tasks were designed for the pilot study. A Bayesian classifier with two features, one energy feature and one duration feature, was trained and tested in both tasks. Table 1 shows the results of the tests. In the first task, disjoint training and testing sets were generated by randomly selecting 9 of the 10 repetitions of each of the 22 target words for the training set, and testing on the 22 words not used in training. The experiment was repeated 100 times. Performance for this *in-vocabulary task* is given by the average word classification error rate.

In the second task, we train on all repetitions of 10 of the 11 stress-minimal word pairs, and test on the 10 repetitions of the remaining word pair. Performance for this *out-of-vocabulary task* is given by the average error over the 11 tests. In both tasks, we compare the performance on the control and natural sentence sets. The results are shown in Table 1.

(A)

	Control sentence set (% error)			Natural sentence set (% error)		
Features	NEng1	NEng2	NEng3	NEng1	NEng2	NEng3
NDur1	0.45	0.18	13.09	14.77	11.05	37.00
NDur2	1.95	0.05	4.68	13.05	10.14	13.86
NDur3	3.13	0.00	14.14	23.73	10.77	42.91

(B)

	Control sentence set (% error)			Natural sentence set (% error)		
Features	NEng1	NEng2	NEng3	NEng1	NEng2	NEng3
NDur1	0.00	0.45	17.27	16.82	14.55	39.55
NDur2	5.00	0.00	7.72	15.00	11.36	18.18
NDur3	8.18	0.00	15.00	26.78	13.18	43.64

Table 1: Test results, showing % error. (A) In-vocabulary task; (B) Out-of-vocabulary task.

A number of conclusions can be drawn from the study. As expected, the error is much lower (essentially error-free) on the control sentence set, reflecting both the more careful pronunciation and the fixed context. The relative effectiveness of the various normalization methods was for the most part the same in both tasks. The least powerful normalization was sufficient for the control sentences. The combination of NEng2 and NDur2 proved most effective on the natural sentence set, with similar performance being achieved using NEng2 and NDur3.

In the two feature model, across all of the tests, normalization by the sum of the vowel durations (NDur2) was the most effective form of duration normalization, and normalization by the sum of the vowel energies (NEng2) was the most effective form of energy normalization. The best performance on the out-of-vocabulary task was only slightly lower (1%) than the performance on the in-vocabulary task, indicating that the features are quite robust for general syllable stress classification.

Three-feature classifier: The number of misclassifications for each target word shows that different classifiers have better performance on different words. For example, for the word "atTRIbute", the [NEng2 NDur2] classifier doubles the number of misclassifications compared to the [NDur3 NEng2] classifier, but reverses the result when classifying the word "COMbine". Based on this observation, a three-feature Bayesian classifier was developed. The hypothesis is that a three dimensional feature space will allow a more consistent separation of the feature space of stressed vowels and unstressed vowels. A three-feature classifier was trained and tested on the combination of three features, [NDur2 NDur3 NEng2]. The performance of the in-vocabulary task experiment improved significantly. The misclassification rate is reduced from 10% to less than 5%. The standard deviation of the error rate remains about the same as previously.

Table 2 shows a comparison of the results of the pilot study to previous work.

5. PERFORMANCE EVALUATION

The 4.37% error rate in the pilot study is the lowest among all the systems listed in Table 1. Because the databases, training, and testing conditions are different in each experiment, it is difficult to make any direct comparisons. However, we suspect that the use of a single speaker contributed to the low error rate in the pilot study. A new experiment was developed in which we added three more male speakers' utterances and one female speaker's utterances to the database. Therefore, for both natural and control sets, each target word contains of 50 different utterances. The same segmentation procedure was applied to this experiment. A different HMM phone recognizer, also trained on the TIMIT database, was used for this experiment [2]. Because of the high classification rate of the words from the control sentence set in pilot study, the control sentence set was not evaluated further.

	Continuous or Discrete	Syllable Segmentation	Vocabulary Type	Classification Method	Decision Level	Number of Features	Number of Speakers	% Error
Aull & Zue	Discrete	Automatic	General Multisyllabic	Template Matching	Word	5	11	13.00
Waibel	Continuous	Automatic	General Multisyllabic	Bayesian	Syllable	4	10	12.27
Freij et al.	Continuous	Hand Labeled	Stress-Minimal Bisyllabic Word Pairs	HMM with frame based features	Syllable	9	3	10.30
Ying et al. (pilot)	Continuous	Automatic	Stress-Minimal Bisyllabic Word Pairs	Bayesian	Word	2	1	10.14
						3	1	4.37
Ying et al.	Continuous	Automatic	Stress-Minimal Bisyllabic Word Pairs	Bayesian	Word	2	5	8.80
						3	5	2.27

Table 2 : Comparison of several lexical stress detection schemes.

In-vocabulary task: In this task, disjoint training and testing sets were generated by selecting 3 of the 5 speakers in the database for training and using the data from the other two speakers for the testing set. Ten different classifiers were generated and the results of the average error rate on syllable stress pattern classification are shown in Table 3. Similar to the pilot study, [NDur2 NEng2] was chosen for building the two-feature classifier. However, instead of using [NDur2 NDur3 NEng2] as in pilot study, [NDur2 NDur3 NEng2] were chosen to build the three-feature classifier for the evaluation.

In-vocabulary Task	Number of Speakers	Two-Feature Classifier	Three-Feature Classifier
Pilot Study	1	10.14%	4.37%
Evaluation	5	8.80%	2.27%

Table 3: Comparison between results from pilot study and new experimental evaluation. The score represents the syllable classification error.

Out-of-vocabulary task The out-of-vocabulary task in the pilot study was also evaluated in this experiment. In addition, we reduced the number of the word pairs to build the classifier and increased the number of the word pairs for testing. The word pairs in training and testing remained disjoint. The results are shown in Table 4.

(M,N)	Two-Feature Classifier	Three-Feature Classifier
(10,1)	11.45%	5.45%
(9,2)	11.56%	5.36%
(8,3)	16.56%	5.32%
(7,4)	10.60%	5.28%
(6,5)	10.22%	5.27%

Table 4: Results of out-of-vocabulary task, where (M,N) indicates that the classifier is trained on M word pairs and tested on N word pairs, with $M + N = 11$. The score represents the syllable classification error.

6. Conclusion and Discussion

The vowels in stressed syllables tend to have higher energy and longer duration than the same vowels in unstressed syllables. However, stressed vowels do not necessarily have longer duration than the unstressed vowel in the same word. For instance, in the word "COMbine", the vowel /aa/ has lower average energy and shorter duration than the vowel /ay/, even though vowel /aa/ is stressed. Vowel /ay/ has much longer duration and stronger average energy than most of other vowels, even when it is in the unstressed syllable. Therefore it is predictable that the classifier trained on the combination of features in first and second normalization pairs will have high misclassification for the word "COMbine".

The third normalization pair, [NEng3 NDur3], was designed initially to avoid this situation. This normalization is mainly aimed at the unstressed vowels which are still *full vowels*, such as /ay/ in "COMbine", /ae/ in "COMpact", and /er/ in "CONvert". The hypothesis is that after normalization by the vowel specific average energy and duration, the vowel in the stressed syllable will have a larger ratio for both features after normalization than the unstressed full vowel in the same word. To our disappointment, the [NDur3 NEng3] classifier has worse performance in classifying stressed vowels than most of the other two-feature classifiers.

However, introducing NEng3 or NDur3 to the [NDur2 NEng2] pair to form a three dimensional Gaussian space has reduced the error rate by more than 50% in both the pilot study and the 5-speaker evaluation. This suggests that removing the vowel-specific effects will be important for the energy-duration based lexical stress classifier.

As shown in Table 2, researchers have used different features for lexical stress detection. Among the listed systems, our approach uses the least number of the features to build the classifiers and is the only system that doesn't include fundamental frequency related information as a basic feature. Waibel [8] pointed out that fundamental frequency related features could have a negative effect on stress detection accuracy. However, it is certainly the case that F_0 is one of the main acoustic correlates of stress. Correctly representing the pitch contour and applying different normalization techniques will directly affect the performance of lexical stress detectors. Incorporating the fundamental frequency related information and expanding the target words to multisyllabic words will be the main focus in our follow-up work.

7. REFERENCES

1. A. M. Aull and V. W. Zue. Lexical stress determination and its application to speech recognition. In *Proc. ICASSP*, pages 1549–1552, 1985.
2. R. Chen and L.H. Jamieson. Explicit modeling of coarticulation in a statistical speech recognizer. In *Proc. ICASSP*, pages I.463–I.466, 1996.
3. G.J. Freij, F. Fallside, C. Hoequist, and F. Nolan. Lexical stress estimation and phonological knowledge. *Computer Speech and Language*, 4(1):1–15, 1990.
4. A. House and G. Fairbanks. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *JASA*, 25:105–113, 1953.
5. Entropic Research Laboratory. *User's Manual for waves+ (Version 5.0)*, 1993.
6. P. Lieberman. Some acoustic correlates of word stress in American English. *JASA*, 32:451–454, 1960.
7. C.D. Mitchell and L.H. Jamieson. Modeling duration in a hidden markov model with the exponential family. In *Proc. ICASSP*, pages II.331–II.334, 1993.
8. A. Waibel. Recognition of lexical stress in a continuous speech system-A pattern recognition approach. In *Proc. ICASSP*, pages 2287–2290, 1986.