Lexical Stress Modeling for Improved Speech Recognition of Spontaneous Telephone Speech in the JUPITER Domain[1]

Chao Wang and Stephanie Seneff

Spoken Language Systems Group, Laboratory for Computer Science
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA
{wangc,seneff}@sls.lcs.mit.edu

Abstract

This paper examines an approach of using lexical stress models to improve the speech recognition performance on spontaneous telephone speech. We analyzed the correlation of various pitch, energy, and duration measurements with lexical stress on a large corpus of spontaneous utterances, and identified the most informative features of stress using classification experiments. We incorporated the stress models into the recognizer first-pass Viterbi search and obtained modest but statistically significant improvements over a state-of-the-art real-time performance on the JUPITER domain.

1. Introduction

Lexical stress is an important property for the English language. It has been suggested in [10] that stressed syllables provide *islands of phonetic reliability* in speech communication. In addition, lexical studies have demonstrated that stressed syllables are more informative to word inference [7], and knowing the stress pattern of a word can greatly reduce the number of competing word candidates [2]. Clearly, lexical stress contains useful information for automatic speech recognition.

Early work on lexical stress modeling has focused on the recognition of stress patterns to reduce word candidates for large-vocabulary isolated word recognition [2, 15], or to disambiguate stress-minimal word pairs [3]. More recently, there have been attempts at utilizing stress information to improve *continuous* speech recognition. In [1, 11], the lexical stress property was used to separate phones during training to obtain more accurate acoustic models. In [6], stress-dependent phonological rules were applied for phone to phoneme mapping. In [9], hidden Markov models for "weak/strong" and "stressed/unstressed" syllables were applied to resort the recognizer N-best outputs. [8, 14] also examined stress classification in continuous speech; however, no speech recognition experiments were performed using the stress models. In general, previous research on using stress models in continuous speech recognition has been limited, and we have not found any work on spontaneous English speech reported in the literature.

In this paper, we test the approach of scoring the lexical stress patterns of recognizer hypotheses to improve automatic speech recognition performance. We expect that substitution, insertion and deletion errors sometimes result in mismatched stress characteristics between the hypothesized syllable nucleus and its acoustics. By scoring the stress pattern of a hypothesis, the additional constraints from stress models will improve over a system which uses segmental constraints only. However, the acoustic manifestations of English lexical stress are quite obscure. Although it has been found that prosodic attributes, i.e., energy, duration, and pitch, correlate with the stress property of a vowel, these features are also highly dependent on its segmental aspects (intrinsic values). To complicate things further, not all lexically stressed syllables are stressed in continuous speech, e.g., mono-syllabic function words are often not stressed; and a subset of lexically stressed syllables in a sentence also carry the pitch accents of the spoken utterance. Although pitch accentedness has been argued to be a more appropriate indication of "stress" in continuous speech, their occurrences can not be predicted from orthographical transcriptions, and hence, they are less useful to a recognizer. On the other hand, lexical stress can easily be encoded in the lexicon of a segment-based recognizer. However, the question remains whether it can be reliably determined from the acoustics in spontaneous speech to benefit recognition.

We address two research issues in this study: 1) how well can the stress property of a vowel be determined from the acoustics in spontaneous speech, and 2) can such information improve speech recognition performance. To answer these questions, we will study the correlation of various pitch, energy, and duration measurements with lexical stress on a large corpus of spontaneous utterances, and identify the most informative features of stress using classification experiments. We will also develop probabilistic models for various lexical stress categories, and combine the stress model scores with other acoustic scores in the recognition search for improved performance. We experimented with prosodic models of various complexity, from only considering the lexical stress property to also taking into account the intrinsic differences among phones. We found that using prosodic models improved over the baseline performance on the JUPITER domain. However, the gain by using prosodic models seems to be achieved mainly by reducing implausible hypotheses, rather than by distinguishing the fine differences among various stress and segmental classes; thus, there is no additional gain by utilizing more refined modeling.

In the following sections, we first provide some background knowledge for the experiments, including the JUPITER corpus and a baseline JUPITER recognizer which incorporates stress markings in its lexicon. After that, we study the correlation of various pitch, energy, and duration related measurements with lexical stress and identify the best feature set using classification experiments. Finally, we present speech recognition experiments using the basic lexical stress models and other prosodic models of varying complexity.

[1]This research was supported by DARPA under contract N66001-99-1-8904, monitored through the Naval Command, Control and Ocean Surveillance Center.

Data Set	Train	Development	Test
# Utterances	84165	1819	3028

Table 1: Summary of data sets in the JUPITER corpus.

2. Experimental Background

2.1. JUPITER Corpus

The JUPITER system [17] is a telephone-based conversational interface to on-line weather information developed at the Spoken Language Systems group of the MIT Laboratory for Computer Science. A user can call the system via a toll-free number and ask weather-related questions using natural speech. JUPITER has real-time knowledge about the weather information for over 500 cities, mostly within the United States, but also some selected major cities world-wide. The system also has some content processing capability, so that it can give specific answers to user queries regarding weather acts, temperature, wind speed, pressure, humidity, sunrise/sunset times, etc.

A tremendous amount of spontaneous telephone speech has been collected since the system was made publicly available via a toll-free number. There have been over 180,000 utterances from over 30,000 phone calls recorded over a two year period [17], and the data are still coming in. We use about 80,000 orthographically transcribed utterances in our experiments. Table 1 summarizes the number of within-vocabulary utterances in the training, development, and test sets.

2.2. Baseline JUPITER Recognizer

The baseline recognizer was adapted from an existing JUPITER recognizer, configured from the SUMMIT recognition system [5]. Lexical stress markings were added to the 2005-word lexicon to facilitate lexical stress modeling experiments. The initial stress labels were obtained from the LDC PRONLEX dictionary, in which each word has a vowel with primary stress and possibly a vowel with secondary stress. However, the vowels of mono-syllabic function words are likely to be unstressed or even reduced in continuous speech, such as in "*a*", "*it*", "*is*", etc. The JUPITER recognizer uses a few specific reduced vowel models as alternative pronunciations to account for them. Initially, the full vowels in mono-syllable function words were marked with primary stress, as in PRONLEX. However, too many vowels (more than 60%) in the forced transcriptions derived with this lexicon were labeled with primary stress. We thus labeled the full vowels in mono-syllabic function words as unstressed, with exceptions for a few wh-words such as "what", "when", "how", etc., because they are likely to be stressed in the JUPITER utterances. We realize that this is only a coarse approximation, because function words can be stressed in continuous speech, while stressed syllables in content words are not necessarily always stressed in spoken utterances. The following example illustrates the stress labeling of a JUPITER utterance (stressed syllables are indicated by capital letters):

WHAT is the WEAther in BOSton ?

It is unclear if secondary stress should be grouped with primary stress or be treated as no stress in terms of acoustic similarity. We decided to defer the decision until after data analysis, so primary and secondary stress were marked distinctively in the lexicon. The reduced vowels were also distinguished from unstressed full vowels in our data analysis for more detailed comparison.

Set	Sub.	Del.	Ins.	WER	SER
Development	4.3	1.6	1.7	7.6	20.2
Test	5.8	2.9	2.2	10.9	24.8

Table 2: Speech recognition error rates (in percentage) on the development data and test data. "WER" is the word error rate, which is the sum of the substitution, insertion, and deletion error rates. "SER" is the sentence error rate.

The baseline system uses only boundary class models, because it was found that adding segment models did not improve recognition performance unless context-dependent segment models were used, in which case the speed of the recognizer was significantly slower [13]. This is possibly because both models use features that are based on Mel-frequency Cepstral coefficients (MFCCs); thus the context-independent segment models are somewhat redundant when boundary models are used. Our approach is to focus on the prosodic aspects in the segment models. We hope that prosodic features can provide independent information to complement the boundary models to achieve improved recognition performance. Therefore, we did not try to retrain boundary models; the diphone labels in each boundary model class were simply expanded to cover variations in lexical stress. A bigram language model was used by the forward Viterbi search, and trigram probabilities are applied during the backward A^* search. The modified recognizer achieved the same performance as the original recognizer, which is the state-of-the-art real-time performance in JUPITER. The detailed results on the development and test data are summarized in Table 2. Various weights in the recognizer have been optimized to achieve the lowest overall error rates on the development data.

3. Lexical Stress Classification

The primary acoustic correlates of stress for English include all three prosodic attributes: energy, duration, and pitch. Stressed syllables are usually indicated by high sonorant energy, long syllable or vowel duration, and high and rising F_0 [10]. Some previous studies have also used spectral features such as subband spectral energy and MFCCs [9, 12, 14]. In this section, we study the distributions of various prosodic measurements for each lexical stress category, and determine the "best" features for stress using classification experiments. Some spectral features will also be included in the classification experiments.

Forced recognition is used to generate phonetic transcriptions (with stress marks on nucleus vowels) for the training and development data. These automatically derived stress labels will serve as the reference for both training and testing the stress models. In practice, the forced alignment process is iterated, once the stress models are trained and incorporated into the recognizer, to improve the quality of the transcriptions. The stress models can also be better trained using more "distinctive" tokens of each lexical stress category, as described in [14]. The results shown in this section are based on iterated forced transcriptions. We observed that the phone boundaries and alternative pronunciations appeared to be more accurately determined in the forced alignments after one iteration with stress models.

3.1. Prosodic Features

The energy signal used in our analysis is the root mean square (RMS) energy, which is computed by taking the square root of the total energy in the amplitude spectrum from the short time Fourier analysis of the speech. To reduce variance due to "vol-

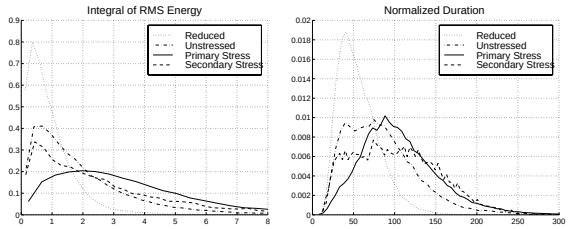

Figure 1: Distributions of energy integral (left) and normalized duration (right) features for different stress classes in the JUPITER data.

ume" differences, the raw RMS energy contour is scaled so that the average energy of each utterance in non-silence regions is roughly equal. Three energy measurements are extracted from each syllable nucleus vowel: the average, maximum, and integral of the RMS energy over the vowel duration.

The F_0 contour of each utterance is obtained using a robust pitch tracking algorithm described in [16]. Each F_0 contour is normalized by a sentence-level average to reduce variances due to speaker pitch differences. Four F_0 related measurements are included in our analysis, including the maximum, average, and slope of the F_0 contour of the nucleus vowel, and the average probability of voicing, which is available via the voicing estimation module of our pitch detection algorithm [16]. We expect the average voicing probability to be higher for stressed vowels than for unstressed and reduced vowels.

The duration is also measured for the syllable nucleus vowel. We tried to normalize the raw duration measure with a sentence-level speaking rate to reduce the variance due to different speaking rates. This is for data analysis only, because speaking rate information is usually not available during first-pass recognition. The speaking rate is estimated from the forced transcription of an utterance as follows:

$$Speaking\ Rate = \frac{\sum \mu_{Dur}(V_i)}{\sum Dur(V_i)} \qquad (1)$$

where $Dur(V_i)$ is the measured duration of the i^{th} vowel (V_i) in the sentence, and $\mu(V_i)$ is the expected duration of V_i, computed from the entire corpus.

We found that the statistics of most prosodic features differ for different lexical stress classes; however, the extent of overlap among classes is also severe. Figure 1 shows the histogram distributions of two prosodic features as examples. Generally speaking, the energy features have the best separation and F_0 features have the poorest separation in our data. Vowels with secondary stress seem to be closer to unstressed full vowels, especially by energy cues.

3.2. Classification Experiments

In addition to prosodic features, we also included the spectral tilt and MFCC features in our classification experiments, following the examples in [9, 12]. The *spectral tilt* is characterized by the average logarithmic spectral energy in four frequency bands (in Hz): [0 500], [500, 1K], [1K, 2K], and [2K, 4K]. The MFCC features include 6 MFCCs averaged over the vowel.

For each stress feature vector, a principle component analysis is first applied, and mixtures of multi-variant diagonal Gaussians are used to model the distributions. Because there seem to be some differences among all classes, and there are plenty of training data for each class, we trained models for all four lexical stress categories described in the previous section. We obtained both 4-class and 2-class classification accuracies for comparison. The 2-class results are obtained by mapping the reduced, unstressed, and secondary stress classes into one "unstressed" class. Maximum likelihood (ML) classification is used, because we are interested to know how well the features can perform without the help of *priors*.

Feature	Accuracy (4-class)	Accuracy (2-class)
(1) energy integral	**47.4**	**71.0**
(2) maximum energy	47.6	69.9
(3) average energy	45.7	70.3
(4) normalized duration	37.2	62.4
(5) raw duration	36.6	62.9
(6) log duration	41.8	61.1
(7) maximum pitch	32.8	56.2
(8) average pitch	33.1	52.9
(9) pitch slope	35.4	64.0
(10) avg. prob. voicing	43.9	62.2

Table 3: Classification accuracy (in percentage) of each individual prosodic feature on the development data.

Feature Combination	Accuracy (4-class)	Accuracy (2-class)
(1)+(5)+(9)+(10)	48.5	73.0
(1-3)+(5-10)	49.4	72.6
(11) sub-band energy (4 features)	44.0	68.3
(12) MFCCs (6 features)	51.4	73.9
(1)+(5)+(9)+(10)+(11)	52.4	74.6
(1)+(5)+(9)+(10)+(12)	**55.9**	**77.0**
(1)+(5)+(9)+(10)+(11)+(12)	55.9	76.9

Table 4: Classification accuracy (in percentage) of various combinations of features on the development data. The combinations of features are described by feature indices as defined in Table 3 and this table.

Table 3 summarizes the classification accuracy using each individual prosodic feature. As expected from the data analysis, the energy features performed the best, while the maximum and average pitch yielded the poorest results. We notice that the normalized duration did not outperform the unnormalized durations at stress classification, possibly due to intrinsic duration interferences. We will discuss this in detail in the next section.

Based on the results of individual features, we tried classification experiments using various combinations of features, including both the prosodic and the spectral measurements, as summarized in Table 4. The best set of *prosodic features* for stress classification consists of the integral of energy, raw duration, pitch slope, and the average probability of voicing. Adding spectral features improved stress classification performance, possibly because they capture the correlations between lexical stress and broad phone class. The highest accuracy was achieved by combining MFCC features with the best prosodic feature set.

4. Speech Recognition Experiments

We incorporated the four-class stress model into the first-pass Viterbi search to improve recognition performance. Only syllable nucleus vowels are scored by the lexical stress models: for segments that do not carry lexical stress, such as consonants and

System	Sub.	Del.	Ins.	WER	SER
Baseline	4.3	1.6	1.7	7.6	20.2
+ Stress	4.1	1.6	1.5	7.2	19.6

Table 5: Speech recognition error rates (in percentage) on the development data. "WER" is the word error rate, which is the sum of the substitution, insertion, and deletion error rates. "SER" is the sentence error rate.

System	Sub.	Del.	Ins.	WER	Significance
Baseline	5.8	2.9	2.2	10.9	
+ Stress	5.6	2.7	2.0	10.3	< 0.001

Table 6: Speech recognition error rates (in percentage) on the test data. The significance level between the baseline performance and the improved performance is also listed.

silences, the stress scores are simply ignored. A weight is used with each applied stress score to avoid bias toward hypothesizing fewer stressed segments. We found that this simple model improved the baseline performance on the development data. In addition, the gain using only prosodic features in the model is greater than when MFCC features are also used, even though the stress classification results implied otherwise. This is likely due to redundancy with the boundary models, in which MFCC features are already used. The optimized baseline word error rate was reduced from 7.6% to 7.2%, a 5.3% relative reduction. The details are summarized in Table 5.

We tried to refine the models by taking into account the intrinsic prosodic differences among vowels for further improvements. This is motivated by the observation that prosodic differences among phones are significant compared to stress-related differences. For example, the duration of the vowel "/ih/" (as in c**i**ty) is inherently shorter than that of "/ey/" (as in Mond**ay**), regardless of the stress properties. By grouping all vowels into a few stress categories, the intrinsic values contribute to large variances in the models, causing extensive overlap among the distributions. There are two approaches to improving the models: 1) normalizing the prosodic measurements by vowel intrinsic values, and 2) building separate models for different vowels. We experimented with the second approach, because there are plenty of training data in our corpus. One extreme is to build prosodic models for the complete inventory of vowels with different stress properties. However, the recognition performance with the new set of models (of much larger size) was virtually unchanged. We also tried less refined categories, by grouping vowels with similar intrinsic durations into classes. However, the changes to recognition results were also negligible.

Puzzled by these results, we performed an experiment in which all vowels were mapped into one class to form a single model. The recognition performance was virtually the same as using the four-class models. This seems to suggest that the gain by using prosodic models in our system is achieved mainly by eliminating implausible hypotheses, rather than by distinguishing the fine differences among various stress and segmental classes.

We applied the prosodic models on the test data and obtained similar improvements. The detailed recognition results are summarized in Table 6. The significance level of the *matched pairs segment word error test* [4] is less than 0.001. This implies that the improvements by using prosodic models, although small, are statistically significant.

5. Summary and Future Work

In this paper, we achieved small but statistically significant improvements over a state-of-the-art performance on the JUPITER domain by using simple prosodic models. In this particular task, it seems that the gain was achieved mainly by eliminating implausible hypotheses, rather than by distinguishing the fine differences among various stress and segmental classes; thus, there is no additional gain by more refined modeling. It is not clear if additional gain can be achieved by a post-processing approach, in which context-dependency and more careful normalization can be explored. We also plan to exploit prosodic features from a different angle, i.e., as indications of recognizer confidence.

6. References

[1] M. Adda-Decker and G. Adda, "Experiments on stress-dependent phone modeling for continuous speech recognition," in *Proc. ICASSP'92*, pp. 561–564, San Francisco, 1992.

[2] A. M. Aull and V. W. Zue, "Lexical stress determination and its application to large vocabulary speech recognition," in *Proc. ICASSP'85*, pp. 1549–1552, Tampa, FL, 1985.

[3] G. J. Freij and F. Fallside, "Lexical stress estimation and phonological knowledge," *Computer Speech and Language*, 4:1–15, 1990.

[4] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP'89*, pp. 532–535, Glasgow, Scotland, 1989.

[5] J. Glass, J. Chang and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. ICSLP'96*, pp. 2277–2280, Philadelphia, PA, 1996.

[6] J. L. Hieronymus, D. McKelvie, and F. McInnes, "Use of acoustic sentence level and lexical stress in HSMM speech recognition," in *Proc. ICASSP'92*, pp. 225–227, San Francisco, 1992.

[7] D. P. Huttenlocher, *Acoustic-phonetic and lexical constraints in word recognition: lexical access using partial information*, Master's thesis, MIT, 1984.

[8] K. L. Jenkin and M. S. Scordilis, "Development and comparison of three syllable stress classifiers," in *Proc. ICSLP'96*, pp. 733–736, Philadelphia, PA, 1996.

[9] M. Jones and P. C. Woodland, "Modeling syllable characteristics to improve a large vocabulary continuous speech recogniser," in *Proc. ICSLP'94*, pp. 2171–2174, Yokohama, Japan, 1994.

[10] W. A. Lea, "Prosodic aids to speech recognition," in *Trends in Speech Recognition* (W. A. Lea, ed.), pp. 166–205. Prentice-hall, 1980.

[11] K. Sjölander and J. Högberg, "Using expanded question sets in decision tree clustering for acoustic modeling," in *Proc. 1997 ASRU Workshop*, pp. 179–184, 1997.

[12] AMC. Sluijter and VJ. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *The Journal of the Acoustic Society of America*, 100(4):2471–2485, 1996.

[13] N. Ström, L. Hetherington, T. J. Hazen, E. Sandness, and J. Glass. "Acoustic modeling improvements in a segment-based speech recognizer," in *Proc. 1999 ASRU Workshop*, Keystone, CO, 1999.

[14] D. van Kuijk and L. Boves, "Acoustic characteristics of lexical stress in continuous telephone speech," *Speech Communication*, 27(2):95–111, 1999.

[15] A. Waibel, *Prosody and Speech Recognition*, Pitman, London, 1988.

[16] C. Wang and S. Seneff, "Robust pitch tracking for prosodic modeling of telephone speech," in *Proc. ICASSP'00*, pp. 1143–1146, Istanbul, Turkey, 2000.

[17] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Trans. on Speech and Audio Processing*, 8(1):100–112, 2000.