LEXICAL STRESS CLASSIFICATION FOR LANGUAGE LEARNING USING SPECTRAL AND SEGMENTAL FEATURES

Luciana Ferrer, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, Kristin Precoda*

Speech Technology and Research Laboratory, SRI International, California, USA

ABSTRACT

We present a system for detecting lexical stress in English words spoken by English learners. The system uses both spectral and segmental features to detect three levels of stress for each syllable in a word. The segmental features are computed on the vowels and include normalized energy, pitch, spectral tilt and duration measurements. The spectral features are computed at the frame level and are modeled by one Gaussian Mixture Model (GMM) for each stress class. These GMMs are used to obtain segmental posteriors, which are then appended to the segmental features to obtain a final set of GMMs. The segmental GMMs are used to obtain posteriors for each stress class. The system was tested on English speech from native English-speaking children and from Japanese-speaking children with variable levels of English proficiency. Our algorithm results in an error rate of approximately 13% on native data and 20% on Japanese non-native data.

Index Terms— Stress classification; Computer-aided language learning; Gaussian Mixture Models

1. INTRODUCTION

Lexical stress is an important component of English pronunciation. To understand spoken words, native speakers of English rely not only on the pronunciation of sounds, but also on the stress patterns. Using an incorrect stress pattern can greatly reduce a speaker's intelligibility. Appropriately using stress patterns poses a big challenge for English learners, especially for the native speakers of languages that have more consistent lexical stress patterns or have different ways of incorporating timing and rhythm. This difficulty is especially true for Japanese speakers learning English: in Japanese the rhythm is more regular than in English, and the syllables are more similar in prominence. English language learners can then benefit from a system that provides automatic feedback on their stress patterns.

Several automatic stress detection systems have been proposed in the literature. Most of these systems are based on pitch, energy and duration features, extracted over the syllable nucleus and normalized in different ways to make the features independent of the speaker's baseline pitch, the channel volume, the speech rate, and so on. Examples of these kinds of segmental features can be found in [1, 2, 3, 4, 5]. Spectral features, on the other hand, have been rarely used for stress detection. Chaolei [6] and Lai [7] both propose the use of spectral features, but they test their systems only on native English data. It is reasonable to assume that, given the phonetic pronunciation mistakes made by language learners, spectral features would fail to carry robust stress information for such speakers. The modeling techniques used for stress detection vary widely and include decision trees [3]; GMMs [1, 4]; support vector machines [3, 2]; and hidden Markov models [7, 6, 8].

We propose a system for stress detection for language learners that uses features based on duration, pitch, energy, spectral tilt, and spectral measurements, successfully integrated at a deep level using GMMs. These GMMs are trained using a large amount of data from native English speakers. This approach is convenient because manual annotation is not needed for this data. Instead, a dictionary of stress pronunciation is used to determine the label for each syllable.

In many cases, the task of stress detection is defined as the problem of locating the single primary stressed syllable in a word [2, 4, 1]. In our work, we do not assume that exactly one syllable in every word has primary stress, because English learners will not necessarily adhere to this rule. In fact, according to our phonetician's annotations, in our Japanese children database, approximately one third of the incorrectly stressed words have primary stress on at least two syllables. For this reason, our system makes decisions at the syllable level without enforcing acceptable English stress patterns at the word level.

The proposed system results on error rates of approximately 13% on native data and 20% on Japanese non-native data. Our results show that native English data can be successfully used to learn stress models for English learners with only a 10% relative degradation with respect to a system that takes advantage of matched non-native data. Finally, we show that spectral features can be successfully used for stress detection even when models are learned on native speakers.

2. SYSTEM DESCRIPTION

This section describes our word-independent stress detection system designed to predict three levels of stress (unstressed, primary and secondary stress) for each syllable in a word.

2.1. Features

Features are extracted over the vowel of each syllable. Five types of segmental features are defined based on duration, pitch, energy, spectral tilt and MFCCs. All features go through some type of normalization to make the features as independent as possible of the characteristics that might confound the classification of stress, such as the channel, the speech rate, the base-

*Luciana Ferrer is currently with CONICET, Argentina, and Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

line pitch of the speaker, and so on. We perform all normalizations at the word level. This way, syllable-level features are all relative to the mean values found in the word.

Phone-level Alignments: To locate the vowels within the waveform we run EduSpeak [9, 10], SRI International's automatic speech recognizer (ASR) and pronunciation-scoring toolkit for language learning applications. EduSpeak uses a standard Gaussian mixture model hidden Markov model (GMM-HMM) speech recognizer. For this experiment, recognition is run in forced alignment mode, where the output is constrained to the words in the transcription, using a single forward pass. A 39-dimensional acoustic speech feature is used, consisting of energy and 12 Mel-frequency cepstral coefficients (MFCCs), plus their deltas and double deltas. The cepstrum is normalized using CMS (cepstral mean subtraction) with the normalization coefficients computed over the entire sentence. The models are trained using data from native English-speaking children. For recognition of Japanese data, the models are adapted to a small amount of data from Japanese children speaking English.

Log of Normalized Duration: The duration of the vowel in the syllable is first normalized by dividing it by the mean vowel duration for all syllables of the same type. The syllable type is given by concatenating two sub types: (1) the *next consonant type*, given by whether the following consonant is unvoiced, voiced, or there is not following consonant (either another vowel follows, or the vowel is the last one in the word); and (2) the *pause type*, given by whether the word is followed by a pause longer than 0.1 seconds or not and, if it is, whether the syllable is the last one in the word or not. The duration normalized by syllable type is further normalized by speech rate by dividing it by the mean of the syllable type-normalized duration for all the vowels within the same word. Finally, the logarithm of the final normalized value is computed.

Polynomial Coefficients of Pitch, Energy, and Spectral Tilt: Pitch, energy, and spectral tilt measurements are extracted every 10 milliseconds over the full waveform. Pitch is approximated by the fundamental frequency (F0), and energy is approximated by the mean RMS value (Eg). F0 is estimated using the algorithm described in [11]. The spectral tilt (ST) values are computed as the slope of the FFT, extracted over a window of 20ms that is shifted every 10ms. Below, F0 and Eg refer to the log of the corresponding raw signals, while ST is not transformed. The exact same processing is done for the F0, Eg and ST signals, as follows. First the F0, Eg, and ST values that correspond to unvoiced frames, as indicated by a missing F0 value, are considered undefined. Undefined values are ignored during the computation of the polynomial approximation. Second, for each word, the mean of these signals over the frames with a defined value corresponding to the vowels is subtracted from the signals. Finally, for each vowel in each word, the Legendre polynomial approximation of order 1 is computed for the three signals resulting in two coefficients for each signal. For details on the Legendre polynomial computation, see [12].

MFCC log Posteriors: MFCCs extracted during speech recognition are also used as features to predict stress. MFCCs over the vowels are modeled at the frame level using one GMM for each stress class. These GMMs are obtained by adaptation to a single GMM trained using samples from all stress classes in the same way as for segmental features (see Section 2.2). Given a test utterance, the likelihood of each of these three GMMs is computed for each frame over each vowel. The geometric mean of the likelihoods over all frames in a vowel is computed for each stress class, resulting in three vowel-level likelihoods, one for each stress class. These likelihoods are transformed into posteriors using Bayes rule, assuming equal priors for the stress classes. Finally, the log of the posteriors for stress classes 0 and 1 are used as segment-level features. The posterior for class 2 is redundant given the other two and, hence, it is discarded.

2.2. Gaussian Mixture Modeling

The five types of segmental features are concatenated into a single feature vector per vowel of size 9: two polynomial coefficients each for pitch, energy, and spectral tilt; plus log normalized duration; plus two log MFCC posteriors. These feature vectors are then modeled with one GMM for each stress class. This modeling is done in two steps. First, a single model for all stress classes is trained. Then, the model is adapted to the data from each stress class. This procedure enables training robust models even for the secondary stress class, for which very little data is available compared to the other two stress classes. The adaptation is done using a maximum a posteriori (MAP) approach commonly used for speaker recognition [13]. This method introduces a regularization parameter, the relevance factor, that controls how much the global means, weights, and covariances, should be adapted to the data from each class.

Given a new utterance, we compute the likelihood of the GMM for each of the three stress classes for each vowel. The likelihoods are converted into posteriors using Bayes rule and a set of priors. These priors should be computed from data as similar to the test data as possible.

3. NATIVE AND NON-NATIVE DATASETS

Experiments were run on a dataset of Japanese children reading English phrases. A set of 959 multisyllable words was selected from this dataset to be labeled by three annotators for stress level. These words came from 668 randomly chosen phrases from 168 distinct speakers from both genders. The chosen speakers were those with the larger numbers of stress pronunciation errors as judged by an initial, quick annotation of the data from the full set of 198 speakers in which stress pronunciation quality was judged at the word level as either correct or incorrect. The multisyllable words from the remaining 30 "better" speakers were used to compute the syllable-type statistics employed to normalize vowel duration for this data.

The annotators were instructed to label each syllable in each selected word from the 168 chosen speakers with a label of "unstressed" (0); "primary stressed" (1); or "secondary stressed" (2). The annotators were allowed to label more than one syllable with primary or secondary stress. The average disagreement between annotators was 21%. The words for which the number of pronounced syllables did not correspond to the number of syllables in the canonical pronunciation according to at least one annotator were discarded. This resulted in 848 words that were labeled by the three annotators, which corresponded to 1776 syllables (most words were disyllabic words). The results reported in this paper were computed on the set of syllables for which all three annotators agreed on the same

stress label. The selection was done at the syllable level in order to preserve as much data as possible. The resulting data contained 1240 syllables; 22% unstressed, 67% primary stressed, and 11% secondary stressed.

A separate dataset of native English-speaking children was used for training the models. The data consisted of read speech from 366 children with a total of 41,022 phrases. The multisyllable words for which a single stress pronunciation is listed in our lexical stress dictionary were selected. The canonical stress found in the dictionary was then assigned as the label for each of these words. We assumed that native speakers pronounce stress as listed in the dictionary in the vast majority of cases for these words. This assumption enabled using a large amount of data for training the models without the need for manual annotations. This database contained 74,206 words with a total of 157,888 syllables; 48.3% unstressed, 47.2% primary stressed, and 4.5% secondary stressed. The syllable type statistics used to normalize vowel durations for the native data were computed on the native data itself.

4. SYSTEM PERFORMANCE

We show results for different systems and subsets of features. The results in this section were obtained with 256-component GMMs for MFCC modeling and 64-component GMMs for segmental feature modeling. These GMMs are trained on the native English data described above. Although larger GMMs give an improvement on native data, no significant gains (or degradations) are observed on non-native data when increasing model size. Since the focus of this work is on non-native performance, we choose to keep the models small for run-time expedience.

We report error rates that are computed as the number of samples in which the detected label disagrees with the annotated label divided by the total number of samples. Error rate is computed on hard decisions. To go from GMM posteriors to decisions, we choose the stress class for which the posterior is highest. For the most important comparisons of two systems, we report the p-value obtained with the McNemar matched-pairs significance test.

4.1. System Comparison

Table 4.1 shows the results for four different systems for the task of classifying stress into three levels: 0 (unstressed), 1 (primary stress) or 2 (secondary stress). We call this task "$0|1|2$". We also show results for the task of classifying syllables as having non-primary stress (0 or 2) or having primary stress (1) for a subset of the systems. We call this task "$02|1$". In this case, the posterior for the 02 class is computed as the sum of the posteriors for classes 0 and 2 output by the system.

The systems shown in the table use priors computed either on native or non-native data for converting GMM likelihood into posteriors (nat-p versus nn-p). For the first case we show a comparison when class-specific GMMs are trained separately or with the proposed adaptation technique. The two nn-p systems use adaptation to obtain the class-specific GMMs. In all cases, when class adaptation is performed, a relevance factor of 0 is used. This value was tuned on native data. Finally, the fourth system also adapts the class-specific models learned on native data to the non-native data in a second step of adaptation. This adaptation is done with the MAP approach described in Section 2.2. In this case, though, given the small amount of non-native data available, only means and weights are adapted.

The table shows the error rate results for native and non-native speakers. For the last two systems, we only show the results on non-native data because these systems are only meant to optimize performance on that data. For the native results, we performed 10-fold cross-validation, training the system in nine folds, then testing it on the held-out fold, and finally collecting posteriors from all folds to compute the shown performance. The priors used for posterior computation were computed on the nine folds used for training and then applied on the test fold. When using non-native priors and when performing the adaptation to non-native data, the same 10-fold cross-validation approach was used. The relevance factor used for the adaptation to non-native data, on the other hand, is selected to optimize the performance on the full set, which means these results are slightly optimistic. The optimal relevance factor was 80, which is a large value compared to the one used in speaker recognition (usually 16), and it results in very little adaptation of the parameters. For non-native data, the table shows the results on two additional subsets of data consisting of only the words that were labeled as correctly or incorrectly stressed by the three annotators. The numbers of words labeled as correct and incorrect were 220 and 191, respectively.

We can see that the adaptation technique for learning class-dependent models gives significant reductions in error rate of around 10% relative on both native and non-native data. Using priors learned on non-native data gives significant gains on this data. Finally, adapting the segmental GMM parameters to the non-native data gives only marginal and not statistically-significant improvements. Note that the last two systems require some amount of labeled non-native data, while the first two systems only use native data for model creation.

We also see that the performance on non-native speakers is significantly worse than on native speakers. This degradation comes from the incorrectly stressed words, because the performance for correctly stressed words when using native priors is comparable to that obtained for native speakers. This suggests that the system has more difficulty classifying words with incorrect stress patterns. This can be due to both issues with the ASR alignments (although even correctly stressed words might be misaligned due to phonetic rather than stress mispronunciations) and to the fact that incorrectly pronounced stress might be labeled as such because it was pronounced in a non-native manner with an unusual combination of segmental or spectral patterns. These patterns would not have been seen under any stress class in the native data. This result could suggest that using non-native data for training or adaptation should give a performance improvement for these words. Nevertheless, the table shows that adapting models to non-native data does not bring much improvement for these words. We believe that the lack of a significant gain from adaptation to non-native data is due both to the small amount of available non-native data and to the high disagreement between annotators, which results in too much noise in the samples used for adaptation.

Note that a system that simply picks the majority class (unstressed for natives and primary stress for non-natives) would

Task	System Setup	Native	Non-Native		
			all	cor	inc
0|1|2	nat-p sep-trn	13.8	24.3	14.4	44.9
	nat-p	13.3***	22.9*	11.8	46.4
	nn-p	-	20.3**	18.6	35.7
	nn-p adapt	-	20.2	19.0	35.7
02|1	nat-p	11.3	16.2	8.5	29.2
	nn-p adapt		14.6*	12.9	24.9

Table 1. Error rates for native and non-native children data on two tasks, 0|1|2 and 02|1, for four different systems: **nat-p sep-trn**, where native priors are used for posterior computation and class-dependent GMMs are trained independently; **nat-p**, where native priors are used and class-dependent GMMs are obtained through adaptation to a class-independent model; **nn-p**, a system identical to nat-p where non-native priors are used for posterior computation instead of native priors; and **nn-p adapt**, a system identical to nn-p where an additional step of adaptation to non-native data is done on the class-dependent GMMs. For non-natives we show the results on the full set of words; on a subset of words that was labeled as correctly stressed (cor); and on a subset of words that was labeled as incorrectly stressed (inc). For the first two columns, we show the significance level between the system corresponding to the line and the one in the previous line within the same block. Symbols *, **, and *** indicates a p-value smaller than 0.05, 0.01, and 0.001, respectively. No symbol indicates a p-value larger than 0.05.

result in an error rate of 51.7% for natives and 33% for non-natives. The error rates achieved by our system are significantly better than those of that naive system. Further, as expected, results on the simpler and more standard task of classifying stressed versus unstressed syllables (for which the naive system results are 47.2% and 33% for natives and non-natives, respectively) are significantly better than those on the three-way classification task, especially for non-native speakers.

4.2. Feature Selection Results

The proposed system uses five types of features based on pitch, energy, spectral tilt, duration, and MFCC information. Figure 1 shows results for the individual features and for systems modeling four feature types, leaving one type out at a time. The system setup is kept identical to the one used for the nat-p system in Table 4.1.

For natives, we see that leaving any feature type out degrades performance, with the 5-feature result being 13.26% which is better than the best 4-feature result of 13.57% at a significance level smaller than 0.001. From this, we can conclude that all five feature types are needed to achieve the best performance. Given the leave-one-out results, one could order the features in terms of their importance for stress classification on native speakers as follows: (1) MFCCs, (2) duration, (3) energy, (4) pitch, and (5) spectral tilt.

For non-native data, the order of importance of feature types is slightly different. In this case, the best single feature type is energy, followed by MFCCs, duration, pitch and, finally, spectral tilt. In fact, discarding pitch and spectral tilt gives a slight (not statistically significant) improvement in performance. However, the performance loss when discarding energy, and MFCC feature types is statistically significant, with p-values smaller than 0.001, and 0.01, respectively.

We chose to present feature selection results using native priors because we believe that this approach gives a more direct assessment of the usefulness of the feature itself. The non-native priors bias all systems toward detecting more primary stressed syllables washing out differences across feature types. Despite this, using non-native priors for the non-native data still results in energy being the single best feature, now followed by duration and then MFCCs, with all three of them giving gains over the all-feature result.

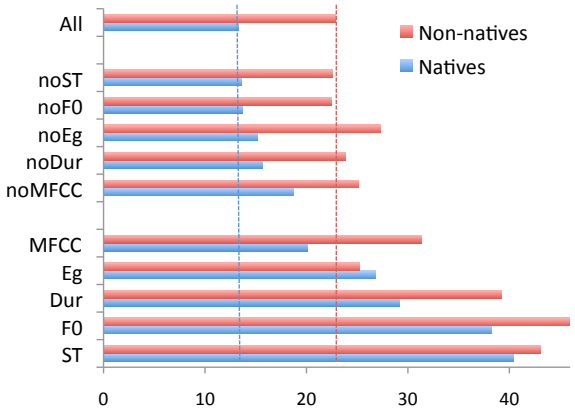

Fig. 1. Native and non-native error rates for different feature combinations. For each block, the systems are sorted based on their native performance. The two vertical lines indicates the all-feature performance for natives and non-natives.

5. CONCLUSIONS

We propose a system for lexical stress classification at the syllable level that uses both segmental and spectral features. We combine the two types of information at a deep level in the system by converting the spectral information into segment-level posteriors. These posteriors are concatenated with the pitch, energy, spectral tilt and duration features at the segment level. The resulting vector is modeled using one GMM for each stress class. We show that the most useful feature types for stress classification for both natives and non-natives are MFCCs, energy, and duration.

The proposed systems provide more detailed and general information than most systems in the literature, enabling multiple stressed syllables in a word and giving syllable-level feedback with three levels of stress. Finally, one of our proposed systems, nat-p, does not require matched labeled data from speakers with the same L1, making it both cheap to train (requiring only native English data for which labels can be automatically derived) and portable to any population of English learners. We show that this system performs only around 10% worse in terms of error rate relative to a system that takes advantage of matched non-native data for prior computation and model adaptation.

6. ACKNOWLEDGMENT

We wish to thank the phoneticians, Alan Mishler, Rebecca Hanson and Timothy Arbisi-Kelm for their hard work in labeling the Japanese data.

7. REFERENCES

[1] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proc. ICASSP*, Philadelphia, Mar. 2005.

[2] J.Y. Chen and L. Wang, "Automatic lexical stress detection for Chinese learners' of English," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, 2010.

[3] O. D. Deshmukh and A. Verma, "Nucleus-level clustering for word-independent syllable stress classification," *Speech Communication*, vol. 51, no. 12, 2009.

[4] L.-Y. Chen and J.-S. Jang, "Stress detection of English words for a CAPT system using word-length dependent GMM-based Bayesian classifiers," *Interdisciplinary Information Sciences*, vol. 18, no. 2, pp. 65–70, 2012.

[5] A. Verma, K.l Lal, Y. Y. Lo, and J. Basak, "Word independent model for syllable stress evaluation," in *Proc. ICASSP*, Toulouse, May 2006.

[6] C. Li, J. Liu, and S. Xia, "English sentence stress detection system based on HMM framework," *Applied Mathematics and Computation*, vol. 185, no. 2, 2007.

[7] M. Lai, Y. Chen, M. Chu, Y. Zhao, and F. Hu, "A hierarchical approach to automatic stress detection in English sentences," in *Proc. ICASSP*, Toulouse, May 2006.

[8] S. Ananthakrishnan and S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *Proc. ICASSP*, Philadelphia, Mar. 2005.

[9] Horacio Franco, Victor Abrash, Kristin Precoda, Harry Bratt, Ramana Rao, John Butzberger, Romain Rossier, and Federico Cesari, "The SRI EduSpeak^{TM}system: Recognition and pronunciation scoring for language learning," *Proceedings of InSTILL 2000*, 2000.

[10] H. Franco, H. Bratt, R. Rossier, V. R. Gadde, E. Shriberg, V. Abrash, and K. Precoda, "EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, pp. 401–418, July 2010.

[11] D. Talkin, *Robust Algorithm for Pitch Tracking*, Elsevier Science, 1995.

[12] Chi-Yueh Lin and Hsiao-Chuan Wang, "Language identification using pitch contour information," in *Proc. ICASSP*, Philadelphia, Mar. 2005, vol. 1, pp. 601–604.

[13] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.