A HIERARCHICAL APPROACH TO AUTOMATIC STRESS DETECTION IN ENGLISH SENTENCES

Min Lai[2] Yining Chen[1] Min Chu[1] Yong Zhao[1] Fangyu Hu[2]

[1]Microsoft Research Asia, Beijing, China
[2]Department of Electronic Engineering & Information Science, University of Science & Technology of China, Hefei, China
[2]mlai@mail.ustc.edu.cn, [1]{ynchen, minchu, yzhao}@microsoft.com, [2] hufy@ustc.edu.cn

ABSTRACT

This paper proposes a hierarchical framework, which consists of three layers of classifiers, for automatic stress detection in English speech utterances. The top two layers are a linguistic classifier, which assigns stressed labels to all content words and unstressed labels to all functions words, and an acoustic classifier, which assigns stressed and unstressed labels with HMM based models and using only acoustic features such as MFCC, energy and f_0. When there is no manual stressed label available, only the top two layers are activated. The best performance we achieved is 92.9%. The third layer in the framework is an AdaBoost classifier that can improve the accuracy by using more features and manual labels. The best result we obtained is 94.1%, which is approaching to the self-agreement ratio (97.4%) of the same annotator, or the upper bound of the performance.

1. INTRODUCTION

Labeling prosodic events in speech database is important for both speech analysis and speech synthesis. Among many prosodic events, stress is one of the most important one. "What many phoneticians and linguists have called stress, and what most laymen readily understand under this term, refers to nothing more than the fact that in a succession of spoken syllables or words some will be perceived as more salient or prominent than others" [1]. It is money and labor consuming to label stressed syllables manually, especially when the speech database is very large. An efficient and reliable automatic prosody labeler is always desired.

Higher intensity, greater duration and higher fundamental frequency are believed to be the primary acoustic cues for stressed syllables, although how the three factors work together to make a syllable more prominent than the surrounding ones is still not very clear. Therefore, they are used as the main acoustic features in the stress detect task in some studies [2, 3]. Stress is found to be correlated with voice quality as well. Usually, stressed vowels are pronounced more clear and unstressed vowels tend to be reduced. Hence spectral parameters such as Mel-Scale Frequency Cepstral Coefficients (MFCC) are used in some stress detection studies [4]. Both [4] and [5] model the acoustic features of stressed/unstressed vowels or syllables with Hidden Markov Model (HMM) and achieve reasonable results.

When listening to an utterance, people not only use acoustic cues but also syntactic and/or semantic cues to help the location of stresses. Therefore, features that are derived from texts, such as part of speech (POS), N-Grams of POS and the position in phrase are used in stress detections as well [5-7]. [4] uses Bayesian decision and [6] uses ANN to combine the results from text level cues with those from acoustic cues. All these methods need some manually labeled training data for the task.

The word accuracy of these stress prediction algorithm is normally between 80-90%. Since the corpora and the methodology of stress labeling are different, it is difficult to compare these results with each other. Most corpora used for the stress detect task are speaker independent [2-7]. ToBI (Tone and Break Index) are used in some corpora and 3-4 levels of stress are labeled in others [5, 8].

In this paper, a hierarchical approach is proposed to detect stress with spectral features, prosodic features and linguistic features. The approach can reach reasonable results without any manually labeled data and will get pretty high accuracy when small amount of manually labeled training data are provided.

In Section 2, the framework of this hierarchical approach and its main modules are introduced. Evaluations and results are given in Section 3 and conclusions are given in Section 4.

2. A THREE-LAYER CLASSIFIER FOR AUTO STRESS DETECTION

According to Pike [1], content words which carry the major semantic weight of the sentence are usually stressed and function words which have less contribution to semantic content are normally unstressed. When investigating the distribution of stresses in our speech corpus, in which all words are labeled as stressed/unstressed manually by listening, we found that more than 15% of function words are stressed and only 3-4% of content words are unstressed.

Therefore, we propose a hierarchical approach shown as in Figure 1. It has 3 layers. First, all words in an utterance will be classified into two categories, content word and function word, by the linguistic classifier. Layer 2 is the acoustic layer; classifiers are utilized individually to function and content words. In third layer, label refinement modules are employed on two of the four branches outputted by the layer 2. One branch is the acoustic stressed vowels in function words and the other is the acoustic unstressed vowels in content words. The label refinement modules are trained by the manually labeled data. The three classifiers are described in details in the subsections below.

In the figure, S and U mean the stressed and unstressed decision of each classifier. And 1 and 2 in two label refinements module mean they use different models. Similarly, the models in the two acoustic classifiers are the same.

Without manual labels, the hierarchical framework is shrunk to which in the shadow in Figure 1. An acoustic classifier is used to detect stressed vowels in function words and all content words are labeled as stressed. The acoustic classifier can be trained from the speech corpus without manually labeled stresses.

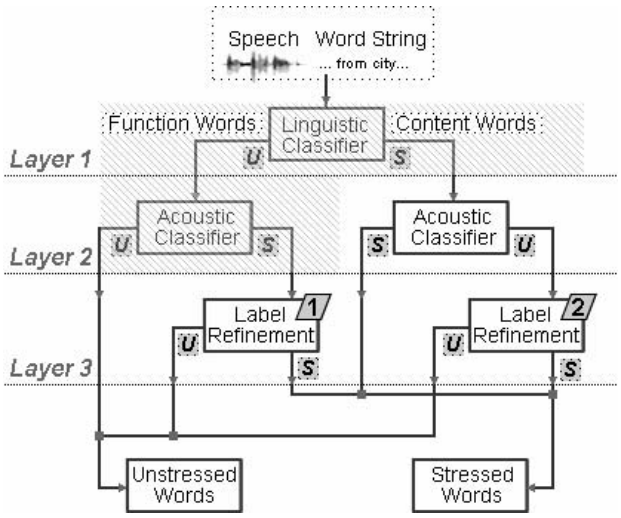

Figure 1: Block diagram of the hierarchical approach for automatic stress detection.

2.1. The linguistic classifier

The linguistic classifier is very straight forward, i.e. POS tagging and classifying words into content/function category by their POS.

2.2. The acoustic classifier

2.2.1. Training stressed/unstressed HMMs for each vowel

In conventional speech recognition task, about 40 phones are defined for English. Stressed vowel and unstressed vowel share the same model. In our stress detection task, stressed vowel and unstressed vowel are treated as two phones, i.e. 56 phones are defined. And, in order to reflect the syllable structure in words, consonants at the onset position of syllables are distinguished from consonants at the coda position, and then 78 phones are defined.

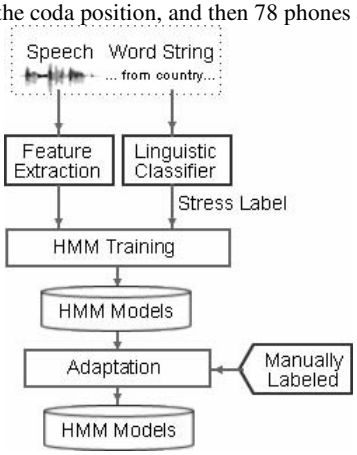

Figure 2: Flowchart of training HMMs for stressed/unstressed vowels.

The flowchart of training stressed/unstressed HMMs is given in Figure 2. In previous studies, manually labeled stress-marks are always required to train stressed/unstressed models. In our approach, the speech corpus is labeled automatically by the linguistic classifier, i.e. the lexical primary stress of content words will be marked as stressed syllable and all other syllables are unstressed. Then, the training process is similar to that in the speech recognition task. When small amount of manually labeled data are available, these data can be used to do adaptation to make the models more accurate. Acoustic features utilized in the model training include fundamental frequency, energy and spectral parameters.

2.2.2. Stress labeling

The stress labeling process is in fact a decoding process within a finite state network shown in Figure 3. The vowels in the syllables that carry the lexical stresses have two nodes, the S (stands for stressed) node and U (stands for unstressed) node and other vowels have only U node. Each consonant has only one node, either O node (stand for consonant at the onset position) or C node (stand for consonant at coda position). For words with multiple syllables, parallel paths will be provided; each has at most one S node (as the word "city" shown in Figure 3). Words aligned with stress vowel are labeled as stressed and other words are labeled as unstressed.

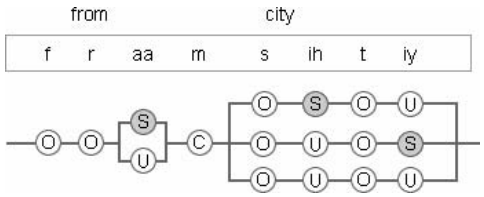

Figure 3: Finite State Network in Decoding.

2.3. The refinement module

AdaBoost [9] is often used to adjust the decision boundaries to reduce false decisions and has achieved good performance in many fields [10]. The advantage of AdaBoost is that it can boost a sequence of weak classifiers, where the weights of each classifier are updated dynamically according to the errors in previous learning.

When we have manually labeled stress marks, AdaBoost is used to reduce the acoustic classification error with the many features besides the output of HMM. In each step of AdaBoost, a one dimensional weak classifier is utilized with one feature whose accuracy is the best.

Three types of features are used. The first type is the likelihood score from the stressed and unstressed model of a given vowel. The second is prosodic features that can not be modeled with HMM directly, such as the duration of the vowels and differences in pitch level of current vowel and its neighbor vowels. The third type is linguistic features beyond POS, such as uni-gram, bi-gram and tri-gram score of a given word because frequently used words tend to be reduced [11].

3. EVALUATION AND RESULTS

3.1. Speech Corpus

Our speech corpus contains 6,412 utterances, recorded by a professional female speaker in American English. Stress marks have been labeled by a well trained annotator in the first 3,000 utterances. The instruction for the annotator is to label the prominent words in the utterances by listening to the waveforms. 1,000 utterances are labeled twice by the same annotator in a time

span of 3 months. The agreement ratio between the two labels is 97.4%. This is the upper bound for auto-labeling.

The first 1000 utterances in the corpus are kept as testing set. The left utterances are used for training.

When performing boosting in label refinement, the testing set is split into two equal parts for developing and testing alternately. Cross validate is done and the average error rate is the final output.

3.2. Accuracy of the linguistic classifier

If all content words are labeled as stressed and function words as unstressed, the agreement ratio between linguistic classifier and human annotator is 91.3% within testing set.

3.3. Accuracy of the acoustic classifier

The accuracy of the acoustic classifier is first optimized without integrating into the hierarchical framework and models are trained without manual stress marks.

3.3.1. Model configuration

In previous studies, since only small amount of manual labels are available, stressed/unstressed models are normally phone independent. [5] uses position dependent stress models. In our approach, since no manual labels are required, the whole training set is used. Three types of models are compared. First, mono-phone models (using the 56 phone definition) are trained and 172 states are obtained. Then mono-phone models are split into tri-phone models (2580 states are obtained). Although tri-phone model can model the context of a vowel, the position of a vowel in syllable may also affect. Hence, each consonant is split into onset and coda categories, which result position based tri-phones (based on the 78 phone definition). The total number of states is 2634. Stress label accuracies with the three types of models are given in Figure 4, from which it is seen that the position based tri-phone performs the best, i.e. the separation of onset consonants from coda consonants does help the detection of stressed syllables. The position based tri-phone models are used in the remaining experiments.

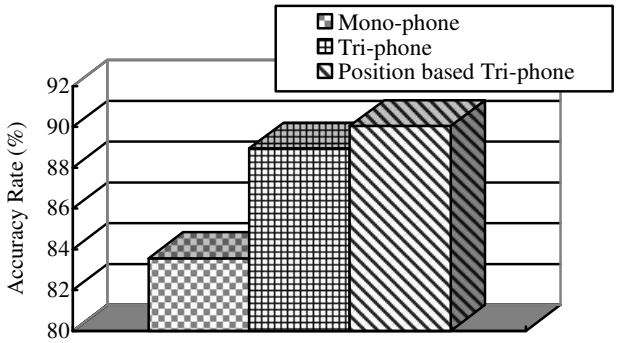

Figure 4: Stress prediction accuracy of different model definition.

3.3.2. Feature selection

HMM models are trained with various combinations of spectral parameters (12 dimension MFCC, their first and second order derivatives), energy (time domain log energy, its first and second order derivatives) and f0 (in log scale and smoothed). The error rates of these models in testing set are compared in Figure 5.

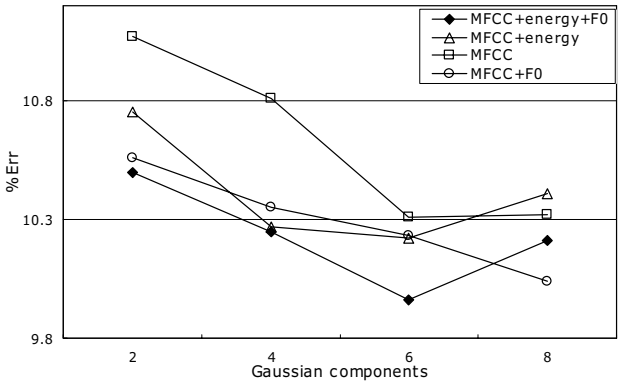

Figure 5: Error rate of different feature sets.

Figure 5 shows that the best performance is achieved by using all features. The feature used in Figure 4 is also MFCC+Energy+F0.

The best error rate of the acoustic classifier is 10.0% when it is used alone, which is worse than the result of using the linguistic classifier alone.

3.4. Accuracy of the hierarchical approach (without manual labels)

Although the accuracy with acoustic classifier is about 10% worse than which of linguistic classifier, the error rates in function category and content category are quit lopsided. In function word category, the error rate of linguistic classifier is 15.7% while that of the acoustic classifier is 11.7%, and in content word category the error rate of acoustic classifier is 8.9% and that of the linguistic classifier is only 3.7%. Hence only adopting acoustic classifier is in function category will improve the accuracy of linguistic classifier.

When the two classifiers are integrated into the hierarchical framework like the shadow part in Figure 1, the overall accuracy is increased to 92.9%. It is worth to note that this accuracy is obtained without any manual labels in model training.

3.5. Accuracy of the hierarchical approach (with manual labels)

When some manual labels are available, they can be used to refine the HMM models and to train the AdaBoosting model.

3.5.1. Accuracy of acoustic models

When manual stress labels are available, they are used in model adaptation to improve the model accuracy. Error rates obtained with different amount of manual labels, tested in a stand alone mode, are shown in Figure 6. It is seen that only when enough manual labels, saying 1000 utterances or above, the error rate will be reduced a little. Adaptation barely works here, so we canceled it in our later experiments.

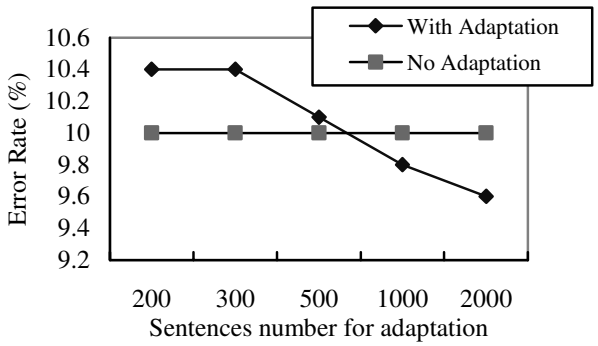

Figure 6: Error rate of different size of the manual labels.

3.5.2. Boosting the results with more features

In function category, a boosting classifier is trained to reduce the mismatch between stressed output of acoustic classifier and manual labels, and in content category, another classifier is trained to reduce the mismatch between unstressed output of acoustic classifier and manual labels. Duration, relative pitch and uni-gram are used as the features besides the differences in likelihood scores by the stressed and unstressed models. The classifiers are trained in the developing set and tested in the testing set. The final accuracy increases to 94.1%.

3.6. Final result

The final results are shown in Figure 7. In that figure, LO means linguistic only. It can be treated as the base line of our system. SA is the self agreement ratio of the human annotator. It can be treated as the upper bound of our system. HNM means hierarchical model without manual labels. The error rate of it decreased 18.4% from which of linguistic only. HM means hierarchical model with manual labels of 500 sentences. The error rate of it decreased 32.2% from which of linguistic only.

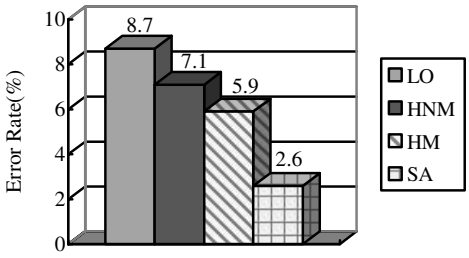

Figure 7: Final results.

4. CONCLUSION

This paper proposed a three layer hierarchical framework for automatic stress detection in English. The first layer is a simple linguistic classifier which separate function/content words, the second layer is a HMM-based acoustic classifier and the third layer contains two AdaBoost-based classifiers.

The manually labeled data are not adopted in training acoustic model. Without manual stressed marks, the hierarchical framework is also worked well by simplifying into 2 layers. Acoustic results are only used in the function category and all the content words are treated as the stressed one.

With the manual labels, we adopted it to refine the output of acoustic classifier in special condition, the stressed output part in function category and unstressed output in content category part. The score of HMM model together with other acoustic and linguistic features are utilized to train a classifier with AdaBoost.

When no manual stress labels available, only the top two layers are activated and the best accuracy we got is 92.9%. When 500 sentences manual labels are provided, the third layer will be activated. The best accuracy we got is 94.1%, which is pretty good compared with the upper bound of 97.4%, the self agreement ratio of the human annotator.

In future work, prosody boundary will be predicted together with stress. Since prosody boundary will affect the acoustic feature like pitch and duration, predicting them together may be better than predicting them individually.

This work will be implemented into other languages like Spanish and French in future.

5. ACKNOWLEDGEMENT

The authors would like to thank Scott Meredith for his great helps on creating the specification of prosody annotation. We have special thanks to Yaya Peng for creating these stress labels.

6. REFERENCES

[1] E. C. Kuhlen, “An Introduction to English Prosody”, Edward Arnold, 1986.

[2] C. W. Wightman and M. Ostendorf, “Automatic labeling of prosodic patterns,” IEEE Trans. on Speech and Audio Processing, 2(4), pp. 469-481, 1994.

[3] I. Bulyko and M. Ostendorf. “A Bootstrapping Approach to Automating Prosodic Annotation for Constrained Domain Synthesis,” in Proc. of the IEEE Workshop on Speech Synthesis, pp 115-118, 2002.

[4] A. Conkie, G. Riccardi, and R.C. Rose “Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events” in Proc. of EUROSPEECH, pp 523-526, 1999.

[5] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji, “Modeling and automatic detection of English sentences stress for computer-assisted English prosody learning system”, in Proc. of ICSLP, pp 749-752, 2002.

[6] K. Chen, and M. Hasegawa-Johnson, “An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model,” in Proc. of ICASSP, pp 509-512, 2004.

[7] S. Arnfield, “Prosody and syntax in corpus based analysis of spoken English,” Ph.D. dissertation, University of Leeds, Dec. 1994.

[8] P.C. Bagshaw. “Criteria for labelling prosodic aspects of English speech,” In Proc. 4th. Australian International Conference on Speech Science and Technology, 1992.

[9] Y. Freund and R.E. Schapire, “A decision-theoretic generalization of online learning and an application to boosting,” J. Comp. & Sys. Sci 55(1), pp 119-139, 1997.

[10] D. Wang, L. Lu, H.J. Zhang. "Speech Segmentation without Speech Recognition,” in Proc. of ICASSP 2003, pp 468-471, 2003.

[11] S. Werner, etc, “Toward Spontaneous Speech Synthesis—— Utilizing Language Model Information in TTS”, IEEE Transactions on speech And Audio Processing, 12(4), pp 436-444, 2004.