Automatic Lexical Stress Detection Using Acoustic Features for Computer-Assisted Language Learning

Junhong Zhao*, Hua Yuan† , Jia Liu† and ShanHong Xia*
* State Key Laboratory on Transducing Technology, Institute of Electronics,
Chinese Academy of Sciences, Beijing
E-mail: zhaojunhong09@mails.gucas.ac.cn, shxia@mail.ie.ac.cn
† Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing
E-mail: yuanh08@mails.tsinghua.edu.cn, liuj@tsinghua.edu.cn

***Abstract*—This paper proposes an English lexical stress detection approach using acoustic features. The approach classifies the vowels of English words into two patterns: primary stress and unstress. We firstly choose the frame-averaged basic feature set of the individual syllable nucleus in polysyllabic words as the baseline to decide the stress pattern. This feature set includes the semitone, the duration, the loudness and the emphasis feature. Furthermore, we introduce the pitch-variation feature set and the context-aware feature set to describe the inside variation characteristic and outside contextual characteristic of the syllable nucleus. By combining the three feature sets, the accuracy rate is improved by $7\%-8\%$. Besides, we train support vector machines (SVMs) classifier for each vowel phoneme respectively. The results show that the phoneme-dependent models performance better than only one shared model. Finally, our system achieved an accuracy of 88.6% compared with human-tagged labels.**

I. Introduction

In recent years, Computer-Assisted Language Learning (CALL) has been paid a lot of attention because of its helpfulness for second language (L2) learners. The stress detection technology in CALL is developed to help the learners on the stress problem. In the stress-timed languages such as English, the stress plays one of the most important roles. Different syllabic stress position may express different parts (e.g., *import*, *increase*) or different meanings (e.g., *conduct*, *desert*) of the words. Besides, misplacing the stress may make the words sound like nothing sometimes. So it's very important to pronounce lexical stress properly.

In the past few years, many lexical stress detection methods have been explored. In these studies, many basic acoustic features are used, such as the duration, the pitch, the energy, MFCC and so on[1][8]. Besides these basic features, many other acoustic features are explored. Reference [1] use the slope- and range-related statistical features to represent the change characteristics of the syllable coarsely. In [2], the syllable contextual information is exploited and is viewed as the features of the word where the syllable located. They extract the differential features among adjacent syllables and combine them together to be one word-level feature. Moreover, various acoustic models, such as Bayes[2], SVM[2][8], have been used. However, most of these studies neither consider the stress-related features completely nor consider the difference of the feature distribution between vowel phonemes. So in this work, we investigate the complementarity between different features and combine them effectively. At the same time, the acoustic model is refined in order to deal with the different feature distribution between different vowel phonemes.

In this paper, we extract the frame-averaged basic features as the baseline of the system to detect lexical stress. Then in order to make up for the deficiency of the averaged basic features, both the semitone variation features and the differential features are extracted to represent the property of the vowels. The semitone variation features are from TILT parameter set, while the differential features are on the basis of the averaged basic features. After that, the phoneme-dependent SVM models are trained to classify the vowels as stressed or unstressed. At last, the final stress pattern results are compared to the human-tagged annotation and the prompt is given out to the learners.

The rest of the paper is organized as follows: Section 2 describes the extraction of the acoustic features that were used in our system and Section 3 specify the detection procedure, including the classifier and post-processing. Section 4 states the experiments that have been implemented and the result they achieved. Finally, we draw the conclusions of our work.

II. The Extraction of Acoustic Features

Orthographically, the syllable is the basic unit of the lexical stress. According to the linguistic rules, each syllable is comprised of only one vowel and one or more consonant(s). But as described in [1], when the speech rate and rhythmic flow of pronunciation are different, the partitioning results of the syllables don't always keep the same even though using the same syllable parser. So this paper chooses the syllable nucleus, which is the essential vowel center of a syllable, as the basic unit to extract features and discriminate stress

This work is supported by the NSFC and RGC scheme (Project 60931160443), National High Technology Research and Development Program of China (Project 2008AA040201), and National Natural Science Foundation of China (Project 90920302).

pattern. The boundary of the syllable nucleus is obtained by the alignment of the speech and the transcriptions.

A. The Frame-Averaged Basic Features

In contrast to the unstressed syllables, the stressed ones often have longer duration, higher intensity and higher F0. In our work, we choose the following four features as the baseline:

Duration

It has been proved in [9] that the syllable duration and the syllable nucleus duration are almost the same in classification performance. So we use the boundary information that is produced by the forced alignments to obtain the vowel's duration. The duration is normalized by the mean duration over all the syllable nuclei in the word which is called ROS(Rate-Of-Speech)[10] standard technique.

Loudness

As one of the prosody event, the stress has close relationship with human's auditory characteristic. To better match with the human perception scale, we substitute the energy for the loudness here. Firstly, we calculate the energy of each frame (8 ms). Then the energy is passed through a set of triangle filters. These filters have uniform distribution in Mel frequency scale. The output energy of each filter is calculated by (1), the final energy of each frame is calculated by (2):

$$M(i) = \ln[\sum_{n=F_{i\text{-}1}}^{F_i} \frac{n - F_{i\text{-}1}}{F_i - F_{i\text{-}1}} E(n) + \sum_{n=F_i}^{F_{i+1}} \frac{F_{i+1} - n}{F_{i+1} - F_i} E(n)], (i{=}1,2,...,K), \tag{1}$$

$$L = \sum_{i=1}^{K} 0.048M(i)^{0.6}, \tag{2}$$

where $M(i)$ is the log energy that comes from the i_{th} triangle filter, $E(n)$ is the energy value of the n_{th} frequency point that is calculated by FFT. F_i is the i_{th} central frequency. F_{i+1} and $F_{i\text{-}1}$ are the upper cut-off frequency and the low cut-off frequency respectively. K is the number of the triangle filters, the value of which is 26 in our work. L is the sum of loudness for one frame.

Semitone

Based on the same point of view with loudness, we choose the semitone rather than the F0 to better approximate the human's perception[5]. The transformation formula between semitone and frequency is (3).

$$S = 69 + 12 \log_2(\frac{f}{440}), \tag{3}$$

where S is the semitone and f is the frequency.

Spectral emphasis

The previous study [7] has shown that the mid-frequency energy is more powerful in the stress classification than the energy covered all frequencies. The mid-frequency is referred to the frequency around 500-2000 Hz. Here we use a FIR filter (Kaiser window) to obtain the energy within this bandwidth as the spectral emphasis feature.

Moreover, the loudness L, the semitone S and the spectral emphasis E are all first extracted in the unit of frame, and then are averaged over all the frames of the vowels in order to reduce the negative impact caused by different speakers and speech rate following (4):

$$X = \frac{\sum_{n=i}^{j} x_n}{j\text{-}i{+}1}, \tag{4}$$

where X can be one of the L, S and E, i and j are the number of the begin frame and the final frame of the vowel respectively. x_n is the feature value of each frame. All the four features above are called "the averaged-basic feature" here.

B. The Pitch-Variation Features

In [1], the study points out that the pitch-related features are effective in detecting the stress position, especially in the case of English learning that the learner's mother language is tonal language such as Mandarin. Because in this kind of language the pitch dictates not only word meaning but also syllabic stress, and the mother-tongue's negative transfer in the cross-cultural learning is unavoidable. Our work follows the rise/fall/connection (RFC) model proposed by Taylor[3] and use the TILT parameter[4] set as features to describe the shape of pitch contour.

The RFC model is one of the intonation models. It tries to label the F0 contour as R (rise), F (fall) and C (connection). Instead of the F0 contour, we use the semitone contour of each syllable nuclei to extract the parameters. It's more reasonable since semitone is more suitable for the human's auditory perception. Firstly, the linear interpolation is implemented to smooth away the outlier and perturbations of each frame (0.064 s).Then the frame is marked with one of the three kinds of shapes according to its slope. If the marks keep the same in sequential frames, we will merge them together. After the marking process, the amplitude, duration and tilt are measured in these rises and falls. According to [4], the transformations that produce these parameters are as follows:

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|}, \tag{5}$$

$$tilt_{dur} = \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}}, \tag{6}$$

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{2(|A_{rise}| + |A_{fall}|)} + \frac{D_{rise} - D_{fall}}{2(D_{rise} + D_{fall})}, \tag{7}$$

where A_{rise} and A_{fall} are the sum of magnitude and D_{rise} and D_{fall} are the sum of duration of the rises and falls. F0 position and F0 height are calculated directly. In our work, we choose the amplitude($tilt_{amp}$),duration($tilt_{dur}$),tilt($tilt$) and F0 height($maxf0$) as our features, and called these four features as "the pitch-variation features". These features are all extracted using the Edinburgh Speech Tools Library(EST) tools.

C. The Context-Aware Features

In this paper, stress syllable is referred to “the ones which are perceived as standing out from their environment” [1]. So the lexical stress judgment of the syllables can’t be independent of the contextual in the word. In our work, we consider the contextual information as one of the vowel’s features. We extract the differential value on the basis of the frame-average basic features mentioned above as:

$$\triangle BX_p = X_p - X_{p\text{-}1}, \tag{8}$$

$$\triangle AX_p = X_p - X_{p+1}, \tag{9}$$

where X can be one of the duration D, the semitone S, the loudness L and the spectral emphasis S. p is the index of the vowel number in the word. BX_p represents the p_{th} vowel differential value with its preceding one. AX_p represents the p_{th} vowel differential value with its subsequent one. If the vowel is the first or the last one in the word, it will always lack one neighbor. In this case the relative value to the mean value over all the syllable nuclei in the word will be calculated as the differential value.

III. Lexical Stress Detection

A. The Acoustic Classifier

Our method is implemented on the premise that the stress of syllable has two patterns: primary stress and unstress, not taking account of secondary stress. So in our work, the stress detection is a two-class problem. We choose the support vector machines (SVMs) to train the model. The SVMs is widely used for its excellent ability of learning. This learning algorithm is based on the statistical learning theory. It maps the input vectors to a high-dimensional separable space using various kernel functions and tries to find a hyperplane that has the maximum margin between the support vectors in this space as a discriminant boundary. Here the LIBSVM[6] is used to train the phoneme-dependent models for the vowels, and the Radial Basis Function (RBF) kernel is chose. Because the probabilistic information is needed in the post-processing, we set the probability estimation mode by making the parameter $b{=}1$. Besides, the feature vectors are all scaled to $[-1, 1]$ interval before the training and testing procedure. Then at last, the output of SVMs is stress classification results and probability pairs (the probability of the stressed pattern, and the probability of the unstressed pattern) for each vowel.

B. Post-Processing

By Linguistics definition, there is only one primary stress in each word. Following [1], regarding the word as the basic unit, we reassign the stress pattern after the detection procedure using the probability estimated by the LIBSVMs like this:

1) If more than one syllable was detected as stressed pattern in the word, we search among the probability of the stressed pattern for all the syllables in the current word, and only choose the syllable with the maximum probability as the stressed one.

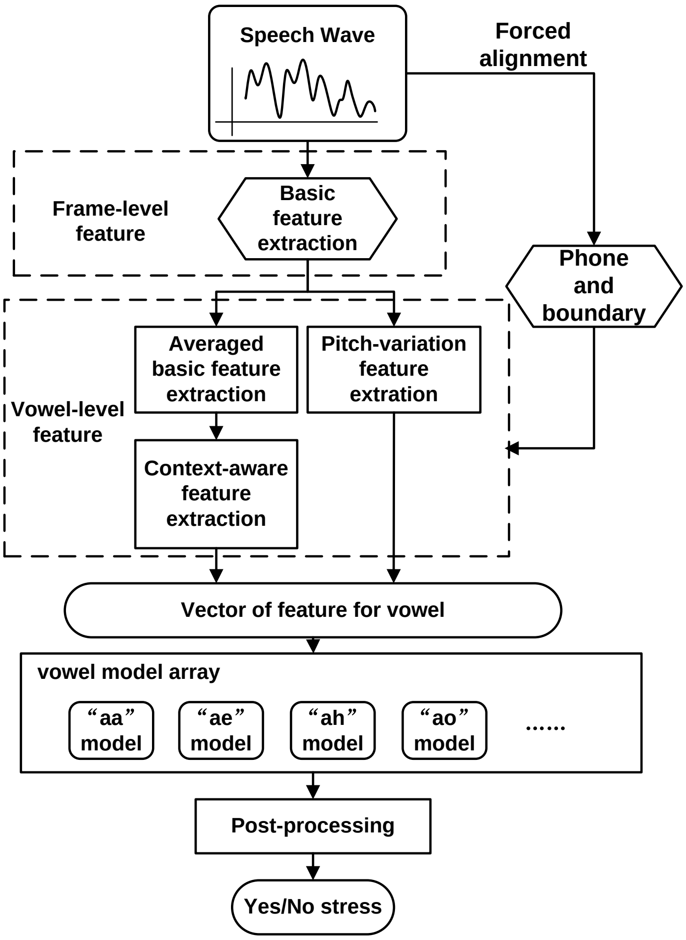

Fig. 1. Flow of the lexical stress detection system

2) If no syllable was detected as stressed pattern in the word, we search among the probability of the unstressed pattern for all the syllables in the current word, and just choose the syllable with the minimum probability to be stressed.

C. Framework of The Proposed System

Fig. 1 describes the framework of our system. All the phoneme segmentation results are obtained by Automatic Speech Recognition (ASR) with the learners’ speech and the canonical transcriptions. Here we don’t take the phoneme pronunciation errors into consideration and just focus on the stress errors. We use the CMU phonemes set, which includes fifteen vowels altogether. So we train fifteen models for each vowel phoneme.

IV. Experiments and Result

A. Database and Evaluation

In the experiment, we use the MIR-SD (Multimedia Information Retrieval lab, Stress Detection) database[11]. This database is designed specialized for lexical stress detection of multi-syllable English words. It is recorded by 22 Taiwanese speakers with middling English level. Each speaker records about 200 utterances. Each utterance contains only one multi-syllable word, which is selected from the English spelling contest for university students in Taiwan. There are 3668 words in all, 3000 words of which are chosen to train models, and the remaining 668 words to test. The distribution of vowels is listed in Table I. To evaluate the performance of our system, we employ the detection accuracy, which is the ratio of the correctly detected syllables to the sum of the syllables in the whole test database.

TABLE I
THE DISTRIBUTION OF THE DATABASE

	$TrainNumber$	$TestNumber$
$TotalWord$	3000	668
$TotalSyllable$	10274	2230
$StressSyllable$	2970	668
$UnstressSyllable$	7304	1562

B. Features Performance Analysis

In order to verify the distinguish ability of each feature mentioned above, we firstly train a single classifier for all the vowel phonemes, ignoring the difference between them. The detection results with the single SVMs model using only one feature are shown as Fig. 2. It can be seen that all the features are effective, especially the frame-averaged loudness, maxf0 and all the contextual-aware features. It's also proved in some respects that the contextual information is indispensable as a property of individual vowels on stress. But in general, each individual feature can't achieve a satisfactory result due to the limited ability. We should consider how to combine them together effectively.

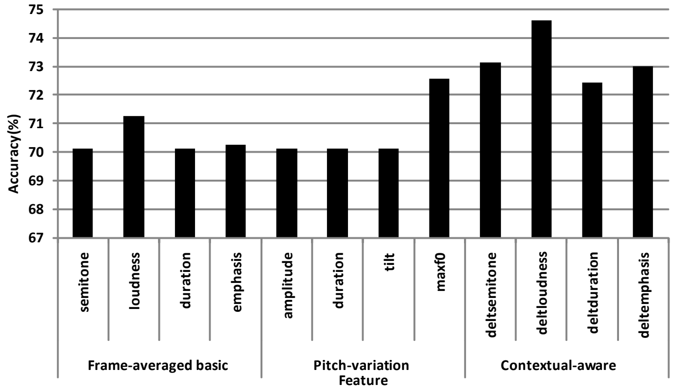

Fig. 2. The lexical stress detection accuracy of different features with single classifier

To investigate the combination characteristics of these features, we use the frame-averaged basic feature set, the pitch-variation feature set and the contextual-aware feature set as three units and combine them arbitrarily. The result is as Table II.

TABLE II
THE PERFORMANCE OF FEATURE COMBINATION

$Feature$	$Accuracy(\%)$
$Basic$[a]	78.30
$TILT$[b]	80.22
$Delt$[c]	81.35
$Basic + TILT$	83.95
$Basic + Delt$	81.44
$TILT + Delt$	84.75
$Basic + TILT + Delt$	85.47

[a] *Basic* represent the frame-averaged basic feature set .
[b] *TILT* represent the pitch-variation feature set.
[c] *Delt* represent the contextual-aware feature set.

From this table, we can see that both the pitch-variation features and the contextual-aware features performance better than the frame-averaged basic features. And these three ones are all pairwise complemented. In detail, as the contextual-aware features are extracted on the basis of the frame-averaged features and the information between them crosses highly, the complementarity of this pair is least remarkably. On the contrary, there is strongest complementarity between the contextual-aware features and the pitch-variation features. The reason is that the pitch-variation features reflect the variation characteristic inside the vowel while the contextual-aware features reflect variation characteristic outside the vowel. The two information are independent and indispensable. As a result they can perform well as a whole. At the same time, both of them can make up for the deficiency of the frame-averaged basic features since the latter ones can reflect the change information neither inside vowels nor outside. So finally, we combined all the three feature sets and achieved a promising result.

C. Detection with Phoneme-Dependent Classifiers and Post-Processing

All the experiments above are implemented with single classifier. But by investigating the detection performance of this single classifier for each vowel phoneme individually, we find that its detection performance varies significantly for different vowel phonemes. The reason for this is that the distributions of different vowel phonemes vary greatly. For example, the features of some stressed syllable may approximate to the features of another unstressed syllable. Since our task is to detect the stress pattern of vowel phoneme rather than identifying the vowel phoneme itself, we train the classifier for each vowel phoneme, replacing the single classifier for all the vowel phonemes. Furthermore, the post-processing is used to improve the performance. The result is showed in Fig. 3. By using the phoneme-dependent classifiers, the performance is improved from 85.47% to 86.46% in the case of using the three feature units. And the accuracy is enhanced to 88.57% by adding the post-processing.

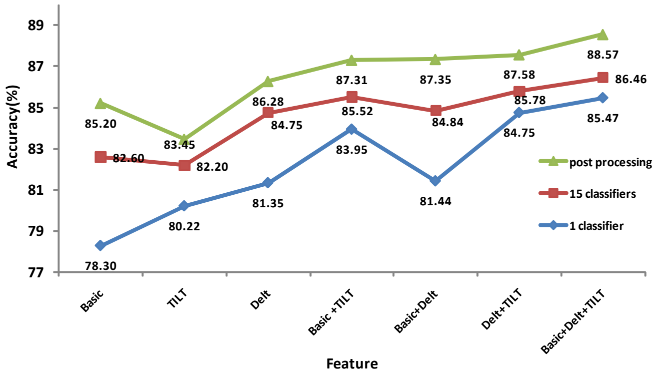

Fig. 3. Comparison of the performance with two different classifier composition

V. Conclusions

This paper presents a lexical stress detection approach. Firstly, the complementarities between different features are investigated and the results show that combining the contextual-aware features with pitch-variation features and even the frame-averaged basic features is quite necessary because they represent the different aspects of the stress information respectively. These three kinds of feature sets are indispensable and should be considered as a whole when detecting the stress. Then we use phoneme-dependent classifiers instead of single one. The detection accuracy is improved obviously by this method. In addition, with the post-processing, our system achieves the accuracy of 88.57% finally.

References

[1] Joseph Tepperman and Shrikanth S. Narayanan, “Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners,”in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA,*, pp. 937-940, Mar 2005.

[2] Jin-Yu Chen and Lan Wang, “Automatic lexical stress detection for Chinese learners’ of English,”in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on,* pp. 407-411, 29 2010-dec. 3 2010.

[3] Paul Taylor, “The rise/fall/connection model of intonation,”in *Speech Commun.,* vol. 15, pp. 169-186, October 1994.

[4] Paul Taylor, “The tilt intonation model, ”in *Proc. ICSLP 98,* 1998, pp. 1383-1386.

[5] Sieb Nooteboom, “The prosody of speech: Melody and rhythm, ”in *The Handbook of Phonetic Sciences, Nr. 5 in Blackwell Handbooks in Linguistics, chap,* 1997, pp. 640-673.

[6] Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin, “Working Set Selection Using Second Order Information for Training Support Vector Machines,”in *J. Mach. Learn. Res.,* vol. 6, pp. 1889-1918, December 2005.

[7] A. M. C. Sluijter, and V. J. van Heuven, “Acoustic correlates of linguistic stress and accent in Dutch and American English,”in *Fourth International Conference on Spoken Language, 1996. ICSLP 96. Proceedings.,* vol. 2, pp. 630-633, Oct 1996.

[8] Jhing-Fa Wang, Gung-Ming Chang, Jia-Ching Wang and Shun-Chieh Lin, “Stress Detection Based on Multi-class Probabilistic Support Vector Machines for Accented English Speech,”in *Computer Science and Information Engineering, 2009 WRI World Congress on,* vol. 7, pp. 346-350, 31 2009-April 2 2009.

[9] Fabio Tamburini, “Automatic Prosodic Prominence Detection in Speech Using Acoustic Features: an Unsupervised System,”in*Proceedings of Eurospeech 2003,* 2003, pp. 129-132.

[10] L. Neumeyer, H. Franco, M. Weintraub and P. Price, “Automatic text-independent pronunciation scoring of foreign language student speech,”in*Fourth International Conference on Spoken Language, 1996. ICSLP 96. Proceedings.,* vol. 3, pp. 1457-1460, Oct 1996.

[11] Chen Liang-Yu and Jyh-Shing Roger Jang, “Automatic pronunciation scoring using learning to rank and DP-based score segmentation,”in*INTERSPEECH-2010,* 2010, pp. 761-764.