Prominence Model for Prosodic Features in Automatic Lexical Stress and Pitch Accent Detection

Kun Li, Shuang Zhang, Mingxing Li, Wai-Kit Lo and Helen Meng

Human-Computer Communications Laboratory, The Chinese University of Hong Kong, Hong Kong

{kli, zhangs, mxli, wklo, hmmeng}@se.cuhk.edu.hk

10.21437/Interspeech.2011-528

Abstract

A prominence model is proposed for enhancing prosodic features in automatic lexical stress and pitch accent detection. We make use of a loudness model and incorporate differential pitch values to improve conventional features. Experiments show that these new prosodic features can improve the detection of lexical stress and pitch accent by about 6%. We further employ a prominence model to take into account of effects from neighboring syllables. For pitch accent detection, we achieve a further performance improvement from 80.61% to 83.30%. For lexical stress detection, we achieve performance improvements in (i) classification of primary, secondary and unstressed syllables (from 76.92% to 78.64%), as well as (ii) determining the presence or absence of primary stress (from 86.99% to 89.80%).

Index Terms: stress, pitch accent, loudness model, prominence model

1. Introduction

Suprasegmental phonology plays an important role in the perceived proficiency of the second language (L2) spoken by a learner [1]. Our previous study [2] has identified several aspects in suprasegmental phonology that deserve attention from a Chinese learner of English, such as lexical stress, narrow focus, reduction / non-reduction of function words, intonation of a sentence, as well as prosodic disambiguation. This paper focuses on the detection of *lexical stress in a word* and *pitch accents in an Intonation Phrase* (IP).

Lexical stress is associated with the prominent syllable of a word. Faithful production of lexical stress is important for the perceived proficiency of L2 English. In some cases, it also serves to disambiguate lexical terms by proper placement of *primary* stress, e.g., “'insert” vs. “in'sert”. Pitch accent is associated with the prominent syllable within an IP, which usually carries important information and needs attention from the listeners.

Previous research has presented various features and approaches on the automatic detection of stress and pitch accent. In the study of syllable stress detection for German and Italian, Tepperman [3] used the mean values of fundamental frequency (F0), syllable nucleus duration, energy and other features related to F0 slope and RMS energy range. Imoto [4] developed Hidden Markov Models (HMMs) to detect stress in English sentences read by Japanese students. Based on a time-delay recursive neural network, Ren [5] developed a gender-dependent pitch accent detector. Sun [6] used ensemble learning methods (bagging and boosting) to predict pitch accents. Tamburini [7] combined the detection of lexical stress and pitch accents into a task of prominence detection. Pitch accent detection was based on F0 movements and overall syllable energy, while stress detection was based on syllable nucleus duration and high-frequency features.

In this work, we adopt a loudness model and a new differential pitch value for lexical stress and pitch accent detection in a speaker-independent and text-independent scenario. To further improve detection performance, we also propose a prominence model that incorporates effects of neighboring syllables.

2. Loudness model

Loudness is the human perception of the strength of sound energy. In speech production, we produce stressed or accented syllables with higher energy. Yet, there is a complex relationship between human perception of loudness and sound energy. We follow Zwicker’s loudness model for a precise estimation of loudness [8, 9].

2.1. Critical-band level

The critical-band rate scale models the human hearing mechanism, which analyzes broad spectrum of sound signals in a non-linear scale corresponding to the critical bands [8, 9]. The critical-band rate scale, as measured in Bark, is related to the frequency scale in Hertz (Hz), as given by Eq. (1) [8]:

$$z = 13\tan^{-1}(0.00076f) + 3.5\tan^{-1}(\frac{f}{7500})^2 \qquad (1)$$

where f is the linear frequency in Hz, and z is the critical-band rate in Bark. The human audible range is usually divided into 24 bands. The critical-band intensity I_G and the critical-band level L_G can be obtained by Eq. (2) and (3) respectively [8]:

$$I_G(z) = \int_{z-0.5}^{z+0.5} \frac{dI}{dz} dz \qquad (2)$$

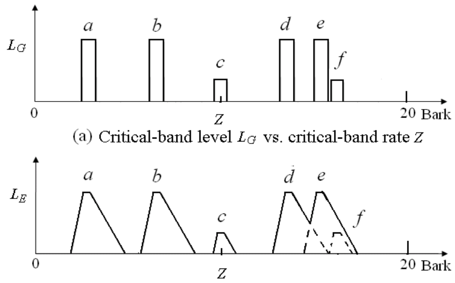
$$L_G(z) = 10\log\frac{I_G(z)}{I_0} dB \qquad (3)$$

where I is the intensity of sound signals, and $I_0 = 10^{-12}$ W/m^2 (standard threshold of hearing at 1kHz).

2.2. Excitation

A single tone excites a range of human hearing elements along the critical-band rate scale (see Fig. 1). Excitation provides an approximation to the frequency selectivity by taking into consideration that the auditory response to the sound intensity levels at a particular frequency, as well as those in the vicinity [8, 9].

For example, upon hearing the single tone signal a (see Fig. 1), a range of hearing elements in the vicinity of the centre frequency of a will be excited. These are shown in the triangular shape, which corresponds to the shape of the masked thresholds [8]. When multiple signals are presented in the sound, they will excite their corresponding elements individually, e.g., a, b and c, if widely separated. In case they are closed in frequency, their excitations to the human hearing system will be combined, e.g., d and e has partial spectral masking, while e and f has total spectral masking. When the signals a to f are presented simultaneously, the overall perceived intensity is given by the total area under the solid line envelope of Fig. 1(b). Following that, the critical-band level L_G can be transformed to the excitation level L_E, i.e. intensity I can be transformed to the excitation E.



(a) Critical-band level L_G vs. critical-band rate Z

(b) Excitation level L_E vs. critical-band rate Z

Figure 1: *(a) Positions and intensity levels of the sound components. (b) Each component excites a range of the human hearing elements in addition to the centre frequency of the component, where the slopes correspond to the shape of the masked thresholds [8]. Excitations from multiple components also affect one another.*

2.3. Specific loudness and loudness summation

Based on the Stevens' power law [10] (i.e., equal intensity ratios yield equal loudness ratios), we compute the specific loudness based on the perceived loudness per Bark, as shown in Eq. (4), which uses excitation instead of intensity [8, 9].

$$N'(z) = 0.08(\frac{E_{TQ}(z)}{E_0})^{0.23}[(0.5+0.5\frac{E(z)}{E_{TQ}(z)})^{0.23}-1] \quad (4)$$

where $N'(z)$ is the specific loudness in the z^{th} Bark, $E(z)$ is the excitation value in the z^{th} Bark, $E_{TQ}(z)$ is the threshold of excitation in the z^{th} Bark in quiet, and E_0 is the excitation corresponding to the reference intensity I_0 in Eq. (3).

The total loudness N, as measured in sone, is given by the summation of the specific loudness $N'(z)$ over the 24 critical bands in Bark scale, as given in Eq. (5).

$$N = \int_{z=0}^{z=24} N'(z)dz \quad (5)$$

3. Syllable-based prosodic features

Stressed or accented syllables usually exhibit greater loudness, longer duration and higher pitch [7]. We incorporate these by adopting three prosodic features: syllable nucleus duration, maximum syllable loudness, and differential pitch value in a syllable. Following [3, 7, 11], we extract the prosodic features from the syllable nuclei rather than from the whole syllable units.

3.1. Syllable nucleus duration

We identify the syllable nuclei duration by first applying the Maximal Onset Principle [12] to extract the syllables from the phoneme sequence output of the Automatic Speech Recognizer (ASR). For example, the word "introduction" is divided into /ih n/, /t r ax/, /d ah k/ and /sh ax n/. The syllable boundaries may be ambiguous for some words, but the syllable nuclei are still correct.

Using Eq. (5), each frame's loudness can be obtained. Within the time boundaries of every extracted syllable, we treat the frames whose loudness fall above N_{bottom} as the syllable nuclei, where N_{bottom} is the value that for all loudness values larger than 1.5 sone, 80% of them are larger than it. The normalized syllable nucleus duration V_D, as given by Eq. (6), is taken as our feature.

$$V_D = d_D - \overline{d_{IP}} \quad (6)$$

where d_D is the syllable nucleus duration, $\overline{d_{IP}}$ is the mean duration of all syllable nuclei in the IP. V_D, d_D, and $\overline{d_{IP}}$ are measured in frame (10 ms).

3.2. Maximum syllable loudness

Each frame's loudness can be obtained from Eq. (5). The normalized maximum syllable loudness V_N, as given by Eq. (7), is taken as our feature.

$$V_N = N_{\max} - \overline{N_{IP}} \quad (7)$$

where N_{max} is the maximum loudness within the identified syllable, and $\overline{N_{IP}}$ is the mean loudness over all syllables in the IP.

3.3. Differential pitch value in a syllable

To derive the differential pitch value, we first perform pitch detection (using the Snack Sound Toolkit [13]) and process pitch values that fall within the time boundaries of the identified syllable nuclei. We also convert the pitch value to the semitone scale, a logarithm scale that better matches human perception of pitch.

$$f_{st} = 12\log_2(f_0 / f_{bottom}), \quad f_0 > 0 \quad (8)$$

where f_0 is the fundamental frequency in Hz, f_{bottom} is a normalization factor in Hz set by a "90% criterion" (i.e. the value above which 90% of all pitch values in the IP fall), and f_{st} is the pitch value in semitone scale.

Previous work [14] shows that the maximum pitch value in a syllable offers the best discrimination power for prominence detection, compared to the minimum, mean, range, etc. of a syllable. Besides, syllables with rising tones often give an accented perception. Syllables with falling tones, especially whose preceding syllable is accented with rising tone, are often perceived as unaccented. Combining these observations, we propose to use the derived value in Eq. (9) as one of our detector features.

$$V_P = f_{st2} + (f_{st2} - f_{st1}) = 2f_{st2} - f_{st1} \quad (9)$$

where f_{st1} is the first maximum / minimum pitch value (in semitone scale) in the syllable, and f_{st2} is the second maximum / minimum pitch value in the syllable.

We further refine this empirical equation by optimizing a_{p1} in Eq. (10) based on detection accuracies (see Fig. 2).

$$V_P = 2f_{st2} - a_{p1}f_{st1} \quad (10)$$

where f_{st1} and f_{st2} are the same as Eq. (9). We found that the optimal value for a_{p1} is around 0.95 and applied in our experiments.

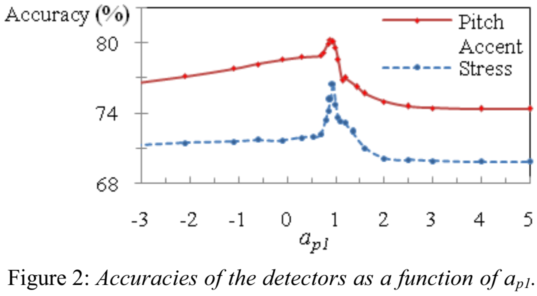


Figure 2: *Accuracies of the detectors as a function of a_{p1}. The solid line is the accuracy for pitch accent detection; the dashed line is the accuracy for stress detection.*

4. Prominence model

Stressed or accented syllables are more prominent than their neighboring syllables [7]. Syllables with loudness, duration, and pitch greater than their neighboring syllables are likely to be stressed or accented, even if their values are not large on average. This means that the differences between the feature values and their neighboring syllables are also important. Taking these effects into consideration, we devise a prominence model that aims to estimate prominence, based on the selected prosodic features from a syllable.

To incorporate the differential values, initially we apply Eq. (11) to all selected prosodic features.

$$P(i)=V(i)+\sum_{k=\pm1,\pm2} a(k)\Delta V_i(k) \qquad (11)$$

where $V(i)$ is the value of the selected prosodic feature (i.e. V_D, V_N or V_P) of the i^{th} syllable, $\Delta V_i(k)=V(i)-V(i+k)$ and $a(k)$ are the factors to be determined. The relationship among $V(i)$,$V(i+k)$ and $\Delta V_i(k)$ are illustrated in Fig. 3.

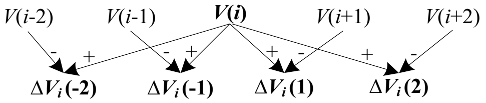


Figure 3: *The relationships among V(i),V(i+k) and ΔV_i(k).*

As the prominence value of the i^{th} syllable is also dependent on its separation in time from the neighboring syllables, the prominence model is improved as in Eq. (12).

$$P(i)=V(i)+\sum_{k=\pm1,\pm2} a(k)\frac{\Delta V_i(k)+b(k)}{\Delta t_i(k)/\overline{\Delta t}+c(k)} \qquad (12)$$

where $a(k)$, $b(k)$, and $c(k)$ are factors for the prominence model to be determined, $\Delta t_i(k)=|t(i)-t(i+k)|$, $t(i)$ is the time of the feature (for V_P, $t(i)$ is the time of f_{st2}; for V_D, $t(i)$ is the center of the syllable nucleus), $\overline{\Delta t}=\frac{t(N_c)-t(1)}{N_c-1}$, N_c is the number of syllables in the IP.

Furthermore, we believe that a longer silence before the syllable is helpful to the perception of the prominence. Hence we introduce the last term $d\frac{\Delta t_i(-1)}{\overline{\Delta t}}$ and arrive at Eq. (13):

$$P(i)=V(i)+\sum_{k=\pm1,\pm2} a(k)\frac{\Delta V_i(k)+b(k)}{\Delta t_i(k)/\overline{\Delta t}+c(k)}+d\frac{\Delta t_i(-1)}{\overline{\Delta t}} \qquad (13)$$

Currently, we set d=2. As a syllable's prominence value may be reduced by its preceding syllable, especially when their feature values are almost the same, we set $b(-1)=-2$ and $b(1)=2$ as a compensation for $\Delta V_i(-1)$ and $\Delta V_i(1)$ respectively. As $V(i+k)$ with larger values of $|k|$ has less effect on $V(i)$, we set $b(-2)=-1$, $b(2)=1$, and $c(k)=0.5|k|$ as a compensation for $\Delta t_i(k)/\overline{\Delta t}$. As $V(i+k)$ with negative values of k has greater effect on $V(i)$ than those with positive value of k, we set $a(k)$ as $a(k)=\begin{cases}1, & k<0\\ 0.8, & k>0\end{cases}$.

Finally, by applying the prominence model to the selected prosodic features (V_N, V_D and V_P), we have the corresponding prominence features: P_N , P_D and P_P.

5. The Classifiers

We built a stress detector and a pitch accent detector separately, both of which are two-category Gaussian mixture models (GMMs). Two sets of features are investigated: the three prosodic features (V_N, V_D and V_P), and the set of prominence features (P_N , P_D *and* P_P).

5.1. The lexical stress classifier

We trained a GMM for the stressed syllables with two mixture components, where one mixture is trained with syllables carrying primary stress (PS) and the other mixture is trained with syllables carrying secondary stress (SS). We also trained another GMM with one mixture component based on the unstressed syllables (NS).

Three examples are shown in Fig. 4. During the classification process, we assume that there is at least one primary stressed syllable in the word under consideration. We first classify all syllables in a word as either stressed or unstressed, and compute the probabilities of being stressed or unstressed for all syllables. The syllable with highest probability of being stressed is considered as PS, regardless of whether it is classified as stressed or not, as shown in case (b). Then, the remaining syllables that are classified as stressed are labeled as SS. The exception is that, certain syllables treated as SS may be classified as PS, if the probabilities of being stressed are close (e.g. with 10%) to that of the most prominent PS, e.g. as illustrated in case (c).

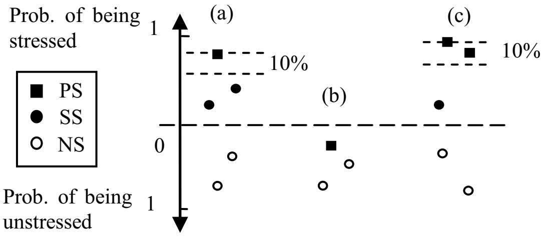


Figure 4: *Examples illustrating our approach for stress classification based on the probabilities of a syllable being stressed or unstressed. The most probable syllables to be stressed are classified as primary stress as shown in the three cases, (a) to (c). In particular (b), even if a syllable is unlikely to be stressed, the most probable one is still classified as primary stress.*

5.2. The pitch accent classifier

For the data used in the pitch accent detection experiments, we annotated the syllables with pitch accents as in H*, L*, !H*, L+H*, L*+H, L+!H*, L*+!H, or H+!H* as in [11].

The GMM for accented syllables has two mixture components: one mixture is trained on high or rising pitch accents using the syllables labeled as H*, !H*, L+H*, L*+H, L+!H*, or L*+!H; the other mixture is trained on low or falling pitch accents using the syllables labeled as L* or H+!H*. The GMM for unaccented syllables has one mixture component and is trained on unaccented syllables.

6. Experiments

6.1. Experimental setup

We adopted the 5-fold cross-validation method to evaluate the lexical stress and pitch accent detection performance. Prompted speech data from 100 Mandarin speakers (50 males and 50 females), and 100 Cantonese speakers (also 50 males and 50 females) [15] are used in our experiments.

For the data on lexical stress, each speaker reads 35 words / compound words embedded in carrier sentence (e.g., "I said xxxx five times"). All syllables in the 35 words are manually annotated for PS, SS, or NS.

For the experiments on pitch accent, we use 29 specially designed utterances that are recorded by all speakers. These utterances include questions (e.g., "Do you need any money?"), declaration sentences, and sentences with continuation rises. All data are transcribed by a single annotator.

6.2. Lexical stress detection

We carried out two sets of lexical stress detection experiments – one using the three prosodic features (V_N, V_D and V_P) and the other using the three prominence features (P_N, P_D and P_P). Results are shown in Table 1, which summarizes the total confusions from all runs in the 5-fold cross-validation. In Table 3, rows R2 and R4 include the Equal Error Rate (EER) averaged from the 5-fold cross-validation and the accuracies under different criteria:

- **PSN**: Identify the syllables carrying primary stress, secondary stress or no stress;
- **SN**: Classify the syllables as either stressed or unstressed;
- **PN**: Determine if the syllables carry PS or not.

As can be seen, although the performance of the prominence model decreases slightly under the SN criterion from 83.52% to 82.90%, improvements are obtained based on the criteria of PSN and PN, i.e. from 76.92% to 78.64%, and from 86.99% to 89.80% respectively.

Table 1 shows that, although the accuracy for NS decreases slightly from 15938 to 15856 (more confusion between NS and SS), the performance for PS and SS is improved from 5520 to 5892 and from 977 to 1189 respectively.

Table 1: *Lexical stress detection results from cross-validation*

Ref. / Det.	V_N, V_D, V_P			P_N, P_D, P_P		
	NS	**PS**	**SS**	**NS**	**PS**	**SS**
NS	15938	1236	1623	15856	1136	1823
PS	634	5520	1051	597	5892	639
SS	1313	875	977	1432	603	1189

6.3. Pitch accent detection

Similarly, the three prosodic features (V_N, V_D and V_P) and the three prominence features (P_N, P_D and P_P) are also used in training the pitch accent detector. Results are shown in Table 2 and Table 3 (rows R2 and R4).

The detector using prosodic features (V_N, V_D and V_P) achieves an accuracy of 80.61%, with EER of 22.97%. By adopting the prominence model, the detection performance is further improved with accuracy of 83.30% and EER of 19.29%.

Table 2: *Pitch accent detection results from cross-validation*

Ref. / Det.	V_N, V_D, V_P		P_N, P_D, P_P	
	Accented	**Unaccented**	**Accented**	**Unaccented**
Accented	14569	7537	15413	6330
Unaccented	7231	46828	6387	48035

7. Analysis

To examine the contributions of different refinements to the feature extraction, we present a set of results in Table 3. Row R0 shows the results of detection using the conventional features [7, 11, 14], i.e. the maximum syllable intensity I_{max}, the syllable nucleus duration V_D and the maximum syllable pitch value V_{Pmax}.

7.1. Prosodic features (V_N, V_D, and V_P)

By Comparing rows R2 with R0, we observe that the prosodic features (V_N, V_D, V_P) outperform the conventional features (I_{max}, V_D, V_{Pmax}) by about 6%.

- ***Loudness model***

By comparing rows R1 with R0, we can see that the loudness model improves the performance of lexical stress detection and pitch accent detection by about 3%~4%.

- ***Differential pitch value***

Fig. 2 shows that use of the differential pitch value gives the best performance when a_{p1} is around 0.95, which is better than use of the mean pitch value (when a_{p1}=-2) by about 3% for pitch accent detection or about 5% for stress detection.

By comparing rows R2 with R1, and R4 with R3, we can see that V_P outperforms V_{Pmax}, which offers better discriminative power than the minimum, mean, range, etc. of a syllable (see section 3.3). This amounts to about 2% for both stress detection and pitch accent detection.

7.2. Prominence model

By comparing rows R4 with R2 in Table 3, we can see that the prominence model improves the performance of pitch accent detection from 22.97% to 19.29% in EER, and from 80.61% to 83.30% in accuracy. For stress detection, although the prominence model gave a slight decrease in performance under the SN criteria, there are performance improvements under the PN and PSN criteria. Similar results can be observed from the comparison of R3 with R1.

Table 3: *Detection results from cross-validation*

		SN		**PN**	**PSN**	**Pitch Accent**	
		EER (%)	**Acc. (%)**	**Acc. (%)**	**Acc. (%)**	**EER (%)**	**Acc. (%)**
R0	I_{max}, V_D, V_{Pmax}	23.51	77.59	80.50	69.36	29.11	75.02
R1	V_N, V_D, V_{Pmax}	19.95	81.62	83.50	73.00	24.79	78.25
R2	$\boldsymbol{V_N}$, $\boldsymbol{V_D}$, $\boldsymbol{V_P}$	**18.30**	**83.52**	**86.99**	**76.92**	**22.97**	**80.61**
R3	P_N, P_D, P_{Pmax}	21.66	80.65	86.48	75.16	21.43	81.24
R4	$\boldsymbol{P_N}$, $\boldsymbol{P_D}$, $\boldsymbol{P_P}$	**19.61**	**82.90**	**89.80**	**78.64**	**19.29**	**83.30**

8. Conclusions

In this work, we adopted loudness (instead of intensity), differential pitch value (instead of maximum pitch value), and the syllable nucleus duration as the syllable prosodic features, which improved the detection of lexical stress and pitch accent by about 6%. We also proposed a prominence model for prosodic features to take into account the effects of neighboring syllables. Using this prominence model, we can achieve performance improvements in pitch accent detection (from 22.97% to 19.29% in EER, and from 80.61% to 83.30% in accuracy). For lexical stress detection, we can also achieve performance improvements in (i) classification of primary, secondary and unstressed syllables (from 76.92% to 78.64%), as well as (ii) determining the presence or absence of primary stress (from 86.99% to 89.80%).

9. Acknowledgements

This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. This projected is funded by the Research Grant Council, Hong Kong SAR government (CUHK 416108).

10. References

[1] Anderson-Hsieh, J. *et al.*, "The Relationship between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentals, Prosody and Syllable Structure," *Language Learning*, vol. 42, 1992.

[2] Meng H. *et al.*, "Studying L2 Suprasegmental Features in Asian Englishes: A Position Paper", *Proc. of INTERSPEECH 2009*.

[3] Tepperman J. and Narayanan S., "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners", *Proc. of ICASSP 2006*.

[4] Imoto K. *et al.*, "Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system", *Proc. of ICSLP 2002*.

[5] Ren Y. *et al.*, "Speaker-independent automatic detection of pitch accent", *Proc. of Speech Prosody 2004*.

[6] Sun X-J, "Pitch accent prediction using ensemble machine learning", *Proc. of ICSLP 2002*.

[7] Tamburini F., "Prosodic prominence detection in speech", *Proc. of Signal Processing and its Applications 2003*.

[8] Zwicker E. and Fastl H., *Psychoacoustics – Facts and Models* 2nd Updated Edition, pp. 61-238, Springer 1999.

[9] Zwicker E. and Scharf B., "A Model of Loudness Summation", *Psychological Review*, vol. 72, pp. 3-26, 1965.

[10] Stevens S. S., "On the psychophysical law". The psychological review, vol. 64, no. 3, pp153-181, 1957.

[11] Li K. *et al.*, "Detection of Intonation in L2 English Speech of Native Mandarin Learners", *Proc. of ISCSLP 2010*.

[12] Pulgram, E., *Syllable, Word, Nexus, Cursus*, Mouton, 1970.

[13] Sjölander K., "The Snack Sound Toolkit", KTH, Online: http://www.speech.kth.se/snack/, accessed on Mar 29, 2011.

[14] Wang D. and Narayanan S., "An Acoustic Measure for Word Prominence in Spontaneous Speech", *IEEE Trans. Speech and Audio Proc.*, vol. 15, no. 2, pp. 690-701, 2007.

[15] Li M., *et al.*, "L2 English Corpus from Chinese Learners Focusing on Suprasegmental Features", *Proc. of ICPhS2011*.