Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks

Kun Li, Helen Meng

Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong SAR, China

Abstract

This paper investigates the use of multi-distribution deep neural networks (MD-DNNs) for automatic lexical stress detection and pitch accent detection, which are useful for suprasegmental mispronunciation detection and diagnosis in second-language (L2) English speech. The features used in this paper cover syllable-based prosodic features (including maximum syllable loudness, syllable nucleus duration and a pair of dynamic pitch values) as well as lexical and syntactic features (encoded as binary variables). As stressed/accented syllables are more prominent than their neighbors, the two preceding and two following syllables are also taken into consideration. Experimental results show that the MD-DNN for lexical stress detection achieves an accuracy of 87.9% in syllable classification (primary/secondary/no stress) for words with three or more syllables. This performance is much better than those of our previous work using Gaussian mixture models (GMMs) and the prominence model (PM), whose accuracies are 72.1% and 76.3% respectively. Approached similarly as the lexical stress detector, the pitch accent detector obtains an accuracy of 90.2%, which is better than the results of using the GMMs and PM by about 9.6% and 6.9% respectively.

Keywords: lexical stress, pitch accent, non-native English, language learning, deep neural networks

1. Introduction

To meet the demands of self-directed language learning, computer-aided pronunciation training (CAPT) has drawn considerable attention. Research in CAPT systems focuses on leveraging speech and language technologies for mispronunciation detection and diagnosis (MDD), which can be implemented at segmental and suprasegmental levels (Meng et al., 2009). The segmental level involves phones (Li and Meng, 2014; Li et al., 2015) and words; while the suprasegmental level includes lexical stress (Li et al., 2011a; Li and Meng, 2012, 2013), pitch accent (Li et al., 2011a; Zhao et al., 2013b), intonation (Li et al., 2010; Arias et al., 2010), rhythm, etc. In this work, we focus on lexical stress and pitch accent.

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

Lexical stress detection is the key module for the MDD of lexical stress. With the canonical lexical stress patterns (extracted from dictionaries) and the results of lexical stress detection, we can determine whether a lexical stress pattern is correctly pronounced (Li and Meng, 2012). A mispronunciation is detected if a lexical stress pattern is deemed incorrect, and appropriate diagnostic feedback can be generated by comparing the detected lexical stress pattern with the canonical one. Similarly, the MDD of pitch accent can also be achieved by leveraging a pitch accent detector, if the canonical pitch accent placement is available.

As the input features of lexical stress and pitch accent detection involve syllable-based prosodic features (modeled by linear units with Gaussian noise) as well as lexical and syntactic features (set as binary vectors), multi-distribution deep neural networks (MD-DNN) (Kang et al., 2013; Li and Meng, 2013, 2014; Li et al., 2015) are used in this work. Similar to traditional DNNs, MD-DNNs are also constructed by stacking up multiple Restricted Boltzmann Machines (RBMs) on top of one

Email addresses: kli@se.cuhk.edu.hk (Kun Li), hmmeng@se.cuhk.edu.hk (Helen Meng)

Preprint submitted to Speech Communication *December 13, 2016*

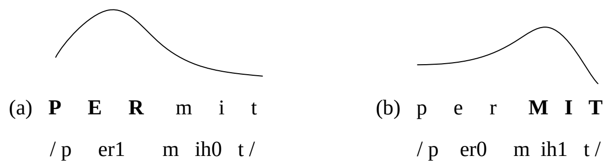

Figure 1: Pitch realization for words *permit* (noun) and *permit* (verb) in citation form (Ladd, 2008).

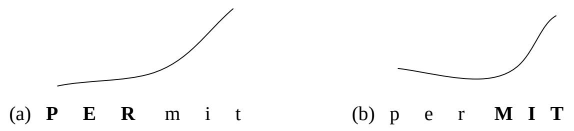

Figure 2: Pitch realization for words *permit* (noun) and *permit* (verb) in question form (Ladd, 2008).

another. Excluding the bottom RBM, all the other ones are traditional Bernoulli RBM (B-RBM), whose hidden and visible units are all binary. The bottom RBM is a type of mixed Gaussian-Bernoulli RBM (GB-RBM), whose hidden units are binary while some visible units are Gaussian distributed and the other visible units are binary.

The rest of this paper is organized as follows: Section 2 introduces previous work related to lexical stress and pitch accent detection. Section 3 discusses the features for lexical stress and pitch accent detection. Section 4 develops a lexical stress detector and a pitch accent detector, both of which use MD-DNNs. Section 5 and 6 present the experimental results and analysis respectively. Section 7 summarizes this paper.

2. Related work

2.1. Phonetic nature of stress — a linguistic view

Ladd (2008) gave simple examples to demonstrate the phonetic nature of stress, as shown in Figure 1 to Figure 3. For the words in citation form, as illustrated in Figure 1, the pitch contours rise to a peak within the stressed syllables and then drop quickly. In addition, there may also be some changes in syllable duration, intensity and vowel quality. Figure 2 shows a possible realization of pitch contours over the words *permit* (noun) and *permit* (verb) in question form; while Figure 3 demonstrates the pitch contours in a context where the words are in the position after the focus of an utterance. In the cases shown in Figure 2 and Figure 3, pitch movement is no longer a cue to stress (Sluijter and van Heuven, 1996). However, these stressed syllables can still be perceived by other cues such as duration, intensity, etc.

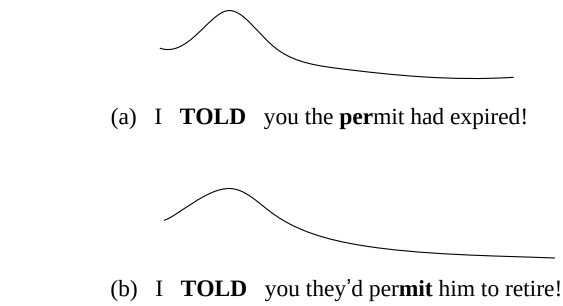
(a) I **TOLD** you the **per**mit had expired!
(b) I **TOLD** you they'd per**mit** him to retire!

Figure 3: Pitch realization for words *permit* (noun) and *permit* (verb) in a context where the words are in the position after the focus of an utterance (Ladd, 2008).

2.2. Acoustic correlates of stress

Prior to automatic stress detection, early research focused in exploring acoustic correlates with stress. In psychology, stress correlates with length, loudness, pitch and quality (Fry, 1958). These psychological factors correspond to duration, intensity, fundamental frequency and formant structure of speech in physical dimensions.

Fry (1958) investigated the influence of changes in duration, intensity and fundamental frequency (F0) to human perception of stress in English. Experiments were conducted by presenting synthesized speech stimuli to a large number of subjects. Experimental results showed that duration, intensity and F0 played important roles in stress judgments. In other words, syllables with longer duration, greater intensity or/and higher F0 were likely to be marked as stressed. In addition, the results also revealed that the direction of a step change of F0 had a strong effect on stress perception, i.e., a syllable with higher F0 was generally heard as stressed; while the magnitude of a step change made little contribution, e.g., the step changes of 5 Hz and 90 Hz might produce similar results if they were in the same direction.

Lieberman (1960) further examined the relevance of fundamental frequency, syllabic duration, relative amplitude and integral of the amplitude over a syllable by devising a binary automatic lexical stress recognition program. Experiments showed that higher fundamental frequency and envelope amplitude were the most relevant features.

Morton and Jassem (1965) investigated acoustic correlates of stress by performing a experiment in which synthetic nonsense syllables (e.g., "sisi", "sasa", etc.) were presented to subjects. The results showed that variations in fundamental frequency were the overriding factors comparing with those in either intensity or duration. A raised fundamental frequency was more efficient to be perceived as stressed. The syllables with more intensity and longer duration tended to be judged as stressed.

Sluijter and van Heuven (1996) investigated the acoustic correlates of stress by conducting experiments with Dutch bisyllabic minimal pairs spoken by ten speakers. It showed that duration was the most reliable correlate of stress; while overall intensity and vowel quality had the poorest performance. Spectral balance which was designed to measure the intensity distribution in different frequency bands was also a reliable cue to stress. The intensity in the high filter bands above 500 Hz had better discriminating ability than the low part intensity. However, Campbell and Beckman (1997) reported that there was no significant difference in spectral balance between vowels in stressed versus unstressed syllables in the unaccented condition.

2.3. Automatic lexical stress detection

As far as we know, Lieberman's work (Lieberman, 1960) is the first to develop a system of automatic lexical stress detection for bisyllabic words using a decision tree method, whose features are described in last subsection. The data for the experiments contained 25 noun-verb pairs of bisyllabic words (e.g., "ˈ*rebel*" vs. "*re*ˈ*bel*") embedded in sentences uttered by 6 female and 10 male native American speakers. Training and testing were performed over the same dataset. The lexical stress detection accuracy was about 99%.

Aull and Zue (1985) used syllable duration, syllable average energy, syllable maximum pitch value and spectral change to identify lexical stress for 1,600 isolated words extracted from continuous speech uttered by 11 speakers. The determination was based on the Euclidean distance from each syllable-based feature vector to the reference vector. Its syllable-based and word-based accuracies were 98% and 87% respectively. Note that 80% of the words had two or three syllables, and the others had four or five syllables.

Waibel (1986) proposed to use a Bayesian classifier assuming multivariate Gaussian distributions for lexical stress detection. Its features included the peak-to-peak amplitude integral over sonorant segments, syllable duration, maximum pitch value and spectral change. Its error rate was about 12%, which was evaluated on a dataset consisting of 50 sentences uttered by 10 speakers. Other similar studies used Bayesian classifiers assuming multivariate Gaussian include Ying et al. (1996), van Kuijk and Boves (1999), Tamburini (2003), etc.

Freij et al. (1990) used two continuous hidden Markov models (HMMs) to construct an automatic lexical stress detector, whose features were based on fundamental frequency, syllabic energy and coarse linear prediction spectra. Experiments were evaluated on a set of bisyllabic words (15 noun-verb pairs of words and 12 tokens for each word) embedded in continuously spoken phrases. The recognition rates for stressed and unstressed syllables were 95% and 93% respectively.

Jenkin and Scordilis (1996) developed three classifiers (neural networks, Markov chains and a rule-based approach) for lexical stress detection. Their features involved mean energy over syllable nucleus, syllable duration, syllable nucleus duration, maximum and mean pitch over syllable nucleus. The neural networks achieved the best performance ranging from 81% to 84%.

Xie et al. (2004) adopted support vector machines (SVMs) to build classifiers for lexical stress detection. Experiments used prosodic features (relating to duration, amplitude and pitch) and vowel quality features (extracted from vowel acoustic features).

Tepperman and Narayanan (2005) applied GMMs in the lexical stress detection for Japanese learners of English. A set of prosodic features were investigated, including the mean values of F0, syllable nucleus duration, energy and other features relating to the F0 slope and the energy range.

Zhao et al. (2011) adopted SVMs to identify the vowels of L2 English speech carrying primary stress or not. The prosodic features included loudness, vowel duration, spectral emphasis, pitch in semitone and pitch variations based on RFC (Taylor, 1995) and Tilt parameters (Taylor, 1998, 2006). Their context-aware features, which were extracted from the differential values from the current vowel and its preceding/succeeding vowels, were also considered. The experiments were conducted on a corpus with 200 utterances spoken by 22 Taiwanese, achieving an accuracy of 88.6%. Other similar studies using SVMs to detect lexical stress of English speech with Taiwanese accent were reported in Wang et al. (2009), Chen and Wang (2010), etc.

Ferrer et al. (2015) introduced a system for lexical stress detection using both prosodic (pitch, energy and duration) and spectral (tilt and MFCC posteriorgrams) features. Three level of stress (primary, secondary or unstressed) were classified using GMMs. Experiments were performed on a corpus containing English speech from L1-English and L1-Japanese children. Most of the words in this corpus had only two syllables. Results showed that the MFCC posteriorgrams helped improve the experimental performance. The error rates of the system for L1-English and L1-Japanese data were 11%

and 20% respectively.

2.4. Pitch accent prediction and detection

The approaches of pitch accent detection only using prosodic features are similar to those of lexical stress detection, e.g., Imoto et al. (2002), Tamburini (2003), Ren et al. (2004), Li et al. (2007), Wang and Narayanan (2007), Zhao et al. (2013b,a), Tamburini et al. (2014), etc. The classifiers for the above efforts involve HMMs (Imoto et al., 2002; Li et al., 2007), Bayesian classifier assuming multivariate Gaussian distributions (Tamburini, 2003), time-delay recursive neural networks (Ren et al., 2004), SVMs (Zhao et al., 2013b), multi-layer perceptrons (Zhao et al., 2013a), latent-dynamic conditional neural fields (a kind of probabilistic graphical models) (Tamburini et al., 2014), etc.

Besides prosodic features, lexical and syntactic features also highly correlate with pitch accents. Ross et al. (1992) investigated several factors influencing pitch accent placement based on a subset of the Boston University Radio News Corpus (BURNC) (Ostendorf et al., 1995). They found that 39% of the words were function words and only about 11% of these function words had pitch accents. For the pitch accents on function words, 42% of them were on negatives or quantifiers. Similarly, Rosenberg (2009) showed that about 76% of the content words in the BURNC were accented, whereas only 14% of the function words were accented.

Due to the correspondence with pitch accents, lexical and syntactic features are widely used in automatic pitch accent prediction — a task that assigns pitch accents from given text and is motivated by synthesizing more natural sounding speech. With a set of lexical and syntactic features from unrestricted text, Hirschberg (1993) proposed a method using classification and regression tree (CART) (Breiman et al., 1984; Lewis, 2000) to predict pitch accent location. Experiments showed that part-of-speech (POS) played an important role in the prediction. To improve the prediction performance, more additional features were adopted, including word class context, dictionary stress, relative distance measures within an IP, etc.

Ross and Ostendorf (1996) applied computational models based on decision trees to predict pitch accent location and relative prominence level. Many kinds of features were investigated, including dictionary stress, POS (e.g., nouns, adverbs, etc.), prosodic phrase structure (e.g., phrase break size, number of syllables/words, etc.), new/given status (e.g., whether word is new or given to the paragraph, etc.), paragraph structure (e.g., the position of phrase within the sentence, etc.), and labels of other units (e.g., types of pitch accent, boundary tone, etc.).

Furthermore, lexical and syntactic information are also used to complement acoustic features in pitch accent detection. Wightman and Ostendorf (1994) proposed a general algorithm based on decision trees and Markov sequence models to automatically label pitch accents as well as other prosodic patterns including prosodic breaks and boundary tones. A large set of features were investigated, such as syllable duration, energy, pitch contour, dictionary stress, a flag indicating whether this syllable is word-final or not, etc.

Conkie et al. (1999) combined acoustic and syntactic modeling to determine whether a word carries a pitch accent or not. The acoustic and syntactic models resulted in accuracies of 83% and 84% respectively; while combing these two models obtained an accuracy of 88%. The experiments were conducted over only one speaker's data from the BURNC.

Sun (2002) used the techniques of bagging and boosting with CART for pitch accent prediction and detection. The accuracies of using only text information and only acoustic features were about 80% and 85% respectively. They were improved to about 87% by incorporating both kinds of features. All these experiments were carried out on one female's data from the BURNC.

Chen et al. (2004) proposed a syntactic-prosodic model based on neural networks for pitch accent prediction and achieved an accuracy of 82.7%. Coupling a GMM-based acoustic-prosodic model with the syntactic-prosodic model obtained an accuracy of 84%. These experiments were also conducted on the BURNC.

Gregory and Altun (2004) used conditional random fields (CRFs) for the task of pitch accent prediction, achieving an accuracy of 75.9% on the Switchboard Corpus (Godfrey et al., 1992). It was slightly improved to 76.4% by further leveraging acoustic features. Similar studies using CRFs for pitch accent prediction and detection include Levow (2008), Qian et al. (2010), Ni et al. (2011), etc.

Ananthakrishnan and Narayanan (2008) also developed a pitch accent detector using acoustic, lexical and syntactic features. The experiments based on neural networks obtained an accuracy of 86.8% over the BURNC. Using similar features and the same corpus for the same task, Sridhar et al. (2008) obtained an accuracy of 86.0% with a maximum entropy framework, and Jeon and Liu (2009) achieved an accuracy of 89.8% with neural networks.

2.5. Prominence model (PM)

Our previous work (Li and Liu, 2010a; Li et al., 2011a) used a set of syllable-based prosodic features (see Section 3) and proposed a prominence model (PM) for lexical stress and pitch accent detection. The PM estimates the prominence values from the syllable in focus as well as the syllables in neighboring contexts. It is based on the observations that syllables with loudness, duration and pitch greater than their neighboring syllables are likely to be perceived as stressed or accented, even if their values are not large on average. Hence, the differences between the feature values of the current syllable and the ones of the neighboring syllables are also considered. With this PM, the syllable-based prosodic features are converted into a set of prominence features. Both the classifiers for lexical stress detection and pitch accent detection are Gaussian mixture models (GMMs). Experiments showed that the PM improved the performance of lexical stress detection by 4.2% (from 72.1% to 76.3%) (Li and Meng, 2013), and pitch accent detection by 2.7% (from 80.6% to 83.3%) (Li et al., 2011a).

3. Features for lexical stress and pitch accent detection

As discussed in Section 2.2, stressed syllables usually exhibit longer duration, greater loudness and higher pitch than their neighbors. Hence, we developed the syllable-based prosodic features (Li et al., 2011a; Li and Meng, 2013): syllable nucleus duration, maximum loudness and a pair of dynamic pitches. Due to the effectiveness of lexical and syntactic features in pitch accent detection, we also use this kind of feature for lexical stress and pitch accent detection.

As stressed/accented syllables are more prominent than their neighbors, the two preceding and two succeeding syllables as well as the syllable in focus are taken into consideration. Shorter context windows, such as the single preceding and single succeeding syllables, may not thoroughly exploit contextual information; while longer context windows, such as the three preceding and three succeeding syllables, may cause over-fitting, since each word in our experiments only has 4.2 syllables for lexical stress detection and 1.5 syllables for pitch accent detection on average (see Table 1).

3.1. Syllable-based prosodic features

The syllable-based prosodic features used in this work include syllable nucleus duration, maximum loudness and a pair of dynamic pitches. More elaboration is provided in the following subsections.

3.1.1. Syllable nucleus duration (V_{dur})

Words can be transcribed into syllables by lexicon lookup and by applying appropriate linguistic rules. However, it is a difficult task to automatically segment an utterance into syllables accurately. Since the syllable nucleus can be extracted more consistently, the syllable nucleus duration is usually used to substitute for the syllable duration (Tamburini, 2003; Tepperman and Narayanan, 2005).

We first apply the Maximal Onset Principle (Pulgram, 1970) to automatically determine the syllable boundaries and extract the syllables from the phoneme sequence output of the automatic speech recognizer described in Li and Meng (2014). For example, the word "*apartment*" uttered by an L2 English learner is divided into / *axr* /, / *p aa t* /, / *m ax n* / and / *d ax* /, as shown in Figure 4. Within the time boundaries of every extracted syllable, we treat the frames with loudness above N_{bottom} as the syllable nuclei, where N_{bottom} is the value above which 50% of all loudness values in the IP lie. The normalized syllable nucleus duration V_{dur}, which is taken as our feature, is given by Equation (1).

$$V_{\mathrm{dur}} = \ln(d_{\mathrm{nucl}}) - \ln(\overline{d_{\mathrm{IP}}}) \tag{1}$$

where d_{nucl} is the syllable nucleus duration, $\overline{d_{\mathrm{IP}}}$ is the mean duration of all syllable nuclei in the IP.

3.1.2. Maximum loudness (V_{loud})

Loudness is the human perception of the strength of sound energy. There is a complex relationship between loudness and sound energy. We follow Zwicker's loudness model (Zwicker and Fastl, 1999) for a precise estimation of loudness, and a simplifying calculation was given in our previous work (Li et al., 2011a), which improved the lexical stress detection and pitch accent detection by about 3%. The normalized maximum syllable loudness V_{loud}, as given by Equation (2), is taken as our feature:

$$V_{\mathrm{loud}} = N_{\mathrm{max}} - \overline{N_{\mathrm{IP}}} \tag{2}$$

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

3.1.3. Pair of dynamic pitches (f_{m1}&f_{m1})

We first perform pitch extraction using a method based on wavelet transformation (Li and Liu, 2010b) and process pitch values that fall within the time boundaries of the identified syllable nuclei. Then we convert the pitch value to the semitone scale, which is a logarithm scale that better matches human perception of pitch. The normalized pitch value in semitone is given by Equation (3).

$$f = 12\log_2(f_0/\overline{f_{\mathrm{IP}}}), \qquad \text{where } f_0 > 0 \tag{3}$$

where f_0 is the fundamental frequency in Hz, $\overline{f_{\mathrm{IP}}}$ is the mean pitch value in the IP.

A differential pitch value in a syllable was derived in our previous work (Li et al., 2011a), as given by Equation (4a). It was proposed based on the following observations: syllables with rising tones often give a stressed perception; while syllables with falling tones, especially whose preceding syllable is stressed with a rising tone, are often perceived as unstressed. Equation (4a) was further improved to Equation (4b), which was used in the experiments of Li et al. (2011a). Results showed that the differential pitch value outperformed the mean and maximum pitch values in a syllable by about 5% and 3% respectively.

$$V_{\mathrm{pitch}} = f_{\mathrm{m2}} + (f_{\mathrm{m2}} - f_{\mathrm{m1}}) = 2f_{\mathrm{m2}} - f_{\mathrm{m1}} \tag{4a}$$

$$V_{\mathrm{pitch}} = 2f_{\mathrm{m2}} - 0.95f_{\mathrm{m1}} \tag{4b}$$

where f_{m1} and f_{m1} are the starting and ending extreme pitch values in the syllable nucleus respectively, as shown in Figure 4.

In this work, we only use a pair of dynamic pitches (f_{m1} and f_{m2}) in a syllable nucleus instead of the differential pitch value (V_{pitch}), as DNNs can optimize the performance by automatically adjusting the relationship between f_{m1} and f_{m2}.

3.2. *Lexical and syntactic features*

Section 2.4 shows that lexical and syntactic information can complement acoustic features in pitch accent detection. Thus, this type of feature is also used in this work to improve the lexical stress and pitch accent detection of L2 English speech.

For lexical stress detection, we use two bits to indicate a syllable carrying primary stress (PS), secondary stress (SS), no stress (NS) or NULL in dictionaries. The NULL means there is no syllable, e.g., for the first syllable in a word, there are no preceding syllables. Since we consider the two preceding and two succeeding syllables as well as the current syllable, there are 10 binary values for each syllable's canonical lexical stress pattern (i.e., lexical and syntactic features).

Take the syllable / *p aa t* / in Figure 4 for example, its canonical lexical stress pattern is (NULL NS $\underline{\text{PS}}$ NS NS). Thus, the 10 binary values are (11 00 $\underline{01}$ 00 00), if we use "00", "01", "10" and "11" to represent NS, PS, SS and NULL, respectively.

For pitch accent detection, an additional bit ($F_{initial}$) is used to indicate an onset of a new word. The lexical and syntactic features (PS, SS, NS, NULL and $F_{initial}$) are encoded with three bits and also used as part of the input features.

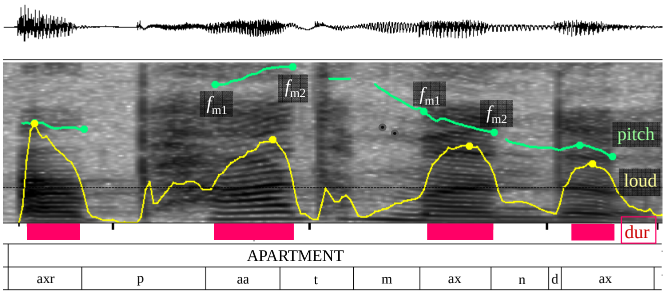

Figure 4: An example of feature extraction for lexical stress detection. The top green curve is the pitch in semitone, the yellow curve is the loudness, and the red bars indicate the syllable nucleus durations. f_{m1} and f_{m2} are also marked for the syllables of / *p aa t* / and / *m ax n* /.

4. Architectures of MD-DNNs for lexical stress and pitch accent detection

4.1. *Lexical stress detection*

Due to the effectiveness of deep learning techniques, we also adopt an MD-DNN for the lexical stress detector, which makes use of the syllable-based prosodic features and the canonical lexical stress pattern as its input features. As described in last section, the syllable-based prosodic features include syllable nucleus duration (V_{dur}), maximum loudness (V_{loud}) and dynamic pitches (f_{m1} and f_{m2}). These features are further scaled to have zero mean and unit variance over the whole corpus.

Since a context window of $(2 + 1 + 2)$ syllables is applied in this work, the bottom layer of the MD-DNN has 20 linear units with Gaussian noise for the syllable-based prosodic features and 10 binary units for the corresponding canonical lexical stress pattern. The diagram of the MD-DNN for lexical stress detector is shown in Figure 5. Above the bottom layer, there are three hidden layers and each has 128 units. For the top output layer, there are only three units generating the posterior probabilities of PS, SS and NS for each syllable.

4.2. *Pitch accent detection*

The pitch accent detector can be approached similarly as the lexical stress detector. The diagram of the MD-DNN for pitch accent detection is shown in Figure 6.

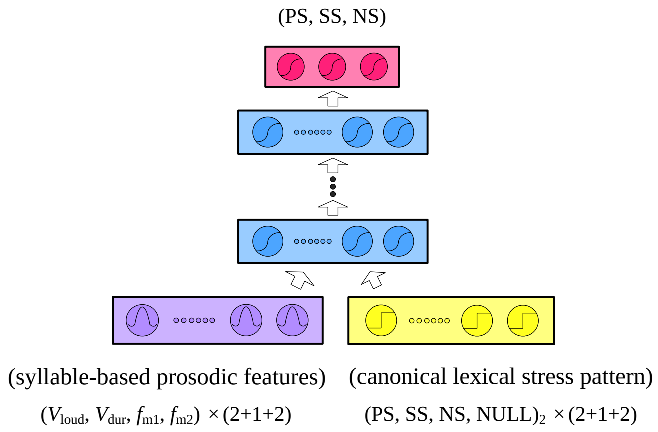
(PS, SS, NS)
(syllable-based prosodic features)
(canonical lexical stress pattern)
$(V_{\mathrm{loud}}, V_{\mathrm{dur}}, f_{\mathrm{m1}}, f_{\mathrm{m2}}) \times (2+1+2)$
$(\mathrm{PS, SS, NS, NULL})_2 \times (2+1+2)$

Figure 5: Diagram of the MD-DNN for lexical stress detection.

The syllable-based prosodic features are the same as those used in lexical stress detection, i.e., V_{dur}, V_{loud}, f_{m1} and f_{m2}. The lexical and syntactic features (PS, SS, NS, NULL and $F_{initial}$) are encoded with three bits. As accented syllables are more prominent than their neighbors, a contextual window of $(2 + 1 + 2)$ syllables is also applied in this work. Thus, there are total 20 linear units with Gaussian noise and 15 binary units in the bottom of the MD-DNN. Above the bottom layer, there are two hidden layers of 256 units. For the top output layer, there are two units generating the posterior probabilities of being accented and unaccented.

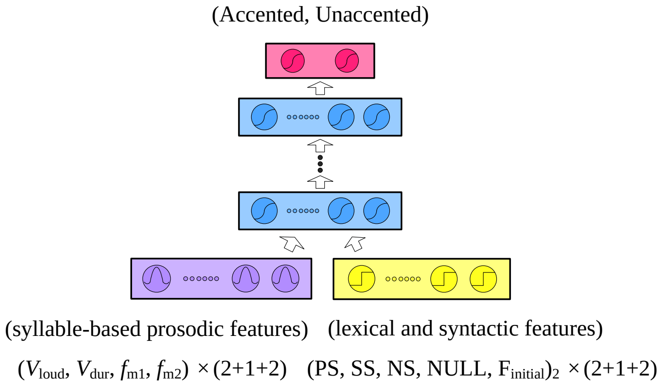
(Accented, Unaccented)
(syllable-based prosodic features)
(lexical and syntactic features)
$(V_{\mathrm{loud}}, V_{\mathrm{dur}}, f_{\mathrm{m1}}, f_{\mathrm{m2}}) \times (2+1+2)$
$(\mathrm{PS, SS, NS, NULL, F_{initial}})_2 \times (2+1+2)$

Figure 6: Diagram of the MD-DNN for pitch accent detection.

5. Experiments

5.1. Supra-CHLOE corpus

The Supra-CHLOE (**Supra**segmental **Ch**inese **L**earners **O**f **E**nglish) corpus (Li et al., 2011b) is used in this work. It contains speech recordings from 100 Mandarin speakers and 100 Cantonese speakers (both groups are gender-balanced). There are five parts in this corpus: *lexical stress*, *utterance-level stress*, *intonation*, *reduced/unreduced function words* and *prosodic disambiguation*.

Only the *lexical stress* part has syllables labeled with PS/SS/NS. In this part, each speaker uttered 30 words embedded in carrier sentences (e.g., "I said *hospital* five times"). The syllables in the other parts are used as unlabeled data for lexical stress detection, as shown in Table 1. Note that Bisyllabic words are excluded from this work due to their simplicity.

Excluding *lexical stress*, syllables from the other parts are labeled with different types of pitch accents, as given in Table 2. These data are further grouped as accented and unaccented syllables, and used for pitch accent detection (see Table 1).

Table 1: Details of data used for lexical stress and pitch accent detection in our experiments.

	Lexical stress		Pitch accent	
	Syllable	Word	Syllable	Word
Unlabeled	91.5 k	29 k	—	—
Labeled	25.4 k	6 k	205.5 k	135.8k

Note: counts of syllables and words are measured in the unit of thousands (k).

Table 2: Annotation results of pitch accents in rates and counts. 'Un' means unaccented.

H*	!H*	L+H*	L+!H*	H+!H*
11.48%	4.05%	9.80%	1.28%	1.88%
(23,594)	(8,311)	(20,143)	(2,627)	(3,868)
L*	L*+H	L*+!H	Un	*?
2.70%	0.73%	0.06%	68.01%	0.01%
(5,555)	(1,489)	(122)	(139,747)	(19)

5.2. DNN training

The DNNs training for lexical stress and pitch accent detection in this work is similar to (Qian et al., 2012; Li and Meng, 2013, 2014). In the pre-training stage, we try to maximize the log-likelihood of RBMs. The one-step contrastive divergence (CD) (Hinton et al., 2006) is used to approximate the stochastic gradient. Twenty epochs are performed with a batch size of 128 syllables. In the fine-tuning stage, the standard back-propagation (BP) algorithm (Rumelhart et al., 1986) is performed. A dropout (Hinton et al., 2012; Deng et al., 2013; Seltzer et al., 2013; Hannun et al., 2014) rate of 10% is used in this work.

5.3. Experimental results

5.3.1. Lexical stress detection

Li and Meng (2012) shows that a word may have more than one lexical stress patterns in dictionaries and the variations are primarily due to the presence or placement of secondary stress. For example, "*ˈautoˌgraph*" versus "*ˈautograph*", "*ˌmisunderˈstand*" versus. "*ˌmisˌunderˈstand*", etc. The lexical stress perceptual test in Li and Meng (2012) also reveals that humans may have difficulty in identifying secondary stress.

Due to the uncertainty of secondary stress, we can merge NS and SS into one group and turn the three-category classification into a two-category classification. Similarly, we can also merge SS and PS into one group. Then we have the following criteria to evaluate the performance of lexical stress detection:

(1) **P-S-N**: Identify the syllables carrying primary stress, secondary stress or no stress;

(2) **S-N**: Classify the syllables as either stressed or unstressed;

(3) **P-N**: Determine if the syllables carry primary stress or not.

The experimental results are shown in Table 3, which summarizes the total confusions from all runs in the 10-fold cross-validation. It shows that the accuracy of the three-category classification is 87.87%. Many errors are due to the secondary stress misclassification: 44.5% or $(667 + 711)/3100$ of the errors is the case that a secondary-stress syllable is misclassified as a syllable with a primary or no stress. On the other hand, 39.7% or $(527 + 703)/3100$ of the errors is the case that a primary-stress or unstressed syllable is mistaken as a secondary-stress syllable. The syllable-based and word-based accuracies under the above three criteria are shown in Table 4.

Table 3: Results of lexical stress detection from 10-fold cross-validation.

Labeled Detected	PS	SS	NS
PS	5,064	667	217
SS	527	2,751	703
NS	275	711	14,669

5.3.2. Pitch accent detection

The evaluation results of the pitch accent detection from 10-fold cross-validation are summarized in Table 5. Among the total 198,600 syllables, 90.15% of them are correctly identified as either accented or unaccented.

Table 4: Performance of lexical stress detection under different criteria.

	P-S-N	S-N	P-N
Syllable	87.87 ± 1.09	92.54 ± 1.00	93.41 ± 0.77
Word	66.76 ± 2.51	74.61 ± 3.02	84.84 ± 1.64

Note: accuracies (%) are shown in the form of ($\mu \pm d$), where μ is the mean value and d is the sample standard deviation.

Among the syllables that are annotated as accented, 83.82% of them are correctly detected.

Table 5: Results of pitch accent detection from 10-fold cross-validation.

Labeled Detected	Unaccented	Accented
Unaccented	125,952	10,250
Accented	9,307	53,091

Note: the syllables in the intonational phrases with only one monosyllabic word are not counted in this experiment.

6. Analysis

In this section, we examine the influence of MD-DNNs with different structures, the contribution of different features and the performance of different approaches for lexical stress and pitch accent detection. Note that the accuracies in this section are all based on syllables unless it is explicitly stated. For lexical stress detection, the P-S-N criterion is used for evaluation.

6.1. Performance of MD-DNNs with different structures

Figures 7 and 8 demonstrate the performance of MD-DNNs with different structures for lexical stress and pitch accent detection. Figure 7 shows that the MD-DNN with three hidden layers performs quite well when the number of hidden units per layer is 16. The performance is further improved if we use 128 nodes per hidden layer for lexical stress detection and 256 nodes for pitch accent detection. These configurations are applied in subsequent experiments. Note that the MD-DNNs with only 8 nodes per hidden layer show poor performance, whose accuracies are 60.2% for lexical stress detection and 67.8% for pitch accent detection.

Figure 8 presents that using more hidden layers improves the performance of lexical stress detection provided it is less than 4. Comparing these two figures, we observe that the MD-DNN with 3 hidden layers of 16 units significantly outperform the one with a single hidden layer of 128 units, whose accuracies are 86.4% and 80.3% respectively. This result verifies the advantage of using a deep architecture. Note that there are only 30 input and 3 output units (see Section 4). In addition, the number of labeled syllables is only about 25,000 (see Table 1).

However, more hidden layers provided limited performance improvement for pitch accent detection. This may be due to the fact that there are only 2 output units and about 205,000 labeled syllables. Thus, training a binary pitch accent detector is simpler than a three-category lexical stress detector in some sense.

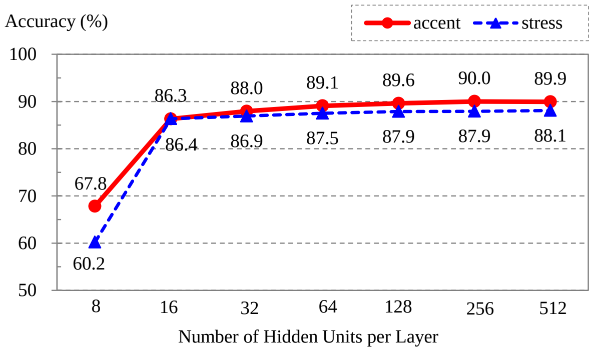

Figure 7: Performance of lexical stress and pitch accent detection by MD-DNNs with different number of hidden units per layer. All these MD-DNNs have three hidden layers.

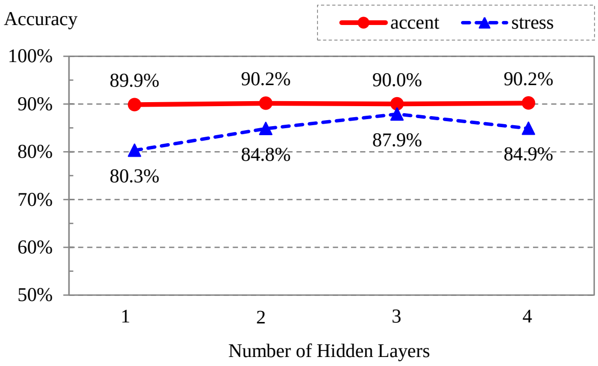

Figure 8: Performance of lexical stress and pitch accent detection by MD-DNNs with different number of hidden layers. Each hidden layer has 128 units for lexical stress detection and 256 units for pitch accent detection.

6.2. *Performance of prosodic features with different length of context window*

As described in Section 3, four syllable-based prosodic features (V_{dur}, V_{loud}, f_{m1}, f_{m2}) are used in this work. Figure 9 shows the performance of these prosodic features with different length of context window. If we only use the prosodic features of the current syllable, the lexical stress detector only obtains an accuracy of about 72.6%. If we leverage the prosodic features of one preceding and one succeeding syllables as well as the syllable in focus, a significant improvement is observed, achieving an accuracy of about 80.2%. The best performance is achieved when the size of context window is five.

Similar performance improvement is observed in pitch accent detection if we vary the length of context window.

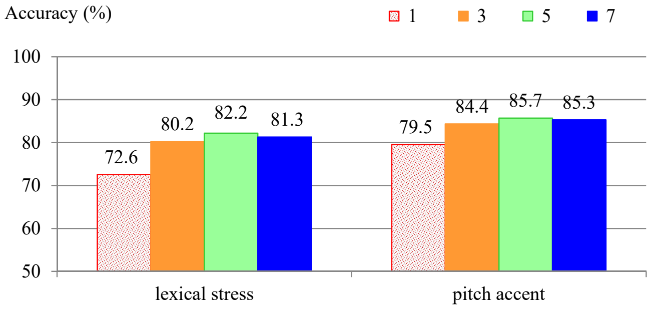

Figure 9: Performance of lexical stress and pitch accent detection using prosodic features with different length of context window.

6.3. *Contribution of different prosodic features*

Figure 10 shows the contribution of the three kinds of prosodic features in lexical stress and pitch accent detection. It shows that the loudness and pitch features have similar performances, both of which are slightly better than the syllable nucleus duration. Combining all these prosodic features gains a performance improvement of 6.5% for lexical stress detection and 7.8% for pitch accent detection. Note that contextual information is used here.

In addition, combining these prosodic features with the differential pitch value (V_{pitch}) in Equation (4b) cannot further improve performance, and the accuracies of lexical stress and pitch accent detection are 80.9% and 84.1% respectively.

6.4. *Contribution of lexical and syntactic features*

Figure 11 shows the contribution of lexical and syntactic features. Only using the syllable-based prosodic features (20 linear units with Gaussian noise) achieves an accuracy of 82.2% for lexical stress detection and 85.7% for pitch accent detection; while only leveraging the lexical and syntactic features obtains an accuracy of 83.7% for lexical stress detection and 83.0% for pitch accent detection. Combing these two kinds of features, the accuracies of lexical stress and pitch accent detection are improved to 87.9% and 90.2% respectively.

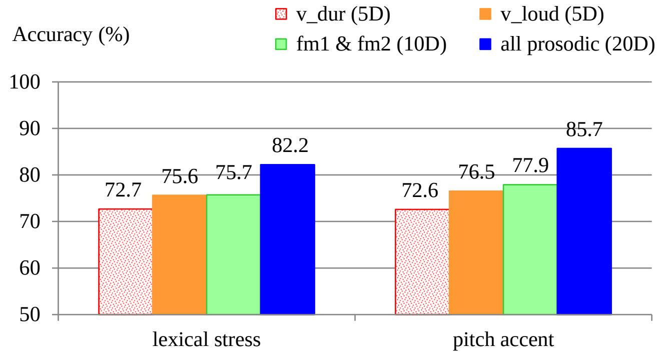

Figure 10: Performance of lexical stress and pitch accent detection using different prosodic features, including syllable nucleus duration (V_{dur}), maximum loudness (V_{loud}) and a pair of dynamic pitches (f_{m1}&f_{m2}).

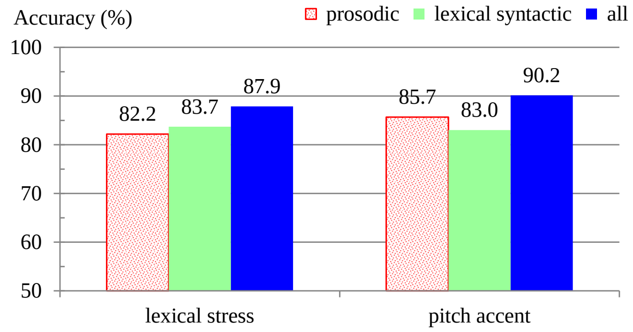

Figure 11: Contribution of lexical and syntactic features towards lexical stress and pitch accent detection.

6.5. *Contribution of unlabeled data*

Table 1 shows that there are only about 25,000 labeled syllables for our experiments of lexical stress detection, which is much smaller than those for pitch accent detection. Since collecting and transcribing L2 English speech are very costly procedures, it is important to be able to leverage unlabeled data, which may be achieved by deep learning techniques. The improvement of MD-DNN using unlabeled data in pre-training is about 1% for lexical stress detection (see Figure 12).

6.6. *Performance of different approaches*

The classifiers for lexical stress and pitch accent detection in Li et al. (2011a) are Gaussian mixture models (GMMs). Two approaches of detection were investigated: one using the syllable-based prosodic features (V_{dur}, V_{loud}, V_{pitch}) and the other using the prominence features from the prominence model (PM). As described in Section 2.5, the PM estimates the prominence values by taking into account the syllable in focus, as well as the syllables in neighboring contexts. Note that both approaches were based on supervised learning. For simplicity in notation, we denote the former approach with GMM and the latter with PM. The results of lexical stress detection using the same data as described in Section 5.1 were updated and reported in Li and Meng (2013).

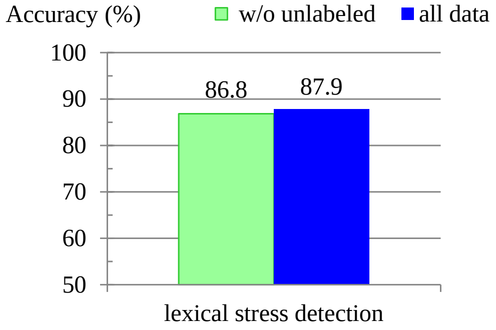

Figure 12: Performance of lexiccal stress detection with and without unlabeled data.

In our previous work (Li and Meng, 2013), the multi-distribution deep belief network (MD-DBN) was proposed for lexical stress detection. The structure of the MD-DBN in Li and Meng (2013) is similar to the one shown in Figure 5. Their main difference is that the top two layers of the MD-DBN form an undirected associative memory (Hinton et al., 2006), while the top two layers of the DNN form an RBM with softmax function. MD-DNN has been shown to have better performance than MD-DBN.

Figure 13 summarizes the performance of using the GMM, PM, MD-DBN and MD-DNN. We observe that the MD-DNN outperforms the PM by 11.6% for lexical stress detection and 6.9% for pitch accent detection, respectively. The MD-DNN outperforms the MD-DBN by about 2.9% for lexical stress detection and 3.1% for pitch accent detection. These are due to that MD-DBN is a generative model, whereas MD-DNN is a discriminative model. MD-DBN only uses the one-step CD in the pre-training and fine-tuning stages; while MD-DNN uses the one-step CD in the pre-training stage and the BP algorithm in the fine-tuning stage (see Section 5.2).

Table 6 shows the word-based accuracies of lexical stress detection using the PM and MD-DNN. For the PM, a syllable-based accuracy of 76.3% is only equivalent to a word-based accuracy of 37.9%. If we evaluate the lexical stress detection based on words, the MD-DNN outperforms the PM significantly by about 29% under the P-S-N criterion.

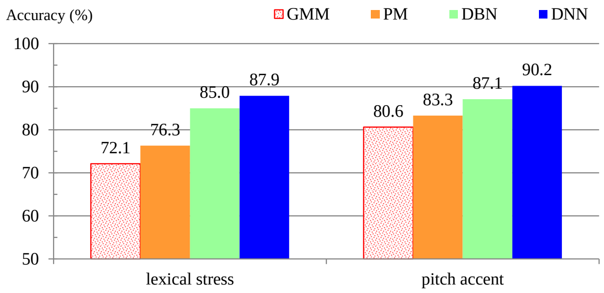

Figure 13: Performance of lexical stress detection and pitch accent detection using the Gaussian mixture models (GMMs), prominence model (PM), deep belief network (DBN) and deep neural network (DNN).

Table 6: Word-based accuracies (%) of lexical stress detection using different approaches under different criteria.

	P-S-N	S-N	P-N
PM	37.88	—	76.88
MD-DNN	66.76	74.61	84.84

7. Conclusions

In this work, we investigate the use of MD-DNNs for automatic lexical stress detection and pitch accent detection, which are useful for suprasegmental mispronunciation detection and diagnosis in L2 English speech. The features used in this work include syllable-based prosodic features (maximum syllable loudness, syllable nucleus duration and a pair of dynamic pitches) as well as their lexical and syntactic features (primary/secondary/no stress in dictionaries). As stressed/accented syllables are more prominent than their neighbors, the two preceding and two following syllables are also taken into consideration. Experimental results show that, for words with three or more syllables, the MD-DNN achieves a syllable-based accuracy of 87.9% under the P-S-N criterion. It outperforms the GMM, PM and MD-DBN by about 15.8%, 11.6% and 2.9% respectively. The pitch accent detector classifies syllables as accented and unaccented with an accuracy of 90.2%, which is better than the results of using the PM and GMM by about 9.6% and 6.9% respectively.

8. Acknowledgements

The work is partially supported by the grant from the HKSAR Government Research Grants Council General Research Fund (project number 14207315).

References

Ananthakrishnan, S., Narayanan, S.S., 2008. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. Audio, Speech, and Language Processing, IEEE Transactions on 16, 216–228.

Arias, J.P., Yoma, N.B., Vivanco, H., 2010. Automatic intonation assessment for computer aided language learning. Speech communication 52, 254–267.

Aull, A., Zue, V.W., 1985. Lexical stress determination and its application to large vocabulary speech recognition, in: Proc. ICASSP.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees. CRC press.

Campbell, N., Beckman, M., 1997. Stress, prominence, and spectral tilt, in: Intonation: Theory, models and applications.

Chen, J.Y., Wang, L., 2010. Automatic lexical stress detection for Chinese learners' of English, in: Proc. ISCSLP.

Chen, K., Hasegawa-Johnson, M., Cohen, A., 2004. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model, in: Proc. ICASSP.

Conkie, A., Riccardi, G., Rose, R.C., 1999. Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events .

Deng, L., Abdel-Hamid, O., Yu, D., 2013. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion, in: Proc. ICASSP.

Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., Precoda, K., 2015. Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. Speech Communication 69, 31–45.

Freij, G., Fallside, F., Hoequist, C., Nolan, F., 1990. Lexical stress estimation and phonological knowledge. Computer Speech & Language 4, 1–15.

Fry, D.B., 1958. Experiments in the perception of stress. Language and speech 1, 126–152.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. Switchboard: Telephone speech corpus for research and development, in: Proc. ICASSP.

Gregory, M.L., Altun, Y., 2004. Using conditional random fields to predict pitch accents in conversational speech, in: Proc. ACL.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., Ng, A.Y., 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567v2 .

Hinton, G., Osindero, S., Teh, Y., 2006. A fast learning algorithm for deep belief nets. Neural Computation 18, 1527–1554.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 .

Hirschberg, J., 1993. Pitch accent in context predicting intonational prominence from text. Artificial Intelligence 63, 305–340.

Imoto, K., Tsubota, Y., Raux, A., Kawahara, T., Dantsuji, M., 2002. Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system, in: Proc. ICSLP.

Jenkin, K.L., Scordilis, M.S., 1996. Development and comparison of three syllable stress classifiers, in: Proc. ICSLP.

Jeon, J.H., Liu, Y., 2009. Automatic prosodic events detection using syllable-based acoustic and syntactic features, in: Proc. ICASSP.

Kang, S., Qian, X., Meng, H., 2013. Multi-distribution deep belief network for speech synthesis, in: Proc. ICASSP.

van Kuijk, D., Boves, L., 1999. Acoustic characteristics of lexical stress in continuous telephone speech. Speech Communication 27, 95–111.

Ladd, D.R., 2008. Intonational phonology. Cambridge University Press.

Levow, G.A., 2008. Automatic prosodic labeling with conditional random fields and rich acoustic features, in: Proc. IJCNLP.

Lewis, R.J., 2000. An introduction to classification and regression tree (CART) analysis, in: Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California, pp. 1–14.

Li, C., Liu, J., Xia, S., 2007. English sentence stress detection system based on HMM framework. Applied mathematics and computation 185, 759–768.

Li, K., Liu, J., 2010a. English sentence accent detection based on auditory features. Journal of Tsinghua University Science and Technology 50, 613–617.

Li, K., Liu, J., 2010b. Pitch extraction based on wavelet transformation and linear prediction. Computer Engineering 36.

Li, K., Meng, H., 2012. Perceptually-motivated assessment of automatically detected lexical stress in L2 learners' speech, in: Proc. ISCSLP.

Li, K., Meng, H., 2013. Lexical stress detection for L2 English speech using deep belief networks, in: Proc. Interspeech.

Li, K., Meng, H., 2014. Mispronunciation detection and diagnosis in L2 English speech using multi-distribution deep neural networks, in: Proc. ISCSLP.

Li, K., Qian, X., Kang, S., Liu, P., Meng, H., 2015. Integrating acoustic and state-transition models for free phone recognition in L2 English speech using multi-distribution deep neural networks, in: Proc. SLaTE.

Li, K., Zhang, S., Li, M., Lo, W.K., Meng, H., 2010. Detection of intonation in L2 English speech of native Mandarin learners, in: Proc. ISCSLP.

Li, K., Zhang, S., Li, M., Lo, W.K., Meng, H., 2011a. Prominence model for prosodic features in automatic lexical stress and pitch accent detection, in: Proc. Interspeech.

Li, M., Zhang, S., Li, K., Harrison, A., Lo, W.K., Meng, H., 2011b. Design and collection of an L2 English corpus with a suprasegmental focus for Chinese learners of English, in: Proc. ICPhS.

Lieberman, P., 1960. Some acoustic correlates of word stress in American English. The Journal of the Acoustical Society of America 32, 451–454.

Meng, H., Tseng, C.Y., Kondo, M., Harrison, A., Viscelgia, T., 2009. Studying L2 suprasegmental features in asian Englishes: A position paper, in: Proc. Interspeech.

Morton, J., Jassem, W., 1965. Acoustic correlates of stress. Language and Speech 8, 159–181.

Ni, C.J., Liu, W., Xu, B., 2011. Automatic prosodic events detection by using syllable-based acoustic, lexical and syntactic features, in: Proc. Interspeech.

Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S., 1995. The Boston University radio news corpus. Linguistic Data Consortium , 1–19.

Pulgram, E., 1970. Syllable, Word, Nexus, Cursus. Mouton.

Qian, X.j., Meng, H., Soong, F., 2012. The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training, in: Proc. Interspeech.

Qian, Y., Wu, Z., Ma, X., Soong, F., 2010. Automatic prosody prediction and detection with conditional random field (CRF) models, in: Proc. ISCSLP.

Ren, Y., Kim, S.S., Hasegawa-Johnson, M., Cole, J., 2004. Speaker-independent automatic detection of pitch accent, in: Proc. Speech Prosody.

Rosenberg, A., 2009. Automatic detection and classification of prosodic events. Ph.D. thesis. Columbia University.

Ross, K., Ostendorf, M., 1996. Prediction of abstract prosodic labels for speech synthesis. Computer Speech & Language 10, 155–185.

Ross, K., Ostendorf, M., Shattuck-Hufnagel, S., 1992. Factors affecting pitch accent placement, in: Proc. ICSLP.

Rumelhart, D.E., Hinton, G., Williams, R.J., 1986. Learning representations by back-propagating errors. Natrue 323, 533–536.

Seltzer, M.L., Yu, D., Wang, Y., 2013. An investigation of deep neural networks for noise robust speech recognition, in: Proc. ICASSP.

Sluijter, A.M., van Heuven, V.J., 1996. Spectral balance as an acoustic correlate of linguistic stress. The Journal of the Acoustical society of America 100, 2471–2485.

Sridhar, V.R., Bangalore, S., Narayanan, S.S., 2008. Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. Audio, Speech, and Language Processing, IEEE Transactions on 16, 797–811.

Sun, X.J., 2002. Pitch accent prediction using ensemble machine learning, in: Proc. ICSLP.

Tamburini, F., 2003. Prosodic prominence detection in speech, in: Proc. Signal Processing and its Applications.

Tamburini, F., Bertini, C., Bertinetto, P.M., 2014. Prosodic prominence detection in Italian continuous speech using probabilistic graphical models, in: Proc. Speech Prosody.

Taylor, P., 1995. The rise/fall/connection model of intonation. Speech Communication 15, 169–186.

Taylor, P., 1998. The Tilt intonation model, in: Proc. ISCLP.

Taylor, P., 2006. Analysis and synthesis of intonation using the tilt model. Journal of the Acoustical Society of America 107, 1697–1714.

Tepperman, J., Narayanan, S., 2005. Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners, in: Proc. ICASSP.

Waibel, A., 1986. Recognition of lexical stress in a continuous speech understanding system-a pattern recognition approach, in: Proc. ICASSP.

Wang, D., Narayanan, S., 2007. An acoustic measure for word prominence in spontaneous speech. IEEE Trans. Speech and Audio Proc. 15, 690–701.

Wang, J.F., Chang, G.M., Wang, J.C., Lin, S.C., 2009. Stress detection based on multi-class probabilistic support vector machines for accented English speech, in: Proc. CSIE.

Wightman, C.W., Ostendorf, M., 1994. Automatic labeling of prosodic patterns. Speech and Audio Processing, IEEE Transactions on 2, 469–481.

Xie, H., Andreae, P., Zhang, M., Warren, P., 2004. Detecting stress in spoken English using decision trees and support vector machines, in: Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation.

Ying, G.S., Jamieson, L.H., Chen, R., Michell, C.D., Liu, H., 1996. Lexical stress detection on stress-minimal word pairs, in: Proc. ICSLP.

Zhao, J., Xu, J., Zhang, W.q., Yuan, H., Liu, J., Xia, S., 2013a. Exploiting articulatory features for pitch accent detection. Journal of Zhejiang University SCIENCE C 14, 835–844.

Zhao, J., Yuan, H., Liu, J., Xia, S., 2011. Automatic lexical stress detection using acoustic features for computer assisted language learning, in: Proc. APSIPA ASC.

Zhao, J., Zhang, W.Q., Yuan, H., Johnson, M.T., Liu, J., Xia, S., 2013b. Exploiting contextual information for prosodic event detection using auto-context. EURASIP Journal on Audio, Speech, and Music Processing 2013, 1–14.

Zwicker, E., Fastl, H., 1999. Psychoacoustics — Facts and Models 2nd Updated Edition. Springer.