Lexical Stress Detection for L2 English Speech Using Deep Belief Networks

Kun Li, Xiaojun Qian, Shiyin Kang and Helen Meng

Human-Computer Communications Laboratory

Department of System Engineering and Engineering Management

The Chinese University of Hong Kong, Hong Kong SAR, China

{kli, xjqian, sykang, hmmeng}@se.cuhk.edu.hk

Abstract

This paper investigates lexical stress detection for L2 English speech using Deep Belief Networks (DBNs). The features of the DBN used in this work include the syllable-based prosodic features (assumed to have Gaussian distribution) and their expected lexical stress (assumed to have Bernoulli distribution). As stressed syllables are more prominent than their neighbors, the two preceding and two following syllables are taken into consideration. Experimental results show that the DBN achieves an accuracy of about 80% in syllable stress classification (primary/secondary/no stress) for words with three or more syllables. It outperforms the conventional Gaussian Mixture Model and our previous Prominence Model by an absolute accuracy of about 8% and 4%, respectively.

Index Terms: lexical stress detection, deep belief network, L2 English speech

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 such as lexical stress, narrow focus, reduction / non-reduction of function words, intonation of a sentence, as well as prosodic disambiguation in suprasegmental phonology that deserve attention from a Chinese learner of English. This paper focuses on the detection of *lexical stress in a word*.

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*".

To develop a Computer-Assisted Pronunciation Training (CAPT) system that can help learners train their lexical stress productions, we need to begin by detecting lexical stress in the L2 learners' speech, i.e. to identify the syllable carrying Primary Stress (PS), Secondary Stress (SS), or No Stress (NS) at all. In [3], lexical stress detection is the key module for lexical stress assessment of L2 English speech—lexical stress needs to be detected before we apply an appropriate criterion to assess the overall word-level stress pattern.

Previous research has presented various features and approaches on the automatic detection of lexical stress. In the study of syllable stress detection for German and Italian, Tepperman [4] used the mean values of fundamental frequency (f_0), syllable nucleus duration, energy and other features related to f_0 slope and RMS energy range. Imoto [5] developed Hidden Markov Models (HMMs) to detect stress in English sentences read by Japanese students. Tamburini [6] combined the detection of lexical stress and pitch accents into a task of prominence detection. Stress detection was based on syllable nucleus duration and high-frequency features. Our work in [7] used a set of syllable-based prosodic features and proposed a Prominence Model for lexical stress detection and pitch accent detection. The Prominence Model estimates the prominence values from the syllable in focus, as well as the syllables in neighboring contexts.

Various approaches have previously been applied to lexical stress detection. Results show that such detection is a challenging task, especially for words with three or more syllables. If we evaluate the lexical stress detection at the word level, 80% syllable-based accuracy equals to about 40% word-based accuracy ($0.8^4 \approx 0.4$, assuming syllables are independent) [3]. Perceptual tests in [3] and [8] show that even humans may not be able to correctly identify the stress patterns in native English speech with high accuracy. These tests were conducted with 58 listeners whose mother tongue is Mandarin, 25 whose mother tongue is Cantonese and 25 whose mother tongue is US English. 30 words covering different stress patterns were recorded by a native American English speaker and were presented to each listener. Results show that the overall average word-based accuracy they achieved is only about 30%. For English words with five or more syllables, the Cantonese and even native US English listeners achieved less than 10% word-based identification accuracies.

Recently, the development of highly effective learning techniques for Deep Belief Networks (DBNs) draws much attention to the neural network research. In [9], Hinton proposed a fast learning algorithm for a DBN model in which the top two hidden layers form an undirected associative memory and the remaining hidden layers form a directed acyclic graph. Due to the effective, DBNs have been applied to speech recognition [10] [11] [12] and synthesis [13], and achieve impressive performance gains.

In this work, we use DBNs to detect the lexical stress of L2 English speech. Generally it is expensive to collect and transcribe the L2 English speech. DBNs offer the advantage of enabling the use of unlabeled data. We present our work with the following organization: Section 2 describes the syllable-based prosodic features for lexical stress detection. Section 3 introduces DBNs and specifies the structure of the DBN in our work. Sections 4 and 5 present our experiments and analysis respectively. Conclusions are given in Section 6.

2. Syllable-based prosodic features

Stressed syllables usually exhibit greater loudness, longer duration and higher pitch than their neighbors [6]. In this section, we introduce the syllable-based prosodic features for DBNs: maximum syllable loudness, syllable nucleus duration and two extreme pitch values. These features were first proposed in [7].

2.1. Syllable nucleus duration (V_{dur})

We first apply the Maximal Onset Principle [14] to extract the syllables from the phoneme sequence output of the speech recognizer. 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 Fig.1

Within the time boundaries of every extracted syllable, we treat the frames whose loudness fall above N_{bot}as the syllable nuclei, where N_{bot} is the value above which 50% of all loudness values in the utterance. The normalized syllable nucleus duration V_{dur} is given as:

$$V_{dur} = d_{dur} - \overline{d_{wd}} \tag{1}$$

where d_{dur} is the syllable nucleus duration, $\overline{d_{wd}}$ is the mean duration of all syllable nuclei in the word.

2.2. Maximum syllable loudness (V_{loud})

Loudness is the human perception of the strength of sound energy. There is a complex relationship between human perception of loudness and sound energy. We follow Zwicker's loudness model [15] for a precise estimation of loudness. We use simplifying calculation of loudness based on Zwicker's model in [7], which works well for stress and pitch accent detection.

The normalized maximum syllable loudness V_{loud}, as given by Eq. (2), is taken as our feature.

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

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

2.3. Extreme pitch values in a syllable (f_{m1} & f_{m2})

We first perform pitch extraction [16] 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 match human perception of pitch.

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

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

A differential pitch value is proposed in [7], as given by Eq. (4a). It is based on the observations: syllables with rising tones often give a stressed perception; while syllables with falling tones are often perceived as unstressed.

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

where $f_{m1}$is the first (in time sequence) extreme pitch value in the syllable nucleus and $f_{m2}$is the second extreme pitch value in the syllable nucleus, as shown in Fig. 1.

Eq. (4a) can be further improved to Eq. (4b), which was used in the experiments of [7]. Results showed that the differential pitch value outperforms the mean or maximum pitch value in a syllable by about 5% or 3% respectively.

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

In this work, we only use the two extreme pitch values (f_{m1} and f_{m2}) in a syllable nucleus instead of the differential pitch value (V_{pitch}), as we believe that DBNs can optimize the performance by automatically adjusting the relationship between $f_{m1}$and f_{m2}.

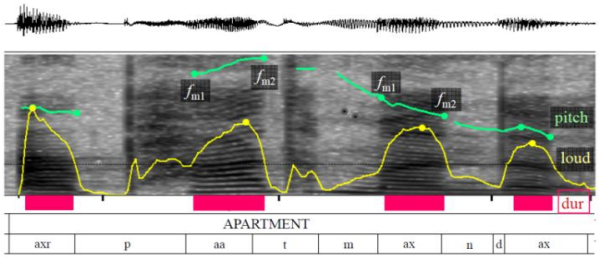

Figure 1: *An example of feature extraction for lexical stress detection. The yellow curve is loudness, the green curve is pitch in semitone and the red bars indicate the syllable nuclei duration. f_{m1} and f_{m2} are also marked for the syllables of /p aa t/ and /m ax n/.*

3. Multi-Distribution Deep Belief Network (MD-DBN)

A Restricted Boltzmann Machine (RBM) is a type of undirected graphical model constructed from a hidden layer and a visible layer. Generally, two types of RBM are commonly used in speech processing: (1) Bernoulli RBMs, whose hidden and visible units are all binary; and (2) Gaussian-Bernoulli RBMs whose hidden units are binary but visible units are Gaussian distributed [10][11][12]. Derived from the above two types of RBM, a type of Mixed Gaussian-Bernoulli RBM [13] is also used in this work.

3.1. Bernoulli RBM (B-RBM)

The energy of the joint configuration of visible and hidden vector (**v**, **h**) is given as:

$$\mathrm{E}(\mathrm{v}, \mathrm{h};\ \boldsymbol{\Theta}) = -\mathbf{h}^{\mathrm{T}}\mathbf{W}\mathbf{v} - \mathbf{a}^{\mathrm{T}}\mathbf{h} - \mathbf{b}^{\mathrm{T}}\mathbf{v} \tag{5}$$

where $\boldsymbol{\Theta} = (\mathbf{W}, \mathbf{a}, \mathbf{b})$ is the set of parameters of an RBM and $\boldsymbol{\Theta}$ will be omitted for clarity hereafter. **W** is the matrix of visible/hidden connection weights, **a** is the hidden unit bias, **b** is the visible unit bias.

The probability is given in term of the energy:

$$\mathrm{P}(\mathbf{v}, \mathbf{h}) = \frac{\mathrm{e}^{-\mathrm{E}(\mathbf{v},\ \mathbf{h})}}{\sum_{\tilde{\mathbf{v}}}\sum_{\tilde{\mathbf{h}}}\mathrm{e}^{-\mathrm{E}(\tilde{\mathbf{v}}, \tilde{\mathrm{h}})}} \tag{6}$$

Since there are no connections within a layer, we can have the following equations [11]:

$$\mathrm{P}(\mathbf{h}|\ \mathbf{v}) = \frac{\mathrm{e}^{-\mathrm{E}(\mathbf{v},\ \mathbf{h})}}{\sum_{\tilde{\mathbf{h}}}\mathrm{e}^{-\mathrm{E}(\mathbf{v},\ \tilde{\mathrm{h}})}} \tag{7a}$$

$$= \prod_{\mathrm{j}} \mathrm{P}(h_{\mathrm{j}}|\ \mathbf{v}) \tag{7b}$$

$$\mathrm{P}(h_{\mathrm{j}} = 1|\ \mathbf{v}) = \sigma(\textstyle\sum_{\mathrm{i}} \omega_{\mathrm{ij}}\, v_{\mathrm{i}} + a_{\mathrm{j}}) \tag{8a}$$

$$\mathrm{P}(v_{\mathrm{i}} = 1|\ \mathbf{h}) = \sigma(\textstyle\sum_{\mathrm{j}} \omega_{\mathrm{ij}}\, h_{\mathrm{j}} + b_{\mathrm{i}}) \tag{8b}$$

where $\sigma(\mathrm{x}) = (1 + \mathrm{e}^{-\mathrm{x}})^{-1}$.

The log probability of a given visible vector $\mathbf{v}^l$ is:

$$\log \mathrm{P}(\mathbf{v}^l) = \log \textstyle\sum_{\tilde{\mathbf{h}}} \mathrm{e}^{-\mathrm{E}(\mathbf{v}^l,\ \tilde{\mathbf{h}})} - \log \sum_{\tilde{\mathbf{v}}}\sum_{\tilde{\mathbf{h}}} \mathrm{e}^{-\mathrm{E}(\tilde{\mathbf{v}},\ \tilde{\mathrm{h}})} \tag{9}$$

To optimize $\log \mathrm{P}(\mathbf{v}^l)$ in a first-order approach, we need the gradient of it with respect to any θ in $\boldsymbol{\Theta}$ [12][13]:

$$\frac{\partial \log \mathrm{P}(\mathbf{v}^l)}{\partial \theta} = \sum_{\mathbf{h}} \frac{\mathrm{e}^{-\mathrm{E}(\mathbf{v}^l,\ \mathbf{h})}}{\sum_{\tilde{\mathbf{h}}}\mathrm{e}^{-\mathrm{E}(\mathbf{v}^l,\ \tilde{\mathrm{h}})}} \frac{\partial -\mathrm{E}(\mathbf{v}^l, \mathbf{h})}{\partial \theta} - \sum_{\mathbf{v}}\sum_{\mathbf{h}} \frac{\mathrm{e}^{-\mathrm{E}(\mathbf{v},\ \mathbf{h})}}{\sum_{\tilde{\mathbf{v}}}\sum_{\tilde{\mathbf{h}}}\mathrm{e}^{-\mathrm{E}(\tilde{\mathbf{v}},\ \tilde{\mathrm{h}})}} \frac{\partial -\mathrm{E}(\mathbf{v}, \mathbf{h})}{\partial \theta}$$

$$= \sum_{\mathbf{h}} \mathrm{P}(\mathbf{h}|\mathbf{v}^l) \frac{\partial -\mathrm{E}(\mathbf{v}^l, \mathbf{h})}{\partial \theta} - \sum_{\mathbf{v}}\sum_{\mathbf{h}} \mathrm{P}(\mathbf{v}, \mathbf{h}) \frac{\partial -\mathrm{E}(\mathbf{v}, \mathbf{h})}{\partial \theta} \tag{10}$$

Take $\theta = w_{ij}$ for example. The first term in Eq. (10) is:

$$\sum_{\mathbf{h}} v_i^l h_j \mathrm{P}(\mathbf{h}|\mathbf{v}^l) = \sum_{\mathbf{h}} \left(v_i^l h_j \prod_k \mathrm{P}(h_k|\mathbf{v}^l)\right)$$
$$= \left(\sum_{h_1} \mathrm{P}(h_1|\mathbf{v}^l)\right)\left(\sum_{h_2} \mathrm{P}(h_2|\mathbf{v}^l)\right)\cdots\left(\sum_{h_j} v_i^l h_j \mathrm{P}(h_j|\mathbf{v}^l)\right)\cdots$$
$$= v_i^l \mathrm{P}(h_j = 1|\mathbf{v}^l) \quad (11)$$

Hence, given the instantiated observation $\mathbf{v}^l$, the expectation of derivatives in the first term in Eq. (10) can be easily computed. Unfortunately, the second term in Eq. (10) involves a summation over all possible $\mathbf{v}$ and is intractable. A widely applied method that approximates this summation is the Gibbs sampler which proceeds in a Markov chain as follows:

$$\mathbf{v}^{(0)} \sim \mathbf{v}^l, \qquad \mathbf{h}^{(0)} \sim \mathrm{P}(\mathbf{h}|\mathbf{v}^{(0)}); \quad (12a)$$

$$\mathbf{v}^{(1)} \sim \mathrm{P}(\mathbf{v}|\mathbf{h}^{(0)}), \qquad \mathbf{h}^{(1)} \sim \mathrm{P}(\mathbf{h}|\mathbf{v}^{(1)}); \quad (12b)$$

…

Given a set of N syllables $\{\mathbf{v}^l\}_{l=1}^N$, the gradient of the log probability of the training data is [9]:

$$\frac{1}{N}\sum_{\mathbf{v}^l} \frac{\partial \log \mathrm{P}(\mathbf{v}^l)}{\partial \omega_{ij}} = \frac{1}{N}\sum_{\mathbf{v}^l}\sum_{\mathbf{h}} v_i^l h_j \mathrm{P}(\mathbf{h}|\mathbf{v}^l) - \sum_{\mathbf{v}}\sum_{\mathbf{h}} v_i h_j \mathrm{P}(\mathbf{v},\mathbf{h})$$
$$= \frac{1}{N}\sum_{\mathbf{v}^l} v_i^l \mathrm{P}(h_j = 1|\mathbf{v}^l) - \sum_{\mathbf{v}}\sum_{\mathbf{h}} v_i h_j \mathrm{P}(\mathbf{v},\mathbf{h})$$
$$= \langle v_i^{(0)} h_j^{(0)}\rangle - \langle v_i^{(\infty)} h_j^{(\infty)}\rangle \quad (13)$$

where $\langle\cdot\rangle$ denotes an average over the sampled states.

In practice, we use the one-step contrastive divergence approximation for the gradient [9]:

$$\frac{1}{\mathrm{N}}\sum_{\mathbf{v}^l} \frac{\partial \log \mathrm{P}(\mathbf{v}^l)}{\partial \omega_{\mathrm{ij}}} \approx \langle v_i^{(0)} h_j^{(0)}\rangle - \langle v_i^{(1)} h_j^{(1)}\rangle \quad (14)$$

where $\langle v_i^{(1)} h_j^{(1)}\rangle$ is the expectation over one-step reconstruction.

3.2. Mixed Gaussian-Bernoulli RBM (GB-RBM)

The GB-RBM has one layer of stochastic hidden binary units and one layer of visible units, some of which are assumed to have Gaussian distribution and the others are binary. The energy of the joint configuration of the visible and hidden vectors $(\mathbf{v}^{\boldsymbol{g}}, \mathbf{v}^{\boldsymbol{b}}, \mathbf{h})$ is given as:

$$\mathrm{E}(\mathbf{v}^{\boldsymbol{g}}, \mathbf{v}^{\mathbf{b}}, \mathbf{h}) = -\mathbf{h}^{\mathrm{T}}\mathbf{W}^{\boldsymbol{g}}\mathbf{v}^{\boldsymbol{g}} + \frac{1}{2}(\mathbf{v}^{\boldsymbol{g}} - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{v}^{\boldsymbol{g}} - \boldsymbol{\mu})$$
$$-\mathbf{h}^{\mathrm{T}}\mathbf{W}^{\boldsymbol{b}}\mathbf{v}^{\mathbf{b}} - \mathbf{b}^{\mathrm{T}}\mathbf{v}^{\mathbf{b}} - \mathbf{a}^{\mathrm{T}}\mathbf{h} \quad (15)$$

where $\mathbf{v}^{\boldsymbol{g}}$and $\mathbf{v}^{\boldsymbol{b}}$are the Gaussian units and the Bernoulli units in the visible layer, $\mathbf{W}^{\boldsymbol{g}}$ and $\mathbf{W}^{\boldsymbol{b}}$ are the respective weight matrices, $\boldsymbol{\mu}$ is the mean of $\mathbf{v}^{\boldsymbol{g}}$, $\mathbf{a}$ and $\mathbf{b}$ are bias terms of $\mathbf{h}$ and $\mathbf{v}^{\mathbf{b}}$. The conditional $\mathrm{P}(\mathbf{h}|\,\mathbf{v}^{\boldsymbol{g}}, \mathbf{v}^{\boldsymbol{b}})$ can be derived as:

$$\mathrm{P}(h_j = 1|\,\mathbf{v}^{\boldsymbol{g}}, \mathbf{v}^{\boldsymbol{b}}) = \sigma(\textstyle\sum_i \mathrm{w}_{\mathrm{ij}}^{g} v_i^{g} + \sum_i \mathrm{w}_{\mathrm{ij}}^{b} v_i^{b} + a_{\mathrm{j}}) \quad (16)$$

And $\mathrm{P}(v_i^b = 1|\,\mathbf{h})$ follows Eq. (8b). The conditional distribution $\mathrm{P}(v_i^g = 1|\,\mathbf{h})$ is:

$$\mathrm{P}(v_i^g = 1|\,\mathbf{h}) = \mathcal{N}(v_i^g;\, \textstyle\sum_{\mathrm{j}} \omega_{\mathrm{ij}} h_{\mathrm{j}} + \mu_{\mathbf{i}}, 1) \quad (17)$$

3.3. Architecture of MD-DBN

We use the following syllable-based prosodic features as described in Section 2: maximum syllable loudness (V_{loud}), syllable nucleus duration (V_{dur}) and two extreme pitch values (f_{m1} and f_{m2}). These features are normalized to zero mean and unit variance. As stressed syllables are more prominent than their neighbors, the two preceding and two following syllables are taken into consideration. Hence there are total 20 Gaussian visible units in the bottom of the DBN, as shown in Fig. 2.

We also include the expected lexical stress for each syllable: four bits to indicate this syllable NS, PS, SS or NULL. The bit of NULL is true when there is no syllable, e.g. for the first syllable in a word, there are no preceding syllables. For the syllable in focus, the bit of NULL is excluded, because it would be always false. Hence there are 19 binary visible units in the bottom of the DBN. Take the syllable /*p aa t*/ in Fig. 1 for example, the 19 binary values are: (0001 1000 010 1000 1000).

The DBN used in this work is shown in Fig. 2. There are four hidden layers, including the top-layer. It is similar to the construction in [9] and [13].

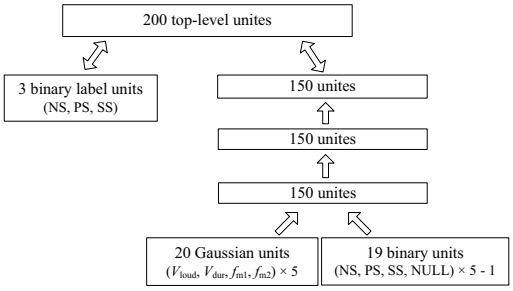

Figure 2: *Architecture of the MD-DBN for lexical stress detection.*

4. Experiments

4.1. Corpus

Our experiments are based on a suprasegmental corpus that we have collected [17]. It contains English speech recordings from 100 Mandarin speakers and 100 Cantonese speakers. There are six parts in this corpus, and only one has syllables labeled with PS/SS/NS. In this part, each speaker utters 28 words, which results in 5,600 words in total. Table 1 shows that the labeled data constitutes about 20% of the entire corpus.

TIMIT is a corpus containing English speech recording from 630 US English speakers. As we try to detect the lexical stress of L2 English speech, we use the TIMIT corpus as unlabeled data for pre-training.

Table 1 summaries the details of the data used in our experiments. Bisyllabic words are excluded from this study due to their simplicity.

Table 1. *Details of corpus used in our experiments.*

	Cantonese		Mandarin		TIMIT	
	Syl.	Word	Syl.	Word	Syl.	Word
Unlabeled	45.7 k	14.5 k	45.8 k	14.5 k	20.0 k	5.8k
Labeled	12.1 k	2.8 k	12.1 k	2.8 k	-	-

Note: Syllable (word) counts are measured in the unit of thousands (k).

4.2. DBN training

In the pre-training stage, we maximize the log-likelihood of RBMs using stochastic gradient ascent for 20 epochs with a batch size of 128 frames. For the GRBM, a learning rate of η = 0.0025 is used for $\mathbf{W}$, $\mathbf{a}$, $\mathbf{b}$. A learning rate of 0.005 is used for all the parameters of BRBMs. Increment in each batch is smoothed by a momentum of $\gamma = 0.9$, which leads to the following update rule for the t^{th} increment of θ: $\Delta\theta^{(t+1)} = \Upsilon\Delta\theta^{(t)} + \eta\frac{\partial\mathcal{L}}{\partial\theta}$, where $\frac{\partial\mathcal{L}}{\partial\theta}$ is the gradient.

In the fine-tuning stage, we also used a 20 epochs with a batch size of 128 frames. The learning rates of η for GRBM and BRBMs are 0.005 and 0.01 respectively.

4.3. Experimental results

The experimental results are shown in Table 2, which summarizes the total confusions from all runs in the 10-fold cross-validation. We use the following three criteria for evaluation [7]:

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

The accuracies under the P-S-N, S-N and P-N criteria are 80.17%, 86.28% and 87.09%, respectively.

Table 2. *Lexical stress detection results from 10-fold cross-validation.*

Annotation / Detection	NS	SS	PS
NS	13440	985	715
SS	695	1585	432
PS	932	1050	4411

5. Analysis

In this section, we examine the influence of the number of hidden units, the number of epochs and the effect of pre-training on the performance of lexical stress detection.

5.1. Number of hidden units

Table 3 shows that the DBN performs quite well when the number of hidden units in each layer is (25, 25, 25, 50). The performance can be further improved if we use (150, 150, 150, 200) hidden units, which are applied in subsequent experiments. Table 3 also shows that further increase in the number of hidden units beyond (200, 200, 200, 300) may cause overfitting.

Table 3. *Performance of DBNs with different numbers of hidden units.*

# of Hidden Units	P-S-N	S-N	P-N
(25, 25, 25, 50)	78.45%	84.93%	86.09%
(50, 50, 50, 100)	79.23%	85.65%	86.33%
(100, 100, 100, 150)	79.23%	85.65%	86.46%
(150, 150, 150, 200)	79.78%	**86.03%**	87.00%
(200, 200, 200, 300)	**79.84%**	85.88%	**87.09%**
(300, 300, 300, 400)	79.28%	85.47%	86.78%

Note: 25 epochs are used for all above experiments.

5.2. Number of epochs

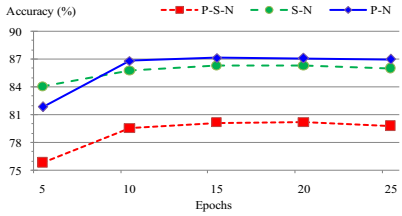

Figure 3. *Accuracies of lexical stress detection as a function of the number of epochs.*

Fig. 3 shows that the performance improves greatly from 5 epochs to 10 epochs. No further improvement can be gained beyond 20 epochs. Hence 20 epochs are used in subsequent experimentations.

5.3. Contribution of pre-training

Table 4 shows the experimental results with and without pre-training. It shows that using unlabeled data for pre-training improves the performance by about 4%.

Table 4. *Results with and without pre-training.*

	P-S-N	S-N	P-N
Without Pre-training	76.17%	82.82%	84.00%
With Pre-training	**80.17%**	**86.28%**	**87.09%**

Note: 20 epochs are used for both experiments, where they achieve best performance.

5.4. Comparing DBN with previous models

The classifiers for lexical stress detection in [7] are Gaussian Mixture Model (GMM). 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). 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 are based on supervised learning. For simplicity in notation, we denote the former approach with GMM and the latter with PM.

Table 5 summarizes the performance of using the GMM, PM and DBN. We observe that the DBN outperforms the PM by about 4% under the P-S-N and S-N criteria, while the PM performs better than the DBN by about 2% under the P-N criterion. These may be due to the fact that the DBN is optimized under the P-S-N criterion, while the PM is optimized under the P-N criterion. By comparing Table 5 with Table 4, we can see that leveraging unlabeled data is the key advantage of the DBN over the PM.

Table 5. *Performance of GMM, PM and DBN.*

	P-S-N	S-N	P-N
GMM	72.11%	78.61%	87.90%
PM	76.31%	80.69%	**89.30%**
DBN	**80.17%**	**86.28%**	87.09%

Note: The accuracies of GMM and PM are slightly different from that in [7], which is due to their different test data.

6. Conclusions

In this paper, we investigate lexical stress detection for L2 English speech using DBNs. The features of the DBN used in this work include syllable-based prosodic features (maximum syllable loudness, syllable nucleus duration and two extreme pitch values) and their expected lexical stress (PS/SS/NS/NULL), which are assumed in Gaussian and Bernoulli distribution, respectively. As stressed syllables are more prominent than their neighboring syllables, the two preceding and two following syllables are also taken into consideration. Experimental results show that, for words with three or more syllables, the DBN achieves an accuracy of about 80% under the P-S-N criterion, which outperforms the GMM and PM by about 8% and 4%, respectively. Experiments also show that using unlabeled data for pre-training can improve the performance by about 4%.

7. Acknowledgements

The work is jointly supported by the Shun Hing Institute of Advanced Engineering and the NSFC/RGC Joint Research Scheme (Project No. N_CUHK 414/09).

8. References

[1] Anderson-Hsieh, J., Johnson, R. and Koehler, K., "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., Tseng, C., Kondo, M., Harrison, A. and, Viscelgia T., "Studying L2 suprasegmental features in Asian Englishes: a position paper", Proc. of INTERSPEECH 2009.

[3] Li, K. and Meng, H., "Perceptually-motivated assessment of automatically detected lexical stress in L2 learners' speech", Proc. of ISCSLP 2012.

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

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

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

[7] Li, K., Zhang, S., Li, M., Lo, W. and Meng, H., "Prominence model for prosodic features in automatic lexical stress and pitch accent detection," in Proc. of INTERSPEECH, 2011.

[8] Zhang, S., Li, K., Lo, W. and Meng, H., "Perception of English suprasegmental features by non-native Chinese learners," in Proc. of Int. Conf. on Speech Prosody, 2010

[9] Hinton, G.E., Osindero, S. and Teh, Y., "A fast learning algorithm for deep belief nets", Neural Computation, vol. 18, 2006.

[10] Mohamed, A., Dahl, G.E. and Hinton G.E., "Acoustic modeling using deep belief networks", IEEE Trans. on Audio, Speech and Language Proc., 2012.

[11] Dahl, G.E., Yu, D., Deng, L. and Acero A., "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition", IEEE Trans. on Audio, Speech and Language Proc., 2012.

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

[13] Kang, S., Qian X. and Meng, H., "Multi-distribution deep belief network for speech synthesis", Proc. of ICASSP 2013.

[14] Pulgram, E., "Syllable, word, nexus, cursus", Mouton, 1970.

[15] Zwicker, E. and Fastl H., "Psychoacoustics: facts and models", 2nd Edition, Springer, 1999.

[16] Li, K. and Liu, J.., "Pitch extraction based on wavelet transformation and linear prediction", Computer Engineering, 2010.

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