The Use of DBN-HMMs for Mispronunciation Detection and Diagnosis in L2 English to Support Computer-Aided Pronunciation Training

Xiaojun Qian[1], *Helen Meng*[1] *and Frank Soong*[1,2]

[1]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Speech Group, Microsoft Research Asia, Beijing, China
{xjqian,hmmeng}@se.cuhk.edu.hk, frankkps@microsoft.com

Abstract

This paper investigates acoustic modeling using the hybrid DBN-HMM framework in mispronunciation detection and diagnosis of L2 English. This is one of the first efforts that compare the performance of DBN-HMM with that of the best-tuned GMM-HMM trained in ML and MWE on the same set of features. Previous work in ASR has also shown the necessity of unsupervised pre-training for DBNs to work well. We explore further the effect of training our ASR engine in an unsupervised manner with additional unannotated L2 data from the test speakers. This is compared with the original ASR that has been trained with annotated data in a supervised manner. Experiments show that DBN-HMM can give significant improvement (between 13-18% relative in word pronunciation error rate) but is computationally more expensive.

Index Terms: mispronunciation detection and diagnosis, restricted boltzmann machine, deep belief network

1. Introduction

Mispronunciation detection and diagnosis (MD&D) in L2 speech, as a typical application in computer-aided pronunciation training, needs to discern acoustically similar but phonetically different phones. Although automatic speech recognition (ASR) technology seems to be a good fit for MD&D, the latter actually presents a higher requirement in acoustic modeling due to the heterogeneous deviations of L2 speech from native productions. Therefore, researchers design dedicated features or classifiers [1][2] to enhance discriminability. As shown in our previous work, we can still exploit the hidden Markov model (HMM) paradigm, which is the predominant technique in state-of-art ASR, by explicitly modeling mispronunciations [3]. The standard HMM formalism uses a separate Gaussian mixture model (GMM) to model the conditional distribution of speech signal spectrum for each state in a phone, which is often considered insufficient in discriminability based on the maximum likelihood (ML) criterion. However, this can be remedied to a certain extent by discriminative training [4].

The deep belief network (DBN) is a probabilistic generative model composing of multiple layers of stochastic latent variables. A layered, unsupervised pre-training algorithm stacks up restricted Boltzmann machines (RBMs) from bottom up to construct a deep neural network [5]. Fine-tuning the pre-trained deep net using back-propagation or other approaches is found to achieve better classification performance than those without pre-training. Deep nets also outperforms shallow nets. Hence, it is re-gaining popularity in the research arena. Compared with a set of GMMs, DBN is inherently centralized. Moreover, in our experience, pre-trained DBNs tend to converge faster to better local optimum points which generalize well. Recently, this technique has been successfully applied to acoustic modeling [6][7] by replacing the GMMs with a DBN which models the state posteriors in HMMs as the output of the network, and this hybrid DBN-HMM framework still allows efficient Viterbi decoding.

The above motivates use to explore the use of DBN-HMM in MD&D as the approach offers the key advantage of leveraging unlabeled L2 speech data. Our previous work in MD&D required annotated L2 speech data, which is limited in quantity as labeling is an expensive procedure. The rest of the paper is organized as follows: We present RBMs which are building blocks of DBNs in Section 2. Then we specify the division of our L2 corpus for comparative evaluation purpose in Section 3. Our approach for MD&D that incorporates acoustic modeling using DBNs will be introduced in Section 4. Experiments and analysis are presented in Section 5. Finally, conclusions are given in Section 6.

2. Deep Belief Network

Restricted Boltzmann machines are the building blocks of DBN. Unsupervised learning of RBMs maximizes the probability of generating data without introducing class labels. Types of RBMs commonly employed in processing speech data include: (i) Gaussian-Bernoulli RBMs, for front-end acoustic representation with assumed Gaussian distributions; and (ii) Bernoulli RBMs, which can encode binary data with high efficiency. DBNs can thus be formed by stacking RBMs on top of one another.

2.1. Gaussian-Bernoulli RBM (GRBM)

The GRBM has one layer of stochastic visible Gaussian units and one layer of stochastic hidden binary $\{0, 1\}$ units. There is no interaction between units in the same layer and is thus "restricted". It assigns joint probability to visible-hidden pair $(\mathbf{v}, \mathbf{h})$ as follows:

$$E(\mathbf{v}, \mathbf{h}; \Theta) = -\mathbf{h}^{\mathrm{T}}\mathbf{W}\mathrm{diag}(\boldsymbol{\sigma})^{-1}\mathbf{v} - \mathbf{a}^{\mathrm{T}}\mathbf{h} + \frac{1}{2}(\mathbf{v}-\mathbf{b})^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\sigma}^2)^{-1}(\mathbf{v}-\mathbf{b}), \tag{1a}$$

$$\Pr(\mathbf{v}, \mathbf{h}; \Theta) = \frac{e^{-E(\mathbf{v},\mathbf{h};\Theta)}}{\int_{\mathbf{v}'}\sum_{\mathbf{h}'} e^{-E(\mathbf{v}',\mathbf{h}';\Theta)}d\mathbf{v}'}, \tag{1b}$$

where $\Theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}, \boldsymbol{\sigma}\}$, and Θ will be omitted to lighten notation wherever necessary. $\mathbf{W} = \{w_{ij}\}$ are weights of the symmetric connections between the hidden unit i and the visible unit j, while a_i and b_j are their bias terms. σ_j is the standard deviation of v_j.

Marginalizing over $\mathbf{h}$ leads to $\Pr(\mathbf{v}) = \sum_{\mathbf{h}} \Pr(\mathbf{v}, \mathbf{h})$. Given a set of F frames $\{\mathbf{v}^f\}_{f=1}^F$, $\sum_f \log \Pr(\mathbf{v}^f)$ is the objective to be maximized. The maximization can be interpreted as an economical representation of $\mathbf{v}$ on $\mathbf{h}$.

Differentiating $\sum_f \log \Pr(\mathbf{v}^f)$ w.r.t. any θ in Θ leads to the following positive and negative phases:

$$\sum_f \left[\sum_{\mathbf{h}} \frac{e^{-E(\mathbf{v}^f,\mathbf{h})}\frac{\partial -E(\mathbf{v}^f,\mathbf{h})}{\partial\theta}}{\sum_{\mathbf{h}'} e^{-E(\mathbf{v}^f,\mathbf{h}')}} \right] \tag{2a}$$

$$-\sum_f \left[\int_{\mathbf{v}} \sum_{\mathbf{h}} \frac{e^{-E(\mathbf{v},\mathbf{h})}\frac{\partial -E(\mathbf{v},\mathbf{h})}{\partial\theta}}{\int_{\mathbf{v}'}\sum_{\mathbf{h}'} e^{-E(\mathbf{v}',\mathbf{h}')}d\mathbf{v}'} d\mathbf{v} \right]. \tag{2b}$$

Substituting θ in $-\frac{\partial E(\mathbf{v}^f,\mathbf{h})}{\partial\theta}$ with w_{ij}, a_i, b_j, σ_j yields:

$$-\frac{\partial E(\mathbf{v}^f,\mathbf{h})}{\partial w_{ij}} = \frac{v_j^f}{\sigma_j}h_i; \quad -\frac{\partial E(\mathbf{v}^f,\mathbf{h})}{\partial a_i} = h_i; \tag{3a}$$

$$-\frac{\partial E(\mathbf{v}^f,\mathbf{h})}{\partial b_j} = (v_j^f - b_j)/\sigma_j^2; \tag{3b}$$

$$-\frac{\partial E(\mathbf{v}^f,\mathbf{h})}{\partial \sigma_j} = -\frac{v_j^f}{\sigma_j^2}\sum_i h_i w_{ij} + \frac{(v_j^f - b_j)^2}{\sigma_j^3}. \tag{3c}$$

Observe that Eqn. (2a) (the positive phase) is the conditional expectation of $-\frac{\partial E(\mathbf{v}^f,\mathbf{h})}{\partial\theta}$, since:

$$\frac{e^{-E(\mathbf{v}^f,\mathbf{h})}}{\sum_{\mathbf{h}'} e^{-E(\mathbf{v}^f,\mathbf{h}')}} = \Pr(\mathbf{h}|\mathbf{v}^f) = \prod_i \Pr(h_i|\mathbf{v}^f), \tag{4}$$

with the last factorization owing to the non-connectivity among h_i. For example, take $\theta = w_{ij}$ and Eqn. (2a) becomes:

$$\sum_f \left[\sum_{h_1} \Pr(h_1|\mathbf{v}^f) \cdots \left(\sum_{h_i} \frac{h_i v_j^f}{\sigma_j} \Pr(h_i|\mathbf{v}^f) \right) \cdots \right]$$

$$= \sum_f \left[\sum_{h_i} \frac{h_i v_j^f}{\sigma_j} \Pr(h_i|\mathbf{v}^f) \right], \tag{5}$$

and basing on the fact that h_i can only be either 0 or 1:

$$\Pr(h_i = 1|\mathbf{v}^f) = \frac{1}{1 + e^{-(\sum_j w_{ij}v_j^f/\sigma_j + a_i)}}. \tag{6}$$

Hence, given the observations $\{\mathbf{v}^f\}$, the expectation of derivatives in Eqn. (2a) can be easily computed. Unfortunately, Eqn. (2b) (the negative phase) involves an integration over the feature space and is intractable. A widely applied method that approximates this integral is the Gibbs sampler which proceeds in a Markov chain as follows:

$$\mathbf{v}^{(0)} \sim \text{a training frame}, \quad \mathbf{h}^{(0)} \sim \Pr(\mathbf{h}|\mathbf{v}^{(0)}); \tag{7a}$$

$$\mathbf{v}^{(1)} \sim \Pr(\mathbf{v}|\mathbf{h}^{(0)}), \quad \mathbf{h}^{(1)} \sim \Pr(\mathbf{h}|\mathbf{v}^{(1)}); \tag{7b}$$

$$\cdots$$

where $\Pr(v_j|\mathbf{h})$ can be derived as:

$$\Pr(v_j|\mathbf{h}) = \mathcal{N}(v_j; b_j + \sigma_j \sum_i w_{ij}h_i, \sigma_j^2). \tag{8}$$

Contrastive divergence (CD) training [5] makes two further approximations: (i) that the chain is run for only k steps; (ii) the integral in Eqn. (2b) is replaced by a single sample. In this work, $k = 1$ as it is fast and empirically works well. Starting from each training frame $\mathbf{v}^{(0)}$, we only sample $\mathbf{h}^{(0)}$ in Eqn. (7a) and use the expectations for Eqn. (6) and Eqn. (8) to replace the random samples $\mathbf{v}^{(1)}$ and $\mathbf{h}^{(1)}$ in Eqn. (7b) for stability. This process in the negative phase can be regarded as a single-round of "instruction and reconstruction". Therefore, Eqn. (2) measures the discrepancy between the model's "beliefs" in the statistics on Eqn. (3a)-(3c) and the actual observed statistics to present a direction for optimization.

2.2. Bernoulli RBM (BRBM)

BRBM differs from GRBM in that all visible and hidden units are binary, with $E(\mathbf{v}, \mathbf{h}; \Theta)$ and $\Pr(\mathbf{v}, \mathbf{h}; \Theta)$ similar to Eqn. (1a) & (1b):

$$E(\mathbf{v}, \mathbf{h}; \Theta) = -\mathbf{h}^{\mathrm{T}}\mathbf{W}\mathbf{v} - \mathbf{a}^{\mathrm{T}}\mathbf{h} - \mathbf{b}^{\mathrm{T}}\mathbf{v}, \tag{9a}$$

$$\Pr(\mathbf{v}, \mathbf{h}; \Theta) = \frac{e^{-E(\mathbf{v},\mathbf{h};\Theta)}}{\sum_{\mathbf{v}'}\sum_{\mathbf{h}'} e^{-E(\mathbf{v}',\mathbf{h}';\Theta)}d\mathbf{v}'}. \tag{9b}$$

The gradient of $\Pr(\mathbf{v})$ can be obtained in a manner similar to that in a GRBM, and the conditional distributions $\Pr(\mathbf{v}|\mathbf{h})$ and $\Pr(\mathbf{h}|\mathbf{v})$ becomes symmetric and trivial.

3. L2 Corpus

The Chinese Learners of English (CHLOE) Cantonese subset consists of 13 hours of recordings by 50 male and 50 female native Cantonese college students reading confusable words, minimal pairs, phonemic sentences and the Aesop's Fable "The North Wind and the Sun". All of the data have been phonetically labeled by trained linguists. Apart from these, we have also collected nearly 40 hours of read speech on TIMIT prompts by the same set of speakers. TIMIT prompts are phonetically-balanced and are very desirable for training acoustic models. However, this large body of speech recordings remain unlabeled due to lack of human resources. The labeled data only constitutes one quarter of the entire corpus, which motivates us to investigate unsupervised methods using DBNs. We split CHLOE by speakers into groups A and B, which leads to four subsets of data as shown in Table 3.

Table 1: *Division of CHLOE.*

	group A	group B
unlabeled	**a** (6.5hr)	**c** (20hr)
labeled	**b** (6.5hr)	**d** (20hr)

4. Approach to MD&D

We explicitly model salient segmental mispronunciation errors in an extended pronunciation lexicon, as described previously in [3]. A ML Viterbi pass using an acoustic model with the extended lexicon outputs a phonetic transcription for the prompted word sequence and achieves MD&D. Note that the acoustic model may be based on GMM-HMMs (as in our previous work [3]) or DBN-HMMs (as in the current investigation).

4.1. GMM-HMM Baseline

We train a tied-state tri-phone GMM-HMM system using the labeled data from training speakers, i.e. subset (**b**) in Table 3, in the ML manner and tune the number of tied states and the number of Gaussians per state on a separate development set. The resulting baseline system has 1.5K tied states and 10 mixtures per state. In addition, we refine the HMM baseline according to the MWE discriminative criterion [4], where we present possible mispronunciations for differentiation in training HMMs.

4.2. DBN-HMM

The unsupervised "pre-training" stage greedily constructs a DBN layer-by-layer from bottom up. We start by treating the two bottom layers as a GRBM. Once the GRBM is trained by CD, either the posteriors of the binary hidden units for every speech frame, or a set of samples drawn from the posteriors, can serve as the input to the upper level BRBM which will also be trained by CD. As we repeat this process by stacking as many BRBMs as desired to develop a deep structure, the outputs of the upper layers are supposed to represent an abstraction of data capturing higher order correlations.

The supervised "fine-tuning" stage further optimizes the parameters of the network according to some criterion. We follow [7] and model the posteriors of HMM states s for each speech frame $\mathbf{o}$ as the output of the network, i.e. $\mathcal{P}(s|\mathbf{o})$. This is achieved by stacking on top of the pre-trained DBN an extra softmax layer with the number of units identical to the number of states in our baseline GMM-HMM system. The frame-state correspondence is obtained by a forced-alignment using this GMM-HMM system, enabling us to minimize the sum of log posteriors over all the training frames $\{\mathbf{o}^f, s^f\}$:

$$\min_{\Theta'} - \sum_f \log \frac{e^{\mathbf{W}^{\text{top}}_{i^f,*} \Pr(\mathbf{h}^f;\mathbf{o}^f,\Theta)+\mathbf{b}^{\text{top}}_{i^f}}}{\sum_i e^{\mathbf{W}^{\text{top}}_{i,*} \Pr(\mathbf{h}^f;\mathbf{o}^f,\Theta)+\mathbf{b}^{\text{top}}_i}} \tag{10}$$

where $\mathbf{W}^{\text{top}}$ and $\mathbf{b}^{\text{top}}$ are the weight matrix and bias vector for the soft-max layer, and $\Theta' = \{\Theta, \mathbf{W}^{\text{top}}, \mathbf{b}^{\text{top}}\}$. The subscript in $\mathbf{W}^{\text{top}}_{i^f,*}$ means the i^fth row corresponding to the aligned state identity for the fth training frame. $\Pr(\mathbf{h}^f; \mathbf{o}^f, \Theta)$ denotes the probabilities of binary units $\mathbf{h}^f$ in DBN's top layer which are recursively propagated from $\mathbf{o}^f$ at the bottom. It is noted that this criterion is conceptually similar to frame discrimination training of Gaussians presented in [8].

During decoding, the GMMs in the GMM-HMMs are replaced by this single fine-tuned DBN, thus becoming a "DBN-HMM". Therefore, instead of evaluation with a GMM, the likelihood of a speech frame given a particular state is approximated by $\mathcal{L}(\mathbf{o}|s) \propto \mathcal{P}(s|\mathbf{o})/p(s)$, where $p(s)$ is the prior estimated from the alignment.

4.2.1. DBN Pre-training

Due to the high volume of data involved, instead of computing the gradients using Eqn. (2a)-(3c) over the entire pre-training data, we maximize the log-likelihood of RBMs using stochastic gradient ascent for 20 epochs with a batch size of 256 frames. Except for the visible layer of the GRBM at the bottom, all layers contain 200 units. For the GRBM, a learning rate of $\eta = 0.004$ is used for $\mathbf{W}$, $\mathbf{a}$, $\mathbf{b}$, while a much smaller value of 0.00001 is used for $\boldsymbol{\sigma}$. A learning rate of 0.1 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 tth increment of θ: $\Delta\theta^{(t+1)} = \gamma\Delta\theta^{(t)} + \eta\frac{\partial\mathcal{L}}{\partial\theta}$, where $\frac{\partial\mathcal{L}}{\partial\theta}$ is the gradient in Eqn. (2). As described in Sec. 4.2, we stack 5 RBMs layer-by-layer upward and yield a 6-layer DBN.

4.2.2. DBN Fine-tuning

We attach a randomly initialized softmax layer on top of the pre-trained DBN. 20 epochs (at which we observe convergence) of fine tuning is performed by stochastic gradient descent with a batch size of 2560 frames. As the partial derivatives of the objective in Eqn. (10) with respect to each individual parameter can be conveniently obtained by the chain rule, we update all the parameters simultaneously in each batch, using the conjugate gradient method with 5 line searches.

5. Experiments

5.1. Preprocessing

Short-time Fourier analysis is performed on a 25-ms hamming window with a 10ms frame shift before a standard MFCC parameterization takes the first 13 cepstral coefficients. The first and second order derivatives are appended to form the 39-dimensional feature vector. Cepstral mean normalization is done on a per utterance basis, but cepstral variance normalization in [7] is not done since the variance is explicitly modeled, as shown in

Eqn. (1a).

5.2. Performance Metric

Performance is evaluated using the Word Pronunciation Error Rate (WPER). Note that a given word with different pronunciations (some of which may be mispronunciations) are treated as different work tokens. Hence, recognition of an acoustic realization given these pronunciation alternatives achieves mispronunciation detection and diagnosis, but may also involve False Acceptances (FA), False Rejections (FR) and Diagnostic Errors (DE) (see [4]) - the sum of which gives rise to WPER. For example, the canonical pronunciation of the prompted word “NORTH” is $[n\ ao\ r\ th]$. If the realization of $[l\ ow\ f]$ (which is common for Cantonese learners) is accepted as a correct pronunciation, it incurs an FA error.

5.3. Results

To investigate the impact of pre-training data on the fine-tuning to follow, we pre-train on different combinations of data subsets. All pre-trained DBNs are fine-tuned on subset (**b**).

Table 2: *GMM-HMM & DBN-HMM Results on test set (*$\mathbf{d}$*). Subsets (*$\mathbf{a}$*)-(*$\mathbf{c}$*) are described in Table 3.*

	criterion	pre-training set	WPER
(1)	ML	N/A	32.73%
(2)	MWE [4]	N/A	31.17%
(3)	N/A	None (random init)	43.39%
(4)	N/A	(**b**)	31.96%
(5)	N/A	(**a**)+(**b**)	27.96%
(6)	N/A	(**b**)+(**c**)	**26.85**%
(7)	N/A	(**a**)+(**b**)+(**c**)	27.08%

Table 2 shows the performance results[1]. The best result on DBN-HMM gives a 18.0% relative improvement over ML training, which shows that introducing unsupervised data from the test speakers during pre-training can move the parameters to a region which is better geared for fine tuning, leading to a better local optimum. Increasing the number of line searches can reduce WPER even further as observed in other preliminary experiments. This may indicate that DBN is less prone to over-training. In addition, there is a 13.8% relative improvement over discriminatively trained HMMs based on the MWE criterion. The performance gain does not come for free, training a DBN using the 39-dimensional features takes more than a week on a 16-core 2.5GHz Xeon machine, which is significantly more expensive than even discriminative training of GMM-HMM by an order of magnitude. Therefore, the iterative fine-tuning using state alignment derived from newer DBN-HMMs [7] is not adopted.

6. Conclusions and Perspectives

We have reported our investigations on using DBN-HMM in discriminative acoustic modeling for mispronunciation detection and diagnosis (MD&D) in L2 English. DBN-HMM allows unsupervised training with unlabeled speech data in the CHLOE corpus. This, together with the discriminative nature of DBN-HMM, give significant gains in word pronunciation error rates (WPER) - an 18% relative improvement over the GMM-HMM trained in an ML fashion and a 13.8% relative improvement over discriminatively trained GMM-HMM. Since the unlabeled data is recorded from the same set of speakers as those in the test set, our scheme can be regarded as a form of unsupervised speaker adaptation. DBN-HMM, however, is computationally much more expensive than GMM-HMM. In the future, we will investigate ways to reduce the computation required in training DBN-HMM, while maintaining discriminability for MD&D.

7. Acknowledgments

This work is affiliated with CUHK MoE-Microsoft Key Laboratory of Human-Centric Computing and Interface Technologies and is partially supported by the Research Grants Council General Research Fund (Project No. 415511).

8. References

[1] Truong, K., Neri, A., Wet, F., Cucchiarini, C. and Strik, H., “Automatic detection of frequent pronunciation errors made by L2-learners”, Proc. of Interspeech, 2005.

[2] Wei, S., Hu, G.P., Hu, Y. and Wang, R.H., “A new method for mispronunciation detection using support vector machine based on pronunciation space models”, Speech Communication, vol. 50, issue 10, pp. 896-905, 2009.

[3] Harrison, A.M., Lau, W.Y., Meng, H. and Wang L., “Improving mispronunciation detection and daignosis of learner’s speech with context-sensitive phonological rules based on language transfer”, Proc. of Interspeech, 2008.

[4] Qian, X., Soong F. and Meng H., “Discriminatively trained acoustic models for improving mispronunciation detection and diagnosis in computer aided pronunciation training (CAPT)”, Proc. of Interspeech, 2010.

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

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

[7] 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.

[8] Povey, D. and Woodland, P.C., “Frame discrimination training for HMMs for large vocabulary speech recognition”, Proc. of ICASSP, 1999.

[1] Later results show the statistically insignificant difference in rows (6) & (7) is due to stochastic optimization. In most cases, (7) is at least at good as (6).