MISPRONUNCIATION DETECTION VIA DYNAMIC TIME WARPING ON DEEP BELIEF NETWORK-BASED POSTERIORGRAMS

Ann Lee, Yaodong Zhang, James Glass

MIT Computer Science and Artificial Intelligence Laboratory,
Cambridge, Massachusetts 02139, USA
{annlee, ydzhang, jrg}@csail.mit.edu

ABSTRACT

In this paper, we explore the use of deep belief network (DBN) posteriorgrams as input to our previously proposed comparison-based system for detecting word-level mispronunciation. The system works by aligning a nonnative utterance with at least one native utterance and extracting features that describe the degree of mis-alignment from the aligned path and the distance matrix. We report system performance under different DBN training scenarios: pre-training and fine-tuning with either native data only or both native and nonnative data. Experimental results have shown that by substituting the system input from MFCC or Gaussian posteriorgrams obtained in a fully unsupervised manner to DBN posteriorgrams, the system performance can be improved by at least 10.4% relatively. Moreover, the system performance remains steady when only 30% of the annotations being used.

Index Terms— mispronunciation detection, dynamic time warping, deep belief networks

1. INTRODUCTION

Computer-aided pronunciation training (CAPT) deals with the problem of detecting pronunciation errors in nonnative speech. Conventional approaches based on automatic speech recognition (ASR) technology may lack the ability of generalizing to different target languages, as the process of training a recognizer for a new language requires a great amount of human effort to annotate data. Currently there are approximately 80 languages having ASR capability [1], taking up less than 2% of the world's languages [2]. For many languages that receive less attention and financial support, we seek to develop a different solution to building a CAPT system.

In our prior work [3], we have demonstrated a comparison-based, word-level mispronunciation detection system which works by analyzing the alignment between a student's utterance and a teacher's utterance. Dynamic time warping (DTW) is carried out between the two utterances, and features describing the degree of mis-alignment are extracted from the aligned path and the distance matrix for classifier training. The speech representations we have explored are Mel-frequency cepstral coefficients (MFCCs) and Gaussian posteriorgrams (GPs), decoded from a Gaussian mixture model (GMM) trained in a fully unsupervised manner [4]. Experimental results have shown that the system that extracts features from the aligned path and the distance matrix outperforms the one that considers alignment scores only.

Recent attempts in applying deep belief networks (DBNs) on speech technology have shown their ability to produce good classification results [5, 6]. One of the attractive characteristics of DBNs is that the pre-training step does not require any annotation of the data, while the back-propagation step allows us to fine-tune the pre-trained generative model with some labels. In other words, the model can be trained in a semi-supervised fashion [5, 7]. One of the challenges we have encountered in our prior work is that some mixtures in the GMM capture the difference in speakers rather than phonetic units, resulting in some parts of the mis-alignment not necessarily corresponding to mispronunciation. The recent success of DBNs shows the potential to improve the discriminability of the posteriorgrams on phones with only a small amount of annotated data.

In this paper, we explore the use of DBN posteriorgrams as the input to our system. We examine various settings where only the information from native speech or both from native and nonnative speech are used. As our goal is to reduce the required human labor as much as possible, we also investigate how system performance would vary with respect to different amount of annotation being used in DBN training.

2. RELATED WORK

Over the past two decades, there has been a great amount of research on mispronunciation detection in nonnative speech [8]. Some of the earliest work adopted template-matching approaches based on vector quantization and DTW. The scores from alignment between native and nonnative speech were used for evaluating pronunciation quality [9, 10]. As ASR technology improved, various likelihood probability-based or posterior probability-based approaches have been proposed. For example, Franco et. al [11] examined the posterior probabilities from acoustic models trained on different levels of nativeness, and goodness of pronunciation (GOP) scores [12] took the ratio between the likelihood scores from

forced alignment and from recognition into account.

Many of the existing pattern classification techniques have also been used to detect mispronunciation. Strik et. al [13] extracted acoustic-phonetic features and applied linear discriminant analysis (LDA), while Wei et. al [14] considered log-likelihood ratios (LLR) between the canonical phone model and a set of pronunciation variation models, and used support vector machine (SVM) for classification. Another direction is to explicitly model the possible mispronunciations based on linguistic knowledge. Meng et. al [15, 16, 17] proposed an extended lexicon where possible mispronunciation patterns were predicted based on language transfer rules.

One common challenge in building a CAPT system lies in the intrinsic difference between the acoustic-phonetic spaces of native and nonnative speech. As a result, some systems make use of training data from nonnative speech to improve system performance [11, 14, 17], although often it is not easy to get a good amount of well-labeled nonnative data. Most recently, Qian et. al [18] demonstrate one of the first efforts that adopts DBN on acoustic modeling for nonnative speech for the task of pronunciation training. Their results have shown the benefit of using DBN to incorporate the unlabeled nonnative speech from the test speakers.

Our recently proposed comparison-based system [3] is a combination of template-matching and classifier-based approaches. As Zhang et. al [7] have reported that using DBN posteriorgrams greatly improves the performance on the task of keyword spotting, which is essentially a comparison task, we would also like to explore the use of DBN posteriorgrams in our system, with various settings for training.

3. DEEP BELIEF NETWORK (DBN)

A deep belief network (DBN) consists of a stack of Restricted Boltzmann machines (RBMs). An RBM contains two layers: a visible layer $\mathbf{v}$ and a hidden layer $\mathbf{h}$. Every node is connected to all nodes in the other layer, while there are no connection between nodes within the same layer. Two types of RBMs are commonly used in speech processing: *Bernoulli* RBMs and *Gaussian-Bernoulli* RBMs. For Bernoulli RBMs, all visible and hidden units are binary, i. e. $\mathbf{v} \in \{0,1\}^D$ and $\mathbf{h} \in \{0,1\}^F$, where D and F are the number of units in each layer, while for Gaussian-Bernoulli RBMs, the visible units take real numbers, i. e. $\mathbf{v} \in \mathbb{R}^D$ and $\mathbf{h} \in \{0,1\}^F$.

The joint probability of $\mathbf{v}$ and $\mathbf{h}$ can be written as:

$$P(\mathbf{v},\mathbf{h};\theta) = \frac{1}{\mathbb{Z}(\theta)} \exp(-E(\mathbf{v},\mathbf{h};\theta)) \quad (1)$$

where $-E(\mathbf{v},\mathbf{h};\theta)$ is an energy function and $\mathbb{Z}(\theta)$ is the normalizing term. For Bernoulli RBMs, we have

$$E(\mathbf{v},\mathbf{h};\theta) = -\sum_{i=1}^{D}\sum_{j=1}^{F} W_{ij} v_i h_j - \sum_{i=1}^{D} b_i v_i - \sum_{j=1}^{F} a_j h_j \quad (2)$$

$$\mathbb{Z}(\theta) = \sum_{\mathbf{v}}\sum_{\mathbf{h}} \exp(-E(\mathbf{v},\mathbf{h};\theta)). \quad (3)$$

The parameters $\theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$ include the symmetric interaction between the units (W_{ij}) and the bias terms (a_j, b_i). On the other hand, for Gaussian-Bernoulli RBMs, we have

$$E(\mathbf{v},\mathbf{h};\theta) = \sum_{i=1}^{D} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{D}\sum_{j=1}^{F} W_{ij} h_j \frac{v_i}{\sigma_i} - \sum_{j=1}^{F} a_j h_j \quad (4)$$

$$\mathbb{Z}(\theta) = \int_{\mathbf{v}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v},\mathbf{h};\theta)) \quad (5)$$

with one more parameter σ, which is the standard deviation.

The pre-training step aims at maximizing the log-likelihood of the data, $\log P(\mathbf{v};\theta)$. Differentiating the log-likelihood with respect to the parameters θ results in a form of expectation over the data distribution minus expectation over the model distribution. Exact computation of the expectation over the model is intractable, so a contrastive divergence-based approach [19] was used to efficiently approximate the gradient of the log-likelihood probability by performing a one-step Gibbs sampling. A DBN can thus be built by stacking a number of Bernoulli RBMs on top of one layer of Gaussian-Bernoulli RBM. The whole structure can be learned in a layer-by-layer manner by treating the hidden activities of one RBM as the input data to a higher level RBM [5].

In the back propagation step, a softmax layer is further added to the top of the pre-trained DBN, and stochastic gradient descent can be carried out if the predicted label (the label with the highest posterior probability) of a training sample is not the same as the label given. With this softmax layer, DBN posteriorgrams can also be decoded. Given the model parameters θ fine-tuned over a pre-defined label set $V = \{l_1, l_2, \cdots, l_V\}$, the DBN posteriorgram for a speech frame x_i can be computed as

$$DBN_{p_{x_i}} = [P(l_1|x_i;\theta), P(l_2|x_i;\theta), \cdots, P(l_V|x_i;\theta)] \quad (6)$$

where $\sum_j P(l_j|x_i;\theta) = 1$.

4. THE COMPARISON-BASED SYSTEM

Given a teacher's utterance $T = (f_{t_1}, f_{t_2}, ..., f_{t_n})$ with n frames and student's utterance $S = (f_{s_1}, f_{s_2}, ..., f_{s_m})$ of m frames, an $n \times m$ distance matrix Φ_{ts} can be built, where $\Phi_{ts}(i,j) = D(f_{t_i}, f_{s_j})$ denotes the distance between two frames of speech representation f_{t_i} and f_{s_j}. The first stage of the system runs DTW between the two sequences by searching for the best path on Φ_{ts} and segments S into words based on the word timing information on T. The second stage of the system extracts phone-level and word-level features that describe the degree of mis-alignment from the shape of the aligned path and the appearance of the distance matrix. The whole problem is then treated as a binary classification problem, and SVM classifiers are trained to detect whether a word is mispronounced or not. More details of the system and feature design can be found in [3].

	Speaker set	Scripts	Duration	# words (# mispronounced)
a	CU-CHLOE, group A	SI	2.1 hr	9,989 (1,523)
b	CU-CHLOE, group B	SX	1.6 hr	6,894 (1,406)
c	TIMIT, all	SI	1.7 hr	N/A
d	TIMIT, all	SX	2.5 hr	N/A

Table 1: Division of the native and nonnative copora

The input to the system, T and S, can be in various speech representations, as long as a proper distance measure $D(f_{t_i}, f_{s_j})$ is defined. In our prior work, we have explored the use of MFCC and GP. For MFCC, $D(f_{t_i}, f_{s_j})$ can be defined as the Euclidean distance, while for GP, $D(f_{t_i}, f_{s_j})$ can be defined as $-\log(f_{t_i} \cdot f_{s_j})$ [4]. In this paper, we investigate the use of DBN posteriorgrams, and similarly, the distance metric can be defined as the inner product distance.

5. CORPORA

The native corpus of the target language (English) comes from TIMIT, and the nonnative speech comes from the Chinese University Chinese Learners of English (CU-CHLOE) corpus [15], which is a specially-designed corpus of Cantonese speaking English, and a subset of which is based on TIMIT prompts. There are 100 nonnative speakers (50 males and 50 females), so we divide them into two disjoint groups A and B, each with 25 males and 25 females. Word-level mispronunciation labels are collected through Amazon mechanical turk (AMT) [20], where there were three turkers labeling the words in each utterance, and we only consider the words that have agreement from all three turkers. Table 1 shows the four subsets of the data we use in experiments.

6. EXPERIMENTS

6.1. Experimental setting

All waveforms are first transformed into 39-dimensional MFCCs every 10-ms frame, including first and second order derivatives. Cepstral mean normalization (CMN) is carried out on a per utterance basis [18].

For SVM training, we use the alignments between set **a** and set **c** as shown in Table 1. For each nonnative utterance in set **a**, there is exactly one matching native utterance of the same gender from set **c**. Set **b** serves as the test set, and on average there are 3.8 matching native utterances of the same gender from set **d**. The parameters of the SVMs with RBF kernels are optimized for different scenarios, respectively.

For DBN training, we set the DBNs to be of 3 hidden layers, each with 512 hidden units. The softmax layer consists of 61 units, the same size as the phone set in the TIMIT corpus. Each MFCC frame is padded with 10 neighboring frames, resulting in a 429-dimensional vector for each frame as input to the DBNs. The samples are first normalized by their global mean and variance, so we can set σ to 1. We run 100 epochs of pre-training for each layer and 50 epochs of back propagation, both over a batch size of 256 frames. Three different scenarios are examined: either using only native data or using both native and nonnative data for pre-training and back propagation, as shown in the top three rows in Table 2. Since the subset of the CU-CHLOE corpus we use is without human phonetic transcription, we run forced alignment to serve as human transcription.

	Pre-training	Back propagation	F-score (%)
DBN	native (**c+d**)	native (**c+d**)	72.2
	both (**a+c+d**)	native (**c+d**)	71.9
	both (**a+c+d**)	both (**a+c+d**)	**72.7**
MFCC			65.1
GP			63.6
ASR			70.0

Table 2: Experimental results (see Table 1 for the details of the datasets used in each DBN training scenario)

The system performance is evaluated by using *precision*, *recall* and *F-score*. *Precision* is the number of words that are correctly detected as mispronounced divided by the total number of mispronunciations in the system output, and *recall* is the number of correctly identified mispronounced words divided by the total number of mispronounced words in the data. *F-score* is the harmonic mean of the two.

6.2. Baseline

We consider an ASR-based baseline. A monophone DBN-HMM recognizer trained on the TIMIT training set is used. The DBN has 2 hidden layers (2048 $\times$ 2048) and a softmax layer of 183 units, and takes 39-dimensional MFCCs stacked with 10 neighboring frames as input. We compute GOP scores [12] by taking the absolute difference between the log-likelihood of a phoneme segment from forced alignment and the log-likelihood score from the recognition pass within that segment, normalized by the duration. Since GOP score is defined for every phone while our application is to detect word-level mispronunciation, we pick the largest GOP score within a word, together with average GOP score over the word, the sum of the log-likelihood scores from forced alignment over the word normalized by duration, and the minimum phone log-likelihood score within the word, to form a feature vector and train an SVM classifier.

This baseline may not be the state-of-the-art ASR-based system for mispronunciation detection. However, it provides us with an idea about how our features extracted from the alignment can capture the characteristics of mispronunciation, compared to the information extracted from the likelihood scores. We also include our previous results based on MFCC and GP alignment for comparison. A 150-mixture GMM is trained on set **c** and set **d** in Table 1 for GP decoding.

6.3. Results

Table 2 shows the system performance under different DBN training scenarios, together with the results from MFCC and GP-based alignment and the ASR baseline. The three DBNs in the table are fine-tuned using all the labels within the corresponding sets. Fig. 1 shows the ROC curves.

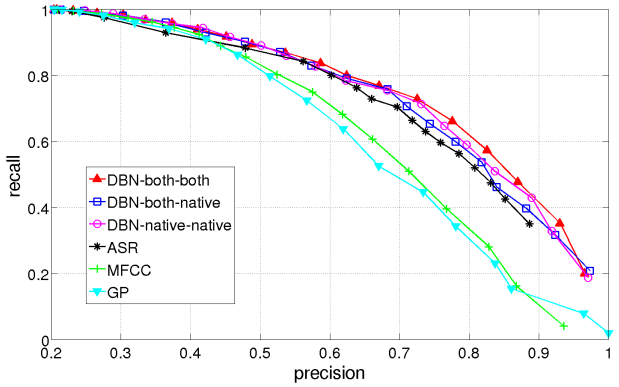

Fig. 1: ROC curves of three DBN training settings, MFCC or GP-based system, and ASR baseline

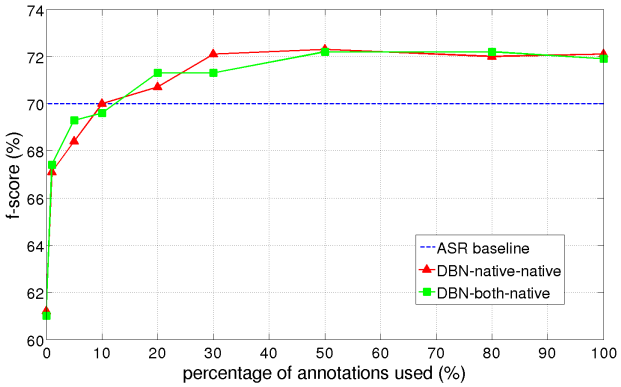

Fig. 2: System performance with respect to different percentage of annotations used for back propagation

By changing the system input from MFCC or GP to DBN posteriorgrams, we can improve the relative performance by at least 10.4%. This difference can be viewed as the gap between having zero resource versus full supervision. Also, if we compare the three DBN training scenarios, we can see that pre-training and fine-tuning with both native and nonnative data performs the best, while pre-training with both data but fine-tuning with only native data comes the third. (All the pairwise differences between the three scenarios are statistically significant with $p < 0.05$ by McNemar's test. This order is more obvious in the high precision region (see Fig. 1), which is in fact the favored manipulating region for a CAPT system, as there is lower chance that the student would be discouraged by a false positive. This order indicates that incorporating data from nonnative speech can benefit the model's discriminability over the phones in nonnative speech. The reason why including nonnative data for the pre-training stage does not help the case where we have native data only is probably because we normalize the input samples by the global mean and variance of the training data before decoding the posteriorgrams. There is still mismatch between the acoustic spaces even after CMN was carried out.

By feeding the DBN posteriorgram as input to our comparison-based system, we can obtain better performance (2.7% absolute) than applying likelihood-based scores as features. The DBN-HMM-based recognizer was trained on native speech only, and we did not perform any particular speaker adaption for the nonnative speech. However, we can still conclude that at least for word-level mispronunciation detection, an SVM can learn as much information from the features extracted based on mis-alignment as what it can learn from the likelihood scores based on an acoustic model.

We also explored how the system performance would vary with respect to different degrees of supervision involved during training. Here we compare only two scenarios: pre-training with either set **c+d** or **a+c+d** and both using set **c+d** for back propagation, as they use the same amount of data for fine-tuning. Fig. 2 shows the result. For the case where we use none of the annotations, we train a 61-mixture GMM first and then assign the labels according to the index of the mixture with the highest posterior probability [7].

We can see that as the amount of annotations decreases, for both scenarios, the system performance remains relatively steady until the point around 30%, which is about 76 minutes of speech. Also note that our recognizer uses around 3.1 hr of data for training, which corresponds to about 75% of the annotations, and the proposed comparison-based system can achieve the same level of performance with only 10% of the annotations, which is around 25 minutes of speech. This finding is encouraging, since our motivation for a comparison-based system was to reduce the human efforts required in preparing training data. With the help of the DBN, we can now improve the quality of the posteriorgrams without too much human labor.

7. CONCLUSION AND FUTURE WORK

In this paper, we have demonstrated the use of DBN posteriorgrams as input to our comparison-based mispronunciation detection system. Compared with an MFCC or GP-based system, DBN posteriorgrams can improve the relative performance by at least 10.4%. We have also shown that incorporating nonnative data with native data during training would benefit the system. While this improvement indeed comes from the trade-off of the human annotations required, experimental results have shown that using only one tenth of the annotations can produce the same level of performance as an ASR-based system.

This is just our initial attempt in applying DBNs. The size of our training data is relatively small, e.g. more than 20 hrs of nonnative speech was used in [17]. In the future, training DBNs from a larger dataset would be a possible way to further improve the system performance, though it would also take much longer time. Also, so far our system performance is on the same level as an ASR-based baseline for detecting word-level mispronunciation, we would also like to investigate if this kind of comparison-based system can detect subword-level mispronunciation, and how the performance would be compared with an ASR-based system.

8. REFERENCES

[1] "Nuance recognizer language availability," http://www.nuance.com/for-business/by-solution/customer-service-solutions/solutions-services/inbound-solutions/self-service-automation/recognizer/recognizer-languages/index.htm.

[2] "Ethnologue: languages of the world," http://www.ethnologue.com/.

[3] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *Proc. SLT*, 2012.

[4] Y. Zhang and J. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. ASRU*, 2009.

[5] R. Salakhutdinov, *Learning deep generative models*, Ph.D. thesis, University of Toronto, 2009.

[6] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[7] Y. Zhang, R. Salakhutdinov, H. Chang, and J. Glass, "Resource configurable spoken query detection using deep Boltzmann machines," in *Proc. ICASSP*, 2012.

[8] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832 – 844, 2009.

[9] D. Kewley-Port, C. Watson, D. Maki, and D. Reed, "Speaker-dependent speech recognition as the basis for a speech training aid," in *Proc. ICASSP*, 1987.

[10] H. Hamada, S. Miki, and R. Nakatsu, "Automatic evaluation of english pronunciation based on speech recognition techniques," *IEICE Transaction on Information and Systems*, vol. 76, no. 3, pp. 352–359, 1993.

[11] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Proc. Eurospeech*, 1999.

[12] S.M. Witt and S.J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.

[13] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[14] S. Wei, G. Hu, Y. Hu, and R.H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, no. 10, pp. 896–905, 2009.

[15] H. Meng, Y.Y. Lo, L. Wang, and W.Y. Lau, "Deriving salient learners' mispronunciations from cross-language phonological comparisons," in *Proc. ASRU*, 2007, pp. 437–442.

[16] A.M. Harrison, W.Y. Lau, H. Meng, and L. Wang, "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer," in *Proc. Interspeech*, 2008.

[17] X. Qian, F. Soong, and H. Meng, "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT)," in *Proc. Interspeech*, 2010.

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

[19] G.E. Hinton, S. Osindero, and Y.W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, 2006.

[20] M. A. Peabody, *Methods for pronunciation assessment in computer aided language learning*, Ph.D. thesis, MIT, 2011.