INVESTIGATION OF TANDEM DEEP BELIEF NETWORK APPROACH FOR PHONEME RECOGNITION

Xin Zheng[⋆†], *Zhiyong Wu*[⋆†‡], *Binbin Shen*[⋆†], *Helen Meng*[⋆‡] *and Lianhong Cai*[⋆†]

[⋆] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems
Shenzhen Key Laboratory of Information Science and Technology
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[†] Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[‡] Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
zhengx11@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, cbb09@mails.tsinghua.edu.cn,
hmmeng@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

ABSTRACT

This paper proposes using tandem DBN approach — a hierarchical architecture that consists of two or more deep belief networks (DBNs) in tandem manner — for phoneme recognition task on TIMIT. First we describe the standard DBN approach applied in phoneme recognition and discuss the motivation of combining it with tandem classifier approach. We then perform series of experiments to find out the best configuration for the DBN in the second level and discover the full potential of this method. The experiments show that for the DBN in the second level, (a) 2048 units in each hidden layer is better than 1024 and 512 units, (b) for sufficient length of temporal context, two hidden layers are better, (c) the one gives best performance on development set shows 4% relative improvement on coretest set.

Index Terms— phoneme recognition, deep belief network (DBN), tandem, Restricted Boltzmann Machine (RBM)

1. INTRODUCTION

Since deep belief network [1] was first proposed as a replacement of Gaussian mixture model (GMM) in the classical Hidden Markov Model (HMM)-GMM architecture for automatic speech recognition, it has shown very strong power in modeling acoustic signals [2][3][4]. Other than phoneme recognition tasks on TIMIT database, some research group have been trying to apply this new architecture to large vocabulary tasks [5] and some real life applications such as Bing voice search and Google Voice Input [6].

Tandem System [7] is an elegant and useful way to combine classifiers hierarchically. In Tandem System, a multilayer perceptron (MLP) is trained with acoustic features to generate posterior probability distribution of phonemes. After that, all the posterior distribution are treated as data (with certain kinds of transformations applied) for a traditional HMM-GMM system. This idea is then extended to an MLP followed by a hybrid HMM-MLP architecture [8][9] (which we call "tandem MLP" approach in the following text) or an MLP followed by a conditional randem field (CRF) [10]. Until recent years some properties of tandem MLP approach were investigated [9]. Since DBN has been shown to be quite different in both the training procedure and effectiveness from a traditional MLP, we wonder if tandem DBN approach might affect some conclusions obtained through tandem MLP approach.

The aim of this paper is to fully investigate the potential capacity of this hierarchical structure of DBNs for phoneme recognition. Our experiments shows that a temporal context of 190-270ms with two hidden layers should be used for the DBN in the second level in tandem DBN architecture to guarantee good results.

2. DEEP BELIEF NETWORK

First of all, we need to make some explanations of the term "deep belief network (DBN)". On one hand, it was referred to as a generative model, which was usually trained as stacked Restricted Boltzmann Machines (RBMs). From the top-down view of the model, it can describe the joint distribution of data and states of each hidden layers. While from the bottom-up perspective, recognition can be performed. And from this

This work is partially supported by the National Basic Research Program (973 Program) of China(2012CB316401), the National Natural Science Foundation of China (60928005, 60805008, 60931160443 and 61003094), the Ph.D. Programs Foundation of Ministry of Education of China (200800031015), the Upgrading Plan Project of Shenzhen Key Laboratory and the Science and Technology R&D Funding of the Shenzhen Municipal.

perspective, it's very much like a feed-forward neural network (or MLP). For this reason, after training all the stacked RBMs, all the units and weights are treated like a neural network to perform fine-tuning to get better discriminative results. And in this way we can get a better feed-forward neural network, which was also referred to as deep belief networks (sometimes also called deep neural networks, DNN).

2.1. Pre-training

The step of training the generative model described above is usually called pre-training. When modeling acoustic model with DBN, to get real valued data into DBN, a Gaussian-Bernoulli RBM is used for the bottom two layers. All the RBMs above are all Bernoulli-Bernoulli RBMs. Since DBN (in this step) is an unsupervised learning method, no label is needed. Unsupervised learning is believed to be able to capture crucial distribution of data and thus can help supervised learning when labels are provided. This opinion has already been proved in practice for speech data [4][11].

2.2. Fine-tuning

After pre-training, the whole DBN (with a softmax layer added) was treated as a feed-forward neural network to perform back-propagation algorithm with all the labels in the dataset to get more discriminative power. After that, a DBN is able to generate the posterior distribution of each phoneme given an input feature vector.

3. MOTIVATION OF TANDEM DBN APPROACH

Tandem classifier approach was first proposed for speech recognition in [7]. Tandem classifier approach is an effective way to combine two (or more) classifiers hierarchically. The input of lowest level classifier are original acoustic features, and the classifier outputs the posterior probability distribution over each phoneme (or state in HMM). The higher level classifiers receive these posterior probability distributions from lower level classifiers and treat them as training data. From these training cases they are trained to generate posterior probability distribution for even higher level classifiers or for a HMM to get recognition results. One important property of this approach is that the posterior classifiers are able to use even longer context information which has been proved to be very critical for decision making in acoustic signals [12].

In [7], an MLP and a GMM was used in tandem and was referred to as the Tandem System. This approach was successfully extended as two or three MLPs [8] and a MLP with a CRF [10]. MLP in the first level was also extended to have a sparse hidden layer in [13]. Tandem classifier approach was also called hierarchical phoneme posterior probability estimator in [9]. There is a difference between Tandem System and other tandem MLP systems in the way they treat posterior probabilities generated from the first MLP: in Tandem System, logarithm and Karhunen-Loeve Transformation are applied, while in tandem MLP system, the posterior probabilities are directly used as features [9].

The aim of this paper is similar to some part of [9], in which extensive discussion was made on the properties of posterior features, but there are 3 major differences between our experiments and [9]:

- *Architecture of neural network*. In all the experiments in [9], MLPs are three layered and numbers of units in the hidden layer is fixed to 1000 for TIMIT tasks. Although a three layered MLP has the capability to approximate the posterior probability distribution when least square error is used, one strong condition must be satisfied, that is, sufficient hidden units must be used. On one hand, four or even more layered MLPs (e.g., DBNs) have shown to be more appropriate for the task of phoneme recognition [2][4]. On the other hand, a network needs exponentially more computational units to represents a function that can be compactly represent by a deeper architecture [14]. For these two reasons, we consider it more appropriate to resort to DBN as a replacement of ordinary three layered MLPs for the classifier in each level.

- *Training procedure*. Unlike traditional MLPs, DBNs are usually trained with two stages, as is described in section 2. This procedure has shown to be able to get better results than only using random weights followed by back-propagation in practice [4][11].

- *Acoustic feature*. 13-order Perceptual Linear Prediction (PLP) coefficients with delta and acceleration are used throughout [9]. But 40-order coefficients (and energy) Mel-scale log filter-bank (fbank) with delta and acceleration are demonstrated to be significantly better than Mel-Frequency Cepstral Coefficients (MFCCs) [2][3]. The reason for this boost in performance, as is stated in [15], was not due to the much more dimensions of feature, but because features that distribute it's information across each dimension evenly are more appropriate for DBNs. Since PLP and MFCC are comparable in the task of phoneme recognition [16], we suspect that the use of fbank features might effect some of the conclusions in [9].

The tandem DBN architecture we used in our experiments is illustrated in Figure 1. For the convenience of our explanation, the classifier whose inputs are original acoustic features is denoted as the "first level (or Lv1)" and the classifier which receives posterior distribution as input data is called the "second level (or Lv2)". We will not be discussing a third level classifier in our study.

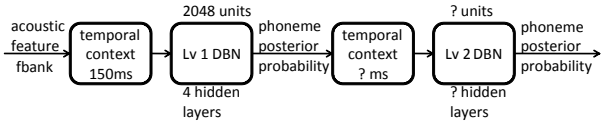

Fig. 1. The tandem DBN architecture used in our work. HMM is omitted. All the question marks are what to be determined in our experiments.

4. EXPERIMENTAL SETUP

4.1. Database

Experiments were performed over TIMIT database for phoneme recognition. All SA records were removed since they might bias the results. The database was devided into three parts: training, development and coretest in the way same as in [17]. For evaluation, 61 phones are mapped to 39 phonemes in the same way as in [18].

4.2. Training of DBN

During pre-training, all RBMs are trained using Contrastive Divergence algorithm [19] with stochastic gradient descent. The mini-batch size is 128. A momentum of 0.9 was used and no weight decay was applied. For Gaussian-Bernoulli RBMs, 225 epochs and learning rate of 0.002 was used. For Bernoulli-Bernoulli RBMs, we trained 100 epoch with learning rate of 0.02. All the hidden layers in DBN have same number of hidden units, and a softmax layer with 183 units was used to output probabilities. These parameters totally follows [4]. For all the number of layers mentioned below, we count only hidden layers.

For fine-tuning, we also used stochastic gradient descent with mini-batch of size 128. Momentum starts from 0.5 and linearly incremented to 0.9 in 10 epochs. Learning rate starts from 0.1 and recognition phoneme error rate (PER) on the development set was used as early-stopping criterion and was calculated at the end of each epoch. Learning rate was halved and all the weights returned to the values at the beginning of this epoch if an increase of substitution error was observed. The process continued until the learning rate was less than 0.0001.

4.3. Decoder

We use HVite, which is part of HTK [20], as our decoder. Word insertion penalty and language model scale factor was fixed to 0.0 and 1.0 separately. All the recognition results are given with the use of a bigram language model, which was estimated from transcriptions in the training set.

5. EXPERIMENTS AND DISCUSSIONS

We first train a DBN for the first level. After that, we fix this DBN and use it to generate input features for all the DBNs in the second level. All the details of the DBN we use in the first level is described in Section 5.1. Section 5.2 focus on investigating the performance of whole tandem DBN architecture on TIMIT phoneme recognition task.

5.1. DBN in the first level

The acoustic feature used to train this DBN was 40 coefficients Mel-scale log filter-bank (with energy) and delta and acceleration. The feature vectors were extracted with Hamming window with 25ms window length and 10ms window shift. Each dimension of input data for the DBN was normalized to have mean 0 and variance 1. A temporal context of 150ms and four hidden layers, each has 2048 units, were used. With this setting, we get 20.96% PER on development set and 22.34% PER on coretest set. All the posterior features used below were generated from this DBN.

5.2. Investigation of DBNs in the second level

Posterior features generated from DBN in the first level was normalized to have mean 0 and variance 1. This step is important as it can remove the effect of prior of phonemes that have already been learned by the DBN in the first level [9].

The main aim of this section is to find out the best configuration of DBN in the second level. There are several things to be determined: number of units in each hidden layer, length of temporal context, and the number of hidden layers.

First of all, we would like to determine the appropriate number of units in each hidden layer. Taking 23 frame context (i.e., 230ms) as an example, the results was depicted in Figure 2. The reason why we are particular interested in 23 frames is that 150ms-230ms has shown to be more appropriate for the MLP in the second level in tandem MLP architectures [9]. From Figure 2, when using 230ms as temporal context, tandem DBN consistently shows superior results over single DBN. We can also notice that 2048 units perform consistently better on development set in this set of experiments, so we used 2048 units in all the following experiments.

One particular interesting thing we can discover from Figure 2 is that *deep does not help*. This is quite surprising because it's very different from what we've observed for the DBN in the first level [4]. To figure out if it's just a coincidence, we need to fix the number of hidden units and investigate the effect of length of temporal context. The results are shown in Figure 3.

The first thing we can see from Figure 3 is that the effect of 150ms-230ms (i.e., 15 frames to 23 frames) temporal context proved in tandem MLP architecture has also been verified in tandem DBN approach. In fact, due to the extra accuracy of DBN over MLP in each level, 190ms-270ms (i.e., 19 frames

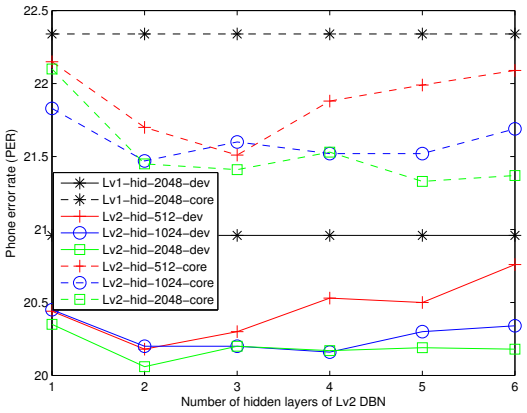

Fig. 2. PER of DBN in the second level on development set and coretest set as a function of number of hidden layers, using 23 frame context (230ms).

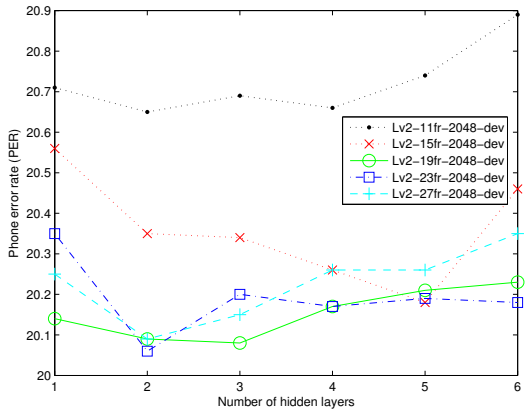

Fig. 3. PER of DBN in the second level on development set as a function of number of hidden layers, using 2048 units per hidden layer.

to 27 frames) is more appropriate. We should also notice that for sufficient length of temporal context, DBN with two hidden layers seems to achieve better performance.

To understand why deeper networks do not help for the DBNs in the second level, we should notice the distinction of input features used between DBN in the first level and second level. Input features for the MLP in the second level were called posterior features in [9]. Posterior features have two very important properties: sparseness and better linear separability (over acoustic features) [9]. These two properties are the key reasons why tandem classifier approach actually works for the task of phoneme recognition. Acoustic features do not have such good properties, so the complexity of their distribution (which leads to highly-varying separation surface) needs much deeper architecture to capture or separate. In other words, the effectiveness of the number of hidden layers is largely determined by the particularity of distribution of data itself. So, to better utilize DBN, it's quite necessary to make clear the effects of the number of hidden layers for specific applications.

5.3. Results

Table 1 compares the PERs between a single DBN that we've used in the first level and the best two-level tandem DBN we get. The best tandem DBN was chosen by it's improvement evaluated on the development set.

Table 1. PER of different architectures

Architecture	dev	coretest
single DBN	20.96%	22.34%
two DBNs in tandem	20.06%	21.45%

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed and investigated tandem DBN architecture for phoneme recognition task. We focused on discovering the best configuration of the DBN in the second level. Our experiments showed that for the DBN in the second level, 230ms temporal context and two hidden layers are appropriate, and the number of hidden units should be no less than 2000. With such configuration, the phone error rate improved 4% on coretest set compared with a single DBN.

The features we used for the DBN in the second level is just normalized posterior probabilities generated from first level DBN. But we have already known that DBN performs significantly better if information tends to spread equally in each dimension (e.g., fbank better than MFCC) [15]. So we are currently seeking some kind of transformations that has such good property (e.g., deep autoencoder [21]) to further improve the performance of tandem DBN architecture.

7. RELATION TO PRIOR WORK

The work presented here proposed a new architecture based on tandem DBN approach. The work in [4] uses DBN as a replacement of traditional MLP in the hybrid HMM-MLP architecture but not in hierarchical manner. While in works that use tandem approach [8][9] only incorporate traditional MLP but not DBN. More information of the relationship between our work and the work in [9] can be found in section 3.

8. REFERENCES

[1] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*,

vol. 18, pp. 1527–1554, 2006.

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

[3] G.E. Dahl, M. Ranzato, A. Mohamed, and G.E. Hinton, "Phone recognition with the mean-covariance restricted boltzmann machine," in *NIPS*, 2012.

[4] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. on Audio, Speech and Language Processing*, 2011.

[5] G. Dahl, D. Yu, L. Deng, , and A. Acero, "Large vocabulary continuous speech recognition with context-dependent dbn-hmms," in *Proc. ICASSP*, 2011.

[6] G. E. Hinton, L. Deng, D. Yu, G.E Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainathand, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, 2012.

[7] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. ICASSP*, 2000, pp. 1635–1638.

[8] P. Schwarz, P. Matějka, and J. Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. of ICASSP*, 2006, pp. 325–328.

[9] J. Pinto, S. Garimella, M. Magimai-Doss, H. Hermansky, and H. Bourlard, "Analysis of mlp-based hierarchical phone posterior probability estimators," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, 2011.

[10] E. Fosler-Lussier and J. Morris, "Crandem systems: Conditional random field acoustic models for hidden markov models," in *Proc. of ICASSP*, 2008, pp. 4049–4052.

[11] D. Yu, L. Deng, and G. E. Dahl, "Roles of pre-training and fine-tuning in context-dependent dbn-hmms for real-world speech recognition," in *NIPS 2010 workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[12] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai.-Doss, "Exploiting contextual information for improved phoneme recognition," in *Proc. of ICASSP*, 2008, pp. 4449–4452.

[13] G.S.V.S. Sivaram and Hynek Hermansky, "Multilayer perceptron with sparse hidden outputs for phoneme recognition," in *Proc. of ICASSP*, 2011.

[14] Y. Bengio, "Learning deep architectures for ai," Tech. Rep. 1312, Université de Montréal, 2007.

[15] A. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. of ICASSP*, 2012.

[16] J. Psutka, L. M üller, and J. V. Psutka, "Comparison of mfcc and plp parameterization in the speaker independent continuous speech recognition task," in *Proc. of Eurospeech*, 2001, pp. 1813–1816.

[17] A. Halberstadt, *Heterogeneous measurements and multiple classifiers for speech recognition*, Ph.D. thesis, MIT, 1998.

[18] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[19] G. E. Hinton, "Training products of experts by minimizingc ontrastive divergence," *Neural Computation*, vol. 14, pp. 1771–1800, 2002.

[20] S. Young et. al., *The HTK Book*, Cambridge University, 2002.

[21] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.