Neurocomputing 170 (2015) 251–256

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A comparative study on selecting acoustic modeling units in deep neural networks based large vocabulary Chinese speech recognition

Xiangang Li, Yuning Yang, Zaihu Pang, Xihong Wu *

Speech and Hearing Research Center, Key Laboratory of Machine Perception (Ministry of Education), Peking University, PR China

ARTICLE INFO

Article history:
Received 9 December 2013
Received in revised form
15 June 2014
Accepted 2 July 2014
Available online 8 May 2015

Keywords:
Deep neural networks
Multi-task learning
Chinese automatic speech recognition
Acoustic modeling units
Syllable

ABSTRACT

This paper compared the performance of different acoustic modeling units in deep neural networks (DNNs) based large vocabulary continuous speech recognition (LVCSR) systems for Chinese. Recently, the deep neural networks based acoustic modeling method has achieved very competitive performance for many speech recognition tasks, and has become the focus of current LVCSR research. Some previous work have studied the context independent and context dependent DNNs based acoustic models. For Chinese, a syllabic language, the choice of basic modeling units under the background of DNNs based LVCSR systems is a very important issue.

Three basic modeling units, syllables, initial/finals, phones, are discussed and compared. Experimental results show that, in the DNNs based systems, the context dependent (CD) phones obtain the best performance, and the context independent (CI) syllables have the similar performance with the CD initial/finals. How the number of clustered states impacts on the performance of DNNs based systems is also discussed, which showed different properties from the GMMs based systems. Besides, through introducing the multi-task learning strategy, these multiple modeling units can be combined in the DNNs training procedure. The experimental results indicate that combining these multiple modeling units using multi-task learning outperforms each individual modeling unit.

1. Introduction

Although the last decades have witnessed significant progress in automatic speech recognition (ASR), the performance of ASR systems in many real usage scenarios still lags far behind human level performance. Many new machine learning algorithms have led to significant advances in ASR. Recently, a major advance has been made in training deep neural networks (DNNs), which contain more than one layer of hidden units between the inputs and outputs [1].

It has long been believed that deep neural network could not bring further performance improvement than the neural network with one or two hidden layers. However, recently, many new algorithms were developed for training deep models, and have been applied successfully in a number of tasks. One of these approaches is the deep belief network (DBN) training algorithms, suggested in [2], in which, the weights of each layer were first initialized by a purely unsupervised way and then fine-tuned with the labeled data. Besides, specifically for the DNN applied in ASR, there are many developments beyond the standard network architectures and learning methodologies in recent years [1,3–5].

Acoustic modeling is a fundamental problem in ASR. Almost all of the state-of-the-art ASR systems are hidden Markov model (HMM) based. The relationship between HMM states and the acoustic input is usually represented by Gaussian Mixture Models (GMMs) or Artificial Neural Networks (ANNs). However, the ANNs were typically trained with only one hidden layer. It has long been suspected that deep networks could model complex higher statistical structure effectively until recently many new algorithms were developed for training deep models. Many researches indicated that DNNs based acoustic models can outperform GMMs in many speech recognition tasks [1]. As first introduced in [4,6], the context independent (CI) pre-trained DNN/HMM hybrid architectures have been proposed for phone recognition. Then, context dependent (CD) pre-trained DNN/HMM for large vocabulary speech recognition is studied and discussed in [7,8]. DNNs based ASR systems achieved very competitive performance, and have become the focus of current ASR research.

Deep belief network pre-training was the first pre-training method to be widely studied, and further research indicated that they could be trained in many different ways, such as the discriminative pre-training method introduced in [9], and generative pre-training with various types of auto-encoder. For fine-tuning, many alternative methods can be applied, such as stochastic gradient descent, nonlinear conjugate-gradient, LBFGS, and "Hessian-free" method. Moreover, there are many neural network

* Corresponding author.
E-mail address: wxh@cis.pku.edu.cn (X. Wu).

http://dx.doi.org/10.1016/j.neucom.2014.07.087
0925-2312/

architectures designed for DNN/HMMs in ASR, such as deep tensor networks [10], deep stacking networks [11], deep convex networks [12]. There are several issues about the use of DNN/HMMs in ASR needing further explorations, including the choice of modeling units for some specific languages, the assessment of the models on large real-world datasets, and the adaptation criteria for this kind of models.

The introduction of DNNs based acoustic models would change many conclusions based on Gaussian mixture models (GMMs), owing to the difference that DNN is a discriminative model and the other is generative model. This paper focuses on the choice of acoustic modeling units in DNNs based large vocabulary continuous speech recognition systems for Chinese. In the GMMs based Chinese ASR systems, there are many researches in the literature discussing the modeling units [13]. Most Chinese ASR systems use initial/finals (IFs) as the basic acoustic modeling units set, which is mainly due to the low complexity of modeling and the fair requirement on the amount of training data. Besides, some researches decompose the finals, in which, phones are adopted as the basic modeling units. Moreover, Chinese is naturally a syllabic language, thus some efforts are made to build the syllable based acoustic models. There are many studies and discussions on the Chinese acoustic modeling units on the background of GMM/HMMs ASR systems. In this work, we will report the study on how the performance of DNNs based ASR systems is affected by the number of different acoustic modeling units: CI IFs, CD IFs, CI phones, CD phones, CI syllable and CD syllable. How the number of clustered states impacts on the performance of DNNs based systems is also discussed. Besides, through introducing the multi-task learning strategy, these multiple modeling units can be combined in the deep neural network training procedure. Some experiments were conducted to demonstrate the performance of combining these three kinds of modeling units.

The remainder of this paper is organized as follow. The next section presents the basic framework of DNN/HMMs acoustic modeling. Section 3 describes the acoustic modeling units for Chinese speech recognition in detail, while section 4 briefly introduces the multi-task learning. The experiments and results are in Section 5. The discussions and conclusions are drawn in the last section.

2. Deep neural network HMMs

A DNN is a feed-forward, artificial neural network that has more than one layer of hidden units between its inputs and outputs [1]. In this section, the details of deep neural network and its integration with HMMs will been presented.

2.1. The deep neural networks

In a DNN, each hidden unit j in layer i typically uses the logistic function to map its total input from the layer below, x_j^i, to the scalar state. For multi-classification, output unit j converts its total inputs, x_j, into a class probability p_j by using "softmax" nonlinearity:

$$y_j^i = logistic(x_j^i) = logistic\left(b_j^i + \sum_k y_k^{i-1} w_{kj}^i\right) \tag{1}$$

$$p_j = \frac{\exp(x_j^N)}{\sum_k \exp(x_j^N)} \tag{2}$$

The estimation of posterior probability can be considered as a two-step process [14]: the first is transforming the observation vector into a feature vectors v^L through the L hidden layers, and the second is estimating the posterior probability using log-linear model with feature v^L. If the first L hidden layers are fixed, the softmax layer would amount to a conditional maximum entropy (MaxEnt) model. Thus, the DNN can be viewed as feature transforming plus conditional MaxEnt model. The development of better feature learning for DNN would lead to the performance improvement of classification.

DNN can be discriminatively trained by back propagating (BP) derivatives of a cost function. When using the softmax output function, the natural cost function is the cross entropy between the target and outputs of softmax. For large training sets, the stochastic gradient descent (SGD) method is always employed with a "momentum" coefficient. On the other hand, to avoid overfitting, a held-out validation set is adopted.

However, DNNs with many hidden layers are always hard to optimize. BP can easily get trapped in poor local optima from a random starting point. This optimization challenge can be somewhat alleviated by introducing the pre-training procedure.

2.2. Pre-training

The most important advance in machine learning for deep neural network is the development of layer-wise unsupervised pre-training methods, which is first provided by [2] and based on restricted Boltzmann machines (RBMs).

In RBMs, the visible units correspond to the input vectors, and the hidden units correspond to the feature detectors. RBMs belong to energy based models, whose join probability is defined via energy function:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i \in visible} a_i v_i - \sum_{i \in hidden} b_i h_i - \sum_{i,j} v_i h_i w_{ij} \tag{3}$$

$$p(\mathbf{v}, \mathbf{h}) = \frac{\exp(-E(\mathbf{v}, \mathbf{h}))}{\sum_{\mathbf{v},\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))} \tag{4}$$

Because there are no direct connection between hidden units in an RBM, the layer-to-layer conditional distribution comes to $P(v_i = 1 \mid \mathbf{h}) = sigmoid(a_i + \sum_i v_i w_{ij})$ and $P(h_j = 1 \mid \mathbf{v}) = sigmoid(b_j + \sum_i v_i w_{ij})$.

RBMs are trained with maximum likelihood criteria, and as an approximate way, Contrastive Divergences training is always employed [15]. The learning rule is given as

$$\Delta w_{ij} = \eta(\langle x_i h_j \rangle_{data} - \langle x_i h_j \rangle_{recon}) \tag{5}$$

$$\Delta a_i = \eta(\langle x_i \rangle_{data} - \langle x_i \rangle_{recon}) \tag{6}$$

$$\Delta b_j = \eta(\langle h_j \rangle_{data} - \langle h_j \rangle_{recon}) \tag{7}$$

where η is a learning rate, $\langle \cdot \rangle_{data}$ is the expectation over $P_\theta(\mathbf{h} \mid \mathbf{d})P(\mathbf{d})$, and $\langle \cdot \rangle_{recon}$ is the expectation over $P_\theta^1(\mathbf{x}, \mathbf{h})$.

As the most important use, RBMs are always composed to form deep belief nets (DBNs). In [2], a greedy layer-wise training algorithm was proposed to train a DBN one layer at a time. The deep belief nets are used as initialization of all except the last layer of a traditional multi-layer neural network. Then the stochastic gradient descent is used to fine-tune the whole network with respect to a supervised training criterion.

The RBMs based pre-training for deep network training has gained a lot of success. The main difference from the classical feedforward neural network is the initialization with training schemes rather than random initialization. The RBMs based pre-training belongs to unsupervised pre-training or generative pre-training, while, there are alternative pre-training method called supervised pre-training or discriminative pre-training, in which a purely supervised but layer-wise procedure is employed [9,14,16]. The discriminative pre-training had shown to give better performance. For discriminative pre-training, the "layer-wise BP" is first

adopted in speech recognition, in which, a one-hidden-layer DNN is firstly trained to full convergence with BP, then replaced the softmax layer by another random initialized hidden layer and a new random softmax layer, again train the network to full convergence, and so on. Further study indicated that the performance can be improved by some simple modifications [9,14]: (1) stopping very early by going through the data only once and (2) using large learning rate.

Some efforts [17] have been made for understanding the reason why the pre-training strategy works so much better than the standard random initialization and gradient-based optimization of a supervised training criterion. These training strategies may hold promise as a principle to solve the problem of training deep networks. The lower layers are supposed to extract the structure of the whole network inputs, while the upper layers focus on classification. Thus, to form a better feature structure extractor would lead to better performance. The deep neural network learning procedure somehow combines the feature learning and classification in a unified framework.

2.3. Interfacing a DNN with an HMM

In the DNN based acoustic models, the DNN outputs the posterior probabilities of the acoustic modeling units over the input acoustic feature. In the HMM framework, the acoustic model is always formulated as

$$p(x|w) = \max_q \pi(q_0) \prod_{t=1}^{T} a_{q_{t-1}} a_{q_t} \prod_{t=0}^{T} p(x_t|q_t) \quad (8)$$

In the GMMs based ASR systems, the observation probability $p(x_t|q_t)$ is directly modeled by GMMs. However, in the DNNs based ASR systems, the observation probability $p(x_t|q_t)$ is converted by $p(x_t|q_t) = p(q_t|x_t)p(x_t)/p(q_t)$, and $p(q_t|x_t)$ is the posterior probability modeled by DNNs.

The DNNs based acoustic models can be trained using the embedded Viterbi algorithm with GMMs seeding. The modeling units of the GMMs are delivered as the modeling units for DNNs. Through the forced alignment, the input acoustic features are labeled, and then, the pre-trained neural net is fine-tuned discriminatively with BP.

3. Acoustic modeling units selection for Chinese speech recognition

Chinese is naturally a syllabic language and each basic language unit can be phonetically represented by a syllable. Among the 1254 distinct syllables, there are 408 toneless base-syllables. In order to conduct speech recognition, the syllable is always decomposed into initial and finals, and furthermore, the finals can be decomposed into medial, main vowel and nasal three parts if necessary.

It is important to select appropriate basic units to represent acoustic information for a specific language in designing ASR systems. For Chinese, the syllable contains the most strong co-articulation, and syllable based model has little problem on constructing the lexicon for new task. However, the syllable based models suffer from the poor coverage and distribution unevenness of training data. Besides, in the IFs based ASR systems, the complexity of model is low and the requirement of data for each unit can be easily satisfied. Thus, this kind of modeling units is widely used in the Asian community. However, some finals may have much more complex acoustic representation than others, such as “iong”, which have the medial part “i”, vowel part “o” and nasal part “ng”. Therefore, many researchers have discussed the phone based systems, in which, “iong” is modeling with three phones: “i”,“o”,“ng”. The acoustic modeling units in phone based systems are very similar with the International Phonetic Alphabet units, which makes these kinds of systems to be easily used for the multi-language ASR systems.

Many efforts have been made to build syllable based Chinese ASR systems. The difficulty of building such a system may come from the following reasons. Firstly, the demand of training data for the large size of modeling units is hard to satisfy, and it is very difficult to make the occurrences of syllables to be even. These facts brought crucial problems in the GMMs based acoustic model training. Secondly, the large size of context dependent models leads to a heavy problem to conduct lextree based Viterbi beam search with these kinds of models. There are more than 400 toneless syllables (when the tone information is not discussed), which may lead to a tri-syllable set containing more than 64,000,000 modeling units. This is a severe problem for generative models training and the HMM based decoding implementation. However, in the DNNs based ASR framework, the data sparsity and unevenness problem get alleviating. Considered these attractive properties employing syllable as modeling units, there is a potential and possibility of performance improvement for Chinese ASR. However, in order to conquer the problem in decoding phase, the key solution is the weighted finite state transducer (WFST) based decoding framework.

4. Combine multiple modeling units with multi-task learning

Different modeling units have different benefits for acoustic modeling. The phone based models decompose the finals, resulting in much easier sharing across different syllables. The syllables based modeling units provide the most strong co-articulation constrains. After an experimental comparison of these candidate modeling units for Chinese speech recognition, we proposed to combine the benefits of different modeling units in the DNNs based speech recognition framework. Thus, the multi-task learning (MTL) strategy is introduced in our experiments, in which, different modeling units were regarded as different supervised tasks for the DNN training.

Multi-task learning is the paradigm of learning several tasks simultaneously for the sake of mutual benefit. In the framework of neural network with MTL, the main learning task is solved jointly with extra related tasks using a shared input representation. Usually, how well the extra tasks are learned is not cared, their sole purpose is to help the main task be learned better. MTL may be beneficial for several reasons. According to [18], the important mechanisms that MTL help for neural networks include

- data amplification (using the same input data for multiple problems may help),
- eavesdropping (a difficult classification can improve if another classification with same data is successfully, and learning some features may be easier in a parallel task),
- attribute selection (may help select better features),
- representation bias (a better optimum might be found),
- overfitting prevention (more reliable feature estimation).

MTL for neural network training has been studied in the literature for years. For the DNN training, MTL can be applied in the fine-tuning procedure. In particular, based on the pre-trained neural network, multiple task output is added and learned at the same time, which is not much different from the traditional neural network with MTL.

For Chinese, the acoustic observations can be labeled with phone, initial/final (IF) and syllable. Though introducing MTL, these different levels of transcriptions can be used at the same

time, thus different information from different modeling units can be merged together to train better acoustic models.

5. Experiments and results

5.1. Experimental setups

We carried out speech recognition experiments on Hub4 Chinese broadcast news database. The training set is 1997 Chinese broadcast news speech corpus (Hub-4NE) training data which contains about 30 h of speech. The test set is Chinese broadcast news evaluation data which consists of about one hour speech. The acoustic model training set was also used to train a 3-gram language model used for these experiments.

For the feature extraction in the experiments, the speech was analyzed using a 25-ms Hamming window with a 10-ms fixed frame rate. In the GMMs based experiments, the speech was represented using 12th-order Mel frequency cepstral coefficients and energy, along with their first and second temporal derivatives. Channel normalization is applied using cepstral mean normalization over each utterance. In the DNNs based experiments, the speech was based on a Fourier-transform-based filter-bank with 21 coefficients distributed on a mel-scale (and energy) together with the corresponding first and second order temporal derivatives. In the experiment, a context of 7 frames was used with current frame, forming a total of 945 (15×63) inputs to the DNNs.

The GMMs based acoustic models were trained using ML criteria, and contain 32 Gaussians. The DNNs used have 4 hidden layers with 2500 nodes in each layer, and the activation function is "sigmoid". The DNNs were trained from the alignments with the GMMs based models. For fine-tuning, we used stochastic gradient descent with mini-batch of 128, the learning rate started at 0.006. At the end of each epoch, if the substitution error on the development set decreased less than 0.1, the learning rate begins to halve. This continued until the substitution error on the development set increased.

The problems of building a syllable based ASR systems come from two aspects, the data sparsity and unevenness for model training and lextree based Viterbi beam search with too many nodes. The WFST based decoding framework was used in the experiments. Under the framework of WFST, although there are too many HMMs caused by the context dependent syllables, the final decoding network is based on the transitions between the clustered states (senones). With the help of decision tree state tying, the context dependent syllable based HMMs have about only thousands of senones. Through the optimization algorithms of WFST, such as determination and minimization, the decoding network can be optimized with similar size as the decoding network for the initial/finals based models or phone based models.

5.2. Comparison of different kinds of CI acoustic modeling units

Firstly, the experiments are conducted to compare the performance of CI phones, CI IFs and CI syllables. All these modeling units are toneless.

In the experiments, the CI phones and CI IFs are 3 state left-to-right HMM, and the states number of each HMM for CI syllables is determined by the corresponding number of phones, for example, "qiong" have "q", "i", "o", "ng" 4 phones, the states number is 7 (3+4); "a" have only 1 phone, the states number is 4(3+1). The experimental results are placed in Table 1.

From Table 1, we can find out that the CI syllables got the best performance. However, the CI syllables contain more than 2000 states in total, the description of acoustic representation is more detailed, which may explain the differences of performance among these three acoustic modeling units.

Table 1
Character error rate of different context independent toneless models.

	CI-phones (%)	CI-IFs (%)	CI-syllables (%)
GMMs	34.27	31.23	29.95
DNNs	22.40	22.71	20.03

Table 2
Character error rate of different context dependent toneless models.

	CD-phones (%)	CD-IFs (%)	CD-syllables (%)
GMMs	24.97	26.27	30.39
DNNs	18.46	20.35	19.81

5.3. Comparison of different kinds of CD acoustic modeling units

Secondly, some experiments are conducted for the comparison of different kinds of CD acoustic modeling units. Just like the method mentioned in [7], the context-dependent acoustic models are based on decision tree based tying. More specifically, the CD phones, CD IFs, CD syllables are modeled by tri-phones with around 4000 shared states (senones). These senones are the labels for DNNs and the modeling units for GMMs. The experimental results are placed in Tables 2 and 3, where one is for toneless models, while the other is for the models with tone information.

Tables 2 and 3 show that the CD phones outperformed the other two types of modeling units. However, the CI phones are much worse than other CI models, but while taking into account the context dependency, the phones become the best modeling units. The introduction of context dependency makes the phones easy to discriminate, while the context independent phones make it easier to share phonetic units across different syllables but lack of the influence caused by different syllable and the information about co-articulation.

Compared with CD toneless IFs and CD toneless syllables, there is remarkable gap between the GMMs system for these two types of modeling units, but performances of DNNs for these two are quite close. However, GMMs model the distributions of each senone, while DNNs care about how to classify around these senones, which makes the GMMs based systems much more easily been influenced by data coverage. In the CD syllables based systems, some syllables may have only less than 10 examples, resulting in serious data coverage problems. Unlike the other two kinds of modeling units, the syllables have not yet been improved from CI to CD. A reasonable explanation to the experimental facts may be the data coverage of these kinds of models. If there are enough data for each syllable, the performance may improve further.

While introducing the tonal information, we can find out that the performances of all kinds of models are improved for the DNNs based systems, which indicates the importance of tonal information for Chinese acoustic modeling. However, as for the GMMs based systems, the performance of CD Syllables degrades, which are caused by the large amount demands of training data of this kinds of modeling units.

Besides, someone may point out that the discriminative training methods, such as MPE [19], Boosted MMI [20], are the start-of-the-art in HMM-GMM systems. However, the discriminative training methods can also been conducted on the DNN-HMM

systems [21,22]. With that in mind, we compare the GMMs and DNNs with the basic model training methods.

5.4. The impact of the number of senones

Nevertheless, the performance is quite affected by the data size, thus, while discussion about the acoustic modeling units, the number of senones should been taken into account. Thus, we have conducted some experiments on the CD phones, in which, GMMs and DNNs based on different number of senones were trained and tested. The results are showed in Fig. 1.

From Fig. 1(a), we can find out that the ASR performance varied with the number of senones. However, for GMMs based acoustic models, the performance becomes worse when the number increases, while for the DNNs based acoustic models, the CER maintains about 18.50%. The frame classification accuracy of the training and developing set is represented in Fig. 1(b), which shows that the classification becomes harder when the number of senones increases. Compared with these two figures, the conclusion is that when the number of senones increases, the classification performance of DNNs and the ASR performance of GMMs will decrease, but the ASR performance of DNNs based acoustic models has been less affected. ASR is a sequences pattern recognition problem, where many knowledge resources are intergraded, such as language models and dictionaries. Due to the constraint of language model, the classification performance decrease of DNNs has not resulted in the significantly decrease of the performance of ASR.

Table 3
Character error rate of different context dependent models with tonal information.

	CD-phones (%)	CD-IFs (%)	CD-syllables (%)
GMMs	21.75	23.49	34.21
DNNs	14.92	15.36	15.27

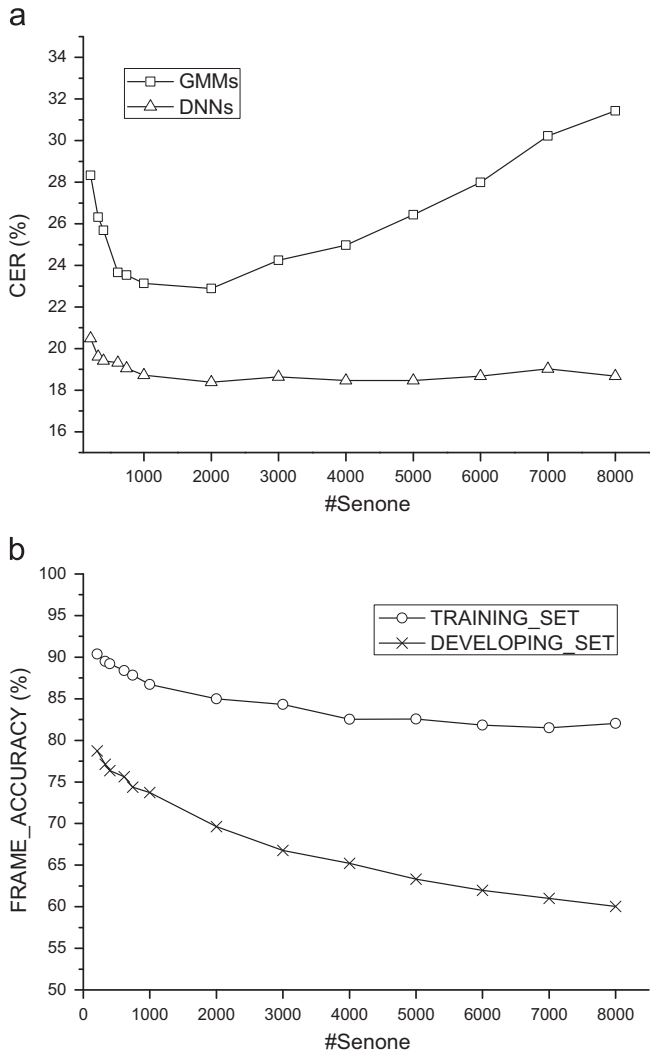

Fig. 1. The performance varying with the number of senones. (a) Compare the character error rate of DNNs and GMMs based systems. (b) Compare the frame classification accuracy on the training and developing set varying with the number of senones.

5.5. Combine multiple modeling units

In order to combine these multiple modeling units for acoustic modeling in Chinese speech recognition, multi-task learning strategy for deep neural networks training was applied. In the experiment, classification task in the phone based acoustic model is selected as the main task, while the classification tasks for IF and syllable were used as the extra tasks. The results are listed in Table 4.

From the experimental results shown in Table 4, we can find out that the applying MTL for DNNs can bring the performance improvement, which indicated that combining these three kinds of modeling units can outperform each one individual modeling units.

6. Discussions and conclusions

This paper presents a systematic performance comparison among various levels of acoustic modeling units for DNNs and GMMs based Chinese speech recognition. The introduction of DNNs based acoustic models would change many conclusions based on GMMs, owing to the difference that DNN is a discriminative model and the other is generative model. For the context independent acoustic modeling units, syllable based models have shown better performance than initial/finals or phone based models, especially in the DNNs based ASR systems. The outstanding performance mainly benefits from the more detailed description of acoustic representation. In addition, the best performance is obtained with the context dependent phones in the DNN systems. When the context dependency information is introduced, the performances of initial/finals and phones have gained remarkable improvement. Besides, for the DNNs based systems, the impact of the number of senones is also discussed. Unlike the GMMs based systems, when the number of senones increases, although the classification performance of DNNs decreases, the ASR performance of DNNs based acoustic models has been less affected. What should be pointed out is that, with DNNs, the context independent syllable based systems have gained the similar performance with context dependent initial/finals based systems. Through introducing the multi-task learning strategy, multiple modeling units can be combined to train a better acoustic model. The information coming from different levels of modeling units can help the feature learning for DNNs.

Table 4
The experimental results of combining multiple modeling units using multi-task learning for DNN.

Main task	Extra tasks	CER (%)
Toneless phone	–	18.46
Toneless phone	Toneless Initial/Finals, Toneless syllables	17.91
Tonal phone	–	14.92
Tonal phone	Tonal Initial/Finals, Tonal syllables	14.26

DNNs based acoustic models showed many important properties. The performance would not decrease seriously facing the distribution unevenness of training data. There is little impact of senones number on the performance of DNNs based ASR systems. In the DNNs based acoustic models, all the senone targets share the same hidden layers' transforms. Thus, when the number of senones increases, only the nodes in last layer in DNN will increase, while, in the GMMs based acoustic model, when the number of senones increases, the number of Gaussian mixtures will increase. Due to the sharing of the hidden layer in DNNs, the data sparsity and unevenness problem also get alleviating. For the syllable based system, the demand of training data for the large size of modeling units is hard to satisfy, and it is very difficult to make the occurrences of syllables to be even. All these properties make the DNNs based modeling method a good suggestion for the syllable based acoustic modeling in Chinese speech recognition. The DNNs based method has shown the potential to obtain better performance for syllable based Chinese ASR.

Through introducing the DNNs into the Chinese speech recognition, the performance has obtained a great improvement. Compared with the best performance of GMMs based systems, the DNNs can obtain more than 20% relative character error rate decrease. We believe that this work on DNNs based Chinese speech recognition is only the first step towards a power Chinese speech recognition system. There are many efforts needed to be done, more specifically, the clustering strategy, the neural networks structures the experiments on a lager dataset, discriminative training methods and so on.

Acknowledgements

The work was supported in part by the National Basic Research Program of China (2013CB329304), the research special fund for public welfare industry of health (201202001) and National Natural Science Foundation of China (Nos. 61121002 and 91120001).

References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Process. Mag. 29 (6) (2012) 82–97.
[2] G. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[3] D. Yu, L. Deng, G. Dahl, Roles of pretraining and fine-tuning in context-dependent DNN-HMMs for real-world speech recognition, in: Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2010.
[4] A. Mohamed, G. Dahl, G. Hinton, Acoustic modeling using deep belief networks, IEEE Trans. Audio Speech Lang. Process. 20 (January (1)) (2012) 14–22.
[5] L. Deng, G. Hinton, B. Kingsbury, New types of deep neural network learning for speech recognition and related applications: an overview, in: Proceedings of ICASSP, 2013.
[6] A. Mohamed, G. Dahl, G. Hinton, Deep belief networks for phone recognition, in: Proceedings of NIPS Workshop Deep Learning for Speech Recognition and Related Applications, 2009.
[7] G. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pretrained deep neural networks for large-vocabulary speech recognition, IEEE Trans. Audio Speech Lang. Process. 20 (Jan. (1)) (2012) 30–42.
[8] F. Seide, G. Li, D. Yu, Conversational speech transcription using context-dependent deep neural networks, in: Proceedings of Interspeech, 2011, pp. 437–440.
[9] D. Yu, L. Deng, G, Li, F. Seide, Discriminative Pretraining of Deep neural Networks, U.S. Patent Filling, November 2011.
[10] D. Yu, L. Deng, F. Seide, Large vocabulary speech recognition using deep tensor neural networks, in: Proceedings of Interspeech, 2012.
[11] L. Deng, D. Yu, J. Platt, Scalable stacking and learning for building deep architectures, in: Proceedings of ICASSP, 2012, pp. 2133–2136.
[12] L. Deng, D. Yu, Deep convex network: A scalable architecture for speech pattern classification, in: Proceedings of Interspeech, 2011, pp. 2285–2288.
[13] H. Wu, X.H. Wu, Context dependent syllable acoustic model for continuous Chinese speech recognition, in: Proceedings of Interspeech, 2007, pp. 1713–1716.
[14] F. Seide, G. Li, X. Chen, D. Yu, Feature engineering in context-dependent deep neural networks for conversational speech recognition, in: Proceedings of ASRU, 2011.
[15] G. Hinton,A Practical Guide to Training Restricted Boltzmann Machines, Technical Report, Report UTML. TR2010-00, Department of Computer Science, 2010.
[16] Y. Bengio, et al., Greedy layer-wise training of deep networks, in: Advances in Neural Information Processing Systems, 2007.
[17] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: International Conference on Artificial Intelligence and Statistics, vol. 9, 2010, pp. 249–256.
[18] R. Caruana, Multitask learning, Mach. Learn. 28 (1997) 41–75.
[19] D. Povey, Discriminative training for large vocabulary speech recognition (Ph. D. dissertation), University of Cambridge, Cambridge, UK, 2003.
[20] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, K. Visweswariah, Boosted MMI for model and feature-space discriminative training, in: Proceedings of IEEE ICASSP, 2008, pp. 4057–4060.
[21] B. Kingsbury, Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling, in: Proceedings of ICASSP, April 2009, pp. 3761–3764.
[22] Karel Veselý, Arnab Ghoshal, Lukáš Burget, Daniel Povey, Sequence-discriminative training of deep neural networks, in: Proceedings of INTERSPEECH, 2013.

Xiangang Li is currently a Ph.D. candidate at the Speech and Hearing Research Center, Peking University, PR China. He received the B.S. degree from Tongji University, PR China, in 2010. His research interests include speech recognition, deep learning, deep neural networks.

Yuning Yang is currently a master candidate at the Speech and Hearing Research Center, Peking University, PR China. She received the B.S. degree from College of Computer Science and Technology, Jilin University, PR China, in 2012. His research interests include speech recognition, deep neural networks.

Zaihu Pang is currently a Ph.D. at the Speech and Hearing Research Center, Peking University, PR China. He received the B.S. degree from College of Computer Science and Technology, Jilin University, PR China, in 2006. His research interests include speech recognition and statistical learning.

Xihong Wu received the B.S. degree from Jilin University, PR China, in 1989, the M.S. degree from the Institute of Harbin Shipbuilding Engineering in PR China, in 1992, and the Ph.D. degree from the Department of Radio Electronics, Peking University, PR China, in 1995. He is currently a professor and supervisor of Ph.D. candidates with Peking University. He has been elected a senior member of IEEE in 2009. His areas of research focus include computational auditory models and auditory scene analysis, auditory psychophysics, speech signal processing, and natural language processing.