Deep Neural Networks for Syllable based Acoustic Modeling in Chinese Speech Recognition

Xiangang Li, Caifu Hong, Yuning Yang, and Xihong Wu
Speech and Hearing Research Center,
Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, 100871, China
E-mail: {lixg, hongcf, yangyn, wxh}@cis.pku.edu.cn

Abstract—**Recently, the deep neural networks (DNNs) based acoustic modeling methods have been successfully applied to many speech recognition tasks. This paper reports the work about applying DNNs for syllable based acoustic modeling in Chinese automatic speech recognition (ASR). Compared with initial/finals (IFs), syllable can implicitly model the intra-syllable variations in better accuracy. However, the context dependent syllable based modeling set holds too many units, bringing about heavy problems on modeling and decoding implementation. In this paper, a WFST decoding framework is applied. Moreover, the decision tree based state tying and DNNs based models are discussed for the acoustic model training. The experimental results show that compared with the traditional IFs based modeling method, the proposed syllable modeling method using DNNs is more robust for data sparsity problem, which indicates that it has the potential to obtain better performance for Chinese ASR.**

I. Introduction

Although many successful everyday-life application systems have been deployed in last two decades, the performance of automatic speech recognition (ASR) still remains unsatisfying. As a vital component, acoustic modeling is always the research focus in ASR. Recently, the proposal of context-dependent deep neural networks (DNNs) hidden Markov models (HMMs) for large vocabulary speech recognition have achieved the most competitive performance.

The DNN based acoustic modeling method is a special artificial neural network (ANN) HMM hybrid approach. The ANN used in ASR was typically trained with only one hidden layer, while the DNN has more than one. Recent advances in machine learning led to the development of algorithms to train deep networks. The most famous approach is the pre-training methods with restricted Boltzmann machine (RBM), which is first introduced in [1,2]. Further studies indicated that the pre-training could be realized in many different ways, such as the discriminative pre-training and some generative pre-training with various types of auto-encoders.

DNN/HMMs have achieved remarkable results in many ASR tasks. The monophone pre-trained DNN/HMMs hybrid architecture have been proposed for phone recognition [1]. Afterwards, context-dependent pre-trained DNN/HMMs for large vocabulary speech recognition [3] and real-world data have been reported [4]. DNNs have made a strong impact on both the research and application of ASR.

The introducing 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. As for Chinese speech recognition, most systems use initial and toneless/tone finals as the basic acoustic modeling units. There are many efforts on applying syllable as the basic modeling units. The motivation comes from the fact that Chinese is a syllabic language. during the past decades, several researches have been made on syllable units based Chinese acoustic modeling [5], [6]. Although the implementing context dependent syllable is intuitive and attractive, there are many problems need to overcome. Firstly, there are more than 400 toneless syllables (the tone information is not discussed in this paper), which may leads 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. Fortunately, through introducing the WFST based decoding framework and some acoustic modeling techniques, the problems can be solved. In this paper, the central syllable independent decision tree based tying and the WFST based decoding framework are adopted, which makes the syllable based Chinese speech recognition can get a comparable results on a small training with the GMMs. Moreover, while introducing the DNNs based models, the experimental results showed a further improvement.

The remainder of this paper is organized as follows. In section 2, the machine learning method of deep neural network is reviewed. Section 3 will discuss the syllable based acoustic modeling method in detail. In order to present how the WFST based decoding framework can solve the problems caused by the too many context dependent syllable modeling units, some studied of WFST are discussed in section 4. The experiments are presented in section 5, followed by the conclusions and future work.

II. Review of deep neural networks

The idea of utilizing deep layered architecture for pattern recognition is not new. Unfortunately, training a deep architecture is a challenge problem due to the poor generalization property and suboptimal solutions [7]. Until recently, many new machine learning algorithms were developed for training deep models, which have trigged the applications of many deep structure models.

The most important advance in learning for deep networks has been the development of layer-wise unsupervised pre-training methods, first provided by [8] was based on restricted Boltzmann machine (RBM). RBMs belong to energy based models, whose joint probability is defined as:

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i\in visible} a_i v_i - \sum_{i\in hidden} b_i h_i - \sum_{i,j} v_i h_i w_{ij} \quad (1)$$

$$p(\mathbf{v},\mathbf{h}) = \frac{exp(-E(\mathbf{v},\mathbf{h}))}{\sum_{\mathbf{v},\mathbf{h}} exp(-E(\mathbf{v},\mathbf{h}))} \quad (2)$$

The RBM parameters can be efficiently trained with contrastive divergence algorithm [8]. The learning rule is given as follows:

$$\Delta w_{ij} = \eta(\langle x_i h_j\rangle_{data} - \langle x_i h_j\rangle_{recon}) \quad (3)$$

$$\Delta a_i = \eta(\langle x_i\rangle_{data} - \langle x_i\rangle_{recon}) \quad (4)$$

$$\Delta b_j = \eta(\langle h_j\rangle_{data} - \langle h_j\rangle_{recon}) \quad (5)$$

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

Once an RBM is trained on data, then the hidden activation probabilities of current RBM are used as the training data for the next RBM. Thus each RBM weights can be used to extract features from the output of the previous layer. When the pre-training complete, a randomly initialized softmax output layer will be added. Finally, the whole network can be fine-tune with back Propagation (BP).

This RBM based pre-training method can be viewed as a generative pretraining method, in which, the structure of input data is firstly learned before the multi-classification task.

There are some discriminative pre-training methods [9], [10]. The DNN is trained by starting with a one-hidden-layer neural network. Once the networks has been trained discriminatively, a second hidden layer is interposed between the last hidden layer and the softmax output layer, and the whole networks is again discriminatively trained. This can be continued until the desired deep structure is reached, and then fine-tune to convergence by BP.

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:

$$p(x|w) = max_q \pi(q_0)\prod_{t=1}^{T} a_{q_{t-1}} a_{q_t} \prod_{t=0}^{T} p(x_t|q_t) \quad (6)$$

In the GMMs based ASR systems, 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 [3].

The DNNs based acoustic models can be trained using the embedded Viterbi algorithm with GMMs seeding. The GMMs based system is firstly built, then conduct a forced alignment procedure with the GMM/HMMs. 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.

III. Syllable based acoustic modeling

In Chinese, each basic language unit can be phonetically represented by a syllable. Syllables contain stronger co-articulation and have more stable acoustic realizations than initial/finals [5]. Besides, compared with huge number of English syllables, the number of Chinese syllables is only around 400 (tone is not considered). These properties make syllable become a suitable candidate unit for Chinese acoustic modeling. In all the experiments of this paper, the tone information is not considered.

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 lead to a heavy problem to conduct lextree based Viterbi beam search with these kinds of models.

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.

IV. The WFST based decoding framework

The WFST based decoding framework has long been proposed since 1996 by Mohri et.al.[11]. It is a framework which can integrate the acoustic model, the pronunciation dictionary and the language model into one unified framework, and extend the search space for decoding offline.

The integrating of all the models can be done by using "Compose" operators as (7).

$$S = H \circ C \circ L \circ G \quad (7)$$

where $\circ$ denotes "Compose" operator; H denotes the WFST-form acoustic model, which transduces a senone sequence into a acoustic model unit(normally triphone) sequence; C denotes the WFST which transduces a triphone sequence into a monophone sequence; L denotes the WFST-form pronunciation dictionary which transduces a monophone sequence into a word sequence; G denotes the WFST-form language model which gives a word sequence language-level confidence(a language model score); S denotes the integrated WFST, which transduces a senone sequence into a word sequence, and we name it senone-level WFST afterwards. During building S, "Determinize" and "Minimize" operators will be used to merge equivalence edges and nodes to optimize S, which makes

$$S = \min(\det(H \circ \det(C \circ \det(L \circ G)))) \quad (8)$$

where $\det$ denotes "Determinize" operator and $\min$ denotes "Minimize" operators. Some WFST-based decoder such as Juicer [12] does not build a senone-level WFST offline but a triphone-level WFST, which is $T = C \circ L \circ G$, and integrate H online.

Normally speaking, S or T will be very large when G is large, and under some circumstances too large for a given hardware condition. Therefore, a lextree-based decoder, such as HDecode [13] is also usually used, which equals to build $R = C \circ L$ offline and integrate H and G online. However, when we are dealing with syllable-based acoustic model, there are more than 64,000,000 modeling units, which means C will be very large, and therefor R will become very large.

Noticed that the stage of decision tree based state tying during acoustic model training will map tri-syllable into a senone sequence and the scale of the senone set is usually about thousands which is dramatically smaller than the number of tri-syllables. Therefore most of the tri-syllables are equivalent. When decoding on R, too many equivalent candidate tokens are maintained, which would cause the decoding become extremely slow and inefficient, and the performance drops severely.

Fortunately, introducing senone-based WFST will merge the equivalent tri-syllables and finally get a tolerant-size WFST. The WFST can be constructed offline on a server with large memory to deal with the memory exploding in the stage $\det(C \circ \det(L \circ G))$. After S is constructed, it can be use on PC with normal hardware condition.

V. Experiments and results

A. Experimental setup

The experiments were carried on Hub4 Chinese broadcast news database, in which, the training set is 1997 Chinese broadcast news speech corpus (Hub-4NE) training data which contains about 30 hours of speech and the test set is Chinese broadcast news evaluation data which consist of about one hour speech. A tri-gram based language model was used in the evaluation, which was trained on the acoustic training set transcriptions with SRI-LM tools.

For the feature extraction in the experiments, the speech was analyzed using a 25-ms Hamming window with a 10-ms fixed frame shift 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 corresponding first and second order temporal derivatives. Besides, in these DNNs based experiments, a context of 7 frames were used with current frame, forming a total of 945 (15×63) inputs to the DNNs.

All the GMMs based acoustic models were trained using ML criteria, and contain 32 Gaussians. Based on these GMM/HMMs, the forced alignment procedure is conducted. Then the modeling units of GMMs are delivered to the DNNs based models. The DNNs used in the experiments have 4 hidden layers with 2500 nodes in each layer. In the RBM based pre-training procedure, the momentum term for RBMs was increased from 0.5 to 0.9 after the first twenty epochs. Learning rate was fixed to 0.001 for the first layer of RBM with Gaussian visible units, and fixed to 0.01 for the other layer. The first layer was trained for 200 epochs while the other layer 60 epochs. For the discriminative training, the BP algorithms with stochastic gradient descent was used. The mini-batch is 128, and the initial learning rate is 0.008. At the end of each epoch, if the substitution error rate on the development set decreased less than 0.1, the learning rate begins to halving and then continue until the substitution error on the development set increased.

B. Build the syllable based speech recognition systems

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.

Firstly, despite the training problems, the WFST based decoding framework is used to test the syllable based acoustic models. 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.

The experiments are conducted to test the performance of syllable based models. The Initial/Finals models are 3 state left-to-right HMMs, and the states number of each HMM for 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 $4(3+1)$. In the experiments about the context dependent syllable, central syllable dependent decision tree based state tying (CD-Syl-SDT) are conducted and compared with central syllable independent decision tree based state tying (CD-Syl-SIT). The experimental results are listed in Table I.

TABLE I
Character error rate (CER) of GMMs based acoustic models.

Models	#senone	CER(%)	RTF
CD IFs	2509	25.71	1.42
CI Syllables	-	29.95	1.19
CD-Syl-SDT	4195	30.39	1.57
CD-Syl-SIT	2375	25.54	1.45
CD-Syl-SIT	3561	26.15	1.52

From the experimental results, we can find out that the CD-Syl-SIT outperformed CD-Syl-SDT, which indicates that the central syllable independent decision tree state tying can help ease the distribution unevenness of training data. Besides, CD-Syl-SIT have a slightly better performance than the CD IFs with similar real time fatcor.

C. DNNs based acoustic models

After building the syllable based speech recognition systems, the experiments about DNNs are conducted, and the experimental results are placed in Tabel II.

TABLE II
The comparison of different setups of GMMs and DNNs based acoustic models.

Models	#senone	CER(%)	
		GMMs	DNNs
CD IFs	2509	25.71	20.18
CI Syllables	-	29.95	19.81
CD-Syl-SDT	4195	30.39	19.78
CD-Syl-SIT	2375	25.54	20.03
CD-Syl-SIT	3561	26.15	19.64

From the experimental results, we can figure out that, with DNNS, CI Syllables and CD-Syl-SDT can obtain good performances. Compared with the performance of GMMs, the DNNs based models show a good property of avoiding data sparsity and unevenness problems. However, the readers may argue that, the performance did not get a significantly improvement while converting from the context independent to the context dependent for syllable models. The reasons come from different aspects. Firstly, the context independent syllables already have a strong constraint of the speech observations, and the contextual information may make sense when enough speech training data employed. Secondly, the experimental dataset is a broadcast news database, and the introducing of contextual information for the conversational speech or spontaneous speech would bring about much more improvement.

There are two main aspects which makes the syllable based models having not been adopted widely in Chinese ASR. Firstly, to conquer the decoding problem, the senone-level WFST based decoding framework is the key. As for the data sparsity and unevenness problems, the central syllable independent decision tree based state tying can help the GMMs based models, and the DNNs based models can avoid these problems. Besides, the DNNs based models significantly outperform the GMMs based models. Thus, the reasonable choice of building a syllable based Chinese speech recognition systems is introducing the senone-level WFST based decoding framework and the DNNs based acoustic model.

VI. Conclusions and future work

This paper focuses on how to build a syllable based Chinese speech recognition system. Chinese is naturally a syllabic language and each basic language unit can be phonetically represented by a syllable. The good properties of making syllable as the basic acoustic modeling units have attracted many efforts and researches. The problems to build such a Chinese ASR systems come from the model training and decoding framework. To fix these problems, in the experiments, the WFST based decoding framework is introduced, the decision tree based state tying strategy is discussed, and the DNNs based acoustic modeling method is utilized. Compared with these experimental results, we can find out that, the DNNs based model is the best choice of for syllable modeling in Chinese ASR.

However, the above works are still preliminary and coarse. There are several works need to be done in next phase. Firstly, the performance of utilizing the context dependent syllable models should be further tested in other speech tasks, such as spontaneous speech, spoken dialogue. Secondly, a large speech corpus should be considered. The experiments in this paper have shown the solution to the data sparsity caused by context dependent syllable based models, but we believe that, when a lager speech corpus is used, the context dependent syllable based models can obtain a better performance improvement than the context dependent Initial/Finals based models.

Acknowledgment

The work was supported in part by the National Basic Research Program of China (2013CB329304) and National Natural Science Foundation of China (No. 61121002, No.91120001).

References

[1] A. Mohamed, G. Dahl, and G. Hinton, “Deep belief networks for phone recognition,” Proc. NIPS Workshop Deep Learning for Speech Recognition and Related Applications, 2009.

[2] A. Mohamed, G. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks” IEEE Trans. Audio Speech Lang. Processing. 20(1):14-22, Jan. 2012.

[3] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pretrained deep neural networks for large-vocabulary speech recognition,” IEEE Trans. Audio Speech Lang. Processing, 20(1)30-42, Jan. 2012.

[4] D. Yu, L. Deng and G. Dahl: “Roles of pretraining and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition,” in Proc. NIPS Workshop Deep Learning and Unsupervised Feature Learning, 2010.

[5] H. Wu, X.H. Wu, “Context Dependent Syllable Acoustic model for Continuous Chinese speech recognition,” Proc. Interspeech:1713-1716, 2007.

[6] X.G. Li, Y.N. Yang, and X.H. Wu, “A Comparative Study on Selecting Acoustic Modeling Units in Deep Neural Networks based Large Vocaulary Chinese Speech Recognition,” unpublished.

[7] Y. Bengio, “Learning deep architecture for AI,” Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1-127, 2009.

[8] G. Hinton, “A practical guide to training restricted Boltzmann machines,” Technical Report UTML TR 2010-003, University of Toronto, 2010.

[9] D. Yu, L. Deng, G, Li, and F. Seide, “Discriminative pretraining of deep neural networks,” U.S. Patent Filling, Nov. 2011.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Processing Mag. 29(6):82-97, 2012.

[11] F. Pereira, M. Riley, M. Mohri, “Weighted automata in text and speech processing,” Proceedings of the ECAI Workshop. 1996.

[12] D. Moore, J. Dines, M. M. Doss, et al, “Juicer: A weighted finite-state transducer speech decoder,” Machine Learning for Multimodal Interaction. Springer Berlin Heidelberg, 2006: 285-296.

[13] S. Young, G. Evermann, M. Gales, et al, *The HTK book,* Cambridge University Engineering Department, 2002.