2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Deep Neural Networks for Mandarin Tone Recognition

Mingming Chen, Zhanlei Yang and Wenju Liu

Abstract— **This paper investigates the application of deep models including deep maxout networks(DMNs) to Mandarin tone recognition. Our focus is on the capacity of extracting high-level robust features and fusing different kinds of serially-concatenated features of deep models. Furthermore, Maxout networks have been proposed to integrate dropout naturally and achieve state-of-the-art results. Therefore, we investigate the advantage of DMNs when the training data is limited and imbalanced. Our experiments on the ASCCD corpus show that comparing with shallow models such as one-hidden layer multi-perception (MLP) and support vector machine(SVM), deep models improve Mandarin tone recognition significantly. Among the deep models, DMNs can get better performance comparing with other deep neural networks based on sigmoid units or rectified linear units(ReLU).**

I. Introduction

DEEP Neural Networks(DNNs) have achieved great success in speech recognition tasks in recent years[1]. Deep neural networks have several advantages for getting much better performance: firstly, they can extract high-level features which are robust to the variations of raw input features through multiple non-linear hidden layers[2]; secondly, they can fuse multiple serially-concatenated features more efficiently; finally, they can be prevented from overfitting using dropout as well as other techniques. Furthermore, high-level features can vary with different kinds of non-linearity function even with the same network architecture. Following the recent success of DNNs based on sigmoid units, several kinds of activation functions have been used in deep neural networks. Rectified Linear Unit(ReLU) [14] has become popular recently which is a simple activation function y = *max(0,x)*. Alongside with new kinds of non-linearities, dropout training has been proposed to prevent overfitting in training deep neural networks because wide and deep neural networks are vulnerable to overfitting. Combining ReLU and dropout training, significant gain has been achieved in [3]. More recently, the maxout non-linearity which can be regarded as a generalization of ReLU was proposed. It is a function $y = max_i x_i$ that takes the maximum over groups of inputs which are put in groups of 3. Maxout networks, combined with dropout training, have given state-of-the-art performace in various computer vision tasks[4], and have achieved improvements in speech recognition task[5].

Mandarin Chinese (or Standard Chinese) is a well-known tonal language in which every syllable is assigned a tone, and the tones play an important lexical role. There are four basic lexical tones (referred to as Tone 1, 2, 3, 4, respectively) plus a neutral tone in Mandarin. In Mandarin, a syllable with different tones corresponds to different words. For example, the syllable *ma* can represent different words. *mā*, *má*, *mǎ*, *mà* refer to the syllable with tone 1, 2, 3, 4 correspondingly. They correspond to different chinese characters meaning *mother, hemp, horse and scold* . As a result, tone recognition plays an very important role in mandarin speech recognition. Tone recognition can be improve the performance of Mandarin speech recognition tasks either by rescoring on the word lattice [6], or by directly including the prosodic features in acoustic model [7]. It has been reported for long that the four lexical tones are characterized by perceptually distinctive pitch patterns, while the neutral tone, according to [8], does not have any specific pitch patterns, and is highly dependent on a shorter duration and a lower energy. In recent studies [9], [10] , it is known that Mandarin tone recognition task is difficult because the prosodic pattern realizations of different tones varies under various conditions such as the co-articulation effect from context syllables[11]. Machine learning approaches have been adopted to classify tones under different tone pattern realizations[6], [12]. Traditionally, only shallow models such as SVM are utilized for tone task which cannot extract features robust to variations of tone under different conditions.

In this work, we investigate the application of deep neual networks including deep maxout networks(DMNs) to Mandarin tone recognition. The difficulty of tone recognition is the prosodic pattern realizations of different tones varies under various conditions such as the co-articulation effect from context syllables[11]. Deep neural networks are good at modeling the variations of input features because the hidden layers of deep models can be seen as feature extractors whose output features are robust to variations of input features. Our experiments on ASCCD corpus show that DNN based on sigmoid, ReLU and maxout units produce 2.16%, 2.30%, 3.37% absolute improvement correspondingly over state-of-the-art shallow model—SVM.

The rest of this paper is organized as follows. The prosody labelled corpus used in this paper are described in section 2. Section 3 introduces the acoustic features used in tone recognition in detail. In section 4, deep neural networks with different non-linear units are presented. The experimental results and the analysis of results will be reported in section 5. The conclusion will be summarized in the final section.

Mingming Chen is a PhD student of National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Beijing China. (e-mail: mmchen@nlpr.ia.ac.cn).

Zhanlei Yang is an assistant researcher of National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Beijing China. (e-mail: zhanlei.yang@nlpr.ia.ac.cn).

Wen-Ju Liu, was with the Electrical Engineering Department Institute of Automation, Tsinghua University , Beijing China. He is now with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing China. (e-mail: LWJ@nlpria.ac.cn).

This research was supported in part by the China National Nature Science Foundation (No.91120303, No.61273267 and No.90820011).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

II. THE CORPUS

In this paper, we conduct our experiments on the Mandarin prosodic annotation speech corpusASCCD. ASCCD is designed for TTS and labeled with prosody. The text of ASCCD contains 18 pieces of narration or argumentum. There are 2 5 sections and 500 600 syllables in each piece. The corpus includes 10 speakers, five male and five female who are M001, M002, M003, M004, M005, F001, F002, F003, F004, F005 separately. The tones were labeled as 0,1,2,3,4 where 0 stands for neutral tone and label 1-4 stand for tone 1-4 correspondingly. The speech was annotated based on SAMPA-C system [13] to describe sound variation phenomena. The corpus contains 782 utterances and 79679 words. Each word was labelled by a tone. Table I lists the distribution of tones in the corpus.

TABLE I
THE DISTRIBUTION OF TONES IN ASCCD

total	neutral	Tone1	Tone2	Tone3	Tone4
79679	7936	15012	18046	13060	25625

From table I, we can conclude that the distribution of tones in this corpus is imbalanced and the samples which can be used in training set is somehow limited.

III. THE ACOUSTIC PROSODIC FEATURES

When extracting acoustic prosodic features for each syllable ,we should consider the influences of its context. Most Chinese words are monosyllabic or disyllabic, the previous syllable has more influences than the following syllable on tone[12]. Then we choose the two previous syllables and one following syllable of current syllable as the contextual window. For every syllable, we extract the following acoustic prosodic features:

* Pitch-related features (16-dimensions): To extract pitch-related features, we split log-F0 pitch contour of each syllable into three segments with equal length. We extract the mean and slope of the linear approximation of the pitch contour for the three segments. To use the context syllable information, we use the same above mean and slope of the last segment of the preceding syllable, and the first segment of the following syllable. We also use the first frame and last frame, pitch minimal and maximal pitch of the whole syllable, and the last voiced frame pitch of the preceding syllable, and the first voiced frame pitch of following syllable.
* Duration-related features (4-dimensions): the duration of current syllable(second), the normalized duration of current syllable, the duration ratio between the current syllable and the previous syllable, the duration ration between the current syllable and the following syllable. We use z-score method to normalize duration.
* Energy-related features (6-dimensions): the minimum, maximum, mean, range(maximum minus minimum), standard deviation and root mean squared(RMS) of log energy for current syllable.
* dynamic features (26-dimensions): for the above three kinds of features, we compute dynamic features in contextual window, which means that the current syllable statistics is normalized by corresponding statistics in the contextual window.

Finally, we extract 52 dimensions acoustic prosodic features for each syllable.

IV. DEEP NEURAL NETWORKS

In this paper, we refer to deep neural networks as feed-forward neural networks that contain many (at least two) hidden layers of non-linear hidden units. According to different kinds of non-linear hidden units, DNNs can achieve different performance tasks. Traditionally, sigmoid units are utilized as non-linear units in DNNs. Recently, rectified linear units(ReLU) and maxout units have been proposed to replace the sigmoid units used in DNNs. With the help of dropout training, it has been reported that the new two kinds of DNNs can achieve better performance on speech recognition and object recognition tasks[4], [5], [15], [16].

A. Rectified non-linearities

Rectified linear unit is a half-rectification non-linearity which is linear for positive values and zero otherwise[14]. It can bring several advantages for DNN. First, it is not necessary to do unsupervised pretraining when we train the DNNs based on recitifed linear units. Second, DNN based on ReLU can converge faster than those based on a regular sigmoid unit with the same topology. Third, DNN based on ReLU is piece-wise linear and become a linear convex system if we consider the units that are non-zero. Therefore they are simple to optimize even using the first-order optimizers. Fourth, DNN based on ReLU generalizes better than its sigmoid couterpart because the internal representation produced by them are much more regularized. Rectified linear units often output exact zero instead of small positive values when the input is not aligned with the internal weights. We can interpret improved generalization as the effect of the increased sparsity of the internal representation.

B. Maxout non-linearities

Maxout is a kind of neural network activation function proposed in [4]. In this paper, deep maxout networks are referred to feed-forward neural networks who have more than two hidden layers in which maxout is used as activation functions.

Given an input $x \in \mathbb{R}^d$ which has *d* dimensions, a maxout hidden layer implements the function

$$h_i(x) = max_{j\in[1,k]} z_{ij}$$

where $z_{ij} = x^T W_{\cdot ij} + b_{ij}$, and $W \in \mathbb{R}^{d\times m\times k}$ and $b \in \mathbb{R}^{m\times k}$ are learned parameters[4]. When training with dropout, the elementwise multiplication is performed with the dropout mask immediately prior to the multiplication by the weights in all cases–the inputs are not dropped to the max operator. See Fig.1 for how maxout unit works.

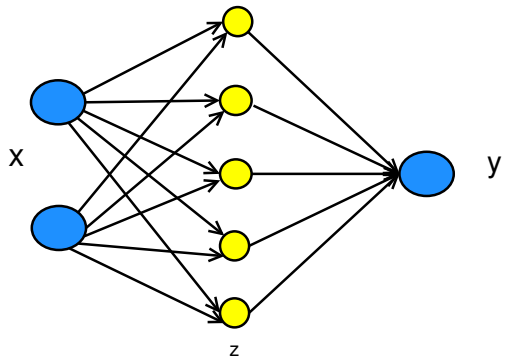

Fig. 1. The maxout unit. X can be input layer or hidden layers, z can be seen as the "hidden hidden layer" whose unit is the weighted sum of the input X, y get the max value of z.

In dropout training, the input X(excluing the input layer) is element-wise multiplied by dropout mask. The input of max operator is not dropped.

V. Experiments

In this part, we apply different classifiers(shallow and deep models) on ASCCD corpus to evaluate their performance on the task of tone recognition. Firstly, we evaluate the performance of two kinds of state-of-the-art shallow models—one-hidden-layer mlp and SVM on this task. Secondly, we use DNNs based on sigmoid units, DNNs based on ReLU units and DMNs to recognise tones.

A. Experiments setup

As stated in Section II, there are 10 speakers in the corpus. Each speaker read about 77 sections. For every speaker, we randomly select 50 sections as the training set, 10 sections as the developing set to tune the model parameter and the remaining sections as the testing set. The ratio between the size of training, developing and testing set at the sentence level is roughly 5:1:2. The training, developing and training set has 50237, 9820 and 16628 syllables correspondingly. The distribution of tones in the training, developing and testing set is listed in Table II.

TABLE II
THE DISTRIBUTION OF TONES IN THE TRAINING, DEVELOPING AND TESTING SET

tone label	total	n	1	2	3	4
training set	53241	5312	10045	12055	8853	16976
developing set	9820	984	1890	2714	1532	3211
testing set	16628	1620	3077	3817	2675	5439

B. Experiments Results and analysis

1) The shallow models: We first use two state-of-the-art shallow models—one-hidden-layer MLP and SVM to do tone recognition. For MLP, all the weights in the model is initised by an guassian distribution $N(0, 0.1)$ and the biases is set to zero. The dimension of the input layer is the same as that of the feature vector and the output layer has 5 units which are correspond to the five different tones. For the hidden layer, we try different number of hidden units. Figure 2 shows the results of tone recognition with different number of hidden units.

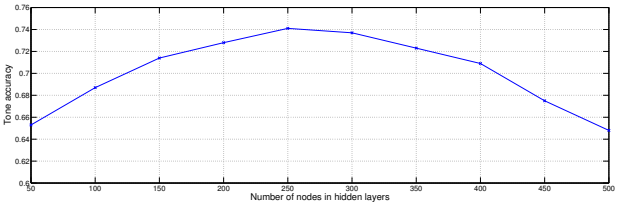

Fig. 2. Tone classification results vary according to the number of the nodes in hidden layer in one-hidden-layer MLP

From this figure, we can see when the number of hidden units increased from 50 to 250, the recognition performance gets better. This is because the capacity of MLP is limited when the dimension of hidden layer is small and the model can not efficiently model the training data. We get the best results—74.19% when the number of hidden nodes is 250. When the dimension of hidden layer is larger than 250, the accuracy of tone recognition gets worse slowly as the dimension increases. It can be explained as when the dimension of hidden layer is larger than 250, the capacity of MLP is enough for the training data. When the dimension increases further, the training data is not enough for the model and the model is underfitting.

We then use support vector machine (SVM) to recognise the tones. We use the LIBSVM [17] to train RBF kernel SVMs. After grid search for the parameter C and gamma, we get the best results when C is set to 32 and gamma is set to 0.125. The best model produces the recognition accuracy 74.75% which is slightly better than best result of one-hidden-layer MLP.

2) The deep models: In this section, we compare the performance of different deep neural networks on tone recognition tasks. We apply deep neural networks based on sigmoid units, rectified linear units and maxout units to recognize tones in Mandarin. To analysis the influence of different non-linear activation units, we use the same network architecture for all the deep models. For all deep models, the dimension of the input layer is the same as the dimension of input feature vector, the hidden layers have the same number of hidden units and the output layer has five output units. Considering the dimension of the input layers, the count of the training data and the performance of one-hidden-layer MLP, we vary the number of hidden layers from 2 to 4 and the dimension of hidden layers from 100 to 300 to get the best network architecture for this tasks.

For deep neural networks based on sigmoid units, we first pretrain Deep Belief Networks with different non-linear units [1,2] and use GPU to accelerate SGD. Since the inputs are real-valued, the first layer is pretrained as a Gaussian RBM. The weights of the first layer are sampled from a Gaussian Distribution N (0,0.01) and the visual biases are initialized

to zero. The pretraining is done using SGD. The SGD uses minibatches of 128 frames. We use momentum to speed up learning. Momentum starts at 0.5 and increases linearly to 0.9 over 20 epochs. A learning rate of 0.01 on the average gradient is used and an L2 weight decay of 0.0002 is used. The model is trained for 80 epochs.

All subsequent layers are trained as binary RBMs. A learning rate of 0.08 is used. The visible bias of each unit was initialized to log(p/(1-p)) where p is the mean activation of that unit in the dataset. All other hyper-parameters are the same as those used in the first layer. Each layer is pretrained for 40 epochs. The pretrained DBN is used to initialize the weights in the deep neural network.

For deep neural networks based on rectified linear units and DMNs, the weights are sampled form Uniform distribution U(-0.1, 0.1) and the biases are set to zero. It has been reported dropout training can improve the performance of the deep models when the models are prune to overfitting. Therefore, for all the three kinds of deep models we use dropout training to get the final models. In our experiments, we get the best tone accuracy for all the deep models when the dropout rate is set to be 0.2.

We present the recognition performance of all the deep models with different structures in Table III. We use DNN-sigmoid, DNN-ReLU and DMN represent deep neural networks using sigmoid, rectified linear and maxout function as their activation function. To compare with shallow models conveniently, we also list the accuracy of MLP and SVM at the bottom of the Table.

TABLE III

TONE ACCURACY OF DEEP MODELS USING DIFFERENT STRUCTURES AND DIFFERENT NON-LINEAR UNITS. $a \times b$ IN THE COLUMN OF CONFIGURATIONS MEANS A DEEP MODEL WITH a HIDDEN LAYERS AND EVERY LAYER HAS b UNITS.

tone models	configurations	developing set	testing set
DNN-sigmoid	2x200	73.52	71.63
	2x250	77.17	74.57
	3x150	**79.01**	**76.91**
	3x200	78.95	76.74
	4x100	78.03	75.97
	4x150	77.65	75.24
DNN-ReLU	2x200	73.79	71.76
	2x250	77.87	74.93
	3x150	**79.24**	**77.05**
	3x200	79.03	76.91
	4x100	78.42	76.19
	4x150	77.62	75.17
DMN	2x200	74.15	72.17
	2x250	78.09	75.69
	3x150	**80.18**	**78.21**
	3x200	79.95	78.02
	4x100	77.87	75.88
	4x150	77.42	75.37
MLP	1x250	76.03	74.19
SVM	-	76.68	74.75

From Table III, we can see all the deep models achieve the best performance with three hidden layers which contains 150 hidden units in each layer. We can conclude that all the deep models improve the performance of tone recognition significantly comparing with the results in Table II. Among the deep models, DMNs achieves the best performance. We think there are two reasons for the deep models achieving better results. Firstly, they can extract a more robust high-level feature through multiple non-linear hidden layers. Secondly, the deep models can fuse the multiple serially-concatenated features more efficiently than shallow models. To prove the above two reasons, we analysis the influence of pitch, duration and energy related features on Mandarin tone recognition. Table IV shows the results.

TABLE IV

THE INFLUENCE OF PITCH, DURATION AND ENERGY RELATED FEATURE ON MANDARIN TONE RECOGNITION

tone models	Mask related features	developing set	testing set
SVM	Pitch	72.95	71.63
	Duration	74.13	73.01
	Energy	73.76	72.64
	no mask	76.68	74.75
DMN	Pitch	75.74	73.72
	Duration	77.79	75.91
	Energy	77.36	75.53
	no mask	80.18	78.21

From Table IV, we can find that: 1) for each kind of features, deep models can achieve better performance which can prove the first reason. 2) when fuse all three kinds of features, deep models produce much better results than shallow models which proves they can fuse the features than the shallow model.

Among the deep models, DMNs achieve better results than others. We think that it is because maxout units can be interpreted as making a piecewise linear approximation to an arbitrary convex function which is more powerful than recitified linear units. Compared to the other two deep networks, deep maxout networks learn not just the relationship between hidden units but also the activation function of each hidden unit. Therefore, they can learn better representation based on the input features and produce better performance when the training data is limited and imbalanced.

VI. CONCLUSION AND DISCUSSION

In this paper, we apply deep neural networks with different kinds of activation function to tone recognition for Mandarin and compare these deep models with shallow models such as one-hidden-layer MLP, SVM. Our experiment results show that deep models improve the performance of tone recognition significantly because they can fuse different kinds of features more efficiently and extract more robust high-level features through multiple non-linear hidden layers. Furthermore, deep maxout models achieve the best performance among deep models because their advantage of preventing over-fitting when the training data is limited and imbalanced. In the future, we plan to use deep neural networks to fuse spectral features(MFCC, PLP) and acoustic prosodic features to improve the accuracy of Mandarin automatic speech recognition.

REFERENCES

[1] G.E. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P, Nguyen, T.N. Sainath and others, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Transactions on Signal Processing Magazine*, vol. 29, no, 6, pp. 82–97, 2012.

[2] Y. Bengio, *Learning deep architectures for AI*, Now Publishers, 2009.

[3] G.E. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR.*,abs/1207.0580, 2012.

[4] I.J. Goodfellow, D. Warde-Farley and M. Mirza and A. Courville and Y. Bengio, "Maxout Networks," *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*,vol.28, pp.1319-1327, 2013.

[5] Y.J. Miao, F. Metze, and S. Rawat, "Deep Maxout Networks for Low-Resource Speech Recognition," *Proceedings of 2013 Automatic Speech Recognition and Understanding Workshop*, 2013.

[6] L.W. Cheng and L.S. Lee , "Improved Large Vocabulary Mandarin Speech Recognition by Selectively Using Tone Information with a Two-stage Prosodic Model," *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, pp. 1137-1140. 2008

[7] H. Wang, Y. Qian, F. K. Soong, J. Zhou, and J. Han , "A Milti-Space Distribution (MSD) Approach to Speech Recognition of Tonal Languages," *Proceedings of the 7th Annual Conference of the International Speech Communication Association*, pp. 125-128. 2006.

[8] Y.R. Chao, "A Grammar of Spoken Chinese," *University of California Press*, Berkeley. 1968.

[9] C.Y. Chiang, H.M. Yu, Y.R. Wang , and S.H. Chen , "An Automatic Prosody Labeling Method for Mandarin Speech," *Proceedings of the 8th Annual Conference of the International Speech Communication Association*, pp. 494-497. 2007.

[10] C-Y. Chiang, S-H. Chen, and Y-R., Wang , "Advanced Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech and Its Application to Prosody Generation for TTS," *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, pp. 504-507. 2009.

[11] J.S. Zhang, K. Hirose , "Tone nucleus modelling for Chinese lexical tone recognition," *Speech Communication*, vol. 42, pp. 447-466. 2004.

[12] Y.B. Wang and L.S. Lee, "Mandarin Tone Recognition using Affine-Invariant Prosodic Features and Tone Posteriorgram," *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pp. 2850-2853. 2010.

[13] X. Chen, A. Li, G. H. Sun, H. Wu, Z.G Yin, "An application of SAMPA-C in standard Chinese," *Proceedings of the Sixth International Conference on Spoken Language Processing(ICSLP)*. 2000.

[14] V. Nair and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proceedings of the 27th International Conference on Machine Learning*, vol. 25, pp. 807-814, 2010.

[15] M.D. Zeiler, M. Ranzato, R. Monga , M. Mao, K. Yang, Q. V. Le, and G.E. Hinton, "On Rectified Linear Units for Speech Processing," *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing*, pp. 3517-3520. 2013.

[16] G.E. Dahl, T.N. Sainath, G.E. Hinton , "Improving Deep Neural Networks for LVCSR using Rectified Linear Units and Dropout," *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing*, pp. 8609-8612. 2013.

[17] C.C Chang and C.J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 27, pp. 1-27, Nov 2011.