A Novel Noise-Robust ASR Method by Applying Partially Connected DNN Model and Mixed-Bandwidth Concept

Lichun Fan, Hongyan Li, Dengfeng Ke, Bo Xu
Interactive Digital Media Technology Research Center
Institute of Automation, Chinese Academy of Sciences
Beijing, China
{lichun.fan, hongyan.li, dengfeng.ke, xubo}@ia.ac.cn

***Abstract*—In recent years, deep neural networks achieve significant improvements in automatic speech recognition. In this paper, we propose a deep structure used for robust ASR. The model has several partially connected layers which can suppress noise in different frequency bands. In order to recognize the speech data which has been distorted by noise seriously, we try to use parts of their frequency bands with a mixed-bandwidth model. The results have shown that the partially connected network could suppress noises in different frequency bands properly. The model's phone recognition on TIMIT corpus outperforms the state-of-the-art DNN model.**

Keywords-robust; ASR; DNN; mixed-bandwidth

I. INTRODUCTION

In recent years, with the development of the multi-layer neural network training techniques, deep neural networks (DNN) have received increasing attention. Due to the outstanding performance of deep neural networks in acoustic modeling [1, 2, 3, 4, 5, 6], CD-DNN-HMM gradually replaces CD-GMM-HMM as the preferred configuration.

Although DNN has greatly improved the speech recognition performance compared with GMMs, the speech recognition accuracy still degrades in noisy environments. Considering the powerful learning ability of DNN, we can overcome this problem by increasing the amount of training data. However, unless we collect training data in all of the noisy cases, this problem cannot be fully resolved. Meanwhile, many noise robust techniques developed under GMM in the past decades are no longer applicable in DNN. Model compensation techniques based on GMM such as VTS [7], will not work without the GMM model. Some robust features such as PNCC [8], which can make significant improvement in noisy condition in GMM, have unsatisfactory performance in DNN.

Neural network models have also been used for robust ASR in [9], but they focus on shallow neural network models. The core idea of deep learning and distributed representation is not expressed in their models. Work in [10] proposed a Deep Recurrent Denoising Autoencoder (DRDA), but the model needs stereo data for training. Convolutional neural networks bring significant improvement to image classification [11], but their improvement in speech recognition is small [12]. The latest research in [13] shows that CNN can get a better results through structure designed appropriately, and research in [14] also extends CNN to large data sets. The pooling operations in CNN handle some typical speech invariance in frequency domain while also suppress the discrimination ability of the network. Thus it will become confused when come across different speech sounds with similar formant frequencies.

In this paper we propose a deep neural network model called partially connected deep neural network. It has several hidden layers partially connected in order to enforce locality of features. Our model has a similar structure with the trap system [15], but our model is deeper. So we use distributed representation to describe such channel characteristics. Such a model could capture the relationship between the speech and noise in different frequency bands. Since the noise may concentrate in particular parts of the spectrum, the partially connected network could deal with the noise well in the individual bands.

When the Signal to Noise Ratio (SNR) is low, some frequency bands are distorted seriously by the noise. We replace these dimensions of the features with global mean or zero and consider them as narrowband data feature with high dimensions padded. Then we construct a mixed-bandwidth model [16]. The partially connected deep neural network can deal with this problem very conveniently.

II. THE PARTIALLY CONNECTED DEEP MODEL

A. Spectrum analysis

Limited to the vocal organ, the fundamental frequency of human's speech is confined to a relatively low frequency band. Accordingly, the energy of a lower band in speech is higher than that of a higher band. However, the energy distributions of most of the noises are nearly uniform. Therefore, the impact of noise on speech is more obvious in high frequency bands.

Figure 1 shows a speech distorted by babble noise. The upper panel shows a section of the speech signal. The noisy speech signal is made by adding 0dB SNR babble noise into the clean speech signal. We take a window of 32ms including 512 points from the 5000th sample in the clean and noisy speech, and then take Fourier transform using a long Hamming window after pre-emphasis. From the spectrum shown in the lower panel of figure 1, we can see that the distortion brought by the noise on voiced speech is more serious in the high frequency band, while the low-frequency

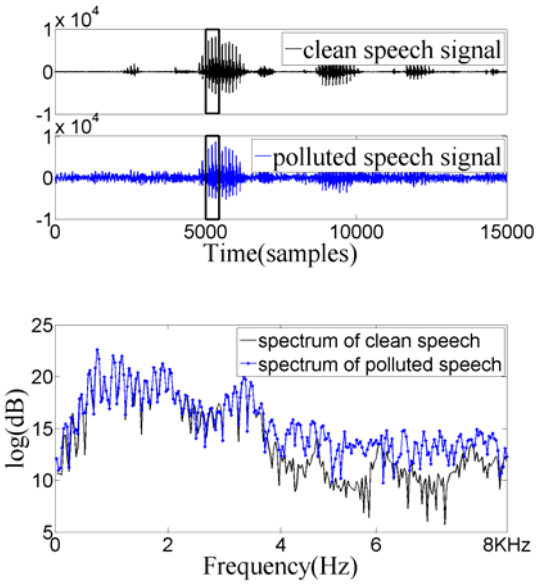

Figure 1. Noisy speech analysis. Upper panel: clean speech signal and noisy speech signal. Lower panel: short-time Fourier transform using a long (512 points) Hamming window on a section (the rectangle in upper panel) of voiced speech.

band distorted by the noise is less obvious. Therefore, taking different noise suppression measures at different frequency bands will achieve better results.

Historically, telephony channels have been band limited to no more than 4 kHz. As a result, speech data collected from telephone is narrowband (8 kHz sampling). The sampling rate of most record equipment is 16 kHz. Therefore, we consider 0-4 kHz as low frequency band and 4-8 kHz as high frequency band. Because the center frequency of the Mel-filter bank is non-linear in linear spectrum, we map the low frequency band and high frequency band to Mel-scale spectrum. We convert 4 kHz and 8 kHz to Mel-scale spectrum and get 932 and 1233 respectively. The results indicate that the ratio of low frequency band and high frequency band approximates to 3:1. Therefore, the first three-quarters dimensions in Mel-scale log filter-bank features represent the low frequency band while the last quarter represents the high frequency band.

B. The partially connected DNN model

Due to the powerful feature selection capabilities of the DNN [17, 18], we use it to learn the nonlinear relationships between noisy speech features and the new representational features.

We propose a partially connected DNN model to capture the distortion brought by complex noise under various environments. This model consists of one or more partially connected layers as shown in figure 2. The partially connected layers are divided into two or more bands which can receive different frequency bands input individually. Every part of the lowest partially connected layer only receives one part of the input features and there is no overlap. The rest of the partially connected layers connect to their own part of previous layer alone. So every part of the network would not affect each other when they suppress the noise on their sole channels. There are several fully connected layers on top of this model to combine all parts of features in the highest partially connected layer. These fully connected layers will produce the full representational features for classification.

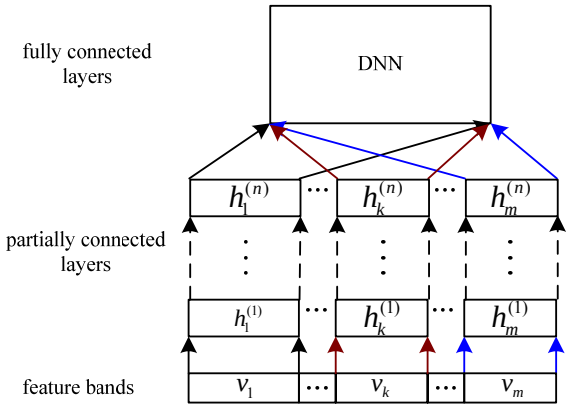

Figure 2. The partially connected DNN.

We assume the input features are divided into m parts frequency bands as $\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_3]$. The partially connected layers are also divided into m parts as shown in figure 2. We assume the partially connected layer's bands are denoted as $\boldsymbol{H} = [\boldsymbol{h}_1\ \boldsymbol{h}_2\ \ldots\ \boldsymbol{h}_3]$. Then the *kth* band activations of the partially connected layer can be computed from the kth band of the lower layer by the following equation:

$$\boldsymbol{h}_k = \theta\,(\boldsymbol{w}_k\boldsymbol{v}_k + \boldsymbol{b}_k)$$

where$\theta(*)$ is the activation function, $\boldsymbol{w}_k$ is the weight matrix, and $\boldsymbol{b}_k$ is the bias of the kth band.

Such a model could achieve better robustness against various noises especially when the noises are concentrated in several bands of the features.

In the training stage, the partially connected DNN will be pre-trained using DBN [19] firstly. The partially connected layers can be seen as several separate networks and should be pre-trained separately. The fully connected layers are pre-trained in the normal way. Then we use back-propagation algorithm to fine-tune the whole network. For a partially connected layer, the error signal from a band is back-propagated only to the lower layer nodes that generate the band's activation.

C. Mixed-bandwidth analysis

Studies in [16] proved that using mixed-bandwidth data can train a better DNN model under the assumption that the narrowband data can be considered as broadband data with some feature dimensions missing. This method not only can improve the recognition performance of the narrowband speech, but also can improve the recognition performance of the broadband speech. The core idea of this model is increasing the amount of training data while making different bandwidth data to use the same model.

In very noisy conditions, the high frequency band is distorted by the noise very seriously. This is shown clearly in figure 1. The high frequency band dimensions in Mel-scale log filter-bank features are so poor that they may not be helpful to the classification. Furthermore, they may add some negative points to the classification.

Considering the narrowband speech data also has a considerable recognition performance and the recognition rate gap between narrowband speech and broadband speech is not very large [16], we ignore the high frequency band in the speech features and only use the low frequency band in the training and test stage. In this way, we construct a "mixed-bandwidth" model just like the model described in [16]. But we use our partially connected DNN to generate the model.

The partially connected layers in this model are divided into two parts while the top layers are fully connected. We use the Mel-scale log filter-bank features and divide them into two parts: the low frequency band and high frequency band. The low SNR noisy speech features are preprocessed. We replace their high frequency dimensions by global mean, so they seem to narrowband speech features with high dimensions padded. Then the clean speech feature and noisy speech feature are treated equally. We input their low frequency band into the first part of the partially connected layer while input their high frequency band into the other part. Such a model is easy to be trained. The first part of the partially connected layers will be optimized by the low frequency band while the other part will be optimized by the high frequency band. Then the connections between fully connected layers will be optimized by all the frequency bands.

In testing, the low SNR speech must be detected firstly. Then they will be processed specially in order to become "narrowband" data. After that, all the data will be processed equally by the model. In our experiments, we manually add gigantic noise to form such a low SNR test set. The low SNR threshold value should be measured if this mixed-bandwidth method can improve the noisy speech recognition performance in low SNR condition.

III. EXPERIMENTS AND RESULTS

The TIMIT Acoustic-Phonetic Continuous Speech Corpus is used to evaluate the effectiveness of the proposed model. We use Mel-scale log filter-bank features as the input to DNN model and we divide the input features into two parts which represent the low frequency and high frequency bands individually. The Mel-scale log filter-bank features are extracted for 25ms speech window with a 10ms fixed frame rate. They have 40 coefficients and the corresponding first and second derivatives. The input layer of the model includes a context window of 11 frames. We divide them into 40 bands where each band includes 3×11 dimensions. Each dimension of the input feature is normalized to mean 0 and variance 1.

We use Kaldi [21] to pre-process the data sets. A learning rate annealing and early stopping strategies are utilized as in [1]. Our model has 7 hidden layers while the first two hidden layers are partially connected. Each of the full connection hidden layers has 1024 nodes. The partially connected layer divides the nodes into two parts, and each part has 1024 nodes. We input the first 30 bands of the feature (the low frequency band) to the first part of the partially connected layer and the last 10 bands of the feature (the high frequency band) to the second part. For comparison, the same DNN architecture is used for the setups.

A. Performance of the partially connected DNN model

The partially connected DNN model is designed to suppress the noise, so we evaluate the model on noisy condition. We corrupt the training data set with different noises and different SNRs. The five noise styles are Babble, Street, Market, Subway and Music. Part of them are from NOISEX92 [22] database. The five kinds of SNRs are 20dB, 15dB, 10dB, 5dB and 0dB. Each kind of the SNRs occupies a ratio of 15%, which means that 25% of the data has not been added any noise. The noise is randomly selected from the 5 styles. The test data set is also artificially corrupted by the five noises with different SNRs. We use phone error rate (PER) to evaluate the recognition performance. The experimental results are shown in table 1.

TABLE I. PER ON TIMIT DATA SET.

	DNN-C	**DNN-N**	**P-DNN-C**	**P-DNN-N**
clean	22.65	27.7	22.25	26.06
20dB	35.4	29.17	34.51	27.89
15dB	43.23	32.26	42.52	30.91
10dB	55	38.02	55.06	36.84
5dB	65.46	47.01	66.34	45.7
0dB	72.18	59.49	72.59	59.09

The "DNN-C" in the second column represents that we train the DNN model with the clean training data set. The results in this column are the baseline. Table 1 show that the recognition performance in clean condition is 22.65%. When the test data are distorted by noise, the recognition performances descend sharply. The PER is as high as 72.18% when the SNR is 0dB. Training the DNN model with noisy data (marked as DNN-N in table 1) could get better recognition performance in noise conditions. We get a relative PER reduction of 17.60% in the 20dB condition.

The partially connected DNN model trained with clean data, which is marked by P-DNN-C, is nearly the same with the DNN model trained with clean data. But P-DNN-N, which means the partially connected DNN model trained with noise, achieves 27.89% PER in 20dB condition. This is a relative 21.21% reduction compared with the baseline. These results prove the validity of the thinking that using separate networks to suppress noise in different frequency bands could get better performance.

We need to pay attention to is the relative reduction in 0dB condition between the best performance and the baseline is smaller than in other noisy conditions. The gap between the DNN model trained with noisy data and the partially connected DNN model trained with noisy data is even smaller in 0dB condition. We consider the high frequency

band is distorted by the noise so seriously that the model gets less useful information from them.

B. *Performance of the mixed-bandwidth model*

To evaluate the "mixed-bandwidth" model, we train the model with mixed-bandwidth data. We get a copy of the training data to make the "narrowband" data. These two data sets are distorted by noises just like the previous section. To generate the narrowband data, we replace the last 10 bands of the Mel-scale log filter-bank features with 0 (We have normalized each dimension of the features, so mean is equal to 0) in one of the training data sets. The narrowband test data sets are generated using the same method. The data sets are used to train and test on the partially connected DNN model. The results are shown in figure 3.

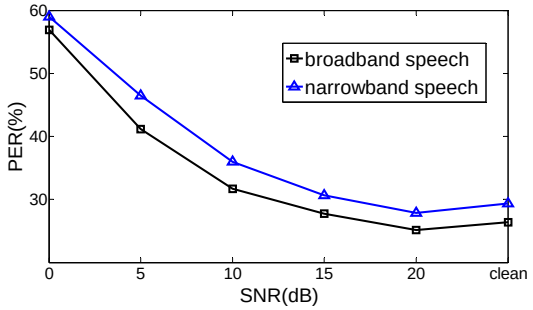

Figure 3. The mixed-bandwidth performance.

From figure 3 we can see that in all the test conditions the broadband data outperforms the narrowband data. This demonstrates the high frequency band in Mel-scale log filter-bank features is helpful even when it is distorted by the noise seriously. We cannot drop the entire high frequency band in any condition. However, the performances of the two speech data sets under 0dB are close. This proves that the effect of the high frequency band is slight when the environment is critical. We may drop some filter banks which are distorted seriously instead of the entry high frequency band.

IV. CONCLUSIONS

In this paper, we have proposed a deep structure used for robust ASR. The results have shown that the partially connected network could suppress noises in different frequency bands properly. The model's phone recognition on TIMIT data sets outperforms the state-of-the-art DNN model.

We tried to recognize the speech data which are distorted by noise seriously using their low frequency band alone and consider them as narrowband data. A mixed-bandwidth model has been constructed using the partially connected network. The experimental results demonstrate the recognition performance will descend if we drop the entire high frequency band. We will try to detect noise in every filter banks and suppress them individually in the future.

V. ACKNOWLEDGMENT

This work was supported by 863 Program in China (No. 2011AA01A207) and 973 Program in China (No. 2013CB329302).

REFERENCES

[1] G. Hinton, L. Deng, D. Yu and etc, "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Processing Magazine, vol. 29, 2012.

[2] A. Mohamed, G. Dahl and G. Hinton, "Acoustic modeling using Deep Belief Networks," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1,pp. 14-22, Jan, 2012.

[3] G. Dahl, D. Yu, L. Deng and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pp. 30-42, 2012.

[4] L. Deng, D. Yu, and J. Platt, "Scalable stacking and learning for building deep architectures," ICASSP, 2012.

[5] D. Yu, L. Deng, and F. Seide, "The deep tensor neural network with applications to large vocabulary speech recognition," IEEE Transactions on Audio, Speech,and Language Processing, vol. 21, no. 2, pp. 388-396, Feb, 2013.

[6] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," Interspeech, 2012.

[7] J. Segura, M. Benítez, A.Torre, S.Dupont and A. Rubio, "VTS residual noise compensation," ICASSP, 2002.

[8] C. Kim and R.M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition," ICASSP, 2012.

[9] S. Tamura and A. Waibel, "Noise reduction using connectionist models," ICASSP, 1988.

[10] A.L. Maas, Q.V. Le, T.M. O'Neil, O. Vinyals, P. Nguyen, and A.Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," Interspeech,2012.

[11] H. Schulz and S. Bchnke, "Learning object-class segmentation with convolutional neural networks," European Symposium on Artificial Neural Networks, 2012.

[12] O. Abdel-Hamid, A.R. Mohamed, H. Jiang and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," ICASSP, 2012.

[13] L. Deng, O. Abdel-Hamid and Dong Yu, "A deep convolutional neural netword using heterogeneous pooling for trading acoustic invariance with phonetic confusion," ICASSP, 2013.

[14] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," ICASSP, 2013.

[15] H. Hermansky and S. Sharma, "TRAPS-classifiers of temporal patterns," ICSLP 1998.

[16] J. Li, D. Yu, J.T. Huan, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth train data in CD-DNN-HMM," IEEE Workshop on Spoken language Technology, 2012.

[17] A. Mohamed, G. Hinton and G. Penn, "Understanding how deep belief networks perform acoustic modelling," ICASSP, 2012.

[18] G. Hinton, "Learning multiple layers of representation," Trends in cognitive sciences, Vol. 11, No. 10, pp. 428-434, 2007.

[19] G. Hinton, S. Osindero and Y.W. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, no. 7, pp. 1527-1554, 2006.

[20] K.F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," IEEE Transactions on Audio, Speech and Language Processing, 1989.

[21] P. Daniel, G. Arnab, B. Gilles and etc, "The Kaldi Speech Recognition Toolkit," IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.

[22] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: Ⅱ. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, vol. 12, pp. 247-251, 1993.