JOINT TRAINING OF FRONT-END AND BACK-END DEEP NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION

Tian Gao[1]*, Jun Du*[1]*, Li-Rong Dai*[1]*, Chin-Hui Lee*[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Georgia Institute of Technology, Atlanta, Georgia, USA

gtian09@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn, chl@ece.gatech.edu

ABSTRACT

Based on the recently proposed speech pre-processing front-end with deep neural networks (DNNs), we first investigate different feature mapping directly from noisy speech via DNN for robust speech recognition. Next, we propose to jointly train a single DNN for both feature mapping and acoustic modeling. In the end, we show that the word error rate (WER) of the jointly trained system could be significantly reduced by the fusion of multiple DNN pre-processing systems which implies that features obtained from different domains of the DNN-enhanced speech signals are strongly complementary. Testing on the Aurora4 noisy speech recognition task our best system with multi-condition training can achieves an average WER of 10.3%, yielding a relative reduction of 16.3% over our previous DNN pre-processing only system with a WER of 12.3%. To the best of our knowledge, this represents the best published result on the Aurora4 task without using any adaptation techniques.

Index Terms— robust speech recognition, deep neural network, feature mapping, joint training, system fusion

1. INTRODUCTION

With the fast development of mobile internet, the speech-enabled applications using automatic speech recognition (ASR) are becoming increasingly popular. However, the noise robustness is one of the critical issues to make ASR system widely used in real world. Historically, most of ASR systems use Mel-frequency cepstral coefficients (MFCCs) and their derivatives as speech features, and a set of Gaussian mixture continuous density HMMs (CDHMMs) for modeling basic speech units. Many techniques [1, 2, 3] have been proposed to address this issues. One category of techniques is the so-called data-driven approach based on stereo-data [4, 5], which is also the topic of this study.

The recent breakthrough of deep learning [6, 7], especially the application of deep neural networks (DNNs) in the ASR area [8, 9, 10], marks a new milestone that DNN-HMM for acoustic modeling becomes the state-of-the-art replacing GMM-HMM. It is believed that the first several layers of DNN play the role of extracting highly nonlinear and discriminative features which are robust to irrelevant variabilities. This makes DNN-HMM inherently noise robust to some extent as verified on the Aurora4 task [11].

In [12, 13], several front-end techniques were shown to yield further performance gains on top of the DNN-HMM system for tasks with small vocabularies or constrained grammars. However for large vocabulary tasks, the conventional enhancement approach as in [14], effective for the GMM-HMM systems, might even lead to a system degradation for DNN-HMM with log mel-filterbank (LMFB) features under the well-matched training-testing conditions [11].

Meanwhile, the data-driven approaches using stereo data via recurrent neural network (RNN) and DNN proposed in [15, 16] can improve the speech recognition accuracy on small vocabulary tasks. More recently, masking techniques [17, 18, 19] were successfully applied to noisy speech recognition. In [19], the approach using time-frequency masking combined with feature mapping via DNN claimed to achieve the best results on the Aurora4 task. Unfortunately, for multi-condition training using DNN-HMM with LMFB features, this approach still resulted in a worse performance similar to those concluded in [11]. In [20], we propose a pre-processing approach via DNN as a regression model to enhance noisy speech for robust speech recognition and was shown to outperform the masking approach [19].

In this study, we report our recent progress to further improve the ASR performance of multi-condition training especially when both additive noise and convolutional distortion are involved in the test data. First, instead of extracting acoustic features from the enhanced speech waveform, DNN is adopted directly as a highly nonlinear mapping function to estimate the clean speech features from observed noisy speech. Second, we employ a hybrid DNN architecture to joint train DNNs for both feature mapping and acoustic modeling. The proposed joint training allows error back-propagation to the feature mapping layers and the input of the hybrid DNN is

This work was supported by the National Natural Science Foundation of China under Grants No. 61305002.

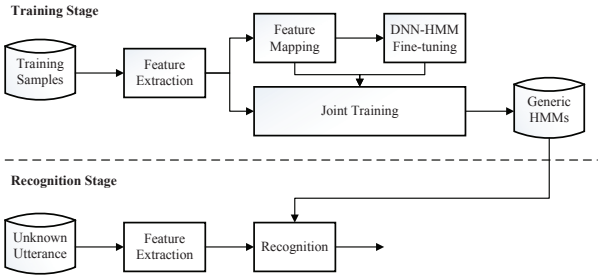

Fig. 1. Overall development flow and architecture.

the original noisy speech feature vectors. Third, an extensive experimental study is conducted to compare different acoustic features, namely LMFB, MFCC, and log-power spectra (LPS). Fusion of multiple DNN pre-processing systems with different features can also be performed.

With the proposed three-step approach, we show that features obtained from different domains of the DNN-enhanced speech signals are strongly complementary. Testing on the Aurora4 noisy speech recognition task our best system with multi-condition training achieves an average WER of 10.3%, yielding a relative WER reduction of 16.3% over our previous DNN pre-processing only system with a WER of 12.3%. To the best of our knowledge, this represents the best published result on the Aurora4 task without using any adaptation techniques. When compared with other enhancement approaches [11, 19, 21], we believe this is the first time to observe significant performance gains when both additive noises and convolutional distortions are involved in the test data on the Aurora4 task, indicating that the proposed front-end DNN can further improve the noise robustness on top of the DNN-HMM systems for large vocabulary tasks.

2. SYSTEM OVERVIEW

The overall flowchart of our proposed ASR system is illustrated in Fig. 1. In the training stage, training samples are firstly processed to extract acoustic features, namely LMFB, MFCC or LPS (with the dynamic features) followed by cepstral mean normalization (CMN). These features are further processed by DNN based feature mapping. Then the enhanced features are adopted to train the generic HMMs. For DNN-HMM system, we use the same procedure proposed in [20] to train DNN acoustic model with enhanced features. First, a reference DNN is trained using original features with clean training labels as our baseline system. Then, on top of this reference DNN as an initialization, the DNN model of enhanced features is fine-tuned by only changing the input of DNN from original noisy features to enhanced features. Meanwhile, the model can be further optimized by integrating feature enhancement with acoustic modeling. This joint training allows error back-propagation to the feature mapping

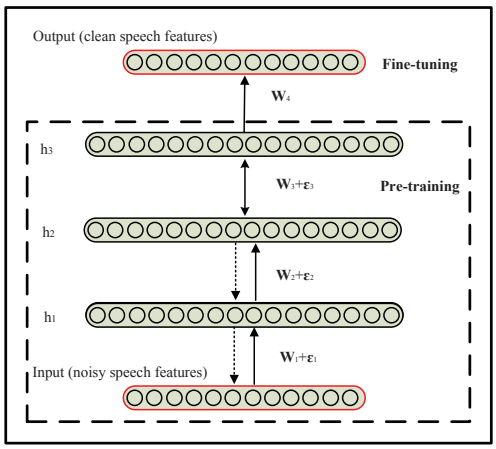

Fig. 2. DNN for feature mapping.

layers. With this joint training procedure, the hybrid DNN can generate better recognition performance. In the recognition stage, the normal recognition is conducted with the hybrid DNN-HMM. In the next section, the details of feature mapping and joint training are elaborated.

3. FEATURE MAPPING AND JOINT TRAINING

3.1. Feature Mapping

In [22], DNN was adopted as a regression model to enhance noisy speech and verified that this DNN-based pre-processing was effective for robust speech recognition [20]. But for test data with both additive noises and channel distortions (Set D) under the multi-condition training using DNN-HMM and LMFB acoustic features, this approach led to a performance degradation. To address this issue, we propose to directly map the input noisy features to the desired clean acoustic features as shown in Fig. 2. The acoustic context information along both time axis (with multiple neighboring frames) and frequency axis (with full frequency bins) can be fully utilized by DNN to improve the continuity of the estimated clean features.

One main difference from the previously proposed pre-processing only DNN is that the DNN output contains the same number of frames as the input instead of just one central frame [22]. Using multiple-frame output in feature mapping matches the input features with the back-end DNN for joint training to be proposed in the next section. As training of this regression DNN requires a large amount of time-synchronized stereo-data with clean and noisy speech pairs, which are difficult and expensive to be collected from real scenarios, the noisy speech utterances are synthesized by corrupting the clean speech utterances with additive noises of different types at various signal-to-noise-ratio (SNR) levels or convolutional (channel) distortions. DNN training al-

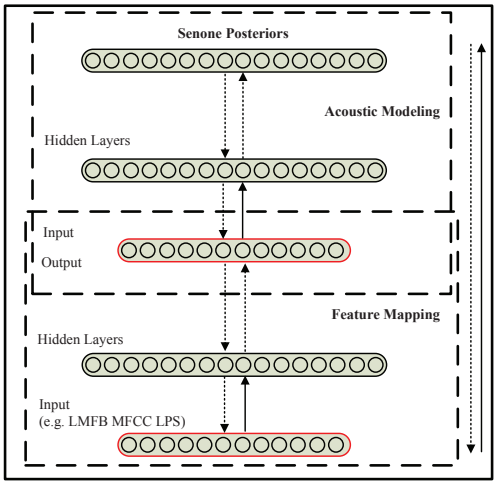

Fig. 3. Joint training of front-end and back-end DNNs.

so consists of unsupervised pre-training and supervised fine-tuning. The pre-training is the same as that in DNN for acoustic modeling. For supervised fine-tuning, we aim at minimizing mean squared error between the DNN output and the reference clean features:

$$E = \frac{1}{N}\sum_{n=1}^{N} \|\hat{\boldsymbol{x}}_{n-\tau}^{n+\tau}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}, \boldsymbol{b}) - \boldsymbol{x}_{n-\tau}^{n+\tau}\|_2^2 + \kappa\|\boldsymbol{W}\|_2^2 \quad (1)$$

where $\hat{\boldsymbol{x}}_{n-\tau}^{n+\tau}$ and $\boldsymbol{x}_{n-\tau}^{n+\tau}$ are the n^{th} $D(2\tau+1)$-dimensional vectors of estimated and reference clean features, respectively. $\boldsymbol{y}_{n-\tau}^{n+\tau}$ is a $D(2\tau+1)$-dimensional vector of input noisy features with neighbouring left and right τ frames as the acoustic context. $\boldsymbol{W}$ and $\boldsymbol{b}$ denote all the weight and bias parameters. κ is the regularization weighting coefficient to avoid over-fitting. The objective function is optimized using back-propagation with a stochastic gradient descent method in mini-batch mode of N sample frames.

3.2. Joint Training

After feature mapping, we used the enhanced features for acoustic modeling [20]. Furthermore, we employed a hybrid DNN framework to perform joint training of DNNs for both feature mapping and acoustic modeling as shown in Fig. 3. In the proposed hybrid DNN, we directly stacked the acoustic modelling layers on top of the feature mapping layers. The output layer of feature mapping became the input layer for acoustic modeling, which was also a hidden layer of the whole network. It is noted that this is a hidden layer with a linear activation function while others are with sigmoid activation functions. Using the same object function as the back-end DNN, all weights were then re-trained. After joint training, the hybrid DNN yielded a better recognition performance which can be explained because the feature mapping network was refined to enable a better phone classification instead of optimizing the original MMSE criterion. In [21], a joint training procedure was also proposed to combine the masking DNN with the back-end DNN. However due to the required middle-stage masking post-processing and dynamic feature calculation operations, fixed layers were used to perform such matching approximations. Instead our proposed joint training can seamlessly connect the front-end and back-end DNNs as the output layer of feature mapping DNN is exactly the input layer of the acoustic modeling DNN.

4. EXPERIMENTS

Aurora4 [23, 24] database was used to verify the effectiveness of the proposed approach for the medium vocabulary continuous speech recognition task. It contains clean speech data from WSJ [25] database. Two training sets were designed for this task. One is clean-condition training set consisting of 7138 utterances recorded by the primary Sennheiser microphone. The other one is multi-condition training set which is time-synchronized with the clean-condition training set. One half of the utterances were recorded by the primary Sennheiser microphone while the other half were recorded using a secondary microphone. Both halves include a combination of clean speech from clean-condition training set and speech corrupted by one of six different noises (street, train station, car, babble, restaurant, airport) at 10-20 dB SNR.

The two training set pairs were also used for feature mapping DNN. For evaluation, the original two sets consisted of 330 utterances from 8 speakers, which was recorded by the primary microphone and a secondary microphone, respectively. Each set was then corrupted by the same six noises used in the training set at 5-15 dB SNR, creating a total of 14 test sets. These 14 test sets were grouped into 4 subsets: clean (Set1), noisy (Set2 to Set7), clean with channel distortion (Set8), noisy with channel distortion (Set9 to Set14), which were denoted as A, B, C, and D, respectively.

As for the front-end, the frame length was set to 400 samples (or 25 msec) with a frame shift of 160 samples (or 10 msec) for the 16kHz speech waveforms. Three acoustic features were adopted, namely 24-dimensional log Mel-filterbank features, 13-dimensional MFCC (including C_0) features, and 257-dimensional log-power spectra features. These features plus their first and second order derivatives were concatenated to form LMFB, MFCC and LPS features and further processed by cepstral mean normalization. The 72-dimensional LMFB features were used to train DNN for feature mapping with the architecture 792-2048-2048-2048-792, which denoted that the size was 792 (72*11, τ=5) at the input layer, 2048 for three hidden layers, and 792 for the output layer. Other parameter settings can be referred to [22, 26]. Similarly, the DNN architectures for MFCC and LPS are 429(39*11)-2048-2048-2048-429, and 3855(771*5)-2048-2048-2048-3855, respectively.

Table 1. Performance (word error rate in %) comparison of the feature mapping systems using LMFB features with different output frames on the testing sets of Aurora4 databases.

System	A	B	C	D	Avg.
Noisy	4.6	8.4	7.8	18.6	12.5
FM_1	4.6	8.5	7.4	18.6	12.5
FM_2	4.6	7.8	7.2	17.5	11.7

Table 2. Performance (word error rate in %) comparison of the feature mapping and joint training systems using different acoustic features (LMFB, MFCC, LPS) under multi-condition training on the testing sets of Aurora4 databases.

System	A	B	C	D	Avg.
LMFB					
Noisy	4.6	8.4	7.8	18.6	12.5
Feature mapping	4.6	7.8	7.2	17.5	11.7
Joint training	4.3	7.8	6.8	17.0	11.4
MFCC					
Noisy	5.4	9.7	9.5	20.6	14.1
Feature mapping	5.1	9.2	9.0	19.4	13.3
Joint training	5.1	9.0	8.5	19.2	13.1
LPS					
Noisy	4.8	8.7	9.8	20.0	13.4
Feature mapping	4.6	7.9	8.6	19.3	12.6
Joint training	4.8	7.8	8.2	18.3	12.1

For acoustic modeling, each triphone was modeled by a CDHMM with 3 emitting states. There were in total 3296 tied states. For the DNN-HMM system, the input layer was a context window of 11 frames of LMFB features (or 11 frames of MFCC features, 5 frames of LPS features). The DNN for acoustic modeling had 7 hidden layers with 2048 hidden units in each layer and the final soft-max output layer had 3296 units, corresponding to the tied states of HMMs. The other parameters were set according to [11]. As for joint training, the learning rate is set to 0.001 and 7 epochs were used.

Table 1 gives a performance comparison of the feature mapping on the test sets of Aurora4 data using the LMFB features with different output frame sizes under multi-condition training. When using one-frame output (denoted as FM_1) it did not bring improvements over the baseline system (denoted as Noisy). However, when using multiple-frame output (denoted as FM_2), our approach could yield a remarkably relative WER reduction of 6.4% in average over the baseline. Comparing to the results in [11, 19, 21], a significant performance gain was achieved by only front-end feature processing under multi-condition training using DNN-HMM and LMFB features when both additive noises and convolutional distortions are involved in the test data (Set D).

Table 2 lists a performance comparison of the feature mapping and joint training using different acoustic features (LMFB, MFCC, LPS) under multi-condition training. For LMFB, joint training can yield a relative WER reduction of 2.6% in average over feature mapping and 8.8% over the baseline. For MFCC and LPS, the relative WER reductions over the baseline are 7.1% and 9.7%, respectively. More interestingly, LMFB achieves the best performance while the LPS features still outperforms the MFCC features. This observation is reasonable as LPS contains the most speech and noise information, contrary to the MFCC case. And LMFB achieves the best tradeoff among them.

Table 3. Performance (word error rate in %) comparison of the system combination with different features or DNN approaches on the testing sets of Aurora4 databases.

System	A	B	C	D	Avg.
DNN-PP	4.5	7.5	7.4	19.3	12.3
Post_Avg_1	4.4	7.5	6.6	16.1	10.9
Post_Avg_2	4.4	6.8	6.4	15.4	10.3
[21]	4.5	7.4	8.1	16.5	11.1

Table 3 shows a performance comparison of system combination with different features or DNN approaches on the testing sets of the Aurora4 task under multi-condition training. Our proposed DNN pre-processing only approach in [20] is denoted as DNN-PP. From Tables 2 and 3, the combined feature mapping with joint training approach can achieve consistent improvements over the pre-processing only approach, especially for Set D. We then used the posterior averaging method [27] to perform system combination. First, the combination of the three joint training systems (namely LMFB, MFCC, and LPS, denoted as Post_Avg_1) yields a WER reduction from 11.4% (the best result in the bottom row of Table 2) to 10.9% in average. Furthermore, the final system (Post_Avg_2) as a combination of Post_Avg_1 and DNN-PP attains the best reported WER (10.3%), which gives a relative WER reduction of 7.2% over the recently reported WER in [21]. These fusion results suggest that the enhanced features from different domains are strongly complementary.

5. CONCLUSION

In this paper, we have presented a novel DNN-based feature mapping approach with joint training for noise robust speech recognition. Compared with our previous work on DNN-based pre-processing, the feature mapping approach can significantly reduce the recognition error in all test conditions, especially when both additive noises and convolutional distortions are involved in the test data. Combined with joint training procedure, additional performance gain could be obtained. The final fusion system using different acoustic features and DNN approaches achieves the best reported WER on the Aurora4 task without any adaptation techniques.

6. REFERENCES

[1] Alejandro Acero, *Acoustical and environmental robustness in automatic speech recognition*, vol. 201, Springer, 1993.

[2] Yifan Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, 1995.

[3] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–33, 2014.

[4] Jasha Droppo, Li Deng, and Alex Acero, "Evaluation of the SPLICE algorithm on the AURORA2 database," in *INTERSPEECH*, 2001, vol. 1, pp. 217–220.

[5] Mohamed Afify, Xiaodong Cui, and Yuqing Gao, "Stereo-based stochastic mapping for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1325–1334, 2009.

[6] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[7] Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[8] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[9] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[10] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[11] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[12] Bo Li, Yu Tsao, and Khe Chai Sim, "An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition," in *INTERSPEECH*, 2013, pp. 3002–3006.

[13] Marc Delcroix, Yotaro Kubo, Tomohiro Nakatani, and Atsushi Nakamura, "Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?," in *INTERSPEECH*, 2013, pp. 2992–2996.

[14] Dong Yu, Li Deng, Jasha Droppo, Jian Wu, Yifan Gong, and Alex Acero, "A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4041–4044.

[15] Andrew L Maas, Quoc V Le, Tyler M O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y Ng, "Recurrent neural networks for noise reduction in robust ASR," in *INTERSPEECH*, 2012, vol. 1, pp. 22–25.

[16] Jun Du, Li-Rong Dai, and Qiang Huo, "Synthesized stereo mapping via deep neural networks for noisy speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1764–1768.

[17] William Hartmann, Arun Narayanan, Eric Fosler-Lussier, and DeLiang Wang, "A direct masking approach to robust ASR," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 1993–2005, 2013.

[18] Arun Narayanan and DeLiang Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7092–7096.

[19] Arun Narayanan and DeLiang Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 4, pp. 826–835, 2014.

[20] Jun Du, Qing Wang, Tian Gao, Yong Xu, Lirong Dai, and Chin-Hui Lee, "Robust speech recognition with speech enhanced deep neural networks," in *INTERSPEECH*, 2014, pp. 616–620.

[21] A. Narayanan and DeLiang Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 2504–2508.

[22] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE SIGNAL PROCESSING LETTERS*, vol. 21, no. 1, pp. 65, 2014.

[23] Guenter Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task," *ETSI STQ Aurora DSR Working Group*, 2002.

[24] N Parihar and J Picone, "DSR front end LVCSR evaluation," *Aurora Working Group*, 2002.

[25] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[26] Geoffrey Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, pp. 926, 2010.

[27] Bo Li and Khe Chai Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 279–284.