Deep Neural Network Based Spectral Feature Mapping for Robust Speech Recognition

Kun Han[1], *Yanzhang He*[2], *Deblin Bagchi*[3], *Eric Fosler-Lussier*[4], *DeLiang Wang*[5]

Department of Computer Science and Engineering[12345]
& Center for Cognitive and Brain Sciences[45]
The Ohio State University, Columbus, OH, USA

hank[1],hey[2],fosler[4],dwang[5]@cse.ohio-state.edu, bagchi.16@osu.edu[3]

Abstract

Automatic speech recognition (ASR) systems suffer from performance degradation under noisy and reverberant conditions. In this work, we explore a deep neural network (DNN) based approach for spectral feature mapping from corrupted speech to clean speech. The DNN based mapping substantially reduces interference and produces estimated clean spectral features for ASR training and decoding. We experiment with several different feature mapping approaches and demonstrate that a DNN trained to predict clean log filterbank coefficients from noisy spectrogram directly can be extremely effective. The experiments show that the ASR systems with these cleaned features perform well under joint noisy and reverberant conditions, and achieve the state-of-the-art results on the CHiME-2 corpus with stereo (corrupted and clean) data.

Index Terms: Robust Automatic Speech Recognition, Deep Neural Networks, Spectral Feature Mapping, Denoising, Deverberation

1. Introduction

Automatic speech recognition (ASR) has been making tremendous progress in the last few years and state-of-the-art systems achieve relatively satisfactory performance under clean conditions. However, ASR systems still suffer from performance degradation in the presence of acoustic interference, such as additive noise and room reverberation. Therefore, increasing attention has been drawn to the robustness of ASR systems.

For acoustic modeling, even the system is trained and tested under the matched noisy conditions, the performance is still lower than that trained and tested under the clean condition. It is primarily because these speech signals are corrupted by interference, leading to less informative extracted features for modeling. Therefore, robust speech recognition can benefit from feature enhancement and it would be important to design a front-end for this process to further improve the ASR performance under noisy and reverberant conditions.

Deep neural networks (DNNs) have shown strong learning capacity [1] and been successfully applied to many speech applications [2, 3]. Our recent studies [4, 5] have used DNNs for speech dereveberation and denoising. The DNN is trained to learn the spectral mapping from corrupted speech to clean speech, which effectively attenuates reverberation and noise in the spectral domain. It has been shown that the resynthesized time-domain signals using the DNN mapped spectral features significantly improve the predicted speech intelligibility as well as automatic speech recognition results [5]. However, to resynthesize time-domain signals, the phases of corrupted speech usually introduce distortion and lead to negative effects on speech spectrograms.

From the ASR standpoint, features are directly computed from spectral magnitudes and phases are not involved during feature extraction. This motivates us to extract acoustic features directly from the DNN-generated spectral features for ASR without time-domain signal resynthesis. In a typical ASR system, instead of directly using spectral magnitudes, the most common features include Mel filterbank features and Mel filterbank cepstral coefficients (MFCC) features. Therefore, it is valuable to employ a DNN to directly produce enhanced features either in the spectrogram domain or the Mel filterbank domain.

In this paper, we propose to use the DNN for spectral feature mapping to generate estimated clean spectral features from the noisy and reverberant signals for robust ASR. We experiment with three spectral feature mapping methods: spectrogram to spectrogram (spec-spec), Mel filterbank to Mel filterbank (fbank-fbank), and spectrogram to Mel filterbank (spec-fbank). Then, the DNN-generated features are used for acoustic modeling. Because DNN mapping substantially attenuates the noise and reverberation in the spectral feature domain, the ASR features are much cleaner than those from the original corrupted speech. We also show that the DNN is effective to learn not only clean representation from corrupted features but also filterbank transformation across different feature domains. The ASR system using the cleaned features achieves better performance under adverse conditions, and performs better than to a state-of-the-art masking-based approach [6].

2. Relation to prior work

To deal with noise, many DNN based methods have been proposed to improve the robustness of ASR systems. Noise-aware training is proposed to improve noise robust ASR in [7], where simple estimates of noise are appended to log Mel filterbank features for DNN acoustic modeling. Recurrent neural networks (RNNs) have been applied in a tandem system [8] or in a hybrid system with deep structure [9] to model temporal dependency of speech and noise for robust ASR. A memory-enhanced deep Long Short-Term Memory RNN is also used in acoustic modeling in combination with a Gaussian mixture model (GMM) system for noise robustness [10]. A joint training approach is used to improve the noise robust ASR using time-frequency masking in [11], which combines a speech separation module and an acoustic modeling module into a single DNN framework. Because the feature plays a critical role in acoustic modeling, some studies focus on feature enhance-

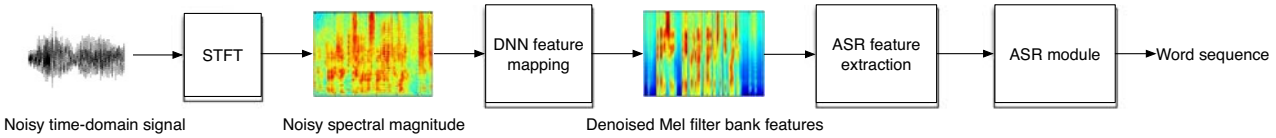

Figure 1: System overview.

ment for robust ASR. Most previous work focused on learning clean version of MFCC using either denoising autoencoders [12, 13] or RNNs [14]. In [15], spectral subtraction and a RNN based denoising auto-encoder are employed to learn Mel filterbank from noise. Some work has proposed to learn mapping directly on spectrogram domain, but the performance on ASR is not clear [16, 17, 18]. To our knowledge, few previous studies focus on the feature mapping across different feature domains, i.e., from spectrogram to Mel filterbank as discussed in this paper.

3. System description

We propose three different methods to learn the spectral mapping using the DNN: log spectrogram to log spectrogram, log Mel filterbank to log Mel filterbank, and log spectrogram to log Mel filterbank. Fig. 1 shows the overview of the system. We use the feature mapping from log spectrogram to log Mel filterbank as an example. A corrupted time-domain signal is first decomposed into the spectrogram domain. The corrupted magnitude spectrum is then fed into the trained DNN spectral feature mapping model to learn the clean representation of Mel filterbank. An enhanced Mel filterbank feature is generated from the DNN, directly followed by an ASR feature extraction module. The enhanced Mel filterbank features are either used directly as ASR features, or used to compute MFCC features and for ASR modeling and decoding. The following subsections discuss the system in detail.

3.1. Spectral feature mapping

We first train a DNN to perform spectral feature mapping for feature denoising and dereverberation. To train this DNN, we extract the spectrogram from corrupted speech as inputs and the spectrogram from clean speech as desired outputs. Specifically, we first divide the input time-domain signals into 25-ms frames with 10-ms frame shift, and then apply short time Fourier transform (STFT) to compute log spectral magnitudes in each time frame. For a 16 kHz signal, each frame contains 400 samples, and we use 512-point Fourier transform to compute the magnitudes, forming a 257-dimensional log magnitude vector $\mathbf{x}(m)$ in the mth frame:

$$\mathbf{x}(m) = [x(m,1), x(m,2), \ldots, x(m,K)]^T \qquad (1)$$

where, $x(m,k)$ is the log magnitude in each frequency bin, and $K = 257$ in this study.

Since temporal dynamics incorporates rich information for speech, we include the spectral magnitude vectors of neighboring frames into a feature vector. Therefore, the input feature vector for the DNN is:

$$\tilde{\mathbf{x}}(m) = [\mathbf{x}(m-d), \ldots, \mathbf{x}(m), \ldots, \mathbf{x}(m+d)]^T \qquad (2)$$

where d denotes the number of neighboring frames on each side and is set to 5 in this study. Therefore, the dimensionality of the input feature vector is $257 \times 11 = 2827$.

The desired output of the DNN is the log Mel filterbank features of clean speech in the current frame, denoted by a 40-dimensional feature vector $\mathbf{y}(m)$. If the DNN is trained to learn the clean spectrogram magnitudes, the output is a 257-dimensional feature vector.

We train a deep neural network to learn the spectral feature mapping from corrupted speech to clean speech. The objective function for optimization is based on mean square error. Eq. 3 is the cost for each training sample:

$$\mathcal{L}(\mathbf{y}, \mathbf{x}; \mathbf{\Theta}) = \sum_{k=1}^{K} (y_k - f_k(\mathbf{x}))^2 \qquad (3)$$

where y_k and $f_k(\cdot)$ are the desired and the actual output of the kth neuron in the output layer, respectively. $\mathbf{\Theta}$ denotes the weights we need to learn. The input features are normalized to zero mean and unit variance over all feature vectors in the training set, and the output is normalized into the range of $[0, 1]$. The activation functions of both the hidden layers and the output layer are the sigmoid functions. We use backpropagation with mini-batch stochastic gradient descent to train the DNN model, and the optimization technique uses adaptive gradient descent along with a momentum term [19].

Fig. 2 shows an example of the spectral features mapping for a sentence in CHiME-2 corpus [20]. Fig. 2(a) shows the log spectrum of the corrupted speech with both living room noise and reverberation, which is used as the input of the DNN. The corresponding corrupted Mel filterbank feature is shown in Fig. 2(b). Figs. 2(c) and (d) show the enhanced Mel filterbank feature produced by the DNN and the Mel filterbank feature of clean speech, respectively. Comparing Fig. (b) with Fig. (c), the energy caused by reverberation and additive noise is largely removed or attenuated, which is a good estimate of the features for the clean speech as shown in Fig. (d). Although the DNN does not perfectly generate the spectral structures of clean speech, the formant information is considerably restored which is essential to speech recognition.

For the DNNs trained to learn Mel filterbank features, we can directly extract MFCC features for ASR GMM modeling and use DNN-generated Mel filterbank features for ASR DNN training. For the DNN trained to learn spectral magnitudes, it is straightforward to apply Mel filterbank to DNN-generated spectrogram and then compute Mel filterbank and MFCC.

Note that, in order to ensure matched features, the DNN is treated as a front-end to enhance the features for both ASR training and test data.

3.2. ASR modeling

We use the Kaldi toolkit [21] to build the automatic speech recognition system for the task. We first build a GMM-HMM system using MFCC features. We concatenate 13-dimensional

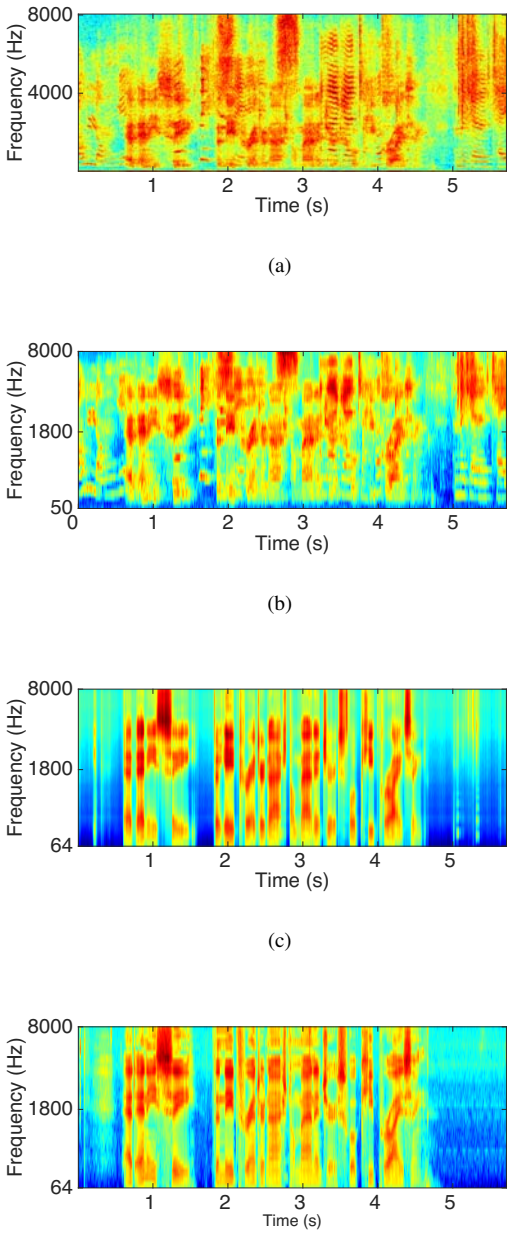

Figure 2: (a) Log magnitude spectrogram of corrupted speech with both noise and reverberation, (b) log Mel filterbank of corrupted speech, (c) DNN outputs, (d) log Mel filterbank of clean speech.

MFCCs from a context window of 7 frames, which are then de-correlated and compressed to 40 dimensions by linear discriminant analysis (LDA). This is followed by maximum likelihood linear transform (MLLT) for further de-correlation. In order to reduce speaker variance, we also apply feature-space maximum likelihood linear regression (fMLLR) on the resulting features, which is estimated by speaker adaptive training (SAT). The HMM is a 3-state cross-word triphone system with around 2000 tied states in the final system.

Then we build a DNN-HMM hybrid system using 40-dimensional log Mel filterbank features with their deltas and double-deltas from a 11-frame context window. We first pretrain the DNN generatively with stacked RBMs, which are then used to initialize the DNN with 7 hidden layers of 2048 sigmoid units. Then we train the DNN with the tied triphone state targets using the alignment obtained from the GMM-HMM system. Following [9], we realign the data with the trained DNN and retrain the DNN using the new alignment. We repeat this process for three times until the performance become saturated. We further improve the system by applying sMBR-based sequence training on the DNN [22]. For faster convergence of the sMBR training, we regenerate the lattices after the first iteration and train for 4 more iterations. For the decoding, we use the standard 5k trigram language models provided by the task.

4. Evaluation

4.1. Task and data description

We evaluate the effectiveness of our proposed neural networks for different spectral features mapping methods on Track 2 of the CHiME-2 challenge [20], which is a medium-vocabulary task for word recognition under reverberant and noisy environments without speaker movements. In this task, three types of data are provided based on the Wall Street Journal (WSJ0) 5K vocabulary read speech corpus: clean, reverberant and reverberant+noisy. The clean utterances are extracted from the WSJ0 database. The reverberant utterances are created by convolving the clean speech with binaural room impulse responses (BRIR) corresponding to a frontal position in a family living room. Real-world non-stationary noise background recorded in the same room is mixed with the reverberant utterances to form the reverberant+noisy set. The noise excerpts are selected such that the signal-to-noise ratio (SNR) ranges among -6, -3, 0, 3, 6 and 9 dB without scaling. The multi-condition training, development and test sets of the reverberant+noisy set contain 7138, 2454 and 1980 utterances respectively, which are the same utterances in the clean set but with reverberation and noise at 6 different SNR conditions.

4.2. Experimental settings

We first choose all sentences from the CHiME-2 clean training set and the reverberant+noisy training set to train the DNN based spectral feature mapping model. With the trained DNN model, we map the corrupted spectral features for all reverberant+noisy training, development and test utterances to estimated clean spectral features. For spectral feature mapping, we choose to use 2×2048 sigmoid units in the hidden layers to train the DNNs. The parameters in the experiments are tuned in the development set to optimize the performance for each method. For acoustic modeling (AM), we use 7×2048 sigmoid hidden units to train the AM DNN for all three feature mapping methods. All the three feature mapping methods use the same training recipe.

4.3. Results and discussions

The CHiME-2 corpus provides two channel signals, and we evaluate systems that take the average of both channel signals as input for DNN spectral feature mapping.

The results are shown in Table 1. For comparison, the baseline (noisy) is trained on the original reverberant+noisy train-

System	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Average
noisy	36.7%	26.5%	21.0%	16.4%	13.1%	11.6%	20.9%
fbank-fbank	34.0%	23.6%	19.0%	14.6%	12.6%	10.9%	19.1%
spec-spec	33.9%	24.0%	19.5%	14.0%	12.2%	11.1%	19.1%
spec-fbank	**30.6%**	**22.5%**	**17.8%**	**12.9%**	**11.3%**	**10.8%**	**17.6%**

Table 1: WER comparisons for three spectral feature mapping methods on the CHiME-2 corpus. "noisy" stands for the ASR baseline using noisy signals without a preprocessing front-end as inputs. "fbank-fbank" stands for the ASR system with the DNN based spectrogram to Mel filterbank features mapping. "spec-spec" stands for the DNN based spectrogram to spectrogram mapping. "spec-fbank" stands for the DNN based spectrogram to Mel filterbank features mapping. The best performance in each condition is marked in **bold**.

System	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Average
noisy	28.2%	20.7%	16.3%	13.1%	**9.49%**	9.10%	16.1%
fbank-fbank	31.2%	22.5%	17.9%	13.2%	10.6%	9.21%	17.4%
spec-spec	29.5%	22.2%	16.2%	12.6%	11.1%	10.1%	16.9%
spec-fbank	28.0%	19.9%	**14.8%**	**11.9%**	10.2%	**8.91%**	**15.6%**
Narayanan[6]	**25.6%**	**19.6%**	16.8%	13.8%	10.7%	10.6%	16.2%

Table 2: WER on the CHiME-2 corpus with the clean alignment.

ing set multi-conditionally without any pre-processing. It is a strong baseline, which by itself already outperforms the best entry in the challenge workshop (26.9%) [23, 20]. The three proposed systems in this paper directly utilize mapped spectral features for acoustic modeling, which achieve considerably better performance than the baseline on all SNR conditions. As is expected, the feature mapping front-end obtains bigger improvement in lower SNRs than in higher SNRs. The average relative improvements are 8.6%, 8.6%, 15.8% for fbank-fbank, spec-spec, and spec-fbank, respectively. In addition, the mapping from spectrogram to Mel filterbank yields the best results, suggesting that the DNN is capable of learning clean features across different feature domains and it is not necessary to perform denoising autoencoders to the same features. It is worth mentioning that all the proposed methods substantially outperform the spectral mapping method in [5], where the learned clean spectral magnitudes are used to resynthesize the time-domain waveforms for ASR modeling.

Note that our ASR system is a very powerful recognizer. We have trained the same ASR system on the clean dataset (WSJ0) that is parallel to the CHiME-2 noisy+reverberant corpus and achieved 2.5% WER on the clean evaluation set. This inspires us to utilize the AM DNN trained on the clean data to produce a better alignment which can be used for AM DNN training on the noisy data. This strategy has been used in previous studies [9, 6] and we present results on using these clean alignments to facilitate comparisons with previous approaches. We evaluate the ASR results using three spectral feature mapping methods with the clean alignment in Table 2.

In Table 2, the baseline system produces substantially better results with clean alignment, which even outperforms two of the spectral feature mapping methods (fbank-fbank and spec-spec). This might suggest that with very accurate clean alignment, a strong AM DNN is able to learn part of the denoising and dereverberation functions. However, the best results still come from the feature mapping from spectrogram to Mel filterbank, where the WER is decreased by 3.1% relative over the baseline. We also directly show the state-of-the-art results on the CHiME-2 corpus reported in [6], which we improve upon by about 0.6%[1]. Therefore, although clean alignment can boost ASR results under noisy and reverberant conditions, the DNN trained to learn the mapping from spectrogram to Mel filterbank can still improve the performance.

5. Conclusion and future work

In this paper, we have proposed to use DNNs for spectral feature mapping to improve ASR performance under noisy and reverberant conditions. The DNNs can be trained to learn the mapping for different feature domains and can produce an enhanced front-end for ASR feature extraction. The experiments show that our proposed approach significantly boosts ASR performance under noisy and reverberant conditions. Especially, the DNN can effectively learn the mapping from the spectrogram domain to the Mel filterbank domain, and the ASR system using this front-end achieves the state-of-the-art results on the CHiME-2 corpus.

In this study, we utilize stereo data to train a front-end DNN for spectral feature denoising and dereverberation, and use the enhanced features to train another DNN for acoustic modeling. It is interesting to train the two DNNs jointly, similar to the mask estimation in [11, 24]. Specifically, our approach provides a good initialization for both DNNs, then we can concatenate both and train a joint DNN for feature enhancement and acoustic modeling togther. This will be our future work.

6. Acknowledgements

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048), and an NSF grant (NSF IIS-1409431).

[1] While the Narayanan number here is presented for comparison, there are a number of distinctions between the systems: Narayanan's system fully trains the masking model jointly with the ASR, while our system includes sMBR discriminative training. Future work will look to isolate differences between these approaches.

7. References

[1] G. E. 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.

[2] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14–22, 2012.

[3] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, 2013.

[4] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. of ICASSP*, 2014, pp. 4661–4665.

[5] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *to appear in IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2015.

[6] A. Narayanan, "Computational auditory scene analysis and robust automatic speech recognition," Ph.D. dissertation, The Ohio State University, Columbus, OH, 2014.

[7] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Prof. of ICASSP*. IEEE, 2013, pp. 7398–7402.

[8] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust ASR," in *Proc. of ICASSP*, 2012, pp. 4085–4088.

[9] C. Weng, D. Yu, S. Watanabe, and B. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc. of ICASSP*, 2014.

[10] J. Geiger, F. Weninger, J. Gemmeke, M. Wollmer, B. Schuller, and G. Rigoll, "Memory-Enhanced Neural Networks and NMF for Robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 6, pp. 1037–1046, 2014.

[11] A. Narayanan and D. L. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. of ICASSP*, 2014, pp. 2523–2527.

[12] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder." in *Proc. of Interspeech 2013*, 2013, pp. 3512–3516.

[13] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. of IEEE ICASSP 2014*. IEEE, 2014, pp. 1759–1763.

[14] A. L. Maas, T. M. O'Neil, A. Y. Hannun, and A. Y. Ng, "Recurrent neural network feature enhancement: The 2nd chime challenge," in *Proc. 2nd CHiME Workshop on Machine Listening in Multisource Environments*, 2013, pp. 79–80.

[15] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *Proc. of ICASSP*, 2014.

[16] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Proc. of Interspeech 2013*, 2013, pp. 436–440.

[17] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[18] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. of Interspeech 2014*, 2014.

[19] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011.

[20] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. of ICASSP*, 2013, pp. 126–130.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, 2011, pp. 1–4.

[22] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proc. of ICASSP*, 2009, pp. 3761–3764.

[23] Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," *Proc. of CHiME-2013*, pp. 19–24, 2013.

[24] Z. Wang and D. Wang, "Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition," in *submission to Proc. of Interspeech 2015*, 2015.