INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Improving mask learning based speech enhancement system with restoration layers and residual connection

Zhuo Chen[1,2], *Yan Huang*[1], *Jinyu Li*[1], *Yifan Gong*[1]

[1]Microsoft Corporation
[2]Electrical and Engineering Department, Columbia University
{zhuc}@microsoft.com

Abstract

For single-channel speech enhancement, mask learning based approach through neural network has been shown to outperform the feature mapping approach, and to be effective as a pre-processor for automatic speech recognition. However, its assumption that the mixture and clean reference must have the correspondent scale doesn't hold in data collected from real world, and thus leads to significant performance degradation on parallel recorded data. In this paper, we first extend the mask learning based speech enhancement by integrating two types of restoration layer to address the scale mismatch problem. We further propose a novel residual learning based speech enhancement model via adding different shortcut connections to a feature mapping network. We show such a structure can benefit from both the mask learning and the feature mapping. We evaluate the proposed speech enhancement models on CHiME 3 data. Without retraining the acoustic model, the best bi-direction LSTM with residue connections yields 24.90% relative WER reduction on real data and 34.57% WER on simulated data.

Index Terms: single-channel speech enhancement, feature mapping learning, mask learning, residue net

1. Introduction

Recently the surge in home smart devices and immersive intelligence has raised many new challenges for far-field automatic speech recognition (ASR) [1, 2]. Even with the latest advances in deep-learning acoustic model [3, 4], speech recognition accuracy degrades severely in far-field ASR due to reverberation and additive noise [5, 6].

Many noise-robustness methods use stereo data to learn the mapping from distorted speech to clean speech. The stereo data consists of speech samples simultaneously recorded in training environments and in representative test environments. Stereo data can also be obtained by digitally introducing (e.g. adding noise) distortion to the clean speech. Data augmentation through far-field data simulation and speech enhancement technologies have been applied to improve far-field ASR with good success [7, 8, 9, 10, 11, 12, 13, 14].

Feature-mapping [15, 16, 17, 18, 19] and mask-learning [20, 21, 22, 23] are two commonly used deep-learning based approaches for single-channel speech enhancement using stereo data. The former directly learns a non-linear mapping to convert the noisy speech to clean speech. In the latter, a ratio mask is first learnt and further applied to the noisy speech to mask the noise interference and recover the clean speech. Due to the fact that the mask learning has constraint dynamic range and typically convergences faster, it usually outperforms the feature mapping approach [22].

Nevertheless, the mask learning system assumes that the scale of the masked signal is the same as the clean target, and the interfering noise is strictly additive, which can be remove by the masking process. If the distorted signal is also affected by the channel distortion or reverberation, an additional feature mapping function has to be learned to composite since the mask is incapable to restore such distortion. This problem is usually more severe in the far-talk scenarios where the reverberation is stronger. Therefore this assumption is usually not applicable for most real recorded stereo data because the varied sound source location and channel difference for each recording microphone breaks the additive relation between the noisy speech and the clean reference.

In this paper, we are interested in re-using a well-developed close-talk acoustic model for the far-field ASR without the need to collect large amount of far-field data and re-train the acoustic model. Specifically, we learn a single-channel speech enhancement model from small amount of recorded parallel speech data. This model can be conveniently plugged in as a front-end module for a practical far-field setup. This is extremely appealing as we move from the close-talking mobile speech application to the far-field speech recognition scenario, given that we already have a well-developed close-talking acoustic model trained from thousands of hours of mobile speech.

We first extend the mask learning approach by integrating two types of restoration layers to address the scale mismatch problem. With this extension, the mask learning can be applied to a much wider range of real scenario. We further propose a novel residual learning [24] based speech enhancement model via adding shortcut connections to a feature mapping network. We show that when the input feature is in logarithmic scale, adding the additive residual connection between layers is equivalent to the masking learning. Such a structure can benefit from both the mask learning and the feature mapping. To the best of our knowledge, this is the first application of residue learning to speech enhancement.

The rest of this paper is organized as follows: Section 2 explains the extended mask learning with restoration layers; Section 3 introduces the residue learning model; Section 4 presents the experiment and results; Section 5 is the conclusion.

2. Extended Mask Learning with Restoration Layers

In this section, we first review the formulation of mask learning and discuss the scale mismatch problem; then introduce the extended mask learning with restoration layers to address the recording mismatch.

This work was done when Zhuo Chen worked as intern in Microsoft

http://dx.doi.org/10.21437/Interspeech.2017-515

2.1. Mask Learning Formulation

There are two ways to learn the mask. The first one is called mask approximation (MA) which directly minimizes the distance between the learned mask and the target mask [20, 21, 22, 23]. The second is called signal approximation (SA) which minimizes the distance between the target signal and the signal constructed by applying the estimated mask to the distorted signal [25, 26].

We choose signal approximation based mask learning in this study. It is shown in [25] that SA is better than MA as its final target is directly related with the source signal. The objective of SA based mask learning for speech enhancement is:

$$\mathcal{L} = \|X - \Phi(Y) \odot M\|_2^2, \quad (1)$$

where X and M are clean speech and noisy speech in the mask learning *output* feature domain; Y is the noisy speech in the mask learning *input* feature domain; $\Phi(\cdot)$ is the mask estimation function learnt from neural networks. $\Phi(Y) \in [0, 1]$ is the learnt soft mask.

Here we use separate notations for noisy speech as the input and the output does not necessarily need to be in the same feature domain in mask learning.

2.2. Recording Mismatch Problem

In mask learning system, when the mixture (M), clean speech (X), and the noise (N) are strictly additive, i.e. $M = X + N$, the clean speech can be perfectly recovered from the ideal mask (i.e. $\frac{X}{X+N}$) learned from the neural network. However, the additivity relation usually only exists in the synthetic data, where the noisy speech is simply synthesized as the summation between the clean speech and the noise. In parallel recording, since the clean reference and the noisy speech are recorded simultaneously through a pair of close-talk and far-talk microphone, several additional distortions are introduced. For example, the channel mismatch between the microphone pair would cause additional channel distortion between the recordings, and since the microphone is close to the speaker, the close-talking recording would usually capture speech-related sound that is not exist in the far-talk recording, such as the spoken wind. Finally, since the relative distance between each sound source to the microphone are different, their relative amplitude would differ in both recordings, and thus cause additional distortion. We refer the mismatch in the real recording as "recording mismatch problem".

To model real parallel recording, we introduce two additional function into the mask learning formulation. The mixture from a far-talk microphone (M) and clean speech (X) is modeled as

$$M = f(g(X) + N), \quad (2)$$

where $f(\cdot)$ is the non-linear transformation introduced by the channel mismatch; $g(\cdot)$ represents the spectral difference between the two recordings. Under this setting, the masking process $\Phi(Y) \odot M$ would fail to recover the clean speech due to the cascaded non-linearities, even with the ideal mask.

2.3. Mask Learning with Restoration

We propose to extend the mask learning with two types of restoration before or after the mask to address the recording mismatch problem, namely pre- and post-restoration. Both restoration consist of neural network layers.

The pre-restoration layers target at the scale and the channel mismatch between microphones. Rectifier linear unit activation(ReLU) is selected as the output non-linearity of the pre-restoration layers since it is unbounded. Through the pre-restoration layer, the dynamic range of the raw noisy speech is re-scaled as the clean reference for the masking processing.

The spectral mismatch between close and far-talking is learnt through the post-restoration layers. The spectral mismatch is generated from the difference recording location in different microphones, which is not directly related with the noise and can not be removed through masking process(e.g. The T-F bins in clean reference has greater amplitude than the noisy mixture.) With the post-restoration, such difference can be fixed after the masking process, which prevents the system from wasting the representation power in producing unachievable target. Figure 1 depicts the architecture of the extended mask learning with pre- or post-restoration layers.

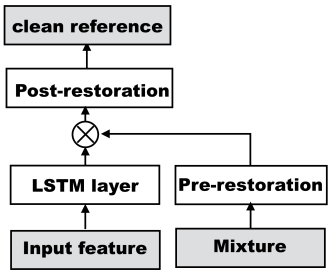

Figure 1: *Architecture of the extended mask learning with pre- or post- restoration layers. Input feature, mixture and clean reference blocks correspond Y, M and X in Equation 1, and the remaining white blocks are neural network parameters($\Phi(\cdot)$ in Equation 1)*

With these two extensions, the mask learning based speech enhancement can be applied to a wide range of real world scenario. We will present results on the mask learning with restoration layers on CHiME3 task in Section 4.

3. Residual Learning Feature Mapping

In this section, we introduce applying the residue learning architecture to speech enhancement.

3.1. Background of Residual Network

Deep residual learning makes use of shortcut connections between neural network layers for fast convergence and the gradient vanishing problem. It was first proposed in [24] for image recognition. Lately its efficacy was also confirmed in large vocabulary speech recognition [27].

In residue network, neural network layers are explicitly reformulated to learn residual functions with respect to the layer inputs. The shortcut connection in deep residue learning effectively addresses gradient vanishing/exploding problem in very deep neural network. Thus it is generally believed that very deep neural networks with residue connections is easier to optimize. The residue learning helps to maintain consistently improved accuracy performance in increasingly deeper and more complicated neural network.

Most previous work in applying residue learning focuses on improving network optimization for very deep network.

3.2. Residual Learning for Speech Enhancement

Unlike solving gradient vanishing/exploding problems and ease of training of very deep network, the motivation behind our work in applying the residue learning in speech enhancement has straightforward physical meaning in signal reconstruction.

Multiplication in linear scale corresponds to summation when performed in logarithm scale. In a direct feature mapping network, when all the representations are in logarithmic scale, the additive residual connection between layers is equivalent to consistently perform the masking process. As discussed in the recent work [28], the residual learning can be viewed as the ensemble of different sub-neural networks. This architecture alternates between the feature mapping and mask learning cross different neural network layers, which could be benefited from both feature mapping and the mask learning architecture. Therefore, it can potentially outperforms speech enhancement with the mask or the feature mapping only.

Based on this observation, we propose a residual learning based architecture for speech enhancement. Two types of residual connection are proposed, namely input residual connection and layer-wise residual connection. In layer-wise residual connection, the shortcut is added between the output of each layer and its previous layer. In the input residual connection, the shortcut connection between input and the output of each layer is incorporated in the network. The intuition behind the two architecture is straightforward. In layer-wise residual connection, a mask is learnt to refine the output from previous layer, and the noise is gradually removed through each masking process. While in input residual connection, since the input is fed to each layer, better mask would be learnt as the network keeps going deeper, until the last layer. Similar to the mask learning system, the recording mismatch also exists in the residual learning network. Therefore, the post restoration is also applied after the last residual connection. Figure 2 presents the architecture of residual learning based speech enhancement using *input* residue connection and/or *layer-wise* residue connection.

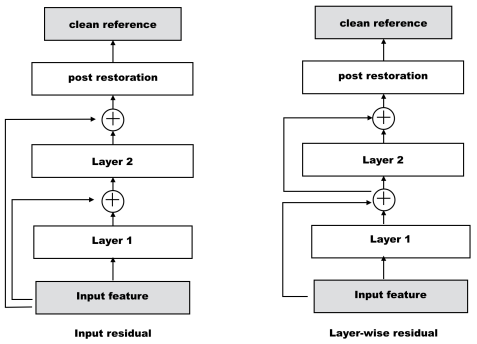

Figure 2: *Architecture of the residue-learning based speech enhancement model. Left: input residual connection, where the input is fed through the shortcut. Right: layer-wise residual, where the previous hidden activation is fed through residual connection*

4. Experiment

In this section, we present our speech enhancement experimental results on the CHiME3 task.

4.1. CHiME 3 and ASR Back-End

CHiME 3 [6] data is recorded using a 6-channel microphone array mounted on a tablet. The training data consists of 1600 real noisy utterances and 7138 simulated utterances. The real data is recorded in different live environments. The simulated data is obtained by mixing clean utterances into different background recordings. In real recording, a simultaneously recorded close talking signal is used as the clean reference, while in synthesized data, the clean utterance is referred as the reference. In all speech enhancement experiment, the real and simulated clean reference are used to form the training target. For both real and simulated data, four environments are selected: caf (CAF), street (STR), public transport (BUS), and pedestrian area (PED). The testing data consists of 1600 real recorded and 1600 simulated noisy utterances under the same four environments. There is no overlapping for both speech and the noise between training and testing set.

We are interested in reusing a well trained acoustic model under noisy environment, which was trained beforehand and didn't change during the experiment. The acoustic model was trained on 8738 clean utterances(1600 close talking recordings+7138 clean utterances). We trained a fully connected deep neural network (DNN) on close-talk clean speech. The DNN has 7-hidden layers, each with 2048 hidden units. The input feature consists of a 2640-dim feature vector formed by 80-dim log filterbank feature with double delta and a context window of 11 frames (80*3*11=2640). The output layer has 3012 senone states. We adopt the restricted Bolzman machine(RBM) pre-training before the fine-tuning of the full network using the cross-entropy criteria. We use Wall Street Journal 5K word 3-gram language model for decoding throughout this paper.

A MVDR beamformer provided by the CHiME 3 challenge was used as the pre-processor to combine the signal from 6 channels[29]. We use the beamformed signal and the single-channel far-field noisy speech(the 5th microphone) in the experiments. The proposed single-channel speech enhancement models are applied as a plug-in module before decoding. It is noted that because the beamforming step breaks the additive assumption between the noisy and clean speech, and will usually introduce additional channel distortion, even the simulated data in the dataset suffer from the recording mismatch problem as discussed in previous chapters.

4.2. Extended Mask Learning with Restoration Layers

The baseline mask learning system consists of two long short-term memory(LSTM) layers, each layer has 300 cells. A fully connected layer with sigmoid activation is build on top of the LSTM layers, to form the mask. We used the feature setting that is similar to a to a previous state-of-art mask learning system[30] where the input feature is 100-dimension log mel-filterbank, calculated with 25ms window and 10ms hop, and the clean 100 dimensional linear filterbank calculated under the same setting is used as clean reference.

Although a more complex architecture would potentially lead to better results, we use one fully connected layer with rectify linear units for both pre and post restoration, with 100 nodes for each. For both systems, the enhanced speech is firstly converted into time-domain representation inverse mel-filterbank and inverse Fourier transform, then the speech recognition feature is calculated and feed into the acoustic model.

In addition, we compare the mask learning based approach with the feature mapping approach. Here the feature mapping is conducted in the acoustic model feature domain, i.e. the net-

Table 1: *Speech recognition word error rate (WER) comparison of mask learning with/without scale restoration.*

method	Real	Simu
Baseline	31.12	15.78
Standard Mask	37.40	15.05
Mask + pre-Restoration	37.40	14.96
Mask + post-Restoration	30.87	13.41
Feature mapping (LSTM)	24.22	11.23

work takes in the noisy acoustic feature and directly target at the feature for acoustic model. Specifically, the input is the 240-dim log mel-filterbank formed by 80-dim log mel-filterbank with double delta, as described in Section 4.1, and the output is the clean 80-dim log mel-filterbank feature. A similar two-layer 300-cell LSTM is used. The result is shown in Table 1.

In Table 1, we see that without restoration, the mask learning completely fails on both real and simulated data. This confirms the recording mismatch problem described earlier. After introducing restoration layers, both pre-restoration and post-restoration layer improve the mask learning result. In particular, we found that the post-restoration outperforms the pre-restoration. This is likely due to the fact that the post-restoration layers have more information from bottom layers and can be better optimized globally.

Also in table 1, the feature mapping learning in the acoustic model feature domain significantly outperforms the mask-learning based approach, even with injected restoration layers. We believe such difference is because of the signal conversion in the mask learning based speech enhancement, i.e. the mask learning method required the enhanced signal to convert in time domain before feeding into the acoustic. This process usually introduced extra distortion.

Compared with the beamformed speech, the system directly trained from signal channel has significantly higher word error rate(WER). This result shows that the signal-channel enhancement is complementary with beamforming, and also suggests that a joint trained multi-signal channel enhancement system would have potentially better performance, as discussed in [5].

4.3. Residual Learning

Our residue learning based speech enhancement is developed based upon the best performed bLSTM feature mapping model as discussed in Section 4.2. Specifically, we compare three residual learning networks with different types of connections: input residual connections only (Res-I); layer-wise residual connections only (Res-L); both input and layer-wise residual connections (Res-B).

Table 2 summarizes the experimental results on residual learning based speech enhancement. First, all three proposed residual networks yield small but consistent additional accuracy gain comparing to the state-of-art bLSTM feature mapping model. This suggests the efficacy of the residue connections in speech enhancement model. In particular, we found that the architecture with input residue connection only (Res-I) performs best with 2.91% additional WER reduction against the baseline bLSTM feature mapping model.

It is worth noting that our architecture is still considered to be a shallow model. As more training data becomes available, we can increase the depth of the enhancement network layers.

Table 2: *Speech recognition accuracy comparison for residue learning based speech enhancement: input residue connection (ResI), layer-wise residue connection (ResL), or both (ResB). The results in the brackets are relative WER reductions from the baseline setup.*

method	Real (WER.R)	Simu (WER.R)
bLSTM (Baseline)	24.07 (NA)	10.81(NA)
bLSTM + Res-I	23.37 (2.91)	10.41 (3.70)
bLSTM + Res-L	23.74 (1.37)	10.88 (-0.65)
bLSTM + Res-B	23.52 (2.29)	10.56 (2.31)

Table 3: *Speech recognition accuracy performance comparison for single-channel far-field speech enhancement using bLSTM with input residue connection. The results in the brackets are relative WER reductions from the baseline setup.*

method	Real (WER.R)	Simu (WER.R)
Noisy (Baseline)	31.12 (NA)	38.06 (NA)
Res-L(Enhanced)	23.37 (24.90)	24.90 (34.57)
Res-I(CH5)	27.47 (18.07)	28.58 (11.73)

4.4. Single-Channel Far-Field Speech Enhancement

In real world far-field applications, microphone array and multi-channel speech enhancement are not always available. We would like to find out how the proposed single-channel speech enhancement would perform on single-channel far-field speech.

To this end, we further apply the proposed extended bLSTM mask learning with post-restoration layers and input residue connection to single far-talk noisy channel speech enhancement. As shown in table 3, the proposed best performed speech enhancement applied to single far-field channel yields 11.73 % WER reduction, comparing to 24.90 % WER reduction when applied to the multi-channel enhanced speech evaluated on the real testing part of CHiME 3. The results suggest the proposed approach is beneficial even with single-channel far-field noisy speech. This makes the proposed single-channel speech enhancement approach practical in a wide rage of real world far-field ASR scenario.

It is worth noting that the state of the art recognition performance the CHiME 3 dataset is around 6% in WER. However, to achieve such performance, many upgrades of the model are required, including advanced beamformer, better and customized language model, system combination, larger acoustic model, and most importantly, the retraining of acoustic model with enhanced speech, all of which were not included in the experiment. In our setting, the acoustic model remains unchanged, since the retraining is usually not available in real world applications, where. By incorporating those add-ons, a much better performance could be expected. Recently, in [31], the author shows that the teacher-student learning could also largely increase the recognition performance on the same dataset, which could also be a potential combination with the proposed models.

5. Conclusion

In conclusion, we relaxed the matching scale constraint in mask learning based speech enhancement model by integrating two types of restoration layer. We proposed a novel residual learning for improving speech enhancement. We evaluated the proposed speech enhancement models on CHiME 3 task. Without retraining the acoustic model, the bi-direction LSTM input residue connection yields 24.90% relative WER reduction on real data and 34.57% relative WER reduction on simulated data.

6. References

[1] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[3] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," in *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 1, 2012.

[4] A. Senior, S. Hašim, F. C. Quitry, T. N. Sainath, and K. Rao, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *ASRU 2015*. IEEE, 2015.

[5] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*. IEEE, 2014, pp. 172–176.

[6] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[7] T. Ko, V. Peddinti, P. Daniel, and K. Sanjeev, "Audio augmentation for speech recognition," in *Interspeech 2015*, 2015.

[8] X. Cui, B. Vaibhava, and K. Brian, "Data augmentation for deep neural network acoustic modeling," *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 9, pp. 1469–1477, 2015.

[9] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. Academic Press, 2015.

[10] T. N. Sainath, R. J. Weiss, K. W. Wilson, N. A., M. Bacchiani, and A. Senior, "Speaker localization and microphone spacing invariant acoustic modeling from raw multi channel waveforms," in *ASRU 2015*. IEEE, 2015.

[11] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, 2016.

[12] T. Yoshioka and et al., "The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *ASRU 2015*. IEEE, 2015.

[13] J. Du and et al., "The ustc-iflytek system for chime-4 challenge,," in *CHiME-4 workshop*, 2016.

[14] H. Erdogan and et al., "Wide residual blstm network with discriminative speaker adaptation for robust speech recognition," in *CHiME-4 workshop*, 2016.

[15] 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," in *Proc. Interspeech*, 2012, pp. 22–25.

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

[17] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. IEEE Intl. Conf. on Acoustic, Speech and Signal Processing*, 2014.

[18] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. IEEE Intl. Conf. on Acoustic, Speech and Signal Processing*, 2014.

[19] J. Du, L. Dai, and Q. Huo, "Synthesized stereo mapping via deep neural networks for noisy speech recognition," in *Proc. IEEE Intl. Conf. on Acoustic, Speech and Signal Processing*, 2014, pp. 1764–1768.

[20] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks," in *Proc. IEEE Intl. Conf. on Acoustic, Speech and Signal Processing*, 2013, pp. 7092–7096.

[21] ——, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. IEEE Intl. Conf. on Acoustic, Speech and Signal Processing*, 2014.

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

[23] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[24] K. He, Z. X., S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[25] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 2014, pp. 577–581.

[26] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[27] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The microsoft 2016 conversational speech recognition system," *arXiv preprint arXiv:1609.03528*, 2016.

[28] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 550–558.

[29] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[30] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Interspeech 2015*. ISCA, 2015.

[31] L.-S. D. A. via Teacher-Student Learn, "Large-scale domain adaptation via teacher-student learn," in *Interspeech 2017*. ISCA, 2017.