PHONEME-SPECIFIC SPEECH SEPARATION

Zhong-Qiu Wang[1], Yan Zhao[1] and DeLiang Wang[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{wangzhon, zhaoya, dwang}@cse.ohio-state.edu

ABSTRACT

Speech separation or enhancement algorithms seldom exploit information about phoneme identities. In this study, we propose a novel phoneme-specific speech separation method. Rather than training a single global model to enhance all the frames, we train a separate model for each phoneme to process its corresponding frames. A robust ASR system is employed to identify the phoneme identity of each frame. This way, the information from ASR systems and language models can directly influence speech separation by selecting a phoneme-specific model to use at the test stage. In addition, phoneme-specific models have fewer variations to model and do not exhibit the data imbalance problem. The improved enhancement results can in turn help recognition. Experiments on the corpus of the second CHiME speech separation and recognition challenge (task-2) demonstrate the effectiveness of this method in terms of objective measures of speech intelligibility and quality, as well as recognition performance.

Index Terms— speech separation, robust ASR, deep neural networks, ideal ratio mask, signal approximation

1. INTRODUCTION

Speech separation and recognition are not two independent tasks. They can clearly benefit from each other. It is intuitive that improved speech separation can boost the performance of robust ASR, and many studies in robust ASR are therefore focusing on improving speech separation [10]. Speech recognition can also elevate speech separation. One can imagine that if the performance of speech recognition is perfect, we can leverage recognized speech to assist speech separation [7]. But currently, the performance of speech recognition is still far from perfect, especially in low SNR conditions and reverberant environments. In some recent studies, first pass recognition results or outputs from acoustic models are used as augmented features to improve speech separation [31][11] or de-reverberation [12]. In [15][5][29], novel frameworks which jointly optimize separation frontends and acoustic models are proposed, while only ASR improvement is observed. In [13], constraints from language models are utilized to restrict separation frontends to produce semantically plausible enhancement.

We believe that high-level information from language models can help speech separation just like in the ASR decoding. However, language models are about the relationships among words, or in a wider sense, among phonemes or states, and speech separation is traditionally done at the signal level or in the time-frequency domain. There is clearly a gap between them that is not easily bridged. To bring the information of language models to bear on speech separation, one has to address the issue of how to improve speech separation if we somehow know the underlying word, phoneme or state identities of a corrupted utterance.

In this study, we propose to train a separate separation frontend for each phoneme based on deep neural networks (DNNs). At the test stage, a strong robust ASR system based on the DNN-HMM hybrid approach is utilized to obtain the phoneme identity of each frame, from which a phoneme-specific model is selected to perform enhancement.

The motivation for training phoneme-specific models is that, if we train a separate separation frontend for each phoneme, the performance of speech separation should be improved since each individual model has fewer variations to model. In addition, speech data is highly unbalanced in terms of phoneme distribution. If only one global model is trained on all the data, under-represented phonemes may not be properly modeled. Obviously, the performance of the ASR system is critical in our method. Although recognizers make errors, especially on utterances in low SNR conditions or corrupted by nonstationary noises and reverberation, modern ASR systems can still recognize most words or phonemes correctly.

We want to mention that a similar phoneme-dependent non-negative matrix factorization (NMF) approach was proposed in [19]. Unlike the proposed approach, it employs NMF models and weaker ASR methods. In addition, no quantitative enhancement results are presented in their paper. The motivation of our study is to utilize the information from ASR systems and language models to help speech separation. From a general viewpoint, our approach is also similar to the mixture of experts method [8] in some sense, where the input space is partitioned into different sub-regions, each of which is processed by a local expert.

The rest of this paper is organized as follows. We describe our system in section 2. Experimental setup and results are presented in sections 3 and 4. We conclude this paper in section 5.

2. SYSTEM DESCRIPTION

Our system is developed in a step-by-step manner. We first build a T-F masking based speech separation frontend using DNNs. We further improve the separation frontend by switching to a better loss function. The refined separation frontend is then used as the initialization for training phoneme-specific models. Finally, we build a robust ASR system to identify phoneme identities at the test stage.

2.1. Mask Estimation

Originated from computational auditory scene analysis [24], supervised T-F masking based methods have shown substantial potential for speech separation [32] and robust ASR [29][14][27]. The key idea is to train a powerful learning machine to estimate an ideal mask at the training stage. With the estimated mask, enhancement results can be obtained by point-wise multiplication at the test stage. In [28], it is shown that the ideal ratio mask (IRM) [16], a mask that represents the ratio of speech energy over sum of speech energy and noise energy within each T-F unit, is likely to be a better target over other ideal masks. Recently, DNNs are employed for mask estimation and have shown promising separation performance in matched or unmatched conditions [32][25]. In this context, we use DNNs to estimate IRMs in this study as our first speech separation baseline.

The IRM in this study is defined in the power spectrogram domain:

$$M(t,f) = \frac{S(t,f)}{S(t,f) + N(t,f)} \quad (1)$$

where M is the ideal ratio mask of a noisy utterance created by mixing a clean utterance with a noise signal at a specific SNR level, S is the power spectrogram of the clean utterance, N is the power spectrogram of the noise signal, and t and f index time and frequency, respectively.

The IRM must be estimated at the test stage. We utilize a DNN with four hidden layers each with 1024 rectified linear units (ReLUs) for mask estimation. Sigmoid units are used in the output layer. The input to the DNN is log compressed power spectrogram with a 19-frame context window, and the output corresponds to the label of the central frame. For signals with 16 kHz sampling rate and 20 ms window length, the input dimension would be 3059 (161*19) and the output dimension be 161. Note that the log power spectrogram feature is globally mean variance normalized before splicing. The network is trained to optimize the mean square error frame-wisely using mini-batch stochastic gradient descent with momentum and Adagrad [3] starting from random initialization. The dropout rates for the input layer and all the hidden layers are set to 0.3. A development set is used for parameter tuning and early stopping.

After obtaining the estimated mask of a noisy utterance, we use the following method to get the enhanced power spectrogram:

$$X^* = (M^*)^{\alpha} \otimes X \quad (2)$$

where X^* is the enhanced power spectrogram, M^* is the estimated mask, X represents the power spectrogram of the noisy utterance, and $\otimes$ stands for pointwise matrix multiplication. A tunable α term is used to scale the estimated mask according to a power law [17]. In this study, we always set α to 1.0 when resynthesizing enhanced time-domain signals. The phase of the noisy utterance is directly utilized for re-synthesis.

2.2. Signal Approximation

One problem of the mask estimation method presented in the previous section is that even if we can estimate the mask perfectly, we cannot reconstruct the exact spectrogram of clean speech using Eq. (2). In addition, direct mask estimation considers all the T-F units equally important, without considering the underlying mixture energy or clean speech energy within each T-F unit. To directly obtain the power spectrogram of clean speech, [33] and [6] propose to directly learn a mapping from corrupted speech to clean speech. However, the output is not naturally bounded in a reasonable range, such as between 0 and 1, which can be well estimated. In [31][4][30], a tradeoff between these two methods called signal approximation is proposed. The key idea is to use the square error between enhanced power spectrogram and target clean power spectrogram as the new loss function, i.e.

$$L^{SA1}(M^*) = \sum_{t,f} (M^*(t,f)X(t,f) - S(t,f))^2 \quad (3)$$

In this study, we use a slightly different loss function as shown in Eq. (4). We think that performing a log compression is necessary since it greatly compresses the dynamic range of the loss function and hence the gradient would not be unfavorably large. In addition, after log compression, the distribution of clean power spectrogram at each channel is more Gaussian-like, and therefore can be reasonably modeled using the square loss function [1].

$$L^{SA2}(M^*) = \sum_{t,f} (\log[M^*(t,f)X(t,f)] - \log[S(t,f)])^2 \quad (4)$$

The parameter initialization strategy of DNNs for signal approximation is important. The performance of signal approximation is worse than mask estimation if both models are trained starting from random initialization. Following [31], only changing the loss function, we train the signal approximation DNN starting from a well-trained mask estimation DNN until convergence. The resulting model gives us much better results than the mask estimation DNN with the same number of parameters. We also use the method in Eq. (2) for resynthesizing time-domain signals.

2.3. Phoneme-specific Speech Separation

After obtaining a global separation frontend based on the signal approximation loss function in Eq. (4), we then use this model as the initialization for training each phoneme-specific model. There are 40 phonemes (including silence) in our system. We cut the training data into 40 pieces based on each frame's phoneme identity, and further train each model using the loss function in Eq. (4) until convergence. The mean and variance of the training data for each phoneme-specific DNN are calculated only from the corresponding frames of each phoneme. The DNN setup and training recipes follow the DNN training in previous sections.

2.4. Acoustic Modeling

The performance of the ASR system is critical in our system. Higher ASR performance leads to better selections of phoneme-specific DNNs. In this study, we use a DNN with 7 hidden layers each with 2048 ReLUs for acoustic modeling. The acoustic models are trained to estimate the posterior probabilities of senone states by minimizing the cross-entropy criteria. Following our previous study [29], in addition to the commonly used log mel filterbank feature, we add more robust features for acoustic modeling. The resulting multi-stream ASR system uses the following features:

TABLE I
ASR PERFORMANCE (% WER) USING MULTI-CONDITION TRAINING WITH DIFFERENT FEATURES FOR ACOUSTIC MODELING

Features for Acoustic Modeling	dev. set Average	test set -6dB	-3dB	0dB	3dB	6dB	9dB	Average
MEL	19.40	26.77	20.49	16.14	12.80	10.67	10.11	16.16
MEL+PNCC+MRCG+Fset	17.93	23.09	17.17	13.32	10.41	8.71	8.07	13.46
Weninger *et al*.[30]	17.87	23.48	17.02	13.71	10.72	8.95	8.67	13.76

TABLE II
ASR PERFORMANCE (% WER) USING ENHANCED MEL+PNCC+MRCG+FSET FEATURE FOR DECODING

Separation Frontends	α	dev. set Average	test set -6dB	-3dB	0dB	3dB	6dB	9dB	Average
Signal Approximation	1.0	18.28	23.65	17.21	13.75	10.46	8.63	8.14	13.64
Signal Approximation	0.5	17.24	21.93	15.09	12.55	9.98	8.11	7.32	12.50
Phoneme-Specific Enhancement (decoding results)	0.5	17.05	21.54	15.15	12.44	10.05	7.86	7.40	12.41
Phoneme-Specific Enhancement (forced alignments)	0.5	13.79	16.68	11.68	9.79	8.24	6.93	6.73	10.01

TABLE III
PERFORMANCE COMPARISON AMONG DIFFERENT SEPARATION FRONTENDS IN TERMS OF STOI AND PESQ SCORES

Separation Frontends	test set -6dB		-3dB		0dB		3dB		6dB		9dB	
	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ
Unprocessed	0.737	2.138	0.778	2.327	0.813	2.492	0.852	2.662	0.881	2.854	0.909	3.049
Mask Estimation	0.834	2.539	0.862	2.693	0.886	2.831	0.909	2.979	0.925	3.138	0.942	3.305
Signal Approximation	0.849	2.699	0.878	2.849	0.899	2.979	0.918	3.114	0.931	3.254	0.946	3.404
Phoneme-Specific Enhancement	0.861	2.731	0.886	2.884	0.905	3.011	0.922	3.146	0.935	3.284	0.949	3.430

- 40-dimensional log mel spectrogram (MEL) together with its deltas and double deltas. Sentence level mean normalization is performed before splicing 11 frames;
- 31-dimensional power normalized cepstral coefficients (PNCC) [9] together with its deltas and double deltas. We further splice 11 frames. The PNCC feature is found to be relatively robust to reverberation and noise;
- 256-dimensional multi-resolution cochleagram (MRCG) [2] together with its deltas and double deltas. The MRCG is shown to be good at handling additive noise and mask estimation;
- A 915-dimensional feature set [26] which combines RASTA-PLP, amplitude modulation spectrogram (AMS), narrowband MFCC and wideband MFCC. This feature set is found to have complementary power for mask estimation [26]. It also leads to improvement for acoustic modeling in [29]. We denote it as "Fset" in this study.

If we concatenate all the features mentioned above for acoustic modeling, the input dimension would be 4026 (40*3*11+31*3*11+256*3+915). All the features are globally mean variance normalized before DNN training.

3. EXPERIMENTAL SETUP

We validate our method on the reverberant and noisy CHiME-2 dataset (task-2) [23][1]. The reverberant utterances are created by convolving the clean utterances in the WSJ0-5k corpus with various binaural room impulse responses measured from a domestic living room. The reverberant utterances are then digitally mixed with a rich set of realistic noises recorded from the same room setup, such as children's laughter, competing speakers, footsteps, background music, distant noises, and sounds from electronic devices, to create reverberant and noisy utterances at six SNR levels linearly spaced from -6dB to 9dB. The multi-conditional training set contains 7138 utterances in total (~14.5h). The development set has 409 utterances at each SNR level (~4.5h). The test set consists of 330 utterances at each SNR level (~4h). With the parallel clean, reverberant noise-free and noisy-reverberant data available, we can readily evaluate the performance of speech separation together with recognition.

Our system is monaural. We average all the binaural signals in the dataset. Note that this is the same as the delay-and-sum beamforming since the speaker is facing the microphones with azimuth approximately 0 degrees in the CHiME-2 setup. We first use the Kaldi toolkit [18] to build a GMM-HMM system on the clean utterances from the WSJ0-5k corpus to obtain the senone label of each frame in the multi-conditional dataset. Then we perform forced alignment on the clean utterances to get the initial clean alignments. The initial alignments are used to train a DNN based acoustic model using the MEL feature extracted from clean utterances, from which better alignments are obtained by performing forced alignment again. The resulting refined alignments are used to train all the other acoustic models in this study. There are 1965 senone states in total in our system. The acoustic models are trained on noisy speech directly since it lets the acoustic models see more variations at the training stage [21]. We think that the high-quality alignments generated from clean utterances can guide the acoustic models to better discriminate senone states even if input features are highly corrupted. We use the CMU pronunciation dictionary that contains 40 phonemes and the official 5k close-vocabulary trigram language model in our experiments. Note that the clean alignments are additionally used to cutting the training data for training phoneme-specific DNNs by simply transforming the state sequences to phoneme sequences.

The separation frontends are trained using the parallel noisy-reverberant and reverberant noise-free data. The mixed noise signals can be obtained by direct subtraction. With these data available, we can train a separation frontend based on mask estimation, a frontend based on signal approximation and

[1]Available at http://spandh.dcs.shef.ac.uk/chime_challenge/chime2013/WSJ0/.

phoneme-specific models using the methods detailed before. Note that these separation frontends are built to only remove additive noise.

4. EVALUATION RESULTS

4.1. ASR Performance

We first report the performance on ASR tasks. As shown in Table I, adding more robust features for acoustic modeling significantly reduces word error rates (WER). The 13.46 percent average WER is already absolute 0.3 percentage point better than the best result [30] reported by other studies to date. We think that it is because we use extra robust features, better DNNs for acoustic modeling, and better clean alignments. In the rest of this paper, we always use the acoustic model trained on the MEL+PNCC+MRCG+Fset feature for decoding.

We then incorporate the separation frontend based on signal approximation into the robust ASR system. We first employ the separation frontend to get the enhanced power spectrogram using Eq. (2), which is then passed to the mel filterbank to obtain the enhanced MEL feature. Together with other robust features, the enhanced MEL feature is passed to the acoustic model for decoding. Note that we do not enhance other robust features since they are considered to be inherently robust. The results are presented in Table II. If we set the α in Eq. (2) to 1.0 when generating enhanced power spectrogram, the performance is even worse than the baseline without any enhancement. This is probably because the separation frontend tries to suppress noise aggressively, therefore some information critical for senone states' discrimination may be totally lost. Following [15], we set the α to 0.5 to preserve more energy in the enhanced power spectrogram. This way, we can obtain 0.96 percent (13.46% vs. 12.50%) average WER improvement on the test set. Since this model obtains the best ASR results, we use it to generate the phoneme sequences for choosing phoneme-specific models in later experiments.

4.2. Speech Separation Performance

We utilize the widely used Perceptual Estimation of Speech Quality (PESQ) [20] metric and Short-Time Objective Intelligibility (STOI) [22] score to measure the objective performance in terms of speech quality and intelligibility, respectively. We use the averaged reverberant noise-free signals as the references when calculating these two scores since the separation frontends in this study only try to remove additive noise. Note that, again, we set the α in Eq. (2) to 1.0 when resynthesizing enhanced signals. The results are shown in Table III. We can see that, compared with unprocessed speech, both mask estimation and signal approximation based methods significantly improve STOI and PESQ scores. The signal approximation method gets better results than the mask estimation method especially in low SNR conditions. We then use the decoding results generated from our current best ASR model (the second entry in Table II) to select phoneme-specific models for enhancement. The results are presented in the last entry of Table III. Compared with using a global separation frontend, phoneme-specific processing consistently improves STOI and PESQ scores in all SNR conditions, even though the ASR system makes errors.

An interesting question to ask is whether the improved separation results can further improve recognition. Note that we use the decoding results to select phoneme-specific frontends to obtain the enhanced MEL feature, which is then fed into the acoustic model for decoding. However, we only get slight improvement on the development set and test set, as shown in the third entry of Table II. This may be due to the false selection of phoneme-specific models resulting from the erroneous first pass decoding results. Nonetheless, we want to point out that if we have the correct phoneme sequences according to which we can select phoneme-specific models perfectly, the ASR performance can be improved significantly, as reported in the last entry of Table II. The perfect phoneme sequences are obtained by performing forced alignment on the clean utterances of the development and test set in the WSJ0-5k corpus.

5. CONCLUDING REMARKS

In this paper, we have proposed a novel phoneme-specific processing method for speech separation. Consistent improvement in objective measures of speech intelligibility and quality, and recognition rate is observed. To incorporate the information from language models into speech separation, one important issue is what to do when we know the word, phoneme or state identities of a test utterance. Training phoneme-specific models is one such method. Future research should focus on other alternative methods.

6. ACKNOWLEDGEMENTS

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NSF grant (IIS-1409431), and the Ohio Supercomputer Center.

7. REFERENCES

[1] C.M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[2] J. Chen, Y. Wang, and D.L. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1993–2002, 2014.

[3] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *The Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

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

[5] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4375–4379.

[6] K. Han, Y. Wang, D.L. Wang, W.S. Woods, and I. Merks, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 982–992, 2015.

[7] O. Hazrati, S. Ghaffarzadegan, and J.H.L. Hansen, "Leveraging automatic speech recognition in cochlear implants for improved speech intelligibility under reverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5093–5097.

[8] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, pp. 79–87, 1991.

[9] C. Kim and R.M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4101–4104.

[10] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 745–777, 2014.

[11] M. Mimura, S. Sakai, and T. Kawahara, "Deep autoencoders augmented with phone-class feature for reverberant speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4365–4369.

[12] M. Mimura, S. Sakai, and T. Kawahara, "Speech dereverberation using long short-term memory," in *Proceedings of Interspeech*, 2015, pp. 2435–2439.

[13] G. Mysore and P. Smaragdis, "A non-negative approach to language informed speech separation," in *Proceedings of Latent Variable Analysis and Signal Separation*, 2012, pp. 356–363.

[14] A. Narayanan, A. Misra, and K. Chin, "Large-scale, sequence-discriminative, joint adaptive training for masking-based robust asr," in *Proceedings of Interspeech*, 2015, pp. 3571–3575.

[15] A. Narayanan and D.L. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 92–101, 2015.

[16] A. Narayanan and D.L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7092–7096.

[17] A. Narayanan and D.L. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 826–835, Apr. 2014.

[18] D. Povey, A. Ghoshal, and G. Boulianne, "The Kaldi speech recognition toolkit," 2011.

[19] B. Raj, R. Singh, and T. Virtanen, "Phoneme-dependent NMF for speech enhancement in monaural mixtures.," in *Proceedings of Interspeech*, 2011, pp. 1217–1220.

[20] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001, vol. 2, pp. 749–752.

[21] M.L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7398–7402.

[22] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2125–2136, Sep. 2011.

[23] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: an overview of challenge systems and outcomes," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 162–167.

[24] D.L. Wang and G.J. Brown, Eds., *Computational auditory scene analysis: principles, algorithms, and applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.

[25] Y. Wang, J. Chen, and D.L. Wang, "Deep neural network based supervised speech segregation generalizes to novel noises through large-scale training," *OSU-CISRC-3/15-TR02*, 2015.

[26] Y. Wang, K. Han, and D.L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 270–279, 2013.

[27] Y. Wang, A. Misra, and K. Chin, "Time-frequency masking for large scale robust speech recognition," in *Proceedings of Interspeech*, 2015, pp. 2469–2473.

[28] Y. Wang, A. Narayanan, and D.L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849–1858, 2014.

[29] Z.-Q. Wang and D.L. Wang, "Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition," in *Proceedings of Interspeech*, 2015, pp. 2839–2843.

[30] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J.R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.

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

[32] Y. Wang and D.L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1381–1390, 2013.

[33] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7–19, 2015.