SPEAKER REINFORCEMENT USING TARGET SOURCE EXTRACTION FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Cătălin Zorilă and Rama Doddipatla

Toshiba Cambridge Research Laboratory, Cambridge, United Kingdom

firstName.lastName@crl.toshiba.co.uk

ABSTRACT

Improving the accuracy of single-channel automatic speech recognition (ASR) in noisy conditions is challenging. Strong speech enhancement front-ends are available, however, they typically require that the ASR model is retrained to cope with the processing artifacts. In this paper we explore a speaker reinforcement strategy for improving recognition performance without retraining the acoustic model (AM). This is achieved by remixing the enhanced signal with the unprocessed input to alleviate the processing artifacts. We evaluate the proposed approach using a DNN speaker extraction based speech denoiser trained with a perceptually motivated loss function. Results show that (without AM retraining) our method yields about 23% and 25% relative accuracy gains compared with the unprocessed for the monoaural simulated and real CHiME-4 evaluation sets, respectively, and outperforms a state-of-the-art reference method.

Index Terms— speaker extraction, SpeakerBeam, speaker reinforcement, automatic speech recognition

1. INTRODUCTION

Recently, the deep neural networks (DNNs) have greatly improved the accuracy of automatic speech recognizers. The current ASR systems have already reached human performance in clean conditions, however, they are still worse than normal-hearing listeners in noisy conditions [1].

To improve ASR robustness in noise, there are at least two distinct strategies in the literature. One strategy relies on large amounts of data to train multi-condition models, and the other strategy is on performing data cleaning using signal enhancement. Although the first approach is simple, it is costly both in terms of computational resources required to train such models and in terms of collecting annotated data. Furthermore, the accuracy of multi-condition systems drops in very challenging environments, such as those with competing speakers [2]. Concerning the data cleaning track, the results reported so far are mixed and they indicate that accuracy gains can be achieved in certain conditions (e.g., strong interfering speech [3]), but there are shortcomings in others. Frequently, distortions introduced by speech enhancement limit the applicability of these methods as standalone front-end for ASR, and the acoustic model has to be retrained with matched distorted data to achieve the best recognition performance [4]. Alternatively, jointly trained enhancement and recognition systems have been proposed to alleviate the distortions [5,6]. However, in real applications with very dynamic acoustic conditions, the latter approach may not work well.

In this paper we show that ASR accuracy in noisy conditions can be boosted by using a *speaker reinforcement strategy* without acoustic model retraining on distorted data. Instead of focusing on fully suppressing the background using state-of-the-art enhancement algorithms, we conjecture that by remixing the enhanced signal with the unprocessed input alleviates the processing artifacts, leading to significant recognition accuracy gains without model retraining. A similar idea has been recently proposed for raw waveform speech denoising in [7], where a dry/wet knob was demonstrated to improve the overall perceived quality of processed samples, but in that work the ASR experiments did not investigate ASR robustness in mismatched noisy conditions. Here, we evaluate whether speaker reinforcement is effective for monoaural speech denoising for ASR on both simulated and real noisy data.

A more powerful speech enhancement method is expected to yield better performance, therefore we have chosen a speaker extraction (SPX) system to perform denoising in this work. Instead of recovering all sources from a mixture (i.e., speech separation), the aim of SPX is to recover only a target speaker from the mix [8–11], which circumvent the requirement to know in advance the total number of sources. Kinoshita et al. [12] have recently proposed a denoiser based on convolutional time-domain audio separation network (Conv-TasNet, [13]) to boost single-channel ASR performance in noisy conditions, achieving very promising results on CHiME-4 data. Inspired by those results and also by the success of the time-domain SpeakerBeam SPX algorithm [14, 15], which also follows Conv-TasNet's architecture, we selected SpeakerBeam SPX to perform our denoising experiments. We also investigate whether adding knowledge from the perceptual studies during network training strengthens the initial model, as other prior speech enhancement studies have shown [16–18].

In summary, our contributions in this paper are the following. We: (i) investigate the ASR performance of time-domain SpeakerBeam for speech denoising of real and simulated data in both matched and mismatched conditions, (ii) propose a new loss function based on an objective intelligibility metric for training time-domain SpeakerBeam denoiser, (ii) suggest speaker reinforcement strategy to improve robustness of ASR models in noisy environments.

The rest of the paper is structured as follows. Section 2 briefly presents the time domain SpeakerBeam algorithm for monoaural speaker extraction, Section 3 describes the proposed modified training loss function and the speaker reinforcement strategy, Section 4 introduces the evaluation methodology, and Section 5 shows the results. The paper is concluded in Section 6.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

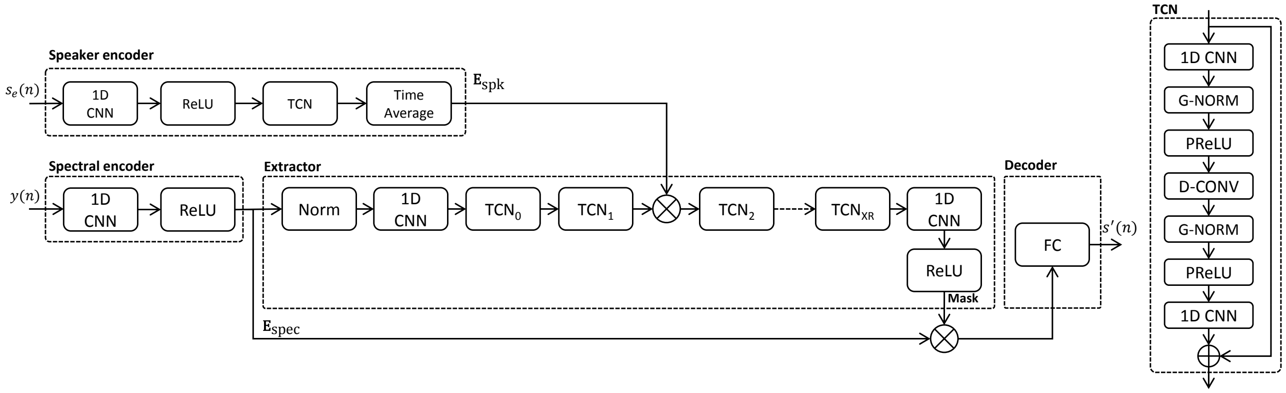

Fig. 1. Block diagram of single-channel time domain SpeakerBeam [14].

2. SINGLE-CHANNEL TIME DOMAIN SPEAKERBEAM

Delcroix et al. [14] have recently proposed a mask-based time domain SPX system whose block diagram is depicted in Fig. 1. Single-channel SpeakerBeam is made of four components: a spectral encoder, a speaker encoder for the target, an extractor network and a decoder. Except for the target speaker encoder, SpeakerBeam mainly follows Conv-TasNet's architecture [13] and it consists of two inputs, one for the mixture and one for the enrolment signal. The later input is used to generate embeddings for the target, which is subsequently employed to bias the extraction. A brief description of each component is given below.

2.1. Spectral encoder

The aim of the spectral encoder is to transform the speech waveform into a higher dimensional space ($\mathbf{E}_{\text{spec}}$) using one CNN layer followed by a rectified linear unit (ReLU) for activation. The CNN has N kernels of size L and stride $\text{L}/2$.

2.2. Extraction network

The extractor is designed to compute masks for the target speaker by exploiting the sparsity of speech in the spectral domain and the long-term dependencies of its temporal features. This is achieved by cascading several blocks of temporal convolution networks (TCNs) with increasing kernel dilation factors which perform multi-resolution analysis of input mixture.

As shown in Fig. 1, a TCN block is formed of three CNN layers, parametric ReLU (PReLU) activations and mean and variance normalization across both time and channel dimensions scaled by trainable bias and gain parameters (G-NORM). The depthwise convolution (D-CONV, H kernels of size P) operates independently on the input channels, and the dilation factors across consecutive TCN blocks is $2^{\text{mod}(\text{i},\text{XR})}$, where i is the block's index, XR is the total number of blocks and mod is the modulo operation. The endpoint 1D CNNs (B kernels) are employed to adjust the channel dimension. A multiplicative adaptation layer is used to combine the target speaker embedding $\mathbf{E}_{\text{spk}}$ with the output of the second TCN.

The spectral representation $\mathbf{E}_{\text{spec}}$ is firstly channel-wise normalized, then processed by a bottleneck 1×1 CNN layer (B kernels) before being fed to the first TCN. Another 1×1 CNN layer with N output channels is used after the last TCN to adjust the mask dimension to that of spectral encoder's output, thus facilitating their pointwise multiplication.

2.3. Speaker encoder

Speaker encoder is made of one CNN layer (N kernels of size L and stride $\text{L}/2$), followed by ReLU activation, a TCN configured as described in the previous section (dilation 1), and a time averaging operator.

2.4. Decoder

The decoder reconstructs the estimated target frames $s'(n)$ using the masked spectral representation and one fully-connected layer (N input and L output dimensions). Overlap-and-add is applied to reconstruct the whole waveform.

2.5. Training criterion

The training objective for SpeakerBeam is to maximize the scale-invariant signal-to-distortion ratio (SISDR), which is defined as:

$$\text{SI-SDR} = 10\log_{10}\frac{\left\|\frac{\langle s',s\rangle}{\langle s,s\rangle}s\right\|^2}{\left\|\frac{\langle s',s\rangle}{\langle s,s\rangle}s - s'\right\|^2} \tag{1}$$

where s' and s denote the estimated and the oracle target speaker signals, respectively.

3. PROPOSED METHOD

Preliminary experiments have shown that the distortions produced by the SPX processing harm the ASR accuracy, especially in mismatched scenarios, therefore we have explored two strategies to alleviate them, as described next.

Firstly, a novel training criterion is proposed that combines the standard SISDR loss with a perceptually motivated term based on the short-term objective intelligibility (STOI) measure [19]:

$$\text{L}_{\text{new}} = \text{L}_{\text{SISDR}}(s,s') + \text{L}_{\text{STOI}}(s,s'), \tag{2}$$

where L_{SISDR} is the standard SpeakerBeam loss defined in Eq. (1), and the second term is the STOI-based loss. STOI is a widely used metric to objectively assess the intelligibility of noisy speech processed by time-frequency weighting, and it was shown to produce a high correlation with human perception, thus it could also help reducing signal distortions for ASR. A DNN based speech enhancement system that maximizes an approximation of STOI has already been proposed in the literature [16], which, however, achieved modest gains compared with classical mean square error DNN systems. Instead of relying exclusively on STOI to train the denoiser, here we combine it with the SISDR loss. To the best of our knowledge, this is a novel approach in the context of SPX enhancement and for robust ASR applications.

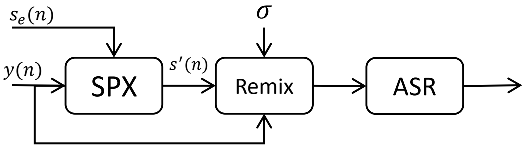

Fig. 2. Proposed target speaker reinforcement method.

Secondly, we propose remixing the enhanced signal with a fraction of the input mixture to mask the processing artifacts that could harm ASR. We denote this strategy as *target speaker reinforcement*. The remixing is controlled by a scaler σ (Fig. 2):

$$\sigma = 10 \log_{10} \frac{\|s'\|^2}{\|\alpha y\|^2}, \tag{3}$$

and the output is computed as $z(n) = s'(n) + \alpha y(n)$. The role of σ on the ASR accuracy of a pre-trained acoustic model has been investigated in the evaluation section.

4. EVALUATION

Experiments have primarily focused on the CHiME-4 data, which contains both simulated and real noisy speech recordings [20]. The CHiME-4 corpus was designed to capture multi-channel speech using a mobile tablet computing device in noisy everyday environments such as cafeteria, on the bus, street or pedestrian areas. The SPX denoiser is trained on clean Wall Street Journal (WSJ) speech artificially mixed with CHiME-4 noise. We report results on the single-channel real and simulated evaluation set (et05) of CHiME-4 (matched conditions with respect to SPX's train set). Additionally, we also report results in mismatched test conditions for the VoiceBank-DEMAND (VBD) and WHAM! sets [21, 22]. The max version of WHAM! test (tt) set was used, and all experiments were performed with 16 kHz resolution data.

Proposed method is compared with two reference methods. One is MetricGAN+ [23], which is a state-of-the-art single channel speech denoiser, and the other one is a TasNet based single-channel speech denoiser (Denoising-TasNet, [12]). The SpeechBrain[1] implementation of MetricGAN+ was employed for our evaluation.

The performance is mainly assessed in terms of word error rate (WER), however, for some preliminary experiments, the signal-to-distortion (SDR) and STOI values are reported as well. The SDR scores are computed using the BSSeval toolkit [24], and the STOI training loss is computed using a freely available PyTorch implementation[2]. More details about the configuration of the denoising network and ASR systems are presented next.

4.1. Speaker extraction based denoiser

Single-channel time-domain SpeakerBeam proposed by Delcroix et al. [14] is used to perform speech denoising, with the same configuration as described in [15], but without the spatial encoder. The spectral encoder consists of 1-D CNN with N=256 kernels of size L=20 and a frame rate of 10 samples. X=8 stacked TCN blocks repeated R=4 times are employed for the extraction network. Each TCN block consists of 1×1 CNNs and 1×3 depthwise convolutions with B=256 and H=512 kernels, respectively. The fully connected layer in the decoder has an input dimension of 256 and an output dimension of 20.

The clean WSJ training list from WSJ0-2mix [25] artificially added with CHiME-4 noise was used to train the SPX system. About 39 hours of data have been generated, the signal-to-noise (SNR) ratio of the mixtures being uniformly sampled from a range of 0 dB to 5 dB, and the audio length randomly varied from 1 s to 6 s. The target enrolment sentences from [9] have been used for training, making sure that the recordings for enrolment and mixture signals were different. The enrolment samples for the simulated CHiME-4, VBD and WHAM! test sets were chosen from the available clean waveforms, while the close-talk microphone recordings were used for enrolment of the real CHiME-4 evaluation set.

Training was performed using the Adam optimizer [26] with the initial learning rate of 0.001, a chunk length of 4 seconds, and a minibatch size of 8. The learning rate halved if no improvement on the cross-validation set was yielded for three consecutive epochs. All competing models were decoded at epoch 20 to avoid overfitting on the training data.

4.2. Automatic speech recognition

Two acoustic models are included in the evaluation. The first model is trained on clean WSJ-SI284 data (WSJ-CLN) and has a 12-layer TDNNF topology [27], while the second model is trained on the standard noisy set from CHiME-4 (C4-ORG) and has a 14-layer TDNNF structure. The later system employs all six channels from the real and simulated training sets of CHiME-4. Both models use 40-dimension MFCCs and 100-dimension i-vectors as acoustic features, and they were trained in KALDI using the lattice-free MMI criterion [28]. Both a standard tri-gram and a more powerful RNN language model (LM) are used for decoding. After 3-fold speed perturbation, WSJ-CLN and C4-ORG have about 246 hours and 327 hours of training data, respectively.

5. RESULTS & DISCUSSION

This section presents the results of our investigation into the effectiveness of proposed target speaker reinforcement approach for improving ASR robustness in matched and mismatched noise conditions.

Firstly, the SPX denoiser is evaluated in mismatched conditions using VBD and WHAM! simulated noisy test data (Table 1). The WER results in Table 1 are with the WSJ-CLN AM and a tri-gram (3-G) LM. Although the Denoising-SPX is trained on simulated noisy CHiME-4 mixtures and, therefore, is mismatched with either test sets, it yields about 14% and 67% relative WER reduction on the VBD and WHAM! test sets compared with the unprocessed case. Notably, Denoising-SPX decisively outperforms the reference system MetricGAN+ (trained on VBD data) in both cases. MetricGAN+ achieved a worse WER than the unprocessed for the WHAM! set, indicating that the current pre-trained system cannot cope with unseen noise conditions. WER results in Table 1 show that the composite SISDR and STOI training loss works better than the standard SISDR loss, although the SDR and STOI values for the vanilla and proposed systems are almost identical. We believe that the additional STOI term is able to help restore some of the temporal modulations

[1] https://speechbrain.github.io

[2] https://github.com/mpariente/pytorch_stoi

Table 1. Performance of proposed method in mismatched noisy conditions (Denoising-SPX was trained on simulated noisy CHiME-4 data). WER (%) are with WSJ-CLN AM and 3-G LM.

Enhancement	Train Loss	σ (dB)	VBD Test			WHAM! tt		
			WER(%)	SDR(dB)	STOI	WER(%)	SDR(dB)	STOI
Unprocessed	-	-	38.6	8.5	0.92	75.7	-2.8	0.76
MetricGAN+ [23]	-	-	35.4	15.0	0.93	77.5	1.8	0.72
Denoising-SPX (vanilla)	SISDR	∞	33.7	16.1	0.92	25.4	11.7	0.92
Denoising-SPX (proposed)	SISDR+STOI	∞	**33.2**	15.8	0.93	**24.6**	11.6	0.92
Clean	-	-	25.1	∞	1.00	9.0	∞	1.00

of speech distorted during enhancement. Our preliminary evaluation showed that training a STOI only denoising system leads to worse performance than by using the pure SISDR loss. A more complete analysis on weighting the STOI and SISDR loss is planned for our future work.

Next set of experiments are performed using the noisy CHiME-4 acoustic model (C4-ORG) and they are assessing the importance of the remixing ratio σ for ASR robustness. Results in Table 2 show that impressive WER reductions can be achieved by decreasing the value of σ from ∞ (no input mixture is added on top of the enhanced signal) to 0 dB for both the simulated and real CHiME-4 test sets. Reducing σ further yields a reversing (increasing) trend for the ASR accuracy, as shown in Table 2. Only by decreasing the remixing ratio, the proposed Denoising-SPX achieves about 28% and 36% relative WER reduction for the simulated and real evaluation set, respectively. These results are remarkable since neither the acoustic nor the SPX models were retrained to achieve these gains. The poor performance of Denoising-SPX for $\sigma = \infty$ compared with the unprocessed case can be attributed to the fact that the system was trained from anechoic simulated CHiME-4 noisy data, while the test sets also contain a small amount of reverberation. Another source of accuracy degradations could be the inherent distortions introduced by SPX, especially with the real data.

Table 2. WER accuracy of proposed speaker reinforcement approach on CHiME-4 for various remixing ratios, σ. All results are with noisy CHiME-4 AM (C4-ORG, 3-G LM).

Enhancement	σ (dB)	WER(%)	
		et05_simu	et05_real
Unprocessed	-	16.4	15.5
Denoising-SPX (vanilla)	∞	18.4	19.4
	20	16.5	17.7
	10	14.9	15.5
	0	13.5	13.0
	-10	13.7	12.2
	-20	14.9	13.2
Denoising-SPX (proposed)	∞	18.4	18.9
	20	16.4	17.2
	10	14.4	15.1
	0	**13.2**	12.7
	-10	13.7	**12.0**
	-20	14.8	13.1
Clean	-	2.5	4.9

Table 3. WER accuracy of proposed method on CHiME-4 using C4-ORG AM and RNN LM.

Enhancement	σ (dB)	WER(%)			
		et05_simu		et05_real	
		3-G	RNN	3-G	RNN
Unprocessed	-	16.4	13.6	15.5	12.3
Denoising-TasNet [12]	-	-	11.9	-	9.8
Denoising-SPX (proposed)	∞	18.4	14.4	18.9	14.9
	0	13.2	**10.4**	12.7	9.6
	-10	13.7	10.8	12.0	**9.2**
Clean	-	2.5	1.5	4.9	3.2

Table 3 compares the recognition accuracy of Denoising-SPX with that of Denoising-TasNet [12] on the single-channel CHiME-4 task. To ensure a fair comparison with the results reported in [12], the standard noisy C4-ORG acoustic model is used to perform ASR, and the 3-G transcriptions are rescored with an RNN-based language model. Notably, the proposed Denoising-SPX with speaker reinforcement has outperformed Denoising-TasNet in both simulated and real evaluation sets by 13% and 6% relative WER, respectively. This is a remarkable outcome since Denoising-SPX is trained on anechoic simulated CHiME-4 noisy data, while the Denoising-TasNet had to be trained on simulated reverberated and noisy recordings because the performance was not sufficient without reverberation [12]. Compared with the Unprocessed for RNN-LM, Denoising-SPX with speaker reinforcement yielded about 23% and 25% relative WER reduction for simulated and real sets, respectively.

As future work, we plan to extend the evaluation of proposed method in reverberant, noisy and multi-talker conditions.

6. CONCLUSIONS

In this paper we have presented a target speaker reinforcement algorithm for improving the ASR accuracy in noisy conditions without acoustic model retraining. Using a denoiser based on a DNN speaker extraction, we show that remixing the noisy input with the enhanced signal achieves about 23% and 25% WER reduction compared with the unprocessed case for the single-channel CHiME-4 simulated and real evaluation sets, respectively. Furthermore, the experiments suggest that adding a perceptually motivated loss on top of a time domain reconstruction loss during training of speaker extraction systems, helps achieve a modest but consistent ASR accuracy gain.

7. REFERENCES

[1] V. A. Trinh and M. Mandel, “Directly comparing the listening strategies of humans and machines,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 312–323, 2021.

[2] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The fifth CHiME speech separation and recognition challenge: Dataset,

task and baselines," in *Proc. Interspeech*, 2018, pp. 1561–1565.

[3] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, T. Moriya, and N. Kamo, "Should we always separate?: Switching between enhanced and observed signals for overlapping speech recognition," in *Proc. Interspeech*, 2021, pp. 1149–1153.

[4] C. Zorilă, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in ASR training and test for Chime-5 dinner party transcription," in *Proc. ASRU*, 2019, pp. 47–53.

[5] S. Settle, J. L. Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *Proc. ICASSP*, 2018, pp. 4819–4823.

[6] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *Proc. ICASSP*, 2020, pp. 6134–6138.

[7] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. Interspeech*, 2020, pp. 3291–3295.

[8] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Learning speaker representation for neural network based multichannel speaker extraction," in *Proc. ASRU*, 2017, pp. 8–15.

[9] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *Proc. ASRU*, 2019, pp. 327–334.

[10] R. Giri, S. Venkataramani, J.-M. Valin, U. Isik, and A. Krishnaswamy, "Personalized PercepNet: Real-Time, Low-Complexity Target Voice Separation and Enhancement," in *Proc. Interspeech*, 2021, pp. 1124–1128.

[11] K. Zmolikova, M. Delcroix, D. Raj, S. Watanabe, and J. Černocký, "Auxiliary Loss Function for Target Speech Extraction and Recognition with Weak Supervision Based on Speaker Characteristics," in *Proc. Interspeech*, 2021, pp. 1464–1468.

[12] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *Proc. ICASSP*, 2020, pp. 7009–7013.

[13] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[14] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *Proc. ICASSP*, 2020, pp. 691–695.

[15] C. Zorilă, M. Li, and R. Doddipatla, "An investigation into the multi-channel time domain speaker extraction network," in *Proc. of IEEE Spoken Language Technology Workshop*, 2021, pp. 793–800.

[16] M. Kolbaek, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *Proc. ICASSP*, 2018, pp. 5059–5063.

[17] M. R. Saddler, A. Francl, J. Feather, K. Qian, Y. Zhang, and J. H. McDermott, "Speech Denoising with Auditory Models," in *Proc. Interspeech*, 2021, pp. 2681–2685.

[18] Z.-Q. Wang, G. Wichern, and J. Le Roux, "On the compensation between magnitude and phase in speech separation," *IEEE Signal Process. Lett.*, pp. 1–1, 2021.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

[20] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Comput. Speech Lang.*, vol. 46, pp. 605–626, 2017.

[21] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," 2017.

[22] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proc. Interspeech*, 2019, pp. 1368–1372.

[23] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: an improved version of MetricGAN for speech enhancement," in *Proc. Interspeech*, 2021, pp. 201–205.

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[25] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[26] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learning Repres.*, 2015.

[27] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. Interspeech*, 2018, pp. 3743–3747.

[28] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, and V. Manohar, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016, pp. 2751–2755.