ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746374

ATTENTION-BASED FUSION FOR BONE-CONDUCTED AND AIR-CONDUCTED SPEECH ENHANCEMENT IN THE COMPLEX DOMAIN

Heming Wang[1]*, Xueliang Zhang*[2]*, and DeLiang Wang*[1,3]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Department of Computer Science, Inner Mongolia University, China
[3]Center for Cognitive and Brain Sciences, The Ohio State University, USA
wang.11401@osu.edu, cszxl@imu.edu.cn, dwang@cse.ohio-state.edu

ABSTRACT

Bone-conduction (BC) microphones capture speech signals by converting the vibrations of the human skull into electrical signals. BC sensors are insensitive to acoustic noise, but limited in bandwidth. On the other hand, conventional or air-conduction (AC) microphones are capable of capturing full-band speech, but are susceptible to background noise. We propose to combine the strengths of AC and BC microphones by employing a convolutional recurrent network that performs complex spectral mapping. To better utilize signals from both kinds of microphone, we employ attention-based fusion with early-fusion and late-fusion strategies. Experiments demonstrate the superiority of the proposed method over other recent speech enhancement methods combining BC and AC signals. In addition, our enhancement performance is significantly better than conventional speech enhancement counterparts, especially in low signal-to-noise ratio scenarios.

Index Terms— bone conduction, speech enhancement, complex spectral mapping, attention-based fusion

1. INTRODUCTION

In real-world applications, speech signals are unavoidably deteriorated by noise interference. Speech enhancement aims to remove background noise and improves speech intelligibility and quality. Recent studies in monaural speech enhancement have demonstrated that deep neural networks (DNNs) based methods perform much better than traditional speech enhancement methods in noise suppression, even for untrained speakers and noise types [1, 2, 3, 4]. However, it remains challenging to produce high enhancement performance in low signal-to-noise ratio (SNR) conditions for non-stationary noises.

Speech enhancement is usually conducted on air-conduction (AC) microphone recordings. Unlike AC microphones, bone-conduction (BC) microphones convert vibrations from the human skull to electrical signals. On the one hand, BC signals are not contaminated by background interference that is acoustic in nature. On the other hand, speech collected from BC sensors suffers from limited bandwidth, as high-frequency components are lost due to the nature of bone conduction [5]. BC speech may sound intelligible but is muffled. AC microphones can record full-band speech with clear and natural sound, but are susceptible to background interference.

Early studies exploit BC signals to extract auxiliary information, like voice activity and pitch, in noisy conditions [6, 7]. Later, researchers attempt to recover clean AC speech by extending the bandwidth of BC signals. Conventional approaches can be divided into two categories. One category assumes that the BC signal can be simulated by passing an AC signal through a low-pass filter, and then attempts to estimate the transfer function to recover the AC signal [8, 9]. The other category is based on an analysis and synthesis model, observing that excitation sources are the same for both AC and BC signals. Previous studies use various features like linear predictive coding [10], linear spectral frequency [11], and mel-frequency cepstrum [12] to predict the spectral envelope of AC signals, and then perform speech synthesis. More recently, DNN based models are introduced to perform bandwidth extension to BC signals [13, 14, 15, 16]. However, the bandwidth of BC speech is very narrow, limited to 1 kHz to 2 kHz, which makes it very difficult to recover high-frequency components. In addition, unvoiced speech sounds are well captured by AC microphones but usually lost by BC sensors, as such sounds produce negligible bone vibrations in the human head. Therefore, performing bandwidth extension to BC speech does not yield satisfactory speech quality.

Recently, earbuds like Apple AirPods are widely adopted by consumers, and such devices feature both AC and BC sensors, making it easier to utilize two kinds of microphone for speech enhancement. Yu et al. [17] propose a time-domain fully convolutional network (FCN), which regards BC speech as another modality. This study demonstrates the utility of fusing AC and BC signals in speech enhancement.

We propose to leverage both kinds of microphone signals, and employ a convolutional recurrent network (CRN) [2] to perform speech enhancement in the complex domain.

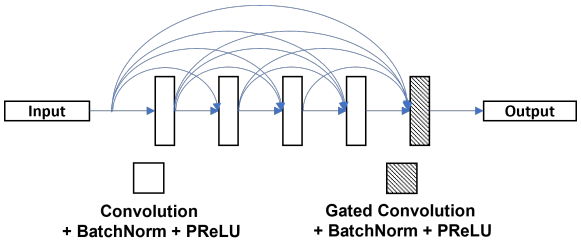


Fig. 1: Diagram of a DC block. The first four layers are standard 2D convolutions, and the last one utilizes the gated convolution.

To fully utilize the information of both AC and BC microphones, we introduce attention-based fusion [18] and dense connectivity [19] into our CRN. We also investigate the effects of different fusion strategies for merging AC and BC signals. Experiments show that our proposed approach substantially outperforms existing methods. In addition, AC-BC microphone fusion offers a clear advantage over conventional speech enhancement in low-SNR conditions.

2. PROPOSED METHOD

We perform speech enhancement using signals collected from both AC and BC microphones. A noisy mixture y_{AC} is collected from the AC microphone, which consists of background noise n and clean target speech s. Meanwhile, we have a noise-insensitive signal y_{BC} recorded from the BC microphone. Our goal is to produce an estimate $\hat{s}$ to recover the target clean speech s with the help of the proposed model f. Our CRN operates in the time-frequency (T-F) domain, so we apply STFT to the signals involved. With the parameters of the model denoted as θ, the problem can be formulated as,

$$\hat{S} = f(\theta, Y_{AC}, Y_{BC}), \tag{1}$$

where $\hat{S}$, Y_{AC} and Y_{BC} are the T-F representations of s, y_{AC} and y_{BC}, respectively.

2.1. Densely Connected Block

A densely connected (DC) network has been shown to be advantageous over the same network without dense connections [19]. In a DC network, one convolutional operation is decomposed into several, each having fewer channels and all convolution layers are directly connected to each other. Such design encourages the reusage of feature maps and alleviates the gradient vanishing problem. In our model, we replace standard convolutional layers with DC blocks. Fig. 1 illustrates the design of a DC block. More specifically, it consists of five convolutional layers. The first four are 2D convolutions with the number of output channels set to 8, and each is followed by a batch normalization and a parametric rectified linear unit (PReLU) activation. The final layer accepts outputs from all previous layers and performs the gated convolution [2] to further facilitate the feature fusion across convolution channels. The kernel size for each convolution layer is (1, 4) in the time and frequency axis, respectively.

2.2. DNN Architecture

As depicted in Fig. 2a, we employ the densely connected CRN (DC-CRN) to perform complex spectral mapping. Our network is based on the CRN architecture [2, 20], which is a complex-domain network built upon the typical encoder-decoder structure, and a recurrent neural network bottleneck is employed to model the temporal dependencies. CRN encoder is essentially a convolutional neural network (CNN) downsampler which reduces the feature dimension along the frequency axis using standard convolutions, and the decoder has a symmetric design that performs upsampling with transposed convolutions. In our case, we replace each convolution layer within the encoder and decoder with a DC block as described in Section 2.1. Pointwise convolutions are employed as skip connections to connect the corresponding layers of the encoder and decoder. Moreover, we use grouped bidirectional long short-term memory (BLSTM) [21, 2] as the bottleneck, allowing the reduction in the computational complexity and model parameters. Finally, the output of the CNN decoder is halved and then reshaped into one-dimensional features. Each halve passes through a linear layer to produce real and imaginary estimates.

2.3. Attention Based Fusion

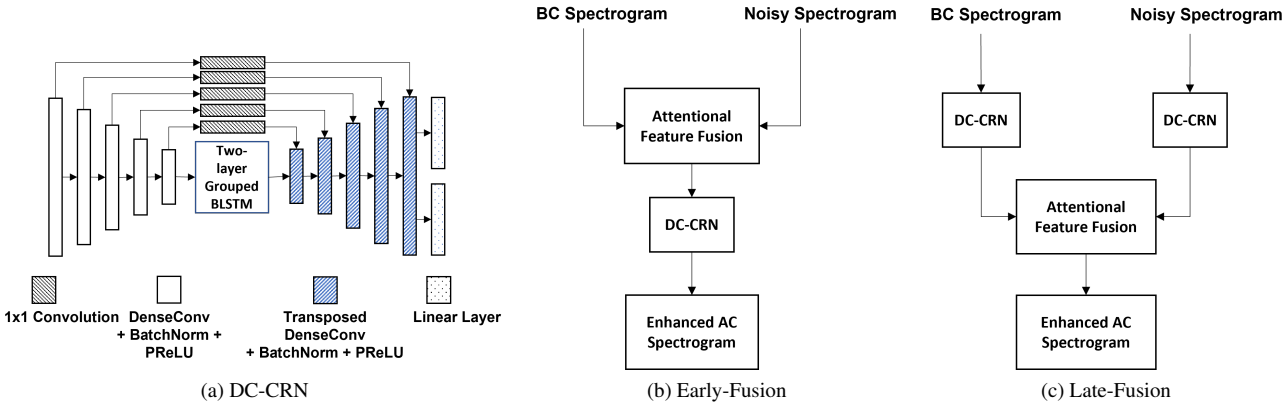
Inspired by [18], we perform attention-based fusion on AC and BC features as illustrated in Fig. 3. First, we aggregate local context and global context which are obtained by pointwise convolutions, and then calculate the attention score M using a sigmoidal activation. Then, we concatenate two features and assign weights M and $1 - M$ to each feature map, respectively. We investigate two fusion strategies, as depicted in Fig. 2b and 2c. Early-fusion merges AC and BC signals before feeding them to the DC-CRN module. For the late-fusion strategy, AC and BC signals are fed to separate DC-CRN modules, and we perform feature fusion on the outputs of the two modules.

2.4. Loss

Our loss function is defined in complex spectrogram. Recent studies [22, 23, 24] show the importance of including a magnitude loss in complex-domain networks and suggest that a well-estimated magnitude implicitly compensates phase estimation. Motivated by that observation, we construct the loss by combining a magnitude difference with a complex representation difference. With the total number of time steps and frequency bins denoted as T and F, the loss is defined as,

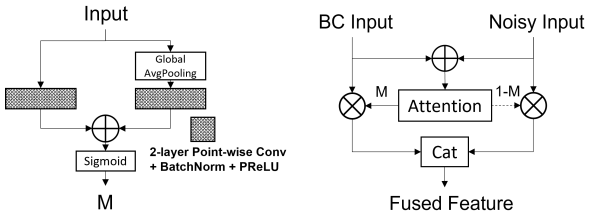
$$L(t,f) = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}[||\hat{S}(t,f)| - |S(t,f)|| + (|\hat{S}_r(t,f) - S_r(t,f)| + |\hat{S}_i(t,f) - S_i(t,f)|)], \tag{2}$$

where t, f index the time step and frequency bin, and the subscripts r and i correspond to the real and imaginary parts of the complex representation, respectively.



(a) DC-CRN (b) Early-Fusion (c) Late-Fusion

Fig. 2: Diagrams of the proposed architecture. (a). DC-CRN, (b). Early-fusion strategy, and (c). Late-fusion strategy.



(a) Attention Module (b) Attention-based Feature Fusion

Fig. 3: Illustration of the AC-BC feature fusion. (a) depicts the process of calculating the score M, and (b) shows we use the score M to perform soft selection and feature concatenation. In the diagrams, $\bigotimes$ represents the element-wise multiplication, and $\bigoplus$ is the symbol for broadcasting summation.

3. EXPERIMENTS

3.1. Dataset

We conduct experiments on the Elevoc Simultaneously-recorded Microphone/Bone-sensor (ESMB) speech corpus[1], which consists of 128 hours of Chinese speech uttered by 131 male and 156 female speakers. During recording, speech is captured by a pair of Elevoc Clear earbuds, each of which consists of a ST 25ba BC sensor located near the ear canal to collect skull vibrations, and an AC sensor outside the ear that serves as a close-talk microphone. Each speaker reads Chinese prompts for around 20 minutes, and a 16 kHz stereo speech is recorded by each earbud. For our experiments, we set aside 4 speakers (2 male and 2 female) for validation and 4 speakers (2 male and 2 female) for testing. We use the same noise set for training and validation, which is extracted from the DNS challenge[2] by randomly picking 20000 files. For each training utterance, we generate a noisy speech utterance by mixing an AC signal with a training noise at an SNR level uniformly sampled from the range {-5, -4, -3, -2, -1, 0} dB. For testing, we select four challenging noises, babble and cafeteria from an Auditec CD[3], and factory and engine from the NOISEX92 dataset [25]. Each testing utterance is mixed with the four noises at three SNR levels -5, 0 and 5 dB. The enhancement performance is evaluated with two standard metrics, perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI). For both metrics, higher values denote better performance.

3.2. Setup

All the recordings are resampled to 8 kHz. For training and validation, we split each recording into non-silent utterances and discard silent sequences with energy 60 dB below the peak power reference. A window length of 32 ms with 50% overlap between adjacent frames is used for STFT operations, which corresponds to a 129-dimensional spectrum. We normalize each noisy AC mixture using the mean-variance normalization (MVN), and the corresponding clean AC utterance is scaled accordingly. BC utterance is first passed through a Butterworth low-pass filter, then normalized with MVN. During training, we use the Adam optimizer and train our model with a batch size of 16 utterances for 30 epochs. An initial learning rate is set to 0.0006, and is halved if the validation loss has not improved for three consecutive epochs.

4. RESULTS AND ANALYSIS

Fig. 4 plots the enhancement performance of the baseline approach FCN [17] and our proposed method on the ESMB dataset. The networks denoted by the subscript AC only utilizes AC signals, and the ones denoted by EF and LF utilize both AC and BC signals with early-fusion and late-fusion strategies, respectively. From the figure, we observe that networks that employ AC-BC fusion always outperform their counterparts that only utilize AC signals. Especially at -5 dB SNR, microphone fusion considerably boosts the enhance-

[1]available at https://github.com/elevoctech/ESMB-corpus
[2]available at https://github.com/microsoft/DNS-Challenge
[3]available at http://www.auditec.com

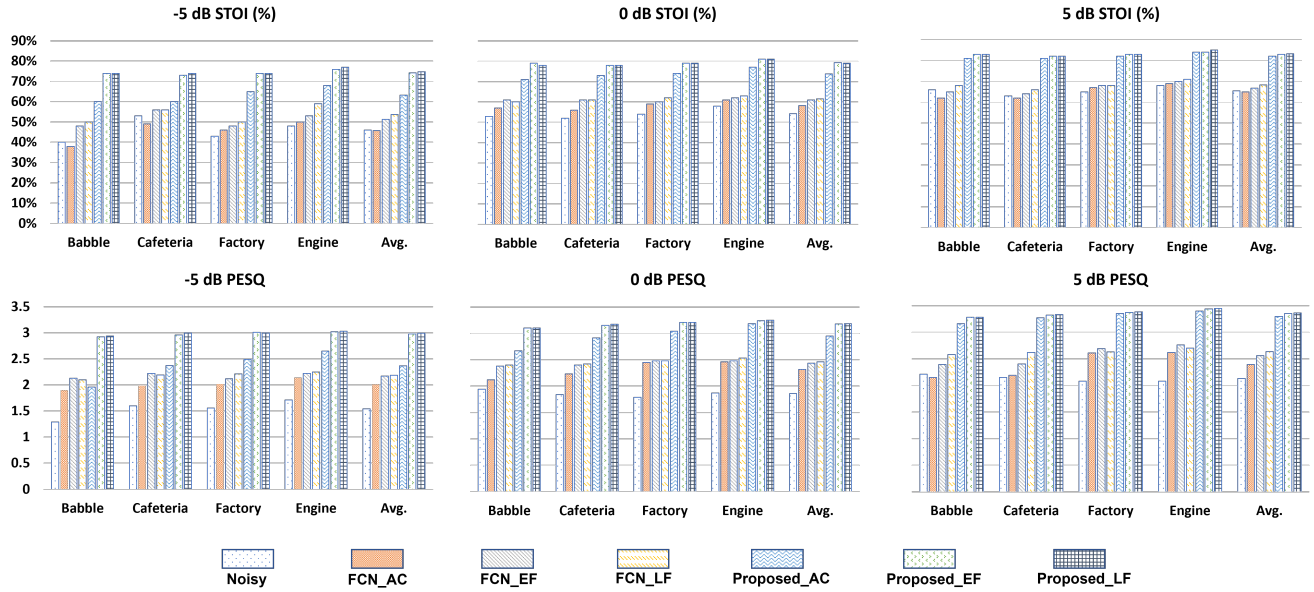


Fig. 4: Enhancement performance of the FCN baseline and our proposed method in terms of STOI (%) and PESQ.

Table 1: Ablation Study of the proposed network at -5 dB SNR

	STOI (%)	PESQ
Proposed_EF	74.4	2.98
- Convolutional skip connections (i)	74.0	2.96
- DC blocks (ii)	70.5	2.71
- Gated convolution (iii)	72.6	2.78
- Attention feature fusion (iv)	73.9	2.92

ment performance. Specifically, for our proposed approach, STOI is improved by over 10% and PESQ by over 0.60. As the SNR level becomes higher, incorporating BC signals is less beneficial. At 5 dB SNR, STOI is merely 1.0% higher, and PESQ is improved by 0.05. In addition, our complex-domain approach shows consistently better enhancement performance in all conditions compared with the time-domain baseline FCN [17]. At -5 dB SNR, we achieve an STOI of 74.8% for the late-fusion version, which is 21.5% higher than the FCN counterpart. Comparing the two different fusion strategies, late-fusion performs slightly better. However, late-fusion has almost twice the model size as it employs two DC-CRN modules. Therefore, there is a trade-off between performance and computational cost.

In Table 1, an ablation study is conducted to investigate the effects of different components within the proposed network. We employ the early-fusion version of our network as the baseline and compare several variants at -5 dB SNR: (i) replacing pointwise convolution-based skip connections with concatenation-based skip connections; (ii) replacing DC blocks with standard convolutions; (iii) replacing the gated convolutions within DC blocks with standard convolutions; (iv) instead of using attention-based fusion, BC and AC features are directly concatenated as the input vector. As shown in the table, these variants all underperform the proposed design. Among all factors, dense connectivity plays a significant role for the final performance, as removing DC blocks degrades STOI by 3.9% and PESQ by 0.27. Furthermore, attention-based feature fusion shows to be more effective compared to the simple concatenation of microphone features.

5. CONCLUSION

In this study, we have proposed a novel attention-based method to fuse AC and BC microphone signals for complex-domain speech enhancement. The proposed method takes advantage of the full bandwidth of AC microphones and the noise insensitivity of BC microphones to obtain high-quality enhanced speech in adverse environments. Experiments have demonstrated that our approach substantially outperforms a previous time-domain baseline. Compared with conventional speech enhancement on AC microphones, our AC-BC fusion significantly boosts enhancement performance, especially in low-SNR conditions. In future work, we plan to introduce semi-supervised learning techniques to utilize AC and BC speech data that are not recorded in parallel, so as to achieve strong performance with a small amount of BC signals.

6. ACKNOWLEDGMENT

This work began when the first author was an intern at Elevoc Co. Ltd. The authors would like to thank Yongjie Yan for his assistance in organizing the ESMB corpus. This research was supported in part by an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

7. REFERENCES

[1] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[2] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.

[3] A. Pandey and D. L. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1179–1188, 2019.

[4] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proceedings of INTERSPEECH*, 2020, pp. 2482–2486.

[5] M. S. Rahman and T. Shimamura, "Intelligibility enhancement of bone conducted speech by an analysis-synthesis method," in *Proceedings of MWSCAS*, 2011, pp. 1–4.

[6] M. Zhu, H. Ji, F. Luo, and W. Chen, "A robust speech enhancement scheme on the basis of bone-conductive microphones," in *Proceedings of IWSDA*, 2007, pp. 353–355.

[7] M. S. Rahman and T. Shimamura, "Pitch characteristics of bone conducted speech," in *Proceedings of EUSIPCO*, 2010, pp. 795–799.

[8] T. Shimamura and T. Tamiya, "A reconstruction filter for bone-conducted speech," in *Proceedings of MWSCAS*, 2005, pp. 1847–1850.

[9] R. E. Bouserhal, T. H. Falk, and J. Voix, "In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension," *The Journal of the Acoustical Society of America*, vol. 141, pp. 1321–1331, 2017.

[10] T. V. Thang, K. Kimura, M. Unoki, and M. Akagi, "A study on restoration of bone-conducted speech with MTF-based and LP-based models," *Journal of signal processing*, 2006.

[11] B. Huang, Y. Gong, J. Sun, and Y. Shen, "A wearable bone-conducted speech enhancement system for strong background noises," in *Proceedings of ICEPT*, 2017, pp. 1682–1684.

[12] P. Singh, M. K. Mukul, and R. Prasad, "Bone conducted speech signal enhancement using LPC and MFCC," in *Proceedings of IHCI*, 2018, pp. 148–158.

[13] C. Zheng, J. Yang, X. Zhang, T. Cao, M. Sun, and L. Zheng, "Bandwidth extension WaveNet for bone-conducted speech enhancement," in *Proceedings of CSMT*, 2020, pp. 3–14.

[14] C. Zheng, T. Cao, J. Yang, X. Zhang, and M. Sun, "Spectra restoration of bone-conducted speech via attention-based contextual information and spectro-temporal structure constraint," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E102.A, pp. 2001–2007, 2019.

[15] H. Q. Nguyen and M. Unoki, "Improvement in bone-conducted speech restoration using linear prediction and long short-term memory model," *Journal of Signal Processing*, vol. 24, pp. 175–178, 2020.

[16] T. Hussain, Y. Tsao, S. M. Siniscalchi, J.-C. Wang, H.-M. Wang, and W.-H. Liao, "Bone-conducted speech enhancement using hierarchical extreme learning machine," in *Proceedings of IWSDS*, 2021, pp. 153–162.

[17] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, "Time-domain multi-modal bone/air conducted speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.

[18] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of WACV*, 2021, pp. 3560–3569.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of CVPR*, 2017, pp. 4700–4708.

[20] K. Tan, X. Zhang, and D. L. Wang, "Deep learning based real-time speech enhancement for dual-microphone mobile phones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1853–1863, 2021.

[21] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T.-Y. Liu, "Efficient sequence learning with group recurrent networks," in *Proceedings of NAACL-HLT*, 2018, pp. 799–808.

[22] Z.-Q. Wang, P. Wang, and D. L. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1778–1787, 2020.

[23] Z.-Q. Wang, G. Wichern, and J. L. Roux, "On the compensation between magnitude and phase in speech separation," *IEEE Signal Processing Letters*, 2021.

[24] J. Zhang, M. D. Plumbley, and W. Wang, "Weighted magnitude-phase loss for speech dereverberation," in *Proceedings of ICASSP*, 2021, pp. 5794–5798.

[25] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.