A DEEP HIERARCHICAL FUSION NETWORK FOR FULLBAND ACOUSTIC ECHO CANCELLATION

Haoran Zhao, Nan Li, Runqiang Han, Lianwu Chen, Xiguang Zheng, Chen Zhang, Liang Guo, Bing Yu

Kuaishou Technology, Beijing, China

ABSTRACT

Deep learning based wideband (16kHz) acoustic echo cancellation (AEC) approaches have surpassed traditional methods. This work proposes a deep hierarchical fusion (DHF) network with intra-network and inter-network fusion to further improve the wideband AEC performance. Meanwhile, this work extends the existing wideband systems to enable fullband (48kHz) AEC while simultaneously ensuring automatic speech recognition compatibility by incorporating with an ASR loss. The proposed system has ranked 2nd place in ICASSP 2022's AEC Challenge.

***Index Terms*—** acoustic echo cancellation, inter-network fusion, intra-network fusion

1. INTRODUCTION

In recent years, deep neural network (DNN) based acoustic echo cancellation (AEC) methods have achieved a significant improvement over the traditional signal processing based methods. While recent systems [1, 2] have performed well for real-time wideband (16kHz) AEC tasks, a deep hierarchical fusion (DHF) system is proposed in this work to simultaneously support fullband (48kHz) AEC and automatic speech recognition (ASR) compatibility. In the remainder of this paper, the sampling frequency is abbreviated to '16k' and '48k' to avoid confusion with the signal frequency.

While less research has been conducted for fullband AEC, existing approaches generally employ psychoacoustically motivated models [3] to reduce the frequency-wise feature dimensions. Compared to the existing fullband AEC approaches, this work extends our recently proposed multi-stage fullband speech enhancement (SE) structure [4] to the AEC task where a computationally efficient highband (16-48k) system is built on top of a wideband (0-16k) system. For the highband system, in addition to the highband AEC input signals, the highband system is also informed by the output of the wideband AEC system and thus achieves better fullband AEC performance compared to the existing systems. For the wideband AEC task, intra-network and inter-network fusions are explored to achieve deep hierarchical fusion. First, a CrossNet structure with cross-connections between two parallel branches of speech and interference is proposed to explore intra-network fusion. A two branch GRU-CrossNet extended from the convolutional recurrent neural network (CRNN) [5] is operating in parallel with a separately trained two branch TCN-CrossNet where the GRU layers are replaced by the temporal convolutional network (TCN) layers. This is motivated by the success of replacing the recurrent layers with the dilated 1-D convolution layers for speech enhancement tasks [6, 7]. The outputs of these two wideband CrossNets are jointly processed by an inter-network fusion module to obtain the final wideband AEC output. It should be noted that the idea of inter-network fusion has been widely used for music source separation [8] and audio scene classification [9], while the idea of intra-network fusion has been explored for speech sentiment analysis in [10]. This work incorporates the inter-network and intra-network fusions and demonstrates their superior performance for the AEC task.

Employing the multi-stage training structure also makes it easier to incorporate the ASR compatibility with the AEC system. The reason is that the common ASR systems also take the wideband input signals. While it has been found that employing a pre-trained back-end ASR system during the SE model training stage can significantly improve the Word Error Rate (WER) of an SE system [11], here, we also employ a multi-loss strategy formed by an AEC loss and an ASR loss obtained by an ASR feature embedding loss generated from the encoder portion of WeNet system [12] to simultaneously ensure good AEC and ASR performance.

The proposed method participated in the ICASSP 2022 Acoustic Echo Cancellation Challenge [13]. With 14.4M parameters, 1.95G FLOPs, and 1.029ms one-frame inference time on Intel Core i7 (2.6GHz) CPU, the proposed system has ranked 2nd place in the final ranking table.

2. OVERVIEW

Figure 1 presents the overview of the proposed hierarchical fusion network. The input 48k near-end and far-end signals are fed to a linear AEC module from the SpeexDSP's[1] implementation with 160ms adaptive filter length, 30ms frame size and 10ms frame shift. Regarding the time delay between near-end and far-end signals caused by the system latency, we use a robust delay estimation method based on cross-correlation [14] to estimate and compensate the causal delay.

[1] https://gitlab.xiph.org/xiph/speexdsp

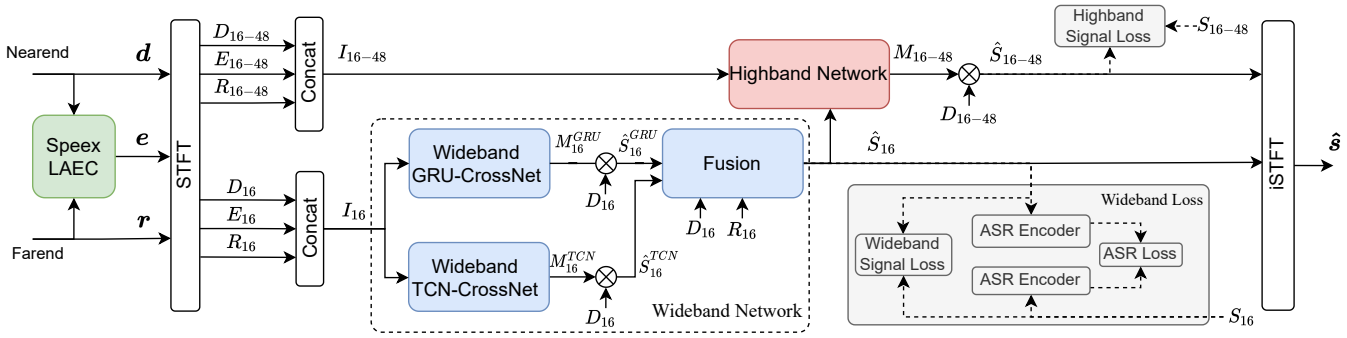

Fig. 1. The proposed deep hierarchical fusion network for acoustic echo cancellation.

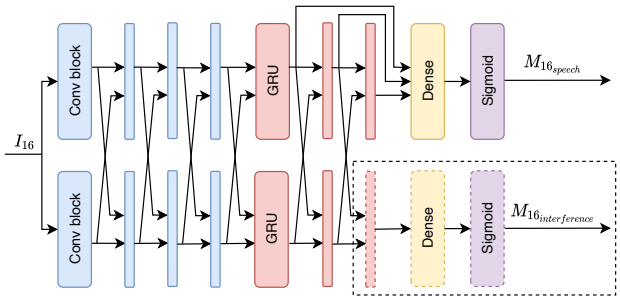

Fig. 2. The proposed wideband GRU-CrossNet architecture.

Table 1. Hyper-parameters for wideband GRU-CrossNet. The size and stride are given in [time, frequency] format.

Layer	CNN			RNN	FC
	In/Out Ch	Size	Stride	Units	Units
Conv2D	3,4	[3,1]	[1,1]		
Conv2D	8,16	[1,3]	[1,2]		
Conv2D	32,16	[3,3]	[1,2]		
Conv2D	32,16	[1,3]	[1,2]		
GRU				384	
Dense					240

The time domain near-end signal $\boldsymbol{d}$, the far-end signal $\boldsymbol{r}$ and the linear AEC output signal $\boldsymbol{e}$ are transformed to the time-frequency (T-F) domain by a 1440-point (30ms) short time Fourier transform (STFT) with 480-point (10ms) stride. These T-F representations are further divided into a wideband $[D_{16}, R_{16}, E_{16}]$ (0-16k) and a highband $[D_{16-48}, R_{16-48}, E_{16-48}]$ (16-48k) parts. The wideband signal D_{16}, R_{16} and E_{16} are concatenated to I_{16} and then fed into two wideband CrossNets. A fusion network is proposed to fuse the estimated wideband speech by these two wideband CrossNets. A joint loss with the wideband signal loss and the ASR loss obtained by a pre-trained ASR encoder is employed for the wideband signal. A highband network is utilized to estimate the highband speech $\hat{S}_{16-48}$ based on the concatenated highband signals I_{16-48} and estimated wideband speech $\hat{S}_{16}$. Finally, the complete magnitude spectrum are obtained from $\hat{S}_{16}$ and $\hat{S}_{16-48}$. The complete magnitude spectrum (combined with the phase of the near-end signal) is transformed back to the time domain using iSTFT.

3. SYSTEM DESCRIPTION

3.1. Intra-network fusion

In this paper, we propose a wideband GRU-CrossNet structure to explore the idea of intra-network fusion, as shown in Figure 2. The proposed GRU-CrossNet structure employs two parallel branches to model the speech and interference (noise and echo) explicitly. The log-power spectrum of the linear AEC out, near-end and far-end signals are used as the inputs to the Conv block containing a 2D convolutional layer to perform the feature extraction. The GRU layers integrate the extracted features over time to model the context information. Finally the dense layers with sigmoid activation produce the magnitude masks for the speech and interference, respectively. Intra-network fusion between speech and interference is achieved using cross-connections added for each of the convolutional and the recurrent layer. In the speech branch, all GRU layer outputs are fed into the dense layer. While the outputs of two branches are used for a joint loss calculation in the training phase, the dashed blocks in the interference branch are discarded in the testing phase. The hyper-parameters of the GRU-CrossNet are listed in Table 1. A parallel separately trained TCN-CrossNet is also employed. The configuration of the TCN-CrossNet is the same as the GRU-CrossNet except the GRU layers are replaced by the TCN layers motivated by [6, 7] for speech enhancement tasks. For each of the TCN layers with channel-wise layer normalization, H, P, X are 150, 3 and 7, using the same notation as in [6].

3.2. Inter-network fusion

To explore the idea of inter-network fusion, a sub-band based fusion module is proposed to fuse the output of the wideband networks. As shown in Figure 3, the near-end wideband signal D_{16}, the far-end wideband signal R_{16}, the wideband estimated speech signals $\hat{S}_{16}^{TCN}$ and $\hat{S}_{16}^{GRU}$ are concatenated in channel dimension to form the input of the fusion network. The fusion network has three conv2D layers, a GRU layer and a Dense layer with softmax activation, the hyper-parameters are shown in Table 2. The sub-band fusion weights for the two AEC systems are estimated by the fusion network. The number of sub-bands is set to 16. To avoid artifacts introduced by fast switching between the two systems, the weights are fur-

Table 2. Hyper-parameters for fusion network.

Layer	CNN			RNN	FC
	Channels	Size	Stride	Units	Units
Conv2D	32	[1,5]	[1,3]		
Conv2D	32	[1,3]	[1,2]		
Conv2D	32	[1,3]	[1,2]		
GRU				240	
Dense					$16 * 2$

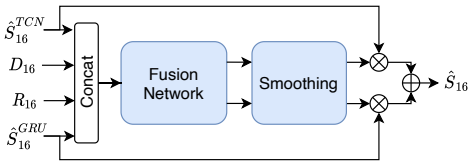

Fig. 3. Illustration of the inter-network fusion architecture.

ther smoothed along the time domain by exponential moving average with decay set to 0.95. The enhanced signals are multiplied with the corresponding weight, and the weighted signals are summed together to generate the fusion output $\hat{S}_{16}$. Since different structures may have complementary strengths and weaknesses, inspired by [15], the motivation of the inter-network fusion is to employ a fusion network to smartly highlight the strengths of the wideband networks.

3.3. Wideband loss fusion

As shown in Figure 1, the wideband loss contains a signal loss to ensure the speech enhancement quality and an ASR loss to optimize the speech recognition accuracy.

For the signal domain loss, Optimal Scale-Invariant Signal-to-Noise Ratio (OSISNR) loss [16] is applied since it can help to preserve the speech with less energy and performed well for speech separation task. In addition, Magnitude power-law Compressed Mean Square Error (MC-MSE) loss function [17] is used to further suppress the residual noise of estimated speech. We combine these two signal loss functions weighted by hyper-parameter γ for each branch:

$$\mathcal{L}_{\text{S}} = \mathcal{L}_{\text{OSISNR}} + \gamma \mathcal{L}_{\text{MC-MSE}} \tag{1}$$

where γ is set to 15 according to our preliminary results. $\mathcal{L}_{\text{S}}$ is used for the loss calculation of speech branch in Figure 2. The identical loss $\mathcal{L}_{\text{S-i}}$ is employed for the interference branch.

In addition, an ASR loss $\mathcal{L}_{\text{ASR}}$ is used to reduce the WER for ASR. More specifically, Wasserstein distance between clean speech ASR embedding and estimated speech ASR embedding is employed as the ASR loss. The ASR embeddings are extracted by the pre-trained WeNet encoder on LibriSpeech with U2++ Conformer structure [12]. The overall wideband loss $\mathcal{L}_{\text{WB}}$ is:

$$\mathcal{L}_{\text{WB}} = \mathcal{L}_{\text{S}} + \mathcal{L}_{\text{S-i}} + \mathcal{L}_{\text{ASR}} \tag{2}$$

3.4. Fullband System

While the effectiveness of feeding the estimated wideband speech signal to the highband network alongside with the high band noisy input signal has been demonstrated in our previous work for speech enhancement [4], here, $\hat{S}_{16}$ is fed to the highband SE network alongside with I_{16-48} for the AEC task. In the highband network, two CNN branches are utilized to extract the wideband and highband features from the estimated wideband speech and the noisy highband speech respectively. The wideband and highband features are then combined and fed into a recurrent layer and a feed-forward layer to estimate the highband mask M_{16-48}. The details of highband network structure and hyper-parameters can be found in [4]. The OSISNR and MC-MSE loss between clean and estimated highband speech are utilized to optimize the highband network:

$$\mathcal{L}_{\text{HB}} = \mathcal{L}_{\text{OSISNR}} + \gamma \mathcal{L}_{\text{MC-MSE}} \tag{3}$$

4. DATASETS, EXPERIMENTS AND RESULTS

4.1. Dataset

The ICASSP 2022 AEC Challenge provides a 48k dataset, in which we randomly selected 14000 and 3000 utterances for training and validation, respectively. Additionally, the LibriSpeech dataset is used for the near-end speech while the noise data from the ICASSP 2022 DNS challenge is used for the background noise. The 16k dataset is downsampled from their 48k counterparts for the wideband networks.

To further ensure the generalization ability, the training data are synthesized online with random parameters in each epoch. The data augmentation steps are done as follows: (1) randomly select the near-end speech, the noise, the far-end reference and the echo from the training dataset; (2) remix the near-end speech and the noise with a random SNR uniformly distributed in [0, 45] dB; (3) remix the near-end speech and the echo with a random SER uniformly distributed in [-15, 15] dB; (4) mimic varying delays by randomly adding silence (0–100ms) in the echo utterances (in 5% of the cases); (5) generate 20000 random room impulse responses using image method [18] and randomly apply to the near-end and the noise utterances; (6) use various equalizers and bandpass filters.

Regarding the test set, we randomly select the utterances from the VoiceBank [19] dataset, including two men and two women. The noise utterances are chosen from the DEMAND [20] dataset. The echo and far-end reference utterances are selected from the single-talk real recordings from the ICASSP 2022 AEC Challenge testset. We generate single-talk near-end, single-talk far-end and double-talk test data with SER randomly chosen from [-10, 10] dB. In 50% cases, a noise sample is mixed with random SNR from [-5, 10] dB.

4.2. Implementation Details

The fullband DHF AEC system in Figure 1 is trained in multiple stages:

- Stage1: The wideband GRU-CrossNet and TCN-CrossNet are separately trained using the signal loss $\mathcal{L}_{\text{S}} + \mathcal{L}_{\text{S-i}}$.
- Stage2: The GRU-CrossNet and TCN-CrossNet are separately fine-tuned by adding $\mathcal{L}_{\text{ASR}}$ to form $\mathcal{L}_{\text{WB}}$. Then these two models are frozen for the following stages.

- Stage3: The fusion network is trained based on the outputs of the pre-trained GRU-CrossNet and TCN-CrossNet with the loss $\mathcal{L}_{\text{WB}}$.
- Stage4: the highband network is trained based on the existing wideband network using the highband loss $\mathcal{L}_{\text{HB}}$.

Batch normalization and parametric rectified linear unit (PReLU) activation are used after each convolution layer. The initial learning rate is set to 0.0002 for Stage2 and 0.0005 for other stages. In each stage, the learning rate decays by a factor of 0.7 once the validation loss does not decrease in successive 8 epochs with a batch size of 8. Gradient norm clipping is used and the max value of the gradients is set to 1.0. The Adam optimizer is employed for all stages.

4.3. Experiment 1: Wideband Hierarchical Fusion

We first evaluate the impact of wideband intra-network fusion and inter-network fusion with signal loss. The evaluation metrics include echo return loss enhancement (ERLE) for single-talk far-end scenario, wideband PESQ [21] and WER for single-talk near-end and double-talk scenario, AECMOS [22] for all scenarios with the same average manner as in ICASSP 2022 subjective test. The results are listed in the upper half of Table 3.

G-CN and T-CN represent the GRU-CrossNet and TCN-CrossNet, respectively. G-speech is the upper part of the G-CN in Fig 2. For a fair comparison, the hidden size of GRU layers is set to 660 instead of 384 to make the model size of G-speech closer to G-CN. The results show that G-CN outperforms G-speech indicating the effectiveness of employing the intra-network fusion mechanism. Comparing G-CN+T-CN with G-CN and T-CN, condition G-CN+T-CN with inter-network fusion has achieved 0.06 and 0.031 improvement on PESQ and AECMOS while reduces the WER by 0.46.

We further evaluate the impact of the ASR loss for the proposed system in the bottom half of Table 3. Models fine-tuned by the ASR loss $\mathcal{L}_{\text{ASR}}$ are able to further reduce the WER and increase the ERLE while maintaining the PESQ score and slight degradation (0.024) on the AECMOS. It should be noted that the results also indicate that the AECMOS may be better correlated with the human subjective ratings in the AEC scenarios than PESQ as the degradation caused by the ASR loss can be captured by AECMOS.

4.4. Experiment 2: Fullband Speech Quality

We also compare different training configurations for the fullband network. We employ PESQ-WB$_{16}$, SISNR$_{16}^{48}$ [23] and SISNR$_{0}^{48}$ to evaluate the wideband (0-16k), highband (16-48k), and the fullband (0-48k) speech quality, respectively. Single-talk near-end and double-talk scenarios are considered for this evaluation. As shown in Table 4, GRU-1-step represents a fullband GRU-CrossNet trained for fullband input in a one-step manner. GRU-2-step is the proposed system, in which the wideband and highband network are trained by two steps as in Figure 1. For a fair comparison, GRU-1-step and GRU-2-step are designed to the similar size and trained using the same signal loss. As indicated in Table 4, the performance of the proposed two-step system outperforms the one-step fullband system (GRU-1-step) in terms of wideband, highband and fullband speech quality.

Table 3. Evaluating the wideband DHF and ASR loss.

Method	PESQ-WB	ERLE(dB)	AECMOS	WER
Noisy	2.28	0	2.768	21.90
G-speech	2.99	50.04	4.214	11.85
G-CN	3.07	55.67	4.263	11.0
T-CN	3.10	53.77	4.252	10.67
G-CN+T-CN	3.16	53.82	**4.294**	10.21
G-CN+$\mathcal{L}_{\text{ASR}}$	3.09	57.85	4.242	9.99
T-CN+$\mathcal{L}_{\text{ASR}}$	3.09	**57.88**	4.244	9.49
G-CN+T-CN +$\mathcal{L}_{\text{ASR}}$	**3.17**	56.30	4.270	**9.11**

Table 4. Comparison of different fullband networks.

Method	PESQ-WB$_{16}$	SISNR$_{16}^{48}$	SISNR$_{0}^{48}$
Noisy	2.28	-4.76	1.20
GRU-1-step	3.03	7.42	16.65
GRU-2-step	**3.07**	**8.53**	**17.85**

Table 5. ICASSP 2022 AEC challenge results.

	Subjective	WAcc	Final
Baseline	4.100	0.659	0.752
Proposed	4.528	0.799	0.866

4.5. Experiment 3: ICASSP 2022 AEC Challenge Results

The proposed method participated in the ICASSP 2022 Acoustic Echo Cancellation Challenge. As shown in Table 5 from the official ranking table, the proposed deep hierarchical fusion network has ranked 2nd place for subjective speech quality evaluation, the word error rate evaluation and the overall score in the ICASSP 2022 AEC Challenge. With 1440-point FFT (corresponding to 30ms for 48k input signal) and 480-point stride (corresponding to 10ms for 48k input signal) with no look-ahead in the network, the total system delay is 30ms + 10ms = 40ms, which satisfies the latency requirement of this challenge. The proposed network has 17.4M parameters in total. It takes 1.029 ms to process a 10ms frame on Intel Core i7 (2.6GHz) CPU, including linear AEC and network inference time.

5. CONCLUSIONS

A deep hierarchical fusion network for fullband acoustic echo cancellation system is proposed. The performance of the intra-network and inter-network fusion has been discussed and evaluated for the wideband AEC with joint signal and ASR losses. Fullband AEC is also achieved by employing a two-step structure and outperforms the one-step structure. The proposed system has ranked 2nd place in ICASSP 2022 Acoustic Echo Cancellation Challenge.

6. REFERENCES

[1] H. Zhang, K. Tan, and D. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions," in *Proc. of Interspeech*, 2019, pp. 4255–4259.

[2] A. Ivry, I. Cohen, and B. Berdugo, "Nonlinear acoustic echo cancellation with deep learning," in *Proc. of Interspeech*, 2021, pp. 4773–4777.

[3] J.-M. Valin, S. V. Tenneti, K. Helwani, U. Isik, and A. Krishnaswamy, "Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet," in *Proc. of ICASSP*, 2021, pp. 7133–7137.

[4] X. Zhang, L. Chen, X. Zheng, X. Ren, C. Zhang, L. Guo, and B. Yu, "A two-step backward compatible fullband speech enhancement system," in *ICASSP 2022*, In Press, https://arxiv.org/abs/2201.10809.

[5] H. Zhao, S. Zarar, I. Tashev, and C. Lee, "Convolutional-Recurrent Neural Networks for Speech Enhancement," in *Proc. of ICASSP*, 2018, pp. 2401–2405.

[6] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM TASLP*, vol. 27, pp. 1256–1266, 2019.

[7] V. Kishore, N. Tiwari, and P. Paramasivam, "Improved speech enhancement using tcn with multiple encoder-decoder layers," in *Proc. of Interspeech*, 2020, pp. 4531–4535.

[8] M. Kim, W. Choi, J. Chung, D. Lee, and S. Jung, "Kuielab-mdx-net: A two-stream neural network for music demixing," in *Proc. of MDX Workshop*, 2021.

[9] S. Seo and J. Kim, "Mobilenet using coordinate attention and fusions for low-complexity acoustic scene classification with multiple devices," Tech. Rep., DCASE2021 Challenge, June 2021.

[10] E. Georgiou, C. Papaioannou, and A. Potamianos, "Deep hierarchical fusion with application in sentiment analysis," in *Proc. of Interspeech*, 2019, pp. 1646–1650.

[11] J. Wu, Z. Chen, S. Chen, Y. Wu, T. Yoshioka, N. Kanda, S. Liu, and J. Li, "Investigation of practical aspects of single channel speech separation for asr," in *Proc. of Interspeech*, 2021.

[12] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "WeNet: Production Oriented Streaming and Non-Streaming End-to-End Speech Recognition Toolkit," in *Proc. of Interspeech*, 2021, pp. 4054–4058.

[13] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sorensen, and R. Aichner, "Icassp 2022 acoustic echo cancellation challenge," 2022.

[14] J Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 6, pp. 998–1003, 1982.

[15] X. Jaureguiberry, E. Vincent, and G. Richard, "Fusion methods for speech enhancement and audio source separation," *IEEE/ACM TASLP*, vol. 24, pp. 1266–1279, 2016.

[16] C. Ma, D. Li, and X. Jia, "Optimal scale-invariant signal-to-noise ratio and curriculum learning for monaural multi-speaker speech separation in noisy environment," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2020, pp. 711–715.

[17] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 9458–9465.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[19] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," *O-COCOSDA/CASLRE*, pp. 1–4, 2013.

[20] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," *Proc. of Meetings on Acoustics*, 2013.

[21] J. Rix, A.and Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. of ICASSP*, 2001.

[22] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "Aecmos: A speech quality assessment metric for echo impairment," *arXiv preprint arXiv:2110.03010*, 2022.

[23] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM TASLP*, vol. 26, no. 4, pp. 787–796, 2018.