Diverse Time-Frequency Attention Neural Network for Acoustic Echo Cancellation

Jinzhuo Yao*, Hongqing Liu*, Yi Zhou*, Lu Gan† and Junkang Yang*
* School of Communications and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China
† College of Engineering, Design and Physical Science, Brunel University London, U.K.
E-mail:s220101190@stu.cqupt.edu.cn

***Abstract*—Acoustic echo cancellation (AEC) aims to eliminate echoes from near-end microphone signals and recover the near-end speech at the same time. In this work, we propose a Diverse Time Frequency Attention Neural Network (DTFAN) for AEC that operates in a full network-based manner. To that end, we first utilize a network aiming at aligning the features of the far-end reference signal and the near-end microphone signal. After that, the diverse time-frequency attentions capturing the intrinsic connections of the features in the time and frequency domains are developed. Since the alignment of the reference signal is conducted by the network, the requirement of traditionally pre-processing the far-end signal is avoided, and the whole network is end-to-end. The experimental results show that the proposed framework performs well and robustly on the synthetic test set and the blind test dataset compared to other recent approaches, especially in double-talk scenarios.**

I. Introduction

In a full-duplex voice communication system, the far-end user will receive a modified version of his/her speech due to the acoustic coupling that occurs between the near-end microphone and the near-end speaker. Acoustic Echo Cancellation (AEC) systems are designed to eliminate echoes from the microphone signal while minimizing distortion of the near-end speaker's voice [1], [2].

In addition to traditional methods [1], [2], recently, there have been many successes in deep learning-based AEC studies due to the great potential demonstrated by deep learning in modeling complex nonlinear problems. In [3], frequency-domain adaptive filters [4] is first used to remove linear echoes, then a complex GCCRN network is utilized to eliminate residual echoes and background noise. It seems that a use of traditional filter on top of network is cumbersome. Some studies that only utilize neural networks for the entire process have also achieved good performance. For example, DTLN [5] uses LSTM [6] networks to learn speech features. F-T-LSTM [7] employs a complex encoder-decoder structure, and after extracting the signal features, F-LSTM and T-LSTM are used for time-frequency domain modeling separately, and the final estimation yields the near-end microphone signal. In DeepVQE [8], the alignment module is designed to estimate the time delay of the far-end reference signal and the near-end microphone as a way to align the features of the two signals. Additionally, the Complex convolving mask block (CCM) module is also used for the final output of the estimated speech, and using three-vector components instead of the conventional two-vector components (real and imaginary) provides more stable output results and prevents low noise and echo leakage.

Self-attention mechanisms [9], [10] are good at capturing global information and contextual relationships as a way to enable better modeling. The authors in [11] propose a lightweight Axial Self-Attention (ASA) module to model along the frequency and time axis separately, which improves the network's ability to capture the global internal relationships between time domain features and frequency domain features separately with a smaller number of parameters and operations. Moreover, PCNN [12] proposes a Self Channel-Time-Frequency Attention (Self-CTFA) Module to capture the connections between signal features from the channel, frequency, and time domain dimensions, respectively, and finally performs feature extraction and fusion. In addition, the study reported after ablation experiments that F-attention has a great impact on the model effect. This shows that for processing speech signals, there is a great potential for a multidimensional self-attention mechanism, leveraging cross domain information.

Noting that current approaches tend to use one time-frequency attention for modeling, we propose a diverse time-frequency attention neural network in designing a AEC system. After initially extracting the features of the far-end reference signal and the near-end microphone signal, we estimate the time delay and align them, and then two different time-frequency attention modules (ASA and Frequency Transformation Block (FTB)) to capture the contextual and global information of the features are developed. We observe that multiple temporal-frequency attention outperforms single temporal-frequency attention in learning the contextual relationships between the time domain and the frequency domain simultaneously, which is particularly important for the model to differentiate between the far-end speaker's voice, the near-end speaker's voice, and the background noise. This is also the reason why our model presents an excellent performance in double-talk scenarios.

II. Problem formulation

The microphone signal $y(n)$ is composed of the echo signal $d(n)$, the near-end speech signal $s(n)$, and the background noise $v(n)$, given by

$$y(n) = d(n) + s(n) + v(n), \tag{1}$$

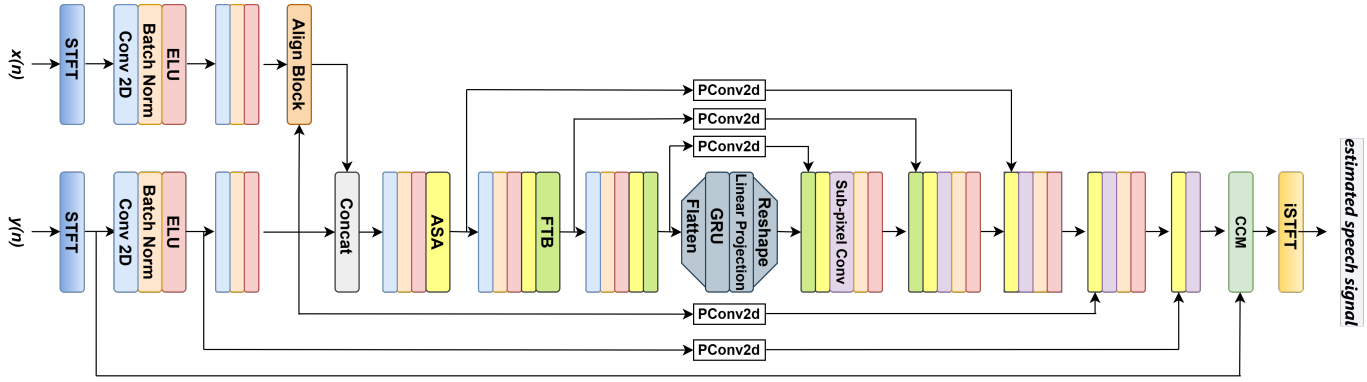


Fig. 1: The proposed DTFAN, where yellow represents ASA and green represents FTB (Best viewed in color).

where n is sample index, $d(n)$ is obtained by convolving the far-end reference signal $x(n)$ with the room impulse response $h(n)$. The task of AEC is to estimate $s(n)$ with $x(n)$ and $y(n)$ provided and the whole process can be formalized as

$$\hat{S}_r, \hat{S}_i = f_\varphi \left(X_r, Y_r, X_i, Y_i \right), \tag{2}$$

where f and φ represent our network and its parameters, and X_r, X_i, Y_r, Y_i are the real and imaginary parts of the complex spectrum obtained by short-time Fourier transform (STFT) of $x(n)$ and $y(n)$. Similarly, $\hat{S}_r$, $\hat{S}_i$ are the real and imaginary parts of the complex spectrum of the estimated clean speech.

A. Network Architecture

The structure of the whole network is depcited in Figure 1. The main body of our model is an encoder and decoder structure with a bottleneck layer. The encoder has two branches to extract features from the far-end reference signal and the near-end speech signal, respectively. After aligning the signal features, the ASA [11] and FTB [13] continue to capture the global information of the features alternatively. After the features passing through the decoder, the real and imaginary parts of the estimated speech spectrum are output by a CCM block [8].

1) Encoder: The encoder consists of a near-end mic branch and a far-end reference signal branch with five and two encoding blocks, respectively. The first two encoding blocks of the two branches are used to initially extract the signal features and feed the features into the Align Block[8], which learns the intrinsic relationship between the two signal features by estimating the delays between the two signals, and then feeds the features into the microphone branch after alignment. The purpose of doing so is that after alignment the network can better cancel out the echos. After that, three encoders further extract the features. The encoders consist of a convolutional block (all have), ASA (last three have), and FTB (last two have), respectively. The convolution block consists of a down-sampled convolutional layer, a batch normalization (BN) layer, and an ELU function. The downsampling convolution layer has a convolution kernel size of 4×3 and a step size of 1×2 to reduce the number of bins along the frequency values. In addition, the convolution is padded to maintain causality.

2) Bottleneck: In the bottleneck layer, we first expand the feature tensor $\mathbf{X} \in \mathbb{R}^{b \times c \times t \times f}$ along the channel and frequency axes as $\mathbf{X} \in \mathbb{R}^{b \times t \times (c \cdot f)}$, where b, c, t, $f \in \mathbb{N}$ denote the batch size, the lengths of the channel, time, and frequency axes. After that, $\mathbf{X}$ is fed into the GRU layer for contextual modeling and then into the linear projection, and the resulting tensor is reshaped back to its original shape. To reduce the number of hidden units, we added a linear injection after the recurrent layer, which additionally improves the training stability and model performance.

3) Decoder: The network contains five decoders and each decoder consists of FTB (first two have), ASA (all have), sub-pixel convolution block [14] (all have). Since the features are down-sampled in the encoder part and the length of the feature tensor is decreasing along the frequency axis, the decoder part needs to be up-sampled. As opposed regular up-scaling method based on transposed convolution, we use sub-pixel convolution instead. For $\mathbf{X} \in \mathbb{R}^{b \times c \times t \times f}$, it is transformed to $\mathbf{X} \in \mathbb{R}^{b \times 2c \times t \times f}$ by a regular convolution with $2c$ filters, and then reshaped to produce $\mathbf{Y} \in \mathbb{R}^{b \times c \times t \times 2f}$. This is done by learning a series of filters to up-sample the feature map. A sub-pixel convolution block consists of a sub-pixel convolution layer, a BN layer, and an ELU function stacked together. The last sub-pixel convolution block has no BN layer and no ELU function. In addition, we use point-wise convolution-based skip connection instead of summing or concatenation. The encoder features are point-wise projected and added to the corresponding decoder outputs. In this way, the number of channels in the encoder can be chosen more flexibly. At the end of the decoder, the CCM module processes the last obtained feature map to reconstruct the spectrum of the estimated clean speech.

4) ASA and FTB: To reduce computation and memory and to better handle long sequence signals represented by speech signals, we leverage the modeling ability of ASA and FTB. The structure of ASA is shown in Figure 2, where Ci denotes the number of input feature channels and C denotes the number of attention channels. ASA computes the attention scores sequentially along the frequency and time axes, which can be abbreviated as F-attention and T-attention. These at-

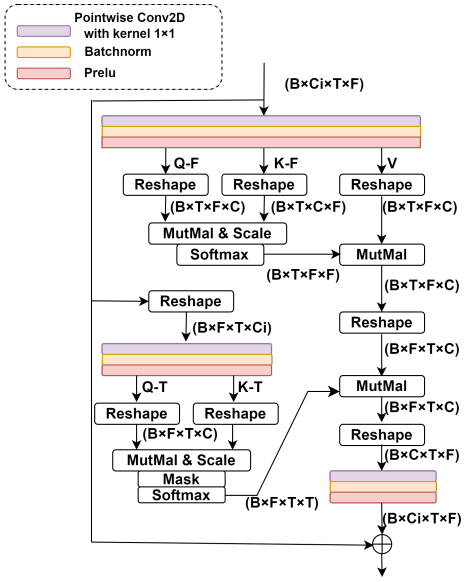


Fig. 2: Axial Self-attention (ASA).

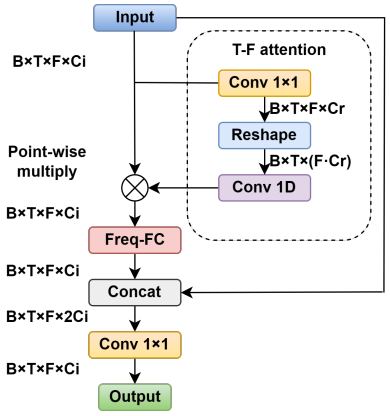


Fig. 3: Frequency Transformation Block (FTB).

tention matrices help the model capture potential relationships between signal feature long distances in both the frequency and time domains. In our network, the ASA module is applied to the last three decoders and all encoders.

In addition, consider fact that ASA is a purely convolutional module, to further diversify the utilization of time-frequency attention mechanisms, we introduce the FTB module into the network. We observe that acquiring time-frequency attention through multiple methods is more conducive to learn contextual information in the time and frequency domains. The structure of FTB is shown in Figure 3, where C and Cr are the number of input channels and the number of attention channels, respectively, and it is worth noting that Freq-FC is a key part of FTB. It contains a trainable frequency transformation matrix for feature map slicing at each time point, which is also an important difference from ASA. FTB solves the problem that 2D convolutional has a small perceptual field. Therefore, we embed FTB into the last two encoders and the first two decoders to produce the feature output of a frequency-aware field to further attend different frequencies.

B. Loss Function

We employ a joint loss function as the optimization objective. Specifically, we use the mean square error (MSE) loss in the frequency domain and the scale-invariant signal-to-noise ratio (SI-SNR) [15] as the loss function in the time domain. The final loss function is the summation of them, given by

$$Loss = -SI - SNR + \log_{10}(MSE(S, \hat{S})), \qquad (3)$$

where S, $\hat{S}$ are the spectrum of the target speech and the estimated clean speech.

III. Experiments

A. Dataset

To train our model, four types of signals need to be prepared: the near-end speech signal, background noise signal, far-end speech signal, and the corresponding echo signal.

We use the synthesized dataset provided by the AEC Challenge [16] for the far-end speech signals and the corresponding echo signals. This dataset consists of 10,000 10-second clips. The first 500 of these clips were used to create the test set and the last 9500 clips were used to create the training set.

For the near-end speech signal and background noise, we used the clean speech signal and background noise signal provided by DNS challenge [17]. Approximately 6.8 hours of clean speech and 5.5 hours of background noise signals were used to create test sets. In addition, we cleaned the background noise using the open-source model sliro-vad [18], cropping out potential segments containing human voices. Finally, we produced a 300-hour dataset for training and validation of the model, with a 4:1 ratio of the training set to the validation set and 10 seconds per audio clip. To improve the model's generalization ability, the signal-to-return ratio (SER) was randomly set from -15 dB to 15 dB and the SNR was randomly set from 5 dB to 25 dB.

B. Experimental setup

All audio sampling rates are 16 kHz. For the network, the STFT window length is 512 points, the window shift is 256, the FFT size is 512, and the Hanning window is used. Our model is trained using Adam's optimizer with an initial learning rate of 1e-4 for 150 epochs. If there is no improvement in performance for 3 epochs, the learning rate is halved.

For the encoder, the down-sampled convolution blocks for the far-end reference signal branch have 64 and 128 filters, respectively, while the number of filters for the near-end microphone signal branch is 32, 128, 128, 128, and 128. The number of filters in the Sub-pixel convolution in the decoder is 128, 128, 128, 64, and 27. The number of attention channels in all ASA modules is one-fourth of the number of input channels. The attention channel of the two FTBs closer to the

TABLE I: Ablation experiment. DT: doubletalk, ST: single-talk, FE: far-end, NE: near-end, T-att: T-attention, F-att: F-attention.

				DT-Noisy		DT-Clean		ST-FE	ST-NE	
CASE	T-att	F-att	FTB	PESQ	STOI	PESQ	STOI	ERLE	PESQ	STOI
1	✓	✓	✓	**2.46**	**0.86**	**2.72**	**0.89**	**46**	**3.27**	**0.95**
2	×	✓	✓	2.29	0.83	2.56	0.87	42	3.10	0.94
3	✓	×	✓	2.33	0.84	2.57	0.87	39	3.15	0.94
4	×	×	✓	2.14	0.81	2.39	0.85	42	3.04	0.94
5	✓	✓	×	2.19	0.82	2.45	0.86	24	3.15	0.94

bottleneck layer is set to 8, and in the remaining two FTBs, the attention channel is set to 15. The hidden unit of the GRU in the bottleneck layer is 320.

C. Performance metrics

We use perceptual evaluation of speech quality (PESQ) [19] and short-time objective intelligibility (STOI) [20] to evaluate the model's performance in double-talk and near-end single-talk scenarios. In addition, we use the echo return loss enhancement (ERLE) [21] to measure the model's performance in far-end single-talk scenarios. Finally, the AEC challenge provides subjective evaluation results based on the average P.808 mean opinion scores [22].

IV. Results and Analysis

To validate the role of diverse attention in the network, we first conducted ablation experiments for five different cases, and the results are shown in Table I. We observe that the model performance improves substantially in all scenarios after adding each type of attention. Notably, when CASE 2 is compared to CASE 3, T-attention is found to have a greater impact on performance than F-attention. We believe this is due to the down-sampling of features by the model in the encoder stage, which leads to a reduction in the feature dimensions along the frequency axis and objectively reduces the F-attention perceptual field. The impact is also highlighted when comparing CASE 4, CASE 5 with CASE 2 and CASE 3. Comparing Case 4 and Case 5, we can see that ASA and FTB have different impacts on performance in different scenarios. The reason is the FTB focuses more on modeling along the frequency axis, while the ASA models along both the time and frequency axes.

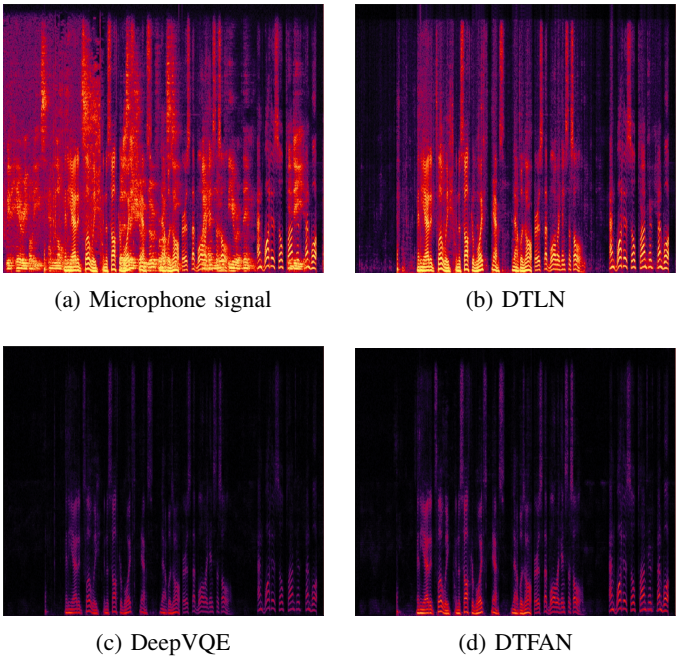
TABLE II: Objective performance metrics on the synthetic test set.

		DT-Noisy		DT-Clean		ST-FE	ST-NE	
Methed	Para	PESQ	STOI	PESQ	STOI	ERLE	PESQ	STOI
Unprocessed	-	1.66	0.70	1.81	0.73	-	2.64	0.92
Baseline23	-	2.05	0.78	2.42	0.84	39	2.77	0.92
DTLN	10M	2.14	0.81	2.37	0.84	33	3.01	0.94
DeepVQE	6.1M	2.40	0.87	2.63	0.90	**50**	3.14	0.95
Ours	5.1M	**2.64**	**0.88**	**2.88**	**0.91**	48	**3.38**	**0.96**

For objective performance evaluation, we compare our network with three other methods, the open-source baseline model of the 2023 AEC Challenge, DTLN, and DeepVQE, where the DTLN and DeepVQE models are retrained with our dataset, and the number of hidden units of the LSTM in the DTLN is 512. From the results in Table II, our method outperforms the other methods in almost all scenarios, especially in the double-talk scenario.

TABLE III: AECMOS of the blind test set in interspeech2021. DT ECHO means more associated with residual echo, DT Other means more related to other degradation.

Method	DT Echo DMOS	DT Other DMOS	ST FE Echo DMOS	ST NE MOS	Overall
Baseline21	4.04	3.45	3.82	4.18	3.87
Baseline23	4.32	3.93	4.41	4.19	4.21
F-T-LSTM	4.44	3.90	4.44	3.78	4.14
GCCRN	4.36	4.23	4.34	4.26	4.29
DeepVQE	4.57	4.15	**4.53**	4.05	4.32
Ours	**4.63**	**4.24**	4.46	**4.27**	**4.40**



(a) Microphone signal (b) DTLN

(c) DeepVQE (d) DTFAN

Fig. 4: A double-talk utterance in synthetic test set.

We also conducted experiments using the blind test set Interspeech2021 and the results are shown in Table III. We compare with four methods, the Interspeech 2021 AEC Challenge Baseline [16], the ICASSP 2023 AEC Challenge baseline [23], F-T-LSTM [7], GCCRN [3] and DeepVQE[8]. The F-T-LSTM employs a complex neural network to better mine the phase information for modeling. At the same time, the F-LSTM and T-LSTM are set up at the bottleneck layer to scan the time and frequency axes, respectively, for

temporal modeling. The GCCRN first estimates the linear echo using the partitioned block frequency domain least mean square (LMS) algorithm [4]. It thereafter performs residual echo cancellation and noise suppression using the GCCRN network. The GCCRN employs gated convolution instead of ordinary convolution, which learns the features of the channel and different spatial locations as a means of selecting the generating mechanism, employs a partitioned LSTM at the bottleneck layer for temporal modeling, and finally decodes the real and imaginary parts for output. In DeepVQE, a self-attention based alignment module is designed to align the near-end microphone signal features with the reference signal features, and subsequently the signal features are fed into the encoder-decoder structure network to finally obtain the estimated near-end speech. Our model performs well in the double talk scenario, maximizing the quality of the near-end speech while eliminating echoes, producing the highest overall score.

Finally, to visually see the differences, we provided the produced spectrum in synthetic dataset, depicted in Figure 4. Compared to DTLN and DeepVQE, the proposed DTFAN produces more details in harmonics and preserves speech signal better in high frequency. This further showcases the benefits using diverse attention mechanism.

V. Conclusions

In this paper, we observed that diverse time-frequency attention is better than single time-frequency attention in learning the contextual relationship between time and frequency domains, which is important for modeling the distinction between far-end speaker speech, near-end speaker speech, and background noise. Based on that, the proposed DTFAN integrates ASA and FTB to diversify learn global information extraction and modeling. Experiments show that diverse time-frequency attention leads to a superior AEC performance compared to single time-frequency attention.

References

[1] J.-S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[2] B. Farhang-Boroujeny, *Adaptive filters: theory and applications*. John wiley & sons, 2013.

[3] R. Peng, L. Cheng, C. Zheng, and X. Li, "Acoustic echo cancellation using deep complex neural network with nonlinear magnitude compression and phase information.," in *Interspeech*, 2021, pp. 4768–4772.

[4] K. Eneman and M. Moonen, "Iterated partitioned block frequency-domain adaptive filtering for acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 143–158, 2003.

[5] N. L. Westhausen and B. T. Meyer, "Acoustic echo cancellation with the dual-signal transformation lstm network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7138–7142.

[6] L. S.-T. Memory, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 2010.

[7] S. Zhang, Y. Kong, S. Lv, Y. Hu, and L. Xie, "Ft-lstm based complex network for joint acoustic echo cancellation and speech enhancement," *arXiv preprint arXiv:2106.07577*, 2021.

[8] E. Indenbom, N.-C. Ristea, A. Saabas, T. Parnamaa, J. Guzvin, and R. Cutler, "Deepvqe: Real time deep voice quality enhancement for joint acoustic echo cancellation, noise suppression and dereverberation," *arXiv preprint arXiv:2306.03177*, 2023.

[9] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Müller, S. Stüker, and A. Waibel, "Very deep self-attention networks for end-to-end speech recognition," *arXiv preprint arXiv:1904.13377*, 2019.

[10] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7867–7871.

[11] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9122–9126.

[12] X. Xu, W. Tu, and Y. Yang, "Pcnn: A lightweight parallel conformer neural network for efficient monaural speech enhancement," *arXiv preprint arXiv:2307.15251*, 2023.

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

[14] W. Shi, J. Caballero, F. Huszár, *et al.*, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[15] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[16] R. Cutler, A. Saabas, T. Pärnamaa, *et al.*, "Interspeech 2021 acoustic echo cancellation challenge.," in *Interspeech*, 2021, pp. 4748–4752.

[17] C. K. Reddy, H. Dubey, K. Koishida, *et al.*, "Interspeech 2021 deep noise suppression challenge," *arXiv preprint arXiv:2101.01902*, 2021.

[18] S. Team, *Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier*, https://github.com/snakers4/silero-vad, 2021.

[19] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, IEEE, vol. 2, 2001, pp. 749–752.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2010, pp. 4214–4217.

[21] S. Theodoridis and R. Chellappa, *Academic press library in signal processing: Image, video processing and analysis, hardware, audio, acoustic and speech processing*. Academic Press, 2013.

[22] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "Aecmos: A speech quality assessment metric for echo impairment," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 901–905.

[23] R. Cutler, A. Saabas, T. Parnamaa, *et al.*, "Icassp 2022 acoustic echo cancellation challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9107–9111.