2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application

Time-frequency Complex Mask Network for Echo Cancellation and Noise Suppression

1st Ning Sun
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China

2nd Hongqing Liu
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China
hongqingliu@outlook.com

3rd Lu Gan
College of Engineering, Design and Physical Science
Brunel University
London, U.K.
lu.gan@brunel.ac.uk

4th Yu Zhao
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China

5th Zhen Luo
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China

6th Yi Zhou
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*—This work investigates the integration of traditional methods and deep learning technique in acoustic echo cancellation (AEC) application. To that end, the generalized cross correlation (GCC) algorithm is explored to align the far-end signal and the echo in the near-end microphone signal, and the echo is estimated in the near-end microphone signal by a use of adaptive filtering. After that, both the error signal and estimated echo are sent to a time-frequency complex mask neural network (TFCN) to suppress residual echo and environmental noise. In TFCN, the dual channel signal of error and estimated echo are processed by LSTMs in frequency domain, and by concatenating the resulting signal, it is converted back to time-domain. Finally, in time domain, a convolutional network is utilized to produce the final target speech. Experimental results show that the proposed framework is robust to blind test set, and effectively removes echo and noise, and achieves an excellent performance in AECMOS scores. The subjective average score of the proposed method is 4.41, which is 0.54 higher than the INTERSPEECH2021 AEC-Challenge baseline.**

***Index Terms*—Acoustic echo cancellation, noise suppression, GCC, adaptive filter, complex mask**

I. Introduction

Due to the coupling between the microphone and the speaker, acoustic echoes will inevitably be generated in occasions such as hands-free voice calls and remote conferencing systems. Due to different reverberation times of the rooms, the acoustic echo may be very significantly, and the system will not function well in severe cases [1]. Acoustic echo cancellation (AEC) aims to eliminate the echo received in the microphone while minimizing the distortion of the near-end voice. The traditional AEC algorithms usually include echo delay estimation, adaptive filtering, nonlinear post-processing, and other modules, but they excessively suppress the near-end speech in the case of doubletalk, thus requiring a good doubletalk detection (DTD) [2]. In recent years, due to the developments of deep learning, it has gradually been evolved into the entire process of using neural networks to replace traditional AEC, for example, DTLN [3] uses LSTM [4] network to learn voice characteristics, and Y^2-Net FCRN [5] network is developed for echo estimation and residual echo suppression, and F-T-LSTM [6] for echo cancellation directly in the time-frequency complex domain. Recent trends indicate that using traditional algorithms of time delay estimation and linear filtering to remove the linear echo and then using neural networks for residual echo cancellation and noise suppression are promising. In [7], frequency domain adaptive filtering [8] is used before complex GCCRN model. The performance improvement of complex network in the field of speech enhancement is also confirmed in DCCRN [9]. In addition, echo cancellation is more like source separation in nature, where the unwanted far-end echo from the mixed near-end signal needs to be eliminated and the near-end voice is required to be kept. Compared with the traditional AEC algorithms, the deep learning echo cancellation demonstrates a better performance and avoids the doubletalk detection and other processes. With those in mind, in this work, the traditional generalized cross-correlation algorithm (GCC) [10] is utilized to estimate the time delay to align the time delay between the far-end signal and the echo in the near-end microphone signal. To remove linear echo, least mean square (LMS) adaptive

This work is supported by the Natural Science Foundation of Chongqing, China (No. cstc2021jcyj-bshX0206).

979-8-3503-1993-4/22/$31.00 ©2022 IEEE

DOI 10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00336

filtering [11] is developed. After that, a neural network that conducts complex mask operations in frequency domain and exploits the time dependencies in time domain to remove the residual echo is developed. We also experiment three combinations of traditional algorithms and it is found that the combination where GCC algorithm is used for alignment first, and then the LMS adaptive filtering is performed to remove the linear echo, and finally the network is utilized to remove the residual echo, produces the best results.

II. Proposed Method

A. Problem Formulation

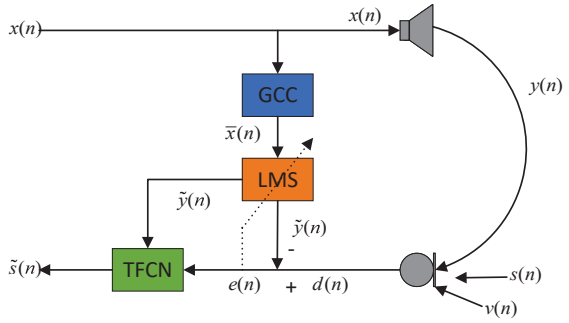


Fig. 1. GLTFCN system framework.

In a typical AEC system, the microphone signal $d(n)$ is composed of near-end speech $s(n)$, acoustic echo $y(n)$, and environmental noise $v(n)$, given by

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

where $y(n)$ is derived from the far-end signal $x(n)$ convolving the room impulse response $h(n)$, and also contains nonlinear distortions caused by the speaker. The task of AEC is to separate $s(n)$ and $y(n)$ on the premise that the reference signal $x(n)$ is known.

As discussed earlier, this work investigates the integration of traditional methods and deep learning, and the overall system framework is depicted in Fig. 1 including GCC delay estimation module, LMS adaptive filter module, a time-frequency complex mask neural network (TFCN) post-processing module. For name conventions, in this work, GTFCN only contains GCC delay estimation part, LTFCN only contains LMS filtering, and GLTFCN contains both traditional algorithms.

B. Time Delay Estimation

In Fig. 1, the far-end signal $x(n)$ is played by the loudspeaker, and then received again by the microphone. There is a natural delay in this process. In practical applications, the frequency domain cross-correlation (CC) between signals is usually used to calculate the time delay, and it is

$$\mathrm{r}_{dx}(m) = d(n) * x^*(-n) = F^{-1}\{D(k,l)X^*(k,l)\}, \tag{2}$$

where $d(n)$ is the near-end microphone signal, $x(n)$ is the far-end signal, $D(k,l)$ and $X(k,l)$ are their short-time Fourier transforms (STFTs), k and l indicate index time frame and frequency bin, respectively, F^{-1} denotes the inverse STFT, and $*$ is the conjugate.

However, the performance of GCC deteriorates due to the low signal-to-noise ratio (SNR). In order to obtain a cross-correlation function with a steeper extreme value, a weighting function is generally used, termed as GCC with phase transform (GCC-PHAT) [10], given by

$$\mathrm{r}_{dx}(m) = F^{-1}\{\frac{D(k,l)X^*(k,l)}{|D(k,l)X^*(k,l)|}\}, \tag{3}$$

$$m_0 = \arg\max_m \mathrm{r}_{dx}(m), \tag{4}$$

where m_0 represents the estimated delay value.

The effect of echo cancellation is strongly related to the delays of the far and near ends, and improper adjustment will bring the risk of unavailability of the algorithm. If the time delay exceeds the estimated range of the linear filter or is adjusted to cause non-causality at the near and far ends, the linear filter might not converge.

C. Adaptive Filter

Adaptive echo cancellation is one of the commonly used traditional methods of echo cancellation. The weight vector of the filter is adjusted by an adaptive algorithm [11] to estimate an expected signal and approximate the echo signal passing through the actual echo path, and then it is subtracted from the mixed signal collected by the microphone to achieve the function of echo cancellation. The choice of adaptive filter plays a very important role in the performance of echo cancellation. The most common linear filtering algorithm is LMS algorithm [11] because of its low complexity and easy implementation. By a use of adaptive filtering to output the echo estimation $\tilde{y}(n)$, filtering out the linear echo produces the error signal $e(n)$, and this error signal serves as the input of the neural network module for post-processing. It is of interest to point out that adaptive filtering also presents a certain delay estimation function, which can partially improve the small range error caused by GCC [10].

D. The TFCN Model

To remove the residual echo, a dual channel time-frequency combined complex mask convolutional neural network is developed, shown in the Fig. 2. The network takes the error signal $e(n)$ and the estimated echo $\tilde{y}(n)$ as inputs and has two modules. The first module uses the frequency domain LSTM [4] network, and the second module uses the dilated convolutional [12] neural network to conduct end-to-end processing. First, the input signals are respectively subjected to STFT to extract complex frequency domain information and then the real parts and the imaginary parts of the inputs are connected as the inputs to the two-layer LSTM [4] network. In Fig. 2, the blue is the real part and the orange represents the imaginary part. After the fully connected layer, the sigmoid activation function obtains the complex mask, which is multiplied by the real part and imaginary part of $e(n)$ in frequency domain to extract the information, and the ISTFT after Concat operation

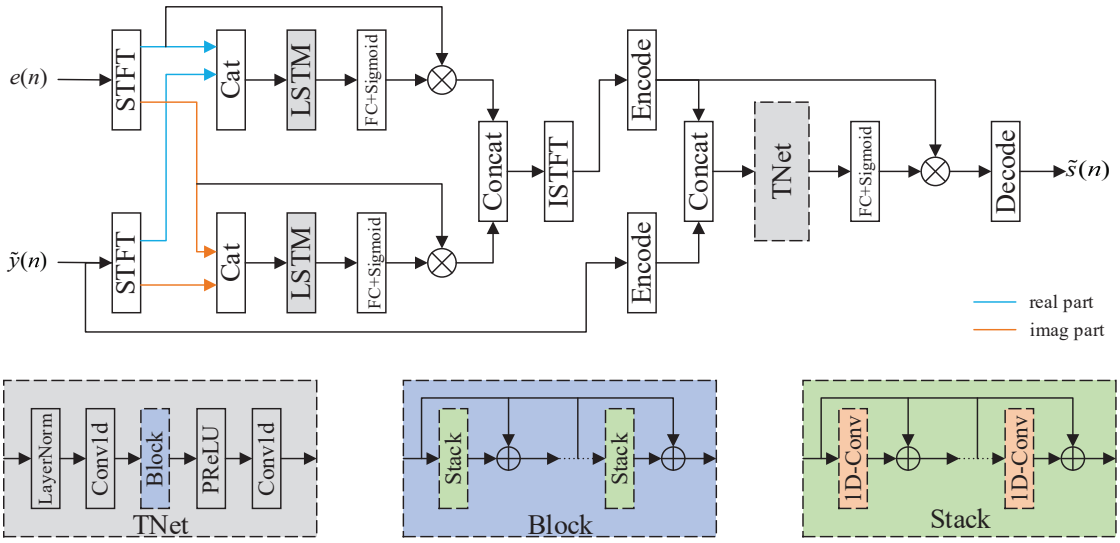


Fig. 2. TFCN structure.

obtains the preliminary near-end speech estimate. For the second module, we perform end-to-end processing. The two inputs are first processed by encoders to obtain time domain information. After concat together, they are sent to the second separation module (TNet), which contains the normalization layer, the Bottleneck layer [13], a block composed of dilated convolution [12] residual network [14], and PReLU and Conv1d layers. After that, the fully connected layer and the sigmoid activation function output mask, which is multiplied with the coding information of the output signal of the first module to further extract the time domain information of the near-end speech. Finally, the output $\tilde{s}(n)$ is obtained by the Decoder.

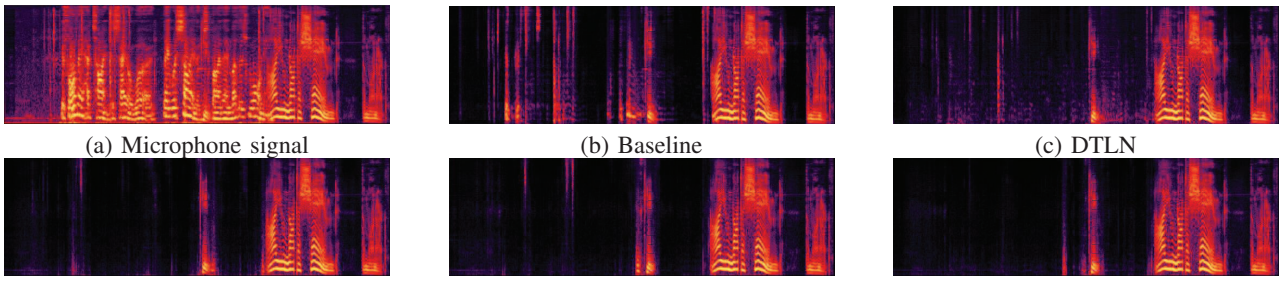
III. Experimental Setup

A. Datasets

AEC-Challenge [15] provides two datasets, which are the collected real recordings and the synthetic dataset. Because the real dataset lacks the clean near-end signals, it cannot be used for model training. In addition, there is automatic gain control (AGC) in the process [15] of generating the synthetic dataset, which may causes the amplitude of the near-end clean signal to be different from the near-end signal in the near-end microphone signal, so we only use the far-end signal and echo signal in the synthetic dataset. For the near-end signal, we have performed data augmentations to expand the datasets by using the reading dataset in DNS-Challenge [16]. We randomly selected and cut voices into 4 to 7 seconds, and randomly added zeros to make them 10 seconds long, as our near-end signal $s(n)$. For the far-end signal $x(n)$ and the echo signal $y(n)$, we use the synthetic dataset to randomly select the corresponding signal and cut them at the same starting positions for 6 to 8 seconds, and pad zeros to 10 seconds. The signal-to-echo ratio (SER) is randomly set from -5 dB to 20 dB. In order to improve the anti-noise performance of the model, we add the noise from the dataset in DNS-Challenge [16] and we randomly cut 4 to 7 seconds and add zeros randomly before and after as additional noise signal $v(n)$ to generate the near-end microphone signal $d(n)$. The signal-to-noise ratio (SNR) is in the range of 5 dB to 25 dB. In addition, 20% of the near-end signal has no background noise and 20% of far-end and echo is zero. Finally, we generate a total of 100 hours of data to train and verify our model. The ratio of training set to validation set is 4:1.

B. Training Setup

The LMS module sets the adaptive filter order to 8 and the update step size to 0.02, and loops once to output the echo estimate and the error signal obtained by subtracting the estimated echo. For the network, the STFT window length is 400 point, the window shift is 100, the FFT size is 512 and the Hanning window is used. For lstm, we set the input size and hidden size are 512 and the number of layers is 2. The kernel size of the encoder is set to 40, and the number of output channels is 256. The kernel size of the decoder is set to 40, and the number of input channels is 256. For TNet, as in [17], we use global normalization [18]. There are 3 layers of stacks inside the block, and each layer of stack contains 8 D-Conv, The dilation setting is from 2^0 to the 2^7, the kernel size is set to 3, and each D-Conv inside the stack and each stack are connected by a residual network [14]. Our model is trained with the Adam optimizer for 100 epochs with an initial learning rate of 2e-4, and the learning rate needs to be halved if there is no improvement for three epochs, and SI-SNR as loss function is utilized. The sampling rate of all modules is 16 kHz.



(a) Microphone signal (b) Baseline (c) DTLN

(d) GTFCN (e) LTFCN (f) GLTFCN

Fig. 3. Comparisons of different models on synthetic test set.

IV. Results

A. Performance Metrics

We use perceptual evaluation of speech quality (PESQ) [19] and short-time objective intelligibility (STOI) [20] to measure the performance of the model in the double-talk scenario. For the far-end single-talk scenario, we use the echo return loss enhancement (ERLE) [21] score to evaluate the echo cancellation capability. At the same time, the AEC-challenge also provides subjective evaluation results based on the average P.808 mean opinion score (MOS) [22].

B. Performance Evaluation - Synthetic Test Set

TABLE I
THE PESQS AND STQIS IN THE CASE OF NOISY DOUBLE-TALK AND ERLES IN THE CASE OF CLEAN FAR-END SINGLE TALK.

	DT-Noisy		FE-Clean
Method	PESQ	STOI	ERLE
Original	1.90	0.79	-
Baseline	1.98	0.82	31.31
DTLN	2.52	0.55	32.69
GTFCN	**2.63**	0.90	45.85
LTFCN	2.48	0.88	45.55
GLTFCN	2.61	**0.90**	**46.19**

In this part, the test set is generated in the same way as the training set, but the test and training sets do not overlap. For the test set, the SER is randomly generated from -10 dB to 15 dB, and the SNR is randomly generated from 0 dB to 20 dB, and the test set is 10 hours. In Table I, the PESQ, STOI and ERLE scores are summarized for different models. In this case, the DTLN [3] is also provided for comparison. It is seen that all the models improves the performance compared with baseline [23]. However, the proposed method outperforms DTLN. To visually see the performance, Fig. 3 shows the effects of different models on the test set. It can be clearly seen that our model better eliminates residual echo and suppresses the environmental noise.

C. Performance Evaluation - Blind Test Set

We now use the blind test sets provided by both AEC-challenge and Interspeech2021 to test our model performance and compare it with the state-of-the-art methods.

TABLE II
AECMOS OF BLIND TEST SET IN AEC-CHALLENGE.

Method	DT Echo DMOS	DT Other DMOS	ST FE Echo DMOS	ST NE MOS	Overall
CDEC	3.78	3.37	4.28	3.75	3.80
GTFCN	4.28	3.74	4.38	3.58	4.00
LTFCN	4.30	3.68	4.48	3.70	4.04
GLTFCN	**4.33**	**3.75**	**4.52**	**3.82**	**4.11**

In Table II, the MOSs are provided for different methods on the blind test set provided by AEC-challenge. It can be seen that our model far exceeds CDEC [24] in overall performance, especially in the DT and FE scenarios. It is also found that the best performance is obtained by the method of first passing GCC [10] time delay estimation, then passing LMS adaptive filtering [11] for echo estimation, and finally sending it to the network for learning.

TABLE III
AECMOS OF BLIND TEST SET IN INTERSPEECH2021.

Method	DT Echo DMOS	DT Other DMOS	ST FE Echo DMOS	ST NE MOS	Overall
Baseline	4.04	3.45	3.82	4.18	3.87
GCCRN	**4.36**	4.23	4.34	4.26	4.30
GTFCN	4.27	4.07	4.35	4.19	4.22
LTFCN	4.30	4.12	4.45	4.39	4.32
GLTFCN	4.35	**4.23**	**4.59**	**4.47**	**4.41**

Finally, the experiment is performed using blind test set Interspeech2021 and the results are provided in Table III. Compared with the baseline, the performance has been improved significantly. Compared with the GCCRN [7], which uses the partitioned block frequency-domain LMS algorithm [8] for echo estimation, GCCRN for residual echo cancellation and noise suppression, the proposed method provides a superior performance on average. The similar conclusion can be drawn that the GLTFCN produces the best performance overall by using both GCC and LMS as traditional methods.

V. Conclusion

This paper proposes a joint framework for echo cancellation and noise suppression, which is divided into three modules. First, GCC delay estimation module for far-end and echo

alignment is performed. Second, LMS adaptive filtering is conducted to output echo estimation and to remove linear echo. Finally, the cascaded TFCN is used for residual echo cancellation and noise suppression. In the AECMOS score, it is 0.54 higher than the baseline model, which demonstrates a strong robustness and superior performance.

REFERENCES

[1] H. Kuttruff, *Room Acoustics, Fifth Edition*. Taylor & Francis, 2009. [Online]. Available: https://books.google.com/books?id=X4BJ9ImKYOsC

[2] J. Benesty, D. Morgan, and J. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 168–172, 2000.

[3] 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)*, 2021, pp. 7138–7142.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[5] E. Seidel, J. Franzen, M. Strake, and T. Fingscheidt, "Y2-Net FCRN for Acoustic Echo and Noise Suppression," in *Proc. Interspeech 2021*, 2021, pp. 4763–4767.

[6] 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.

[7] 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 *Proc. Interspeech 2021*, 2021, pp. 4768–4772.

[8] 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.

[9] 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," 2020. [Online]. Available: https://arxiv.org/abs/2008.00264

[10] M. Cobos, F. Antonacci, L. Comanducci, and A. Sarti, "Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1270–1281, 2020.

[11] B. Farhang-Boroujeny, *Adaptive Filters: Theory and Applications*. Wiley, 2013. [Online]. Available: https://books.google.com/books?id=Fmf8TumgxEYC

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015. [Online]. Available: https://arxiv.org/abs/1511.07122

[13] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *INTERSPEECH*, 2011.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[15] R. Cutler, A. Saabas, T. Parnamaa, M. Loide, S. Sootla, M. Purin, H. Gamper, S. Braun, K. Sorensen, R. Aichner, and S. Srinivasan, "Interspeech 2021 acoustic echo cancellation challenge: Datasets and testing framework," in *INTERSPEECH 2021*, 2021.

[16] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6623–6627, 2021.

[17] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[18] H. Adel and H. Schütze, "Global normalization of convolutional neural networks for joint entity and relation classification," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1723–1729. [Online]. Available: https://aclanthology.org/D17-1181

[19] A. Rix, J. 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[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," *arXiv preprint arXiv:2110.03010*, 2021.

[23] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "Icassp 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," 2020. [Online]. Available: https://arxiv.org/abs/2009.04972

[24] L. Pfeifenberger, M. Zoehrer, and F. Pernkopf, "Acoustic Echo Cancellation with Cross-Domain Learning," in *Proc. Interspeech 2021*, 2021, pp. 4753–4757.