Exploring the Interactions between Target Positive and Negative Information for Acoustic Echo Cancellation

Chang Han[1,†], Xinmeng Xu[1,†], Weiping Tu[1,2,3,✉], Yuhong Yang[1,3], Yajie Liu[1]

[1]NERCMS, School of Computer Science, Wuhan University, China
[2]Hubei Luojia Laboratory, China
[3]Hubei Key Laboratory of Multimedia and Network Communication Engineering,
Wuhan University, China

{changhan, xuxinmeng, tuweiping, yangyuhong, yajieliu07} @whu.edu.cn

Abstract

Acoustic echo cancellation (AEC) aims to remove interference signals while leaving near-end speech least distorted. As the indistinguishable patterns between near-end speech and interference signals, near-end speech can't be separated completely, causing speech distortion and interference signals residual. We observe that besides target positive information, e.g., ground-truth speech and features, the target negative information, such as interference signals and features, helps make pattern of target speech and interference signals more discriminative. Therefore, we present a novel AEC model encoder-decoder architecture with the guidance of negative information termed as CMNet. A collaboration module (CM) is designed to establish the correlation between the target positive and negative information in a learnable manner via three blocks: target positive, target negative, and interactive block. Experimental results demonstrate our CMNet achieves superior performance than recent methods.

Index Terms: acoustic echo cancellation, encoder-decoder architecture, target positive and negative information

1. Introduction

Acoustic echo arises in a full-duplex voice communication system when a near-end microphone picks up audio signals from a near-end loudspeaker and sends it back to far-end participants such that the far-end users receive a modified version of their voice. Acoustic echo cancellation (AEC) aims to suppress the echo from the microphone signal while leaving the target near-end signal least distorted.

The traditional AEC applies adaptive filters [1, 2] to identify the echo path between the loudspeaker and the microphone at the near-end point. The main assumption that underpins conventional AECs is that the echo path is linear [3]. However, the power amplifiers and loudspeakers, incredibly cheap and small, can be the sources of nonlinearity. When nonlinearities occur, conventional AECs performance decreases the overall achievable quality. Although several nonlinear models [4, 5, 6] have been utilized, these traditional methods for convergence may take a lot of work in complex acoustic environments.

Deep learning (DL) has recently gained much attention for its capacity to model complicated nonlinear relationships

---

and has been successfully applied to AEC. DL-based AEC can be formulated as a supervised speech separation/enhancement problem [7, 8, 9, 10], which separates near-end speech from the microphone signal and has shown better performance than their classical counterparts [11]. However, DL-based methods often suffer from near-end speech distortion and interference signals (noise and echo) residual, which suggests some indistinguishable patterns between target speech and interference signals are hard to distinguish.

To alleviate this challenge, some approaches introduce target negative information and capture its correlation with target positive information to make the patterns between target and interference signals more discriminative. However, the definition of target positive and negative information is diverse and can be concretely grouped into two categories. The methods of the first category [12, 13, 14] aim to predict near-end speech by adding the characteristics of echo signals that are used as target negative information, in which a separate algorithm is set to estimate the echo signal before the target signal prediction. These approaches generally consider the echo signal by incorporating it into loss functions or by directly predicting it as prior knowledge. However, it is difficult to accurately estimate echo signals when faced with strong nonlinear distortion or excessively high noise levels [15], which may introduce interfering information for AEC.

The methods of the second category [16] extract features of near-end speech as target positive information and features of interference signals as negative information by utilizing a two-parallel branch network to model target and interference signals separately. Subsequently, cross-connections between two branches are employed to use the information learned from the other branch to improve the target signal modeling. As a result, the target positive and negative information have interacted in a latent representation space, and the interaction makes the simultaneous modeling of two signals feasible and effective. Although the two-branch approach ensures the preservation of both target speech and interference signal information during the modeling process, cross-connections pass information without any selection resulting in some redundant information. Additionally, the two-branch structure requires many parameters, resulting in high computational costs.

In our study, features whose weights are high and regarded as helpful information to predict the target signal by the neural network are considered as target positive information. In contrast, features whose weight is low and regarded as interference information by the network are considered as target negative information. The aforementioned analysis motivates us to design a new approach with fewer parameters to explore the correlation between target positive and negative information to make the patterns between near-end speech and interference signals
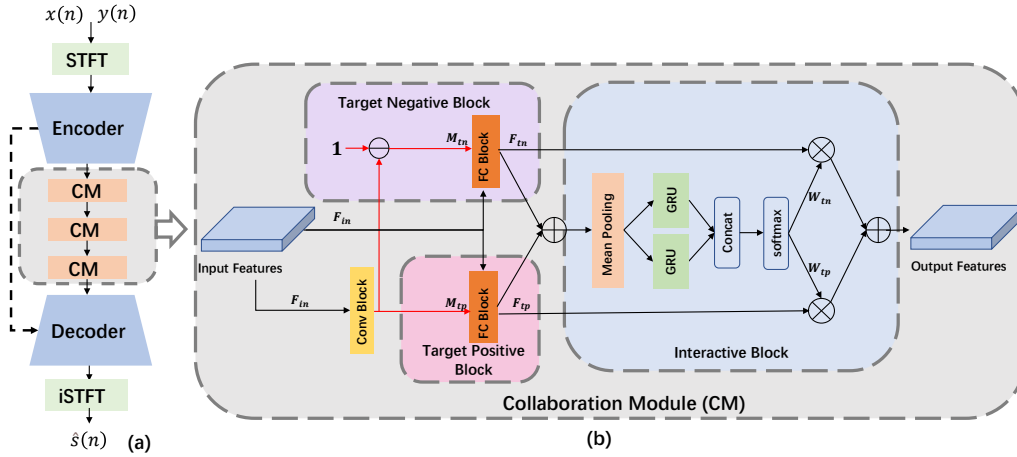
Figure 1: *(a) The proposed CMNet. (b) The proposed collaborative module. "FC" denotes feature catcher block.*

more discriminative. We propose a collaboration module (CM) that includes a target positive block, a target negative block, and an interactive block. The target positive/negative block captures the target positive/negative features at local-level from the global-level features. The interactive block integrates the captured features in a self-adaptive and learnable way to establish a correlation between target positive and negative features. Only a single branch structure is required to build the network. In summary, we propose a novel AEC model, CMNet, which inserts CM into the encoder-decoder architecture to achieve superior performance.

The rest of this paper is organized as follows: In section 2, the problem of the AEC is briefly define. Then the model architecture is presented in Section 3. Section 4 is the dataset and experimental settings. Section 5 demonstrates the results and analysis, and a conclusion is shown in Section 6.

2. Problem formulation

The microphone signal $y(n)$ is a mixture of echo $d(n)$, near-end speech $s(n)$, and background noise $v(n)$:

$$y(n) = d(n) + s(n) + v(n) \qquad (1)$$

where n is sample index, $d(n)$ is obtained by a linear or non-linear transform of the far-end signal $x(n)$. Provided that $x(n)$ and $y(n)$ are known, the task of AEC is to estimate near-end signal $\hat{s}(n)$. A time delay compensation module[17] based on the generalized cross-correlation phase transform method is used to align the microphone and far-end signal. Our overall model can be formulized as:

$$\hat{M} = f_\varphi(X_r, Y_r, X_i, Y_i) \qquad (2)$$

where f and φ denotes CMNet and its network parameters, X and Y denote $x(n)$ and $y(n)$ after short-time Fourier transform (STFT) respectively, r and i represent real and imaginary parts of complex spectrogram, $\hat{M}$ is the estimated complex ratio mask (CRM)[18] optimized by signal approximation, which can be defined as:

$$\mathrm{CRM} = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + j \frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2} \qquad (3)$$

The final predicted mask of the network $\hat{M} = M_r + jM_i$ can be defined in polar coordinates

$$\begin{cases} M_{\mathrm{mag}} = \sqrt{M_r^2 + M_i^2} \\ M_{\mathrm{phase}} = \arctan 2(M_i, M_r) \end{cases}$$

and the estimated clean speech $\hat{s}$ can be expressed:

$$\hat{s} = Y_{\mathrm{mag}} \cdot M_{\mathrm{mag}} \cdot e^{Y_{\mathrm{phase}} + M_{\mathrm{phase}}} \qquad (4)$$

3. Proposed algorithm

In our design, the proposed CMNet model consists of an encoder-decoder network to learn and reconstruct target signal and a collaboration module network to extract and combine target positive and negative features. The entire CMNet architecture is shown in Fig.1(a).

3.1. Encoder and Decoder

The encoder contains three 2-D convolutional layers with a kernel size of (3, 5). The stride is (1, 1) for the first layer and (1, 2) for the following two. The channel numbers are [16, 32, 64]. The decoder consists of three gated blocks followed by one 2-D convolutional layer[19]. The first layer in the gated block is a deconvolutional layer, followed by a 2-D convolutional block learning multiplicative mask for the corresponding feature from the encoder. The purpose of this mask is to suppress undesired parts of the feature. Afterward, the masked encoder feature is concatenated with the deconvolutional feature and inputted into another 2-D convolutional layer to produce the residual representation. After three gated blocks, the final convolutional layer learns CRM. The kernel size for all 2-D deconvolutional layers is (3,5). The stride is (1,2) for the first two gated blocks and (1,1) for the last. The channel numbers are 32, 16, and 2, respectively. All the 2-D convolutional layers in the decoder have a kernel size of (1,1), a stride of (1,1), and a channel number the same as that of their deconvolutional layers. All of the convolutional layers in both the encoder and decoder are preceded by a batch normalization (BN) layer and a parametric ReLU (PReLU) activation function. The padding of the convolution layers is set to keep the causality of the whole system.

3.2. Collaboration Module

Our work presents a crucial idea that target positive and negative information is valuable for AEC to distinguish between target and interference signals. But using these two types of information indiscriminately may result in passive effects. Therefore, we propose a collaborative module (CM) to model and combine target positive and negative features to obtain more discriminative information for AEC. Note that this combination
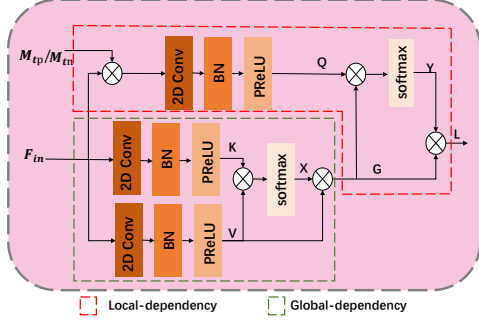
Figure 2: *The structure of feature catcher (FC) block*

is not a simple series or parallel connection but is learnable and self-adaptive.

The proposed CM consists of three parts: target positive block, target negative block, and interactive block. As demonstrated in Figure1(b), given input features F_{in}, a 2D convolutional block is applied to produce the feature map M_{tp}, in which the region of target positive features are highlighted, and can be expressed as the following:

$$\mathbf{M}_{tp} = \text{Sigmoid}\left(\text{Conv}\left(\text{ReLU}\left(\text{Conv}\left(\mathbf{F}_{in}\right)\right)\right)\right) \quad (5)$$

After that, we feed F_{in} and M_{tp} into a feature catcher (FC) block, which is designed for capturing local information from global-level features to extract target positive features F_{tp}. We utilize one minus M_{tp} to obtain the feature map M_{tn} for target negative features. Similarly, we feed F_{in} and M_{tn} into a FC block to extract target negative features F_{tn}. Inspired Li et al.[20], the interactive block is proposed to adaptively integrate F_{tp} and F_{tn}. In detail, we first merge F_{tp} and F_{tn} from two branches through an element-wise addition. Next, mean pooling is utilized to generate global features that guide the adaptive selection between target positive and negative features. After that, two gate recurrent units (GRU) are used to generate two weight vectors w_{tp} and w_{tn} for channel-wise selection between two features. w_{tp} and w_{tn} are content-aware, hence learnable and self-adaptive. The interactive block can be expressed as the following steps, $\mathcal{G}^1$ and $\mathcal{G}^2$ denote two independent GRUs:

$$\begin{cases} F_{gf} = MeanPool\left(F_{tn} + F_{tp}\right) \\ w_{tp}, w_{tn} = \text{softmax}\left(\left[\mathcal{G}^1\left(F_{gf}\right), \mathcal{G}^2\left(F_{gf}\right)\right]\right) \\ F_{\text{out}} = W_{tp} \times F_{tp} + W_{tn} \times F_{tn} \end{cases} \quad (6)$$

Inspired by Xu et al.[21], we propose a FC block to capture speech information in global-level and local-level. As shown in Figure2, we divide the process of computing FC block into two parts: global-dependency and local-dependency. We first capture global dependency by self-attention mechanism. Afterward, we capture the local dependency with the help of global-level features. In global-dependency part, given an input feature F_{in}, two 2-D convolutional layers are adopted to generate new feature maps, $\mathbf{K}$ and $\mathbf{V}$. The global-dependency part follows the non-local operation[22] to compute responses based on relationships between different TF units of the spectrogram. Consequently, the non-local attention map $\mathbf{X}$ can be efficiently calculated by dot product:

$$\mathbf{X} = \text{softmax}\left(\mathbf{K}\mathbf{V}^{\top}\right) \quad (7)$$

The output of global-dependency part $\mathbf{G}$ is defined as:

$$\mathbf{G} = \mathbf{V}\mathbf{X}^{\top} \quad (8)$$

Previously, we extracted feature map M_{tp} at the local level using a convolutional layer with restricted receptive field, which make network pay more attention to target positive features. In local-dependency part, the output of global dependency part $\mathbf{G}$ and the feature map after the local process $\mathbf{Q}$ are treated as the inputs.

$$\mathbf{Y} = \text{softmax}\left(\mathbf{Q}\mathbf{G}^{\top}\right) \quad (9)$$

Finally, the output of local-dependency part is computed by the metrics multiplication between $\mathbf{G}$ and $\mathbf{Y}$.

3.3. Training objectives

We utilize SI-SNR [23] as our loss function:

$$\boldsymbol{s}_{target} = \frac{< \hat{\boldsymbol{s}}, \boldsymbol{s} > \boldsymbol{s}}{||\boldsymbol{s}||^2} \quad (10)$$

$$\boldsymbol{e}_{noise} = \hat{\boldsymbol{s}} - \boldsymbol{s}_{target} \quad (11)$$

$$\text{L}_{\text{SISNR}} = 10\log_{10}\frac{||\boldsymbol{s}_{target}||^2}{||\boldsymbol{e}_{noise}||^2} \quad (12)$$

where s and $\hat{s}$ are the clean and estimated time-domain waveform, respectively. $<,>$ denotes the dot product between two vectors and $||.||_2$ is L2 norm.

4. Experimental setup

4.1. Dataset

Our model is trained with 9500 synthetic files from the database provided by Microsoft for the ICASSP 2022 AEC Challenge[24] that is an open speech corpus. Besides, we also have performed data expansion.

For data expansion, we first prepare four types of signals: near-end speech, background noise, far-end signal and corresponding echo signal. For near-end speech $s(n)$, there are 10,000 near-end speech utterances in the 2022 official synthetic dataset and we select the first 500 utterances as the test set which is unseen in training. The rest 9,500 utterances, together with 10,000 utterances from 2021 ICASSP AEC-challenge synthetic dataset are used for training. For background noise $v(n)$, we randomly select 5000 pieces of noise audio from the DNS[25] dataset for training and 1000 pieces of noise audio for testing. For far-end signal $x(n)$ and echo signal $d(n)$, the first 500 sentences of the 2022 official synthetic dataset are used as the test set and the rest 9,500 utterances for training. In addition, we also use the real far-end single-talk utterances provided by the 2021 and 2022 AEC challenge, which covers a variety of recording devices and signal time delay. And then we combine these four signals together to get microphone signal.

4.2. Implementation details

All audio signals are resampled to 16kHz. The chunk size of our training data is set to 10s. The proposed model uses STFT to extract the spectrum from each utterance. A Hamming window with 512 bins and overlap interval of 256 bins is used. Our model is trained with the Adam optimizer with an initial learning rate of 1e-3. For CM, The filter sizes and strides of convolutional layer in the Conv Block of CM are (3, 7) and (1, 1), and in the FC of CM, they are (1, 1) and (1, 1) in time and frequency dimension. We use masks for self-attention and set the GRU to unidirectional in FC to avoid involving future information. The number of hidden units in GRU is 1. The whole parameters of CMNet are 2.5M.

				DT				ST_NE		ST_FE
	SNR(in dB)			-		5		5		-
CASE	TPB	TNB	IB	PESQ	STOI	PESE	STOI	PESQ	STOI	ERLE
1	✓	✓	✓	1.96	0.88	1.80	0.86	2.06	0.91	34
2	✓	✗	✗	1.94	0.88	1.69	0.84	2.01	0.89	32
3	✗	✓	✗	1.90	0.88	1.65	0.84	1.94	0.89	34
4	✓	✓	✗	1.90	0.88	1.68	0.85	2.00	0.90	21
5	✗	✗	✗	1.87	0.87	1.65	0.82	2.03	0.84	23

4.3. Evaluation metrics

The following three metrics are used to evaluate our model and state-of-the-art competitors. All metrics are better if higher.

- PESQ: Perceptual evaluation of speech quality (from -0.5 to 4.5) [26].
- STOI: Short-time objective intelligibility measure (from 0 to 1) [27].
- ERLE: Echo return loss enhancement for far-end single-talk periods [28], which is defined as:

$$\text{ERLE} = 10 \log_{10} \left[\sum_n y^2(n) / \sum_n \hat{s}^2(n) \right] \quad (13)$$

PESQ and STOI are used for double-talk and near-end single-talk scenarios and ERLE used for far-end single-talk.

5. Results and analysis

We demonstrate the ablation study in Table1 to investigate the effect of different components in CMNet. Note that **Case 1** denotes the basic CMNet with the default setting. In the ablation study, we compare our method with several baseline models: **Case 2**, we only keep target positive block of CM. **Case 3**, we only keep target negative block of CM. **Case 4**, we remove interactive block in CM. **Case 5**, we remove whole CM and add self-attention layers to keep the number of parameters roughly constant. A detailed analysis of the ablation study is presented below.

Table 2: *Echo cancellation performance. SNR is randomly picked up from 5dB or without additional noise in DT and is set to 5dB in NE_ST. The DT scenario randomly picks SER from [-15,15]dB.*

		DT		ST_NE		ST_FE
Method	#Params(M)	PESQ	STOI	PESQ	STOI	ERLE
CRN	4	1.53	0.83	1.71	0.88	23
F-T-LSTM	1.2	1.67	0.85	1.82	0.89	23
MTFAA	2.1	1.79	**0.88**	2.00	**0.91**	25
two-branch	4.8	1.74	0.87	1.97	0.90	28
CMNet	2.5	**1.88**	0.88	**2.06**	**0.91**	**34**

To verify the effect of target negative information, we make the comparison between Case 1 and Case 2. 0.02 STOI gains and 0.11 PESQ gains by Case 1 over Case 2 at a signal-to-noise ratio of 5 dB show the superiority of the use of target negative, especially in a noisy environment. In Case 3, we only apply target negative block. The performance comparison between Case 3 and Case 5 indicates the necessity of involving target negative features for the AEC model. Interestingly, comparing Case 2 with Case 3, we observe that CMNet with only target positive blocks performs slightly better than with only target negative blocks. Figure3 shows the attention feature maps of target positive and target negative blocks. From (a) and (b), we observe that target positive and target negative blocks behave differently in CM. This is reasonable as the two blocks model different features and their focus differs. And in some indistinguishable patterns between speech and interference signals, the weights in the feature map may be about 0.5, so the feature maps are similar. In Case 4, we only remove interactive block. The performance comparison between Case 1 and Case 4 suggests the importance of feature fusion in a learnable and adaptive manner.
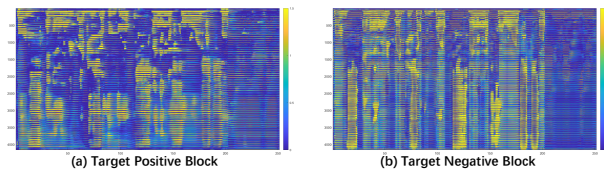

| (a) Target Positive Block | (b) Target Negative Block |

Figure 3: *Visualization of attention feature map from different FC blocks. (a) target positive block. (b) target negative block.*

We compare our CMNet with three other methods, CRN[8], F-T-LSTM[29], and MTFAA[30] trained with our dateset. CRN is a causal convolutional recurrent network for complex spectral mapping. F-T-LSTM adopts an AEC approach using a complex neural network to better model the important phase information and frequency-time-LSTMs, which scan both frequency and time axis, for better temporal modeling. MTFAA is a system that presents a novel backbone for speech dense-prediction called a multi-scale temporal frequency convolutional network with axial self-attention. Table2 shows the comparison results in terms of STOI and PESQ. We highlight the best score under each condition in boldface. Besides, based on our network, a two-branch network is constructed by mimicking the structure in [16], which further validates that our model can achieve better results with fewer parameters.

6. Conclusions

In this paper, we observe that both the target positive information and the target negative information, such as interference signals and features, are valuable in guiding the AEC model training procedure. Therefore, we propose a novel AEC method termed CMNet, encoder-decoder architecture with a collaboration module (CM) inserted. The CM is used to establish the correlation between the target positive and negative information at the global and local levels in a learnable and self-adaptive manner. Experimental results demonstrate that the guidance of both positive and negative information can improve AEC performance.

7. References

[1] H.-C. Shin, A. Sayed, and W.-J. Song, "Variable step-size nlms and affine projection algorithms," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 132–135, 2004.

[2] C. Paleologu, J. Benesty, S. Ciochină, and S. L. Grant, "A kalman filter with individual control factors for echo cancellation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5974–5978.

[3] D. Comminiello, M. Scarpiniti, R. Parisi, and A. Uncini, "A functional link based nonlinear echo canceller exploiting sparsity," in *Proc. Int. Workshop Acoust. Echo, Noise Control (IWAENC'10)*, 2010.

[4] A. Guérin, G. Faucon, and R. Le Bouquin-Jeannès, "Nonlinear acoustic echo cancellation based on volterra filters," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 672–683, 2003.

[5] C. Hofmann, C. Huemmer, and W. Kellermann, "Significance-aware hammerstein group models for nonlinear acoustic echo cancellation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5934–5938.

[6] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-Garcia, and A. Uncini, "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1502–1512, 2013.

[7] H. Zhang and D. Wang, "Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios," in *Proc. Interspeech 2018*, 2018, pp. 3239–3243.

[8] H. Zhang, K. Tan, and D. Wang, "Deep Learning for Joint Acoustic Echo and Noise Cancellation with Nonlinear Distortions," in *Proc. Interspeech 2019*, 2019, pp. 4255–4259.

[9] J.-H. Kim and J.-H. Chang, "Attention Wave-U-Net for Acoustic Echo Cancellation," in *Proc. Interspeech 2020*, 2020, pp. 3969–3973.

[10] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation lstm network for real-time noise suppression," *arXiv preprint arXiv:2005.07551*, 2020.

[11] 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," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 151–155.

[12] A. Fazel, M. El-Khamy, and J. Lee, "Deep multitask acoustic echo cancellation." in *Interspeech*, 2019, pp. 4250–4254.

[13] 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.

[14] J. Li, M. Ge, L. Wang, and J. Dang, "Deep multi-task cascaded acoustic echo cancellation and noise suppression," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2022, pp. 130–134.

[15] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-García, and A. Uncini, "Nonlinear acoustic echo cancellation based on sparse functional link representations," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 7, pp. 1172–1183, 2014.

[16] H. Zhao, N. Li, R. Han, L. Chen, X. Zheng, C. Zhang, L. Guo, and B. Yu, "A deep hierarchical fusion network for fullband acoustic echo cancellation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9112–9116.

[17] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[18] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[19] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, and Y. Lu, "Interactive speech and noise modeling for speech enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 549–14 557.

[20] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[21] X. Xu, Y. Wang, J. Jia, B. Chen, and J. Hao, "GLD-Net: Improving Monaural Speech Enhancement by Learning Global and Local Dependency Features with GLD Block," in *Proc. Interspeech 2022*, 2022, pp. 966–970.

[22] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[23] 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.

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

[25] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," in *INTERSPEECH*, 2021.

[26] I. Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and 4speech codecs," *International Telecommunication Union, CH–Geneva*, 2005.

[27] 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*, 2010, pp. 4214–4217.

[28] D. N. E. Cancellers, "Itu-t recommendation g. 168," 2009.

[29] S. Zhang, Y. Kong, S. Lv, Y. Hu, and L. Xie, "F-T-LSTM Based Complex Network for Joint Acoustic Echo Cancellation and Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 4758–4762.

[30] 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)*, 2022, pp. 9122–9126.