



# Speaker- and Phone-aware Convolutional Transformer Network for Acoustic Echo Cancellation

Chang Han<sup>1</sup>, Weiping Tu<sup>1,2</sup>, Yuhong Yang<sup>1,2</sup>, Jingyi Li<sup>1</sup>, Xinhong Li<sup>1</sup>

<sup>1</sup>National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China

<sup>2</sup>Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China

{changhan, tuweiping, yangyuhong, 2021202110004, jingyi.li}@whu.edu.cn

## Abstract

Recent studies indicate the effectiveness of deep learning (DL) based methods for acoustic echo cancellation (AEC) in background noise and nonlinear distortion scenarios. However, content and speaker variations degrade the performance of such DL-based AEC models. In this study, we propose a AEC model that takes phonetic and speaker identities features as auxiliary inputs, and present a complex dual-path convolutional transformer network (DPCTNet). Given an input signal, the phonetic and speaker identities features extracted by the contrastive predictive coding network that is a self-supervised pre-training model, and the complex spectrum generated by short time Fourier transform are treated as the spectrum pattern inputs for DPCTNet. In addition, the DPCTNet applies an encoder-decoder architecture improved by inserting a dual-path transformer to effectively model the extracted inputs in a single frame and the dependence between consecutive frames. Comparative experimental results showed that the performance of AEC can be improved by explicitly considering phonetic and speaker identities features.

**Index Terms:** acoustic echo cancellation, complex network, contrastive predictive coding, speaker and phonetic characteristics, dual-path transformer

## 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 a far-end participant such that the far-end user receives a modified version of his/her voice. AEC aims to remove the echo from the microphone signal while leaving the near-end speech least distorted.

The traditional AEC methods model the acoustic echo path as a long linear adaptive filter then subtract the echo signal from the microphone observation [1]. The acoustic echo and the far-end speech are assumed to be a linear relationship in the traditional AEC methods. However, nonlinear distortions exist and are caused by electronic devices such as amplifiers and loudspeakers. To overcome this difficulty, several nonlinear models such as the Volterra model [2], the Hammerstein model [3], and

functional link adaptive filters [4] have been utilized. Although these traditional methods are fast and lightweight, their performance and robustness are not reliable in a complex acoustic environment.

Recently, deep learning (DL) has gained much attention for their capacity to model complicated nonlinear relationships and they have been successfully applied to various speech signal processing tasks such as speech enhancement, speech separation, and AEC. DL-based AEC can be formulated as a supervised speech separation problem [5], which separates the echo signal and the near-end signal so that only the latter is transmitted to the far end. And then several AEC methods based on speech enhancement/separation network have been proposed [6]. For example, Zhang et al. [7] propose a causal system based on convolutional recurrent network to estimate the real and imaginary spectrograms of near-end speech from the microphone signal and far-end signal. Westhausen et al. [8] apply the dual-signal transformation LSTM network (DTLN) to the task of real-time AEC by feeding the far-end signal as additional information. Kim et al. [9] propose an attention Wave-U-Net for the AEC, which includes an auxiliary encoder to extract the features of the far-end speech.

Recent studies [10, 11] in speech enhancement have shown significant benefit of using a deep complex network that handles magnitude and phase simultaneously because accurate phase spectrum estimation can achieve considerable improvements in both objective and subjective speech quality [12]. Based on deep complex network, Qiu et al. [13] proposed a self-supervised learning based phone-fortified method for speech enhancement. They explicitly import phonetic characteristics into a deep complex network to improve speech representation learning and speech enhancement performance. In fact, the selective listening ability of humans is considered to make effective use of not only phonetic features but also speaker identity features. Moreover, it also has been proven that the performance of speech separation can be improved by explicitly considering phonetic features and/or speaker identities [14].

To take advantage of phonetic and speaker identities features, we propose a novel AEC method, which imports phonetic and speaker identities features into a modified deep complex network explicitly. In this study, we adopt contrastive predictive coding (CPC) network which is a self-supervised pre-training model and has achieved promising performance in phone and speaker classification [15] for phonetic and speaker identities features extraction. To import features obtained from CPC, we present a feature fusion network to fuse them with the original frequency spectrum features. Moreover, we present a complex dual-path convolutional transformer network (DPCTNet)

This work was supported in part by the National Nature Science Foundation of China (No. 62071342, No.62171326), Hubei Province Technological Innovation Major Project (No. 2019AAA049, 2020B-AB018) and the Fundamental Research Funds for the Central Universities (No. 2042022kf0001).

The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

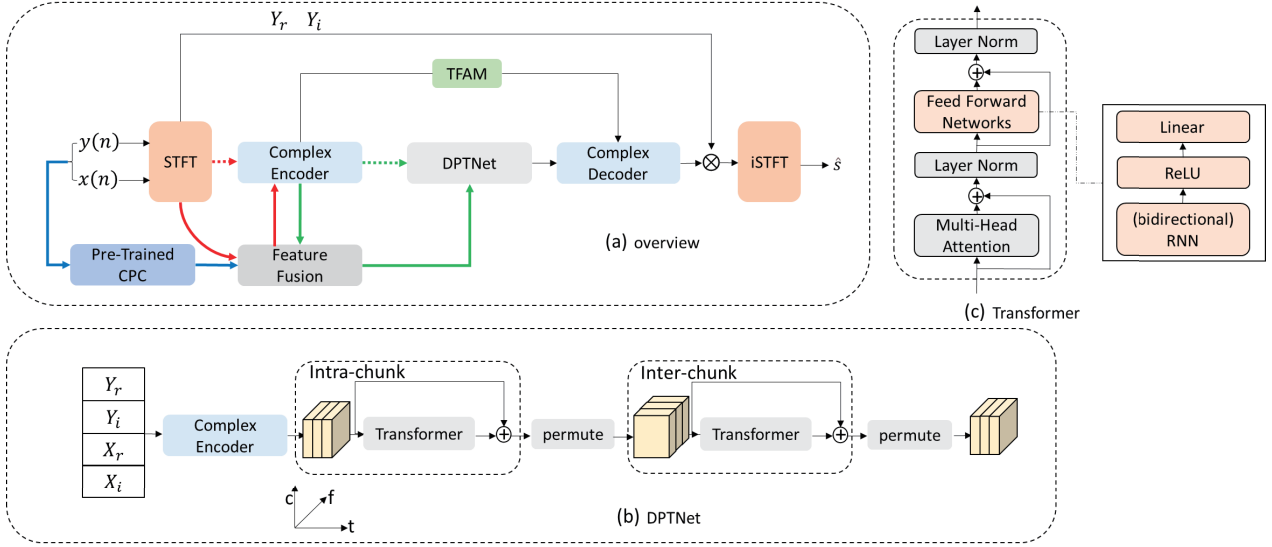


Figure 1: (a) Proposed speaker- and phone-aware DPCTNet model with various options for the feature fusion position. Time frequency attention module (TFAM) is described in detail in section 3.2. **Early fusion** is indicated by removing the red dotted line and adding the blue and red solid lines. **Middle fusion** is indicated by removing the green dotted line and adding the blue and green solid lines. (b) Diagram of the DPTNet module. “f”, “t” and “c” represent frequency, time and channel axis, respectively (c) Structure of transformer in DPTNet

by applying a dual-path transformer module to deep complex network. The dual-path transformer module splits input sequence features into smaller chunks and iteratively processes these chunks through intra-chunk and inter-chunk transformer.

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  $v(n)$ :

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

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 [16] 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:

$$\hat{M} = f_{\varphi} \left( X_r, Y_r, X_i, Y_i, \tilde{X}_{sp}, \tilde{Y}_{sp} \right) \quad (2)$$

where  $f$  and  $\varphi$  denotes DPCTNet and its network parameters,  $X$  and  $Y$  denote  $x(n)$  and  $y(n)$  after STFT respectively,  $r$  and  $i$  represent real and imaginary parts of complex spectrogram,  $\tilde{X}_{sp}$  and  $\tilde{Y}_{sp}$  denote  $x(n)$  and  $y(n)$  after CPC respectively,  $\hat{M}$  is the estimated complex ratio mask (CRM) [17], which can be defined as:

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

## 3. Proposed algorithm

In our design, the proposed speaker- and phone-aware DPCTNet consists of a CPC network to extract representations that contain phonetic and speaker identities information and a complex dual-path convolutional transformer network which is encoder-decoder structure to learn and reconstruct near-end speech. The entire speaker- and phone-aware DPCTNet architecture is shown in Fig.1(a).

### 3.1. Contrastive predictive coding

To obtain representations with phonetic features and speaker identities, we employ a pre-trained CPC model provided in the s3prl open source toolkit [18, 19]. CPC discriminates the correlated positive samples from negative samples with contrastive InfoNCE loss, which maximizes the mutual information between raw data and representations [15]. In the CPC model, the input sequence is mapped to a sequence of latent representations  $z_t$  by a non-linear encoder. Then all  $z_{\leq t}$  are summarized into the latent space by an autoregressive model and produces a context latent representation  $c_t$ . Finally, a probabilistic contrastive loss is used in CPC to induce the latent space to capture information that is maximally useful to predict future samples. Either of  $z_t$  and  $c_t$  could be used as representation for downstream tasks. In this paper, we explore the effects of using  $z_t$  and  $c_t$  respectively on echo cancellation tasks. The latent representation  $c_t$  contains extra context from the past, which might be useful in temporal correlations modeling.

In addition, we also explore two positions to import the representation of CPC as shown in Fig 1(a). One is the early fusion where the representation of CPC is converted into the same dimension with the output of short-time Fourier transform (STFT) by a Conv1D layer and then fused with the output of STFT through a simple feature fusion network. The proposed feature fusion network consists of a Conv2D layer followed by PRelu activation and batch normalization. The other is middle fusion

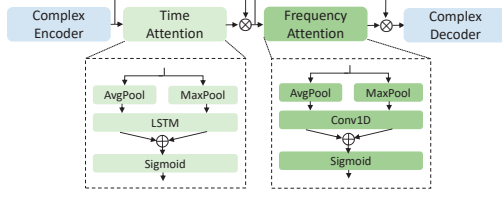


Figure 2: Diagram of the proposed time frequency attention module (TFAM).

where the representation of CPC is projected and fused with the output of the complex encoder.

### 3.2. Dual-path convolution transformer network

DPCTNet consists of a complex encoder, a dual-path transformer, a complex decoder, and a time frequency attention module (TFAM) as the add-skip connections between encoder and decoder. The complex encoder (decoder) consists of multiple (transposed) convolutional layers, batch normalization, and PReLU activation. We integrate DPTNet between encoder and decoder, as depicted in Fig.1(b). Different from the original DPTNet in time domain, the frames in STFT is regarded as the chunks for DPTNet processing. Instead of learning the dependence in the time domain, the intra-chunk transformers are applied to model the spectral patterns in a single frame. The structure of transformer in DPTNet is shown in Fig.1(c), in which the feed forward networks are composed of a RNN layer, ReLU activation and a linear layer. Bidirectional long short-term memory (BiLSTM) is used in the feed forward networks of intra-chunk transformer, which will not influence the causality of the whole system. While LSTM is used in the inter-chunk transformer to avoid involving in future information.

There are skip connections between the encoder and the decoder. Inspired by the channel attention modules in computer version [20], we propose a time frequency attention module (TFAM) embedded into the skip connections to automatically suppress unimportant regions and emphasize the important features as shown in Fig.2. Given an intermediate input  $\mathbf{F} \in \mathbb{R}^{T \times C \times F}$ , TFAM infers a 1D time attention weight  $\mathbf{M}_t \in \mathbb{R}^{T \times 1 \times 1}$  and a 2D frequency attention weight  $\mathbf{M}_f \in \mathbb{R}^{T \times 1 \times F}$ . The overall attention process can be formulized as:

$$\begin{aligned} \mathbf{F}' &= \mathbf{M}_t(\mathbf{F}) \otimes \mathbf{F} \\ \mathbf{F}'' &= \mathbf{M}_f(\mathbf{F}') \otimes \mathbf{F}' \end{aligned} \quad (4)$$

where  $\otimes$  denotes element-wise multiplication.

$$\begin{aligned} \mathbf{M}_t(\mathbf{F}) &= \text{LSTM}(\text{AvgPool}(\mathbf{F})) \\ &+ \text{LSTM}(\text{MaxPool}(\mathbf{F})) \end{aligned} \quad (5)$$

where  $\sigma$  denotes sigmoid activation function. The two LSTMs mentioned above share the same weight for both inputs.

$$\mathbf{M}_f(\mathbf{F}) = \sigma(f^{1 \times 1}([\text{Avg Pool}(\mathbf{F}); \text{MaxPool}(\mathbf{F})])) \quad (6)$$

where  $f$  denotes convolution layer and  $1 \times 1$  is the kernel size.

### 3.3. Training objectives

We explore two cost functions. One is SI-SNR [21]:

$$\begin{cases} s_{\text{target}} & := \langle \tilde{s}, s \rangle / \|s\|_2^2 \\ e_{\text{noise}} & := \tilde{s} - s \\ \text{SI-SNR} & := 10 \log_{10} \left( \frac{\|s_{\text{target}}\|_2^2}{\|e_{\text{noise}}\|_2^2} \right) \end{cases} \quad (7)$$

where  $s$  and  $\tilde{s}$  are the clean and estimated time-domain waveform, respectively.  $\langle \cdot, \cdot \rangle$  denotes the dot product between two vectors and  $\|\cdot\|_2$  is L2 norm. The other cost function is stretched-SI-SNR [22](S-SI-SNR) which outperforms to SI-SNR especially in the case of low SNR in speech enhancement:

$$\text{S-SI-SNR} = 10 \log_{10} \left( \frac{1 + \cos \theta}{1 - \cos \theta} \right) \quad (8)$$

where  $\theta$  represents the angle between  $s$  and  $\tilde{s}$ .

## 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 [23]. Besides, we also trained and evaluated the model using on-the-fly data generation. For online data generation, 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 [24] 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.

During training, there is a 50% chance that the data directly comes from the 2022 official synthetic dataset and a 50% chance that it is generated online. During online generation, we randomly select the near-end speech  $s(n)$ , background noise  $v(n)$ , far-end signal  $x(n)$  and echo signal  $d(n)$  prepared before training and combine these four signals together to get microphone signal. There is 50% probability to set echo signal  $d(n)$  as zeros to simulate the situation of near-end single-talk, in which the far-end signal may be noise or speech with 50% and 50% probability, respectively. And the noise signal  $v(n)$  in the near-end is set to 0 with 50% probability.

### 4.2. Evaluation metrics

The following four 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) [25].
- STOI: Short-time objective intelligibility measure (from 0 to 1) [26].
- AECMOS: A speech quality assessment metric for echo impairment (from 1 to 5) to evaluate call quality degradations in two separate categories: echo and degradations from other sources [27], which are dubbed respectively by "EMOS" and "DMOS" in Tables 2.
- ERLE: Echo return loss enhancement for far-end single-talk periods [28], which is defined as:

$$\text{ERLE} = 10 \log_{10} \left[ \frac{\sum_n y^2(n)}{\sum_n \hat{s}^2(n)} \right] \quad (9)$$

Table 1: Echo cancellation performance. DT: doubletalk, ST: single-talk, NE: near-end, FE: far-end. PESQ and STOI are used for DT and ST-NE scenarios and ERLE used for ST-FE scenario. Signal-to-noise ratio(SNR) is randomly picked up from [5, 20]dB, signal-to-echo ratio (SER) is set to -5, 5, and 15dB in DT scenario. Speaker identities and phonetic features (sp), early fusion (E), original microphone signal (Orig)

Method	SER(in dB)	DT						ST-NE		ST-FE
		-5		5		15		$\infty$		0
		Loss	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
Orig	-	0.815	0.713	1.381	0.847	1.875	0.907	2.255	0.914	-
Baseline-22 [23]	MSE	1.413	0.814	1.823	0.875	2.069	0.902	2.222	0.913	36.185
DPCTNet	SI-SNR	1.312	0.840	1.898	0.903	2.205	0.926	2.346	0.919	34.935
DPCTNet <sub>sp</sub>	SI-SNR	1.531	0.854	2.050	0.913	2.368	0.934	2.475	0.930	42.014
DPCTNet	S-SI-SNR	1.471	0.851	2.030	0.912	2.332	0.932	2.542	0.927	<b>46.681</b>
DPCTNet <sub>sp</sub>	S-SI-SNR	1.517	0.859	2.033	0.918	2.353	<b>0.937</b>	2.553	0.932	45.650
DPCTNet <sub>sp</sub> (E)	S-SI-SNR	1.379	0.833	1.940	0.904	2.298	0.929	2.525	0.932	36.381
DPCTNet <sub>sp</sub> (z <sub>t</sub> )	S-SI-SNR	1.528	0.858	2.061	0.914	2.358	0.935	2.534	0.932	42.421
DPCTNet <sub>sp</sub> (c <sub>t</sub> )	S-SI-SNR	<b>1.573</b>	<b>0.861</b>	<b>2.118</b>	<b>0.918</b>	<b>2.421</b>	0.936	<b>2.582</b>	<b>0.932</b>	44.015

### 4.3. Implementation details

All audio signals are resampled to 16kHz. 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 are used. Our model is trained with the Adam optimizer with an initial learning rate of 1e-4. For complex encoder, the channel number of the convolutional layers is [32,64,128,256,256,256]. The kernel size and the stride are respectively set to (5,2) and (2,1) in frequency and time dimension. The complex decoder is symmetric with the encoder, except that its kernel size is (5,1). We use masks for self-attention and set the LSTM to unidirectional in inter-transformers to avoid involving future information. The hidden LSTM units are 128 in DPCTNet. The whole parameters of DPCTNet are 4.6M, after adding the feature fusion module the whole parameters are 5.7M. Some of the processed audio clips can be found in this page<sup>1</sup>.

## 5. Results and analysis

As shown in Table1, we compare our model with a recurrent neural network that consists of 2 GRU layers with 322 hidden units which is from 2022 AEC Challenge baseline (**Baseline-22**) [23] under various SER conditions. Our model outperforms the baseline in all conditions. In addition, we conduct several ablation experiments: (I) **DPCTNet** consists of a complex encoder, a dual-path transformer, a complex decoder, and TFAM as the add-skip connections between encoder and decoder. (II) **DPCTNet<sub>sp</sub>** is formed by DPCTNet plus speaker identities and phonetic features(sp) which is either z<sub>t</sub> or c<sub>t</sub> or the sum of them. (III) **DPCTNet<sub>sp</sub>(E)** means adding representation in early fusion manner, otherwise in middle fusion manner.

First, the results show that using negative S-SI-SNR as loss function outperforms SI-SNR with respect to all metric scores, which indicates that similar to speech enhancement S-SI-SNR achieve better training effect in our model. Second, we studied the impact of whether and where to import phonetic and speaker identities features. DPCTNet<sub>sp</sub> exceeds DPCTNet in all cases, which implies that the phonetic and speaker identities features have a great potential in improving AEC performance. Then, we investigate two candidate positions for representations fusion as shown in Fig 1(a): middle fusion and early fusion. The

result shows that middle fusion outperforms early fusion, which implies the position to fuse the representations of CPC is an important factor to improve AEC performance. Finally, the effect of representations obtained from different layers of CPC is explored for AEC tasks. As we can see, compared to z<sub>t</sub> whose receptive field might not contain enough information to capture phonetic content, the model fused with c<sub>t</sub> that includes extra context from the past achieve better performance in all scenes, which reveals additional context may be helpful to AEC.

Table 2: DMOS and EMOS scores for the 2021 INTERSPEECH blind test set.

Method	NE	FE	DT	DT
	DMOS	EMOS	DMOS	EMOS
baseline-21 [29]	3.72	3.48	3.02	2.78
Y <sup>2</sup> -Net [30]	3.76	3.49	3.55	3.18
CDEC [31]	3.75	<b>4.28</b>	3.37	3.78
Two-stage AEC [32]	3.71	4.23	3.00	3.90
DPCTNet <sub>sp</sub> (c <sub>t</sub> )	<b>3.79</b>	3.92	<b>3.68</b>	<b>4.28</b>

Table 2 demonstrates that our model produces state-of-the-art results in terms of AECMOS by comparing against four recent proposed methods that show advantageous performance in the Interspeech 2021 AEC Challenge. Results for competing methods are taken from the corresponding papers. Although the competing results are for reference only, our proposed approach outperforms state-of-the-art results on the blind test of INTERSPEECH 2021 AEC Challenge.

## 6. Conclusions

In this paper, we propose a novel speaker- and phone-aware dual-path convolutional transformer network (DPCTNet) for acoustic echo cancellation. In the proposed approach, the phonetic features and speaker identities are imported into the DPCTNet model via a self-supervised learning based CPC model to improve acoustic echo cancellation performance. The DPCTNet applies an encoder-decoder architecture improved by inserting a dual-path transformer for effectively learning representations. Experimental results show that the performance of AEC can be improved by explicitly considering phonetic and speaker identities features.

<sup>1</sup><https://captain2xxx-coder.github.io/>

## 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] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," in *Proc. Interspeech 2020*, 2020, pp. 2477–2481. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2020-2631>
- [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] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 2472–2476.
- [11] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2811–2815.
- [12] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.
- [13] Y. Qiu, R. Wang, S. Singh, Z. Ma, and F. Hou, "Self-Supervised Learning Based Phone-Fortified Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 211–215.
- [14] Y. Du, K. Sekiguchi, Y. Bando, A. Arie Nugraha, M. Fontaine, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech separation based on a phone- and speaker-aware deep generative model of speech spectrograms," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 870–874.
- [15] A. Van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv e-prints*, pp. arXiv-1807, 2018.
- [16] 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.
- [17] 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.
- [18] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERFORMANCE Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.
- [19] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsuper-vised pretraining transfers well across languages," 2020.
- [20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 3–19.
- [21] 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.
- [22] Y. Sun, L. Yang, H. Zhu, and J. Hao, "Funnel Deep Complex U-Net for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 161–165.
- [23] 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.
- [24] 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 *INTER-SPEECH*, 2021.
- [25] I. Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH-Geneva*, 2005.
- [26] 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.
- [27] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "Aec-mos: A speech quality assessment metric for echo impairment," *arXiv preprint arXiv:2110.03010*, 2021.
- [28] D. N. E. Cancellers, "ITU-t recommendation g. 168," 2009.
- [29] 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," in *Proc. Interspeech*, 2021.
- [30] 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.
- [31] L. Pfeifenberger, M. Zoehrer, and F. Pernkopf, "Acoustic Echo Cancellation with Cross-Domain Learning," in *Proc. Interspeech 2021*, 2021, pp. 4753–4757.
- [32] 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.