ISSN: 2229-6948(ONLINE)
DOI: 10.21917/ijct.2026.0562

REAL-TIME JOINT ACOUSTIC ECHO CANCELLATION AND PERSONALIZED SPEECH ENHANCEMENT BASED ON CROSS-ATTENTION ALIGNMENT AND X-VECTOR

Kwon Kim, Yong-Hun Yun and Chol-Nam Om
Institute of Information Technology, High-Tech Research and Development Center, Kim Il Sung University, Democratic People's Republic of Korea

Abstract

Personalized speech enhancement (PSE) is a speech enhancement method to remove interfering speech, background noise, and reverberation based on a speaker embedding extracted from the target speaker such as d-vector and x-vector. In full duplex communication scenarios, when the microphone and far-end signal are coexisted together, it creates acoustic echoes. This echo is one of the major factors to the degradation of the sound quality of online communication systems, including video conferencing. Hence, Acoustic Echo Cancellation (AEC), a technique that can effectively remove these acoustic echoes, has been investigated. For full-duplex communications, which acoustic echoes are exist with background noises and interfering speech together, AEC and PSE must be combined. We study this combination. Our goal is to develop a causal model that can be applied to various model architectures to efficiently handle the tasks of AEC, PSE, and joint AEC-PSE. The features are extracted from the far-end signal and the near-end signal. The cross-attention alignment mechanism is used for feature alignment of the far-end signal and x-vectors are used as speaker embedding features. The proposed method is applied to PSE models such as E3Net and VoiceFilter-Lite. We present extensive experimental results. We demonstrate the effectiveness of the proposed method through the experiments in terms of various evaluation metrics with several standard audio and real recording datasets.

Keywords:

Personalized Speech Enhancement, Acoustic Echo Cancellation, Cross-Attention Alignment, X-Vector

1. INTRODUCTION

. Since the spread of COVID-19 pandemic, online communication systems including video conferencing have been widely used worldwide. In particular, video conferencing systems have played an important role in communicating between people in different regions, given the limited travel to other regions. However, these systems suffer from background noise, interference speech, acoustic echoes, and other factors that cause the degradation of the quality of the speech. To eliminate these factors and improve the quality of conferencing, SE (Speech Enhancement) and AEC (Acoustic Echo Cancellation) techniques are used.

1.1 REVIEW ON SE

Recently, deep learning has been used in the research of speech enhancement to achieve remarkable performance, and has been actively used in real-world applications such as automatic speech recognition, intelligent elevators, and hearing aids. Speech enhancement methods using deep learning can be classified generally in time-frequency (T-F)-domain based and time-domain based approaches. In SE model based on the T-F-domain, the spectral features of the noise are obtained by STFT (short-time Fourier transform) and the spectra of the clean speech are estimated by the denoising system. Meanwhile, the time-domain method directly estimates the clean speech from the noise waveform.

1.1.1 T-F-Domain based SE:

In [1], a deep neural network (DNN) model for SE using mask method is proposed. To provide better nonlinearity, a nonlinear function ReLU is inserted in each layer. The ReLU activation function deactivates the neuron when it is fed negative values [2]. In [3], a CNN architecture that overcomes the drawbacks of DNN is proposed, where CNN uses a small number of parameters due to its local operation and weight sharing properties. A convolutional encoder-decoder network (Convolutional Encoder-Decoder, CED) is proposed for noise cancellation, and the proposed network consists of a symmetric encoder and decoder layer. Recently, many papers have combined CNN and RNN to build a complex structure called convolutional recurrent network (CRN), where CNN-based layers are used to extract low-level features and RNN-based layers are used to extract the contextual information of features. This CRN architecture has shown very good performance in speech enhancement [4,5]. More recently, [6] has designed a complex-valued neural network for speech enhancement, called Deep Complex Convolution Recurrent Network (DCCRN). The DCCRN was built to use complex spectra, which replace the traditional convolution of CRN and LSTM, and proposed complex-valued convolution and complex-valued LSTM.

1.1.2 Time-Domain based SE:

The time-domain based method does not address the problem of phase estimation, which is a challenge for the T-F-domain based method because it directly estimates clean speech waveforms from noisy waveforms. Many time-domain SE systems use the U-Net framework to model long speech waveforms. Here, a convolutional encoder is used to extract high-level time series features reconstructed by a symmetric decoder with a skip connection. In [7], a Wave-U-Net framework for time-domain SE consisting of a one-dimensional convolutional encoder and decoder is proposed to perform downsampling and upsampling.

Contextual information is very important in SE and, to achieve this goal, they increase the depth of convolutional neural networks (CNNs) or extend the kernel size. However, this requires huge computational cost. In [8], a new convolutional network is proposed to solve this problem, and an extended convolutional network is effectively extended to the suitable area without adding parameters. In [9], an extended convolutional U-Net for SE in the time-domain is proposed. In [10], the structure of the cross-

attention conformer is improved so that the noise context can be used in the model. ConvTasNet, proposed in [11], is a DL (deep learning) model for time-domain single-channel speech source separation. They use recurrent neural networks for training and estimate a mask for separating the speech of the target speaker from the mixed speech. In [12], a band-wise recurrent neural network (Band-Split Recurrent Neural Network, BSRNN) is proposed for music source separation. This BSRNN classifies the spectrogram into distinct frequency bands.

1.2 REVIEW ON AEC

Acoustic echo cancellation (AEC) aims to remove echoes from microphone signals while minimizing the speech distortion of the target speaker to obtain a clean near-end signal. The AEC method can be broadly classified into adaptive filtering, deep neural network based methods. AEC based on conventional digital signal processing (DSP) is achieved by estimating the acoustic echo path with an adaptive filter. Recently, deep learning has shown great effect on AEC due to the strong nonlinear model potential.

1.2.1 DSP based AEC:

Generally, the echo canceller consists of three main processing parts : a double-talk (DT) detector, an adaptive filter, and a nonlinear processor. In the conventional conversation, the situation in which the far-end speaker and the near-end speaker speak simultaneously is called the double-talk (DT) situation, and the DT detector detects this DT situation, thus stopping the coefficient adaptation of the adaptive filter. The purpose of using a DT detector is to prevent the divergence of the adaptive filter. The AEC problem is usually solved using an adaptive filter that describes the acoustic echo path well. An adaptive filter is used to estimate the acoustic echo signal and then subtract this estimated echo from the microphone signal to estimate the near-end signal. After the microphone input signal has passed through the adaptive filter, the echo remains, which is removed by the nonlinear processor.

Although many practical studies have been carried out for echo cancellation, it is one of the most difficult problems to estimate the transfer function of echo paths accurately in practical situations. In general, the impulse response of the acoustic path is more than tens of milliseconds, so to implement a good performance acoustic echo canceller, an adaptive filter of large-dimension has to be implemented. However, the large-dimension adaptive filter is difficult to achieve in real time due to the large amount of operation, and it is necessary to implement an adaptive filter with low computational cost and high convergence speed. The most popular adaptive filter algorithms are the least mean square (LMS), recursive least squares (RLS), and multi-delay block frequency domain adaptive filter (MDF) [13, 14]. In addition, various methods have been studied, including affine projection algorithms, Kalman algorithms, etc.

1.2.2 DNN based AEC:

In the neural network-based approach, we usually use the $\log|X(l,k)|$, $\log|Y(l,k)|$ of the far-end signal $x(n)$ and microphone signal $y(n)$ to obtain their time-frequency representations $X(l,k)$ and $Y(l,k)$ first, and then the input of the neural network to the far-end signal, the $\log|X(l,k)|$. The output of the neural network and the microphone signal phase $\arg(Y(l,k))$ are used to estimate the time-frequency representation $S(l,k)$ of the near-end signal and then the inverse time-frequency conversion to estimate the near-end signal $s(n)$.

As a method of echo cancellation using deep neural networks, a number of methods have been investigated using deep learning techniques, including magnetic encoder, restricted Boltzmann machine (RBM), LSTM-recurrent neural network (LSTM-RNN), convolutional neural network (CNN) [15], recurrent neural network (RNN) [16], and generative adversarial network (GAN) [17]. In [18], a frequency-domain mask-based time-convolutional network, defined as a supervised speech separation problem, is proposed for AEC. In [19], the attention-based alignment method is proposed to address the potentially existing time delay problem between microphone and far-end signals, which is a challenging problem in AEC.

1.3 REVIEW ON JOINT AEC-PSE

Personalized speech enhancement (PSE) models have achieved significant improvement in removing background noise and interfering speech [20,21, 22]. The PSE systems depend on a cue representing the target speaker, usually a speaker embedding vector such as an x-vector or a d-vector. These systems can remove all other human voices in the input audio except for the target speaker and remove background noise [21,22].

Many previous studies have proposed effective PSE models. VoiceFilter is an STFT-based convolutional recurrent model using the d-vector extracted from the target speaker. In [23], a new version, VoiceFilter-Lite, that outperforms the original VoiceFilter with less computational cost is proposed. In [24], a Personalized PercepNet is proposed by adding the use of speaker embedding vectors to the original PercepNet. [21] proposed a personalized deep complex convolutional recurrent network (pDCCRN) that outperformed causal VoiceFilter and first proposed the TSOS problem. It also proposed a metric to measure TSOS with a degree of mitigation. In [22], a personalized E3Net, an efficient model for performing PSE in the time domain, is proposed. This model outperformed relatively larger models such as pDCCRN with a much smaller computational cost. However, none of these PSE models have addressed the problem of combining with AEC.

Recently, in [25], a joint AEC-PSE model called personalized gated temporal convolutional neural network (pGTCNN) was proposed, which used the speaker embedding of the target speaker and the far-end speaker. However, they have stayed in the simulation stage and did not consider the TSOS problem. In [26], a series of methods have been proposed to develop a causal model that efficiently handles AEC, PSE, and joint AEC-PSE tasks.

In this paper, we propose a series of methods to develop a causal model that efficiently handles AEC, PSE, and joint AEC-PSE tasks. First of all, we propose a time-delay synchronization method between microphone and far-end signal by using Cross-attention alignment network to improve the AEC performance. We also use x-vectors as speaker embedding vectors in PSE systems to improve the automatic speech recognition (ASR) rate. We mitigate the target speaker over-suppression (TSOS) resulting from over-reliance on speaker embedding vectors by concatenating this x-vector to the output of the first N_1^{th} layer of the proposed network. During training, a bypass path is introduced to focus only on echo cancellation and background

noise removal in the earlier layers while eliminating interfering speech in the latter layers. The models are trained with multi-task learning including AEC, PSE, and joint PSE-AEC. We evaluate the performance of the proposed method with several standard audio and real recording databases.

2. PROPOSED METHOD

A joint AEC-PSE model uses microphone signal, far-end signal and speaker embedding as input. In [25], a neural network architecture is proposed that concatenates extracted features from microphone signal and far-end signal and combines speaker embedding. However, this method is limited to handle the time delay that exists between the microphone signal and the far-end signal. They also did not consider the target speaker over suppression (TSOS) problem arising from over-reliance on speaker embedding.

We propose a series of methods to address these problems and apply them to the typical time-domain and short-time Fourier transform (STFT) domain PSE models, E3Net and VoiceFilter-Lite. Fig.1 shows the structure of the proposed joint AEC-PSE model.

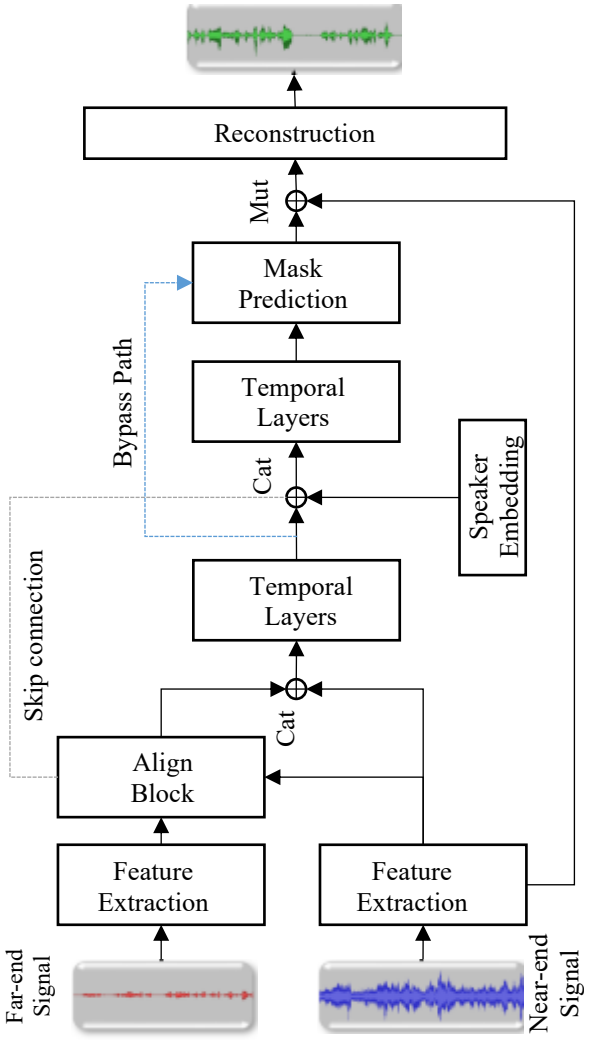

Fig.1. Our proposed joint AEC-PSE model

2.1 FEATURE EXTRACTION

The far-end signal and microphone signal are sampled at 16 kHz and the input feature to the network is selected to be the logarithmic power spectra using Hamming window.

For the far-end signal $x(t)$ and microphone signal y(t), the short-time Fourier transform is applied to calculate the logarithmic power spectra $\log|X(t,f)|$, $\log|Y(t,f)|$. When m is an integer, f_s is the sampling frequency, we assume that the window function $w(\tau)$ is discretized to $\tau=m/f_s$, and when $|\tau|\geq M/(2f_s)$ (M is a positive integer) and $w(\tau)=0$, then the STFT formula is as follows.

$$X\left(\frac{n}{f_s},\frac{kf_s}{2M}\right)=\sum_{|m|<\frac{M}{2}} x\left(\frac{n+m}{f_s}\right)w\left(\frac{m}{f_s}\right)e^{-j\frac{2km}{M}} \quad (1)$$

$X\in \mathrm{R}^{C\times T\times F}$ is called far-end signal feature. The same formula is used to find the microphone signal feature, $X\in \mathrm{R}^{C\times T\times F}$. Here C is the channel number, T is the time dimension and F is the frequency dimension.

One frame length is 320 samples for 20 ms, and the segment of the speech signal processed at a time contains 40 frames and a 320-point discrete Fourier transform is applied. Thus, the channel number was set to 40, the time dimension was set to 320, and the frequency dimension was set to 40.

2.1.1 Learnable Encoder:

In [26], a learnable encoder is applied to the time-domain model E3Net, which uses features extracted from both microphone and far-end signal for AEC. For the microphone signal, it was proved in [22] that the higher the number of learnable encoder's filters, F_{mic}, the higher the performance. However, for the far-end signal, it is stated in [26] that even smaller numbers of learnable encoder's filters, F_{far}, do not affect the speech quality. This is because features extracted from the far-end signal are only used to synchronize the time delay between the microphone and the far-end signal and to estimate the intensity of the acoustic echo that must be removed.

2.2 ALIGNMENT BLOCK

The Cross-attention alignment network computes the alignment based on the neural network internal state and synchronizes the microphone signal and the far-end signal. We implement the attention and calculate the delay distribution using features in time-frequency domain instead of the attention alignment network in time-domain of [37]. For this, we focus on the deep time-frequency feature by computing the active delay distribution used to smoothly align the far-end signal feature.

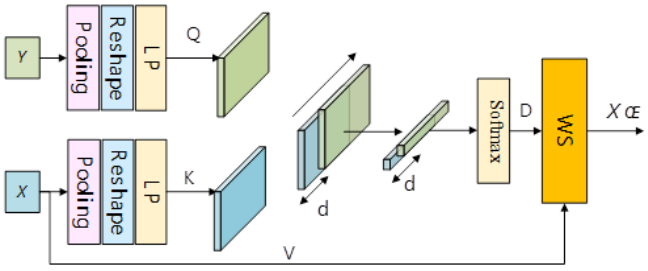

Fig.2. Cross-attention alignment network

The proposed Cross-attention alignment network is shown in Fig.2. In the figure, Pooling denotes the max-Pooling layer, Reshape denotes the dimensional deformation layer, LP denotes the linear projection layer, Softmax denotes the soft maximization function, and WS denotes the weighted sum.

The microphone signal acts as a question (Q), and the far-end signal acts as a key (K) and a value (V).

The principle of alignment is to calculate the similarity between the T^{th} frame of the microphone signal and the T^{th} frame of the far-end signal using the previous T-1 frames to obtain the aligned far-end signal in the T^{th} frame, when the length of the attention window is N. In convolution layer, the feature map is reduced to $\mathrm{R}^{C\times T\times \frac{F}{4}}$ using a maximization convolution operation with a kernel size of 1×4 depending on the frequency dimension. This is a process to reduce the computational cost caused by the Cross-attention alignment network. In the dimensional deformation layer, we transform the feature dimension to be $X, Y \in \mathrm{R}^{T\times\left(\frac{F}{4}\cdot C\right)}$. In the linear projection layer, we map this feature to the query $Q \in \mathrm{R}^{T\times P}$ and the key $K \in \mathrm{R}^{T\times P}$. Here $P \in \mathrm{N}$ is the projective size.

The reason for applying the linear projection layer in designing the Cross-attention alignment network is to match the output feature dimension if the dimensions of the microphone signal feature and the far-end signal feature are different from each other.

Next, we cut the K tensor into the same d values that first zero-fill and finally generate the synthesis delay. To estimate the similarity between the elementary sequences for the obtained Q and K, the time-axis point-wise multiplication is performed as follows.

$$Q=[q_1,q_2,\ldots,q_T], \qquad K=[k_1,k_2,\ldots,k_T]$$

$$r_d=\sum_{t=1}^{T-d} q_{t+d-1}\cdot k_t$$

where, q_t, k_t denotes the elementary sequences vectors of Q and K, respectively, and d=1,2,⋯,d_{max}, and d_{max} is the maximum delay length.

Applying the soft maximization function to $R=[r_1,\ldots,r_{d_{\max}}]$, we compute the delay distribution $D=[w_1,\ldots,w_{d_{\max}}]\in \mathrm{R}^{d_{\max}}$ as follows.

$$w_d=\mathrm{softmax}(r_d)=\frac{\exp(r_d)}{\sum_j \exp(r_j)} \tag{2}$$

Then, applying weighted sum with delay distribution to X can improve the robustness of the delay estimation. The aligned far-end feature $X' \in \mathrm{R}^{C\times T\times F}$ is computed by the weighted sum according to the value of the distribution D of X shifted on the time-axis.

$$X'=\sum_{d=1}^{d_{\max}} w_d X_d \tag{3}$$

where $X_d \in \mathrm{R}^{C\times T\times F}$ is denoted by $X_d=[0,\ldots,0,x_1,x_2,\ldots,x_{T-d}]$.

$0 \in \mathrm{R}^{C\times F}$ denotes the $C\times F$ dimension matrix, and $x_t \in \mathrm{R}^{C\times F}$ is the t^{th} matrix of X along the time-axis.

$$X=[x_1,x_2,\ldots,x_T] \tag{4}$$

Our proposed method can eliminate the need of traditional alignment methods, including GCC-PHAT, and has the advantage of reducing the inference time and delay time and ensuring the real-time requirements of video conferencing systems due to less computational complexity and computational time than previous methods (time-domain alignment method) [38].

2.3 SPEAKER EMBEDDING

The x-vector is used as speaker embedding. The x-vector is based on the DNN embedding proposed in [27]. The features are 24-dimensional filter banks with a frame length of 25ms and are mean-normalized over a sliding window up to 3s. The energy SAD, as used in the reference systems, filters out-nonstandard frames.

Table 1 shows the DNN architecture. Suppose that the input segment has T frames. The first five layers operate on speech frames, with a small temporal context centered on the current frame t. For example, the input to *frame 3* is the output associated to *frame 2* at frames $t-3$, t, $t+3$. This is based on the temporal context of the previous layers, so *frame 3* sees the overall context of 15 frames.

The statistical pooling layer aggregates all T frame level outputs from *frame 5* and calculates its mean and standard deviation. The statistics are 1500 dimensional vectors and are computed once for each input segment. This process aggregates the information through time dimension so that subsequent layers can operate on the entire segment.

In Table 1, this is represented by the layer context of { 0 } and the whole context of T. The mean and standard deviation are concatenated together and propagate through the segment level layers and finally the soft maximum output layer. All nonlinearities are commutating linear units (ReLUs).

The DNN is trained to classify N speakers in the training data. The training example consists of a set of speech features (about 3s on average) and a corresponding speaker label. After learning, the embeddings are extracted from the affine component of *segment 6*. Ther is a total of 4.2 million parameters except the soft maximum output layer and *segment 7* (since they are not required after training).

Table.1. The embedding DNN architecture

Layer	Layer Context	Total Context	Input × Output
frame1	[*t*-2, *t* +2]	5	120 × 512
frame2	{*t* -2, *t*, *t* +2}	9	1536 × 512
frame3	{*t* -3, *t*, *t* +3}	15	1536 × 512
frame4	{*t*}	15	512 × 512
frame5	{*t*}	15	512 × 1500
stats pooling	[0, *T*)	*T*	1500*T* × 3000
segment6	{0}	*T*	3000 × 512
segment7	{0}	*T*	512 × 512

softmax	{0}	T	$512 \times N$

2.4 MULTI-TASK TRAINING

In the original E3Net and VoiceFilter-Lite for PSE, the speaker embedding vector is concatenated with the observed features prior to the fist temporal layers. However, in the AEC-PSE task, this causes an increase in TSOS. We prevent the increase of TSOS by concatenating the speaker embedding vector to the output of the N1th temporal layer.(Bypass Path)

We added a attention weight from the alignment block to the output of the N1th temporal layer using a skip connection, so that the latter time layers can control the noise suppression in the presence of the echo signal.(Skip Connection)

The joint AEC-PSE model is aimed to outperform the AEC-PSE-only model when the microphone signal is simultaneously existed with target speaker, interfering speaker, acoustic echo, and near-end noise, while it is as effective as the AEC and PSE-only model. To do this, the following three mini-batches were performed alternately with different data structures in each iteration of the learning process. AEC mini-batch-which includes the target speakers, near-end noise, and echo signals and does not contain speaker embedding vectors. Therefore, this model is trained using a bypass path to allow the earlier temporal layers to focus on AEC and noise suppression. PSE mini-batch- which includes the near-end noise, target speaker and interfering speaker and speaker embedding vectors, and the far-end signal is zero. This model learns the PSE capability through the full path. AEC-PSE mini-batch-which includes all possible signals. This model learns the AEC-PSE capability through the entire path.

3. EXPERIMENTS AND DISCUSSION

We simulated training and validation data for PSE using the same approach as proposed in [21,22]. From DNS Challenge [28], we obtained clean speech utterances containing 544-hour speech samples based on LibriVox audio [29]. We also used noise samples of AudioSet [30] and Freesound [31]. For each training and validation sample, the target speaker was randomly placed 0-1.3m from the microphone and the interfering speaker was placed more than 2m away using the simulated room impulse responses (RIRs). By allowing only half of the samples to contain the interfering speech, we improved the performance of the SE task of the model. Signal-to-noise ratio (SNR), signal-to-echo ratio (SER) and signal-to-interference ratio (SIR) were varied between 0 and 15dB, -20 to 40dB and 0 to 10dB, respectively. From DNS Challenge [28], we obtained a clean speech source that would be a far-end echo signal.

We evaluated the performance of the model using the simulated long-duration test sets as described in [21, 22] based on the voice cloning toolkit (VCTK) corpus [32]. These test sets encompass five important scenarios:

- TS1: target speaker + interfering speaker + noise
- TS1-echo: TS1 + echo (target speaker + interfering speaker + noise + echo)
- TS2: target speaker + noise
- TS2-echo: TS2 + echo (targer speaker + noise + echo)
- TS3: target speaker only.

TS1 and TS2 evaluate the performance of the model in the PSE and SE scenarios, respectively, while TS3 is used to assess the target speech quality degradation. TS1-echo is the most interesting test set that contains all possible signals, and measures AEC-PSE performance. TS2-echo evaluates the AEC performance in the presence of noise.

We varied SNR and SER between 0 and 15 dB, and SIR between 0 and 10dB. VCTK was used as the clean audio source. To create a single long-duration file for each speaker, files from the same speaker were stitched together. The average duration of individual test samples was 27.5 min. Each test set consists of 109 speakers, all of which are about 50 hours long.

We also tested the model with real recordings. For PSE performance evaluation, we used the DNS Challenge personalized track [28] blind test set, which consists of 859 real recordings, of which 121 contain interfering speakers. For AEC performance evaluation, we used the blind test set of the ICASSP 2022 AEC Challenge [33]. This test set consists of 600 recordings, each of which is 30 to 45s, half of which is a far-end single-talk (FST) and the rest is a double-talk (DT) case. And this test set contains interesting scenarios that are commonly found in real video conferences.

3.1 EVALUATION METRICS AND IMPLEMENTATION DETAILS

We used the evaluation metrics such as ASR (Automatic Speech Recognition) rate, speech quality, and TSOS to evaluate the PSE performance. To measure speech quality, we used DNSMOS P.835 [34], which is a non-intrusive neural network based mean opinion score (MOS), which makes it possible to predict speech quality subjectively. We used the word error rate (WER) used in Microsoft's internal ASR model. And the TSOS metric was calculated by the method proposed in [21]. AEC quality was measured using AECMOS [35], a neural network MOS estimator similar to DNSMOS. For the FST scenario, the echo return loss enhancement (ERLE) was also used as an evaluation metric. For real data without clean reference signals, DNS and AEC Challenge test sets, DNSMOS and AECMOS were used, respectively.

The length of both training and validation samples was 20s. We set the f_{emb} and $f_{emb-hid}$ of E3Net to 128 and 768, respectively. The dimensionality of the input and hidden LSTM blocks was set equal to f_{emb}. The learnable encoder and decoder window (filter size) and hop size (stride) were set to 20ms (320) and 10ms (160), respectively. The dimensions of the learnable encoders for the microphone and far-end signals, f_{mic} and f_{far}, were 2048 and 256, respectively. Number of LSTM blocks, N_1 and N_2 were both 2, totalling 4 LSTM blocks, which are the same as the original E3Net. The baseline E3Net training was performed with the aforementioned parameters and 4 LSTM blocks were used. For all VoiceFilter-Lite models, we used 4 LSTM layers with 512 hidden dimensions and divided them equally as in E3Net. As input, we used power-law compressed STFT magnitudes, where STFT parameters are equal to the window size and hop size as E3Net. We trained all models with power-law compressed phase-aware (PLCPA) loss function (see Eq. (1) of [21]).

3.2 BASELINE MODELS FOR INDIVIDUAL TASKS

We used AlignCruse [19] as a reference model for AEC, which has high AEC performance and low computational cost. This model uses the STFT feature as input and consists of a 2D convolutional encoder and decoder pair, an align block, and a recurrent bottleneck block. We have experimented with the pre-trained model and code provided in [19], which outperformed the one reported in the original paper [19]. And to train the E3Net and VoiceFilter-Lite models for the AEC task, we removed the x-vector input from the model proposed in [20]. When x-vector input is removed, these models have the same structures as the E3Net and VoiceFilter-Lite models for the AEC-PSE task.

We used the personalized E3Net model and VoiceFilter-Lite model using 4 LSTM layers as PSE baseline models. These models do not include far-end signal input. We used a slightly different architecture from that proposed in [22] to ensure a fair comparison between this E3Net-based PSE model and the AEC-PSE model we proposed.

We also trained the AEC-PSE E3Net and VoiceFilter-Lite Naïve models to evaluate the performance of the proposed method. The E3Net Naïve model applied a learnable encoder to both microphone and far-end signals and then concatenated features and speaker embedding vectors prior to the first projection layer. Even in VoiceFilter-Lite Naïve, the STFT features of microphone and far-end signals are concatenated to the speaker embedding vector before being transferred to the first LSTMs. Naïve models do not use our proposed Cross-attention alignment block and bypass path, and do not include the second projection layer in the case of E3Net. We have evaluated the performance improvement of our proposed method compared to [26] by training the model using self-attention alignment blocks and d-vectors proposed in [26].

Finally, to compare our model with the AEC-PSE model, which is more computationally expensive, we trained pGTCNN with the same parameters as described in [25].

3.3 RESULTS AND DISCUSSIONS

The Table.2 and Table.3 show all the experimental results for the simulated and real recording test sets, respectively.

Table.2. Experimental results for PSE and PSE-AEC using VCTK data sets.

	TS1			**TS1-echo**			**TS2**			**TS2-echo**			**TS3**	
	WER ↓	**DNSMOS ↑**	**TSOS ↓**	**WER ↓**	**DNSMOS ↑**	**AECMOS ECHO↑**	**WER ↓**	**DNSMOS ↑**	**TSOS ↓**	**WER ↓**	**DNSMOS ↑**	**AECMOS ECHO↑**	**WER ↓**	**TSOS ↓**
Baseline Systems														
No enhancement	43.03	2.92	0	54.78	1.18	2.42	13.35	2.98	0	33.10	2.0	2.37	7.12	0
AlignCruse [20] – AEC	57.10	3.27	0	62.46	2.39	3.53	21.82	3.41	0	37.29	2.58	3.85	7.47	0
pGTCNN [26] – AEC-PSE	42.21	3.18	1.48	52.29	2.58	4.02	20.41	3.34	0.63	34.22	2.81	4.13	7.95	0.56
Effects of proposed improvements on VoiceFilter-Lite and E3Net														
VoiceFilter-Lite														
-AEC	49.32	3.24	0.47	59.57	2.51	4.05	20.52	3.44	0.14	33.99	2.82	4.23	7.90	0.02
-PSE	40.13	3.25	1.19	53.36	2.54	3.97	19.71	3.44	0.22	34.34	2.69	4.13	7.41	0.12
-AEC-PSE Naïve	48.85	3.17	3.55	55.51	2.46	3.99	22.80	3.35	1.62	34.91	2.76	4.12	8.21	1.44
-AEC-PSE [27]	41.86	3.24	1.56	52.54	2.56	4.02	20.93	3.43	0.62	34.25	2.82	4.15	7.90	0.33
-AEC-PSE proposed	**41.63**	**3.25**	**1.47**	**52.21**	**2.59**	**4.04**	**20.39**	**3.44**	**0.48**	**34.18**	**2.84**	**4.19**	**7.69**	**0.31**
E3Net														
-AEC	43.50	3.51	0.33	58.19	2.80	4.26	18.05	3.71	0.17	30.12	3.06	4.48	7.19	0.05
-PSE	36.91	3.49	2.15	51.40	2.69	4.23	18.35	3.74	0.33	33.51	2.95	4.41	7.54	0.32
-AEC-PSE Naïve	38.70	3.43	1.54	54.06	2.64	4.24	19.76	3.70	0.96	33.85	2.87	4.39	7.82	1.38
-AEC-PSE [27]	38.96	3.46	1.45	51.88	2.63	4.27	19.43	3.68	0.58	32.70	2.91	4.46	7.46	0.34
-AEC-PSE proposed	**38.05**	**3.49**	**1.42**	**49.97**	**2.75**	**4.35**	**19.32**	**3.71**	**0.52**	**31.08**	**2.98**	**4.48**	**7.45**	**0.14**

Table.3. Computational complexities of different models and their results for real recording test sets.

	Complexity		DNS Challenge v4 Blind Test Set			AEC Challenge FST		AEC Challenge DT	
	Parameters (millions)	RTF	SIG↑	BAK↑	OVR↑	AECMOS ECHO↑	ERLE↑	AECMOS ECHO↑	AECMOS DEG↑
Baseline Systems									
No enhancement	-	-	3.71	2.17	2.40	1.98	0	1.81	4.11
AlignCruse [20] – AEC	0.45	0.056	3.58	3.85	3.18	4.12	44.40	4.34	3.91
pGTCNN [26] – AEC-PSE	5.33	0.229	3.42	3.88	3.09	4.39	48.27	4.58	3.69
Effects of proposed improvements on VoiceFilter-Lite and E3Net									
VoiceFilter-Lite									
-AEC	8.56	0.143	3.43	3.86	3.07	4.46	48.97	4.33	3.62
-PSE	8.03	0.134	3.44	3.91	3.11	2.71	18.03	4.11	3.73
-AEC-PSE Naïve	8.36	0.138	3.25	3.87	2.92	4.43	45.41	4.26	3.24
-AEC-PSE [27]	8.36	0.142	3.45	3.89	3.09	4.44	45.63	4.31	3.69
-AEC-PSE proposed	**8.36**	**0.143**	**3.49**	**3.91**	**3.10**	**4.44**	**45.76**	**4.32**	**3.71**
E3Net									
-AEC	3.28	0.054	3.52	4.09	3.24	4.47	47.91	4.65	3.97
-PSE	3.17	0.050	3.53	4.07	3.24	2.20	11.42	4.33	4.04
-AEC-PSE Naïve	3.23	0.054	3.49	4.06	3.19	4.39	45.38	4.61	3.46
-AEC-PSE [27]	3.27	0.054	3.45	4.08	3.16	4.45	49.06	4.61	3.92
-AEC-PSE proposed	**3.22**	**0.054**	**3.52**	**4.09**	**3.23**	**4.46**	**49.23**	**4.62**	**3.98**

In the lower half of the two tables, we show the results of AEC, PSE, and AEC-PSE architectures for the VoiceFilter-Lite and E3Net models. As shown in the simulation results in Table 2, our AEC-PSE model has shown good performance over all evaluation metrics compared to AEC-PSE Naïve and AEC-PSE model which adopted the method proposed in [26]. The results of the real recording tests in Table.3 show that Naïve's structure and the structure proposed in [26] are lower in terms of AECMOS DEG scores than our proposed structure, which shows high distortion in the case of double-talk. Using skip connection slightly improved DNSMOS and AECMOS DEG for real recording data at the cost of FST performance, and lower WER were generated for TS1-echo and TS2-echo for simulated data. Finally, comparing the proposed AEC-PSE architecture with the AEC-only and PSE-only architecture, it can be seen that the joint model achieved good performance for most of the metrics compared to the individual expert models (PSE-only TS1 and AEC-only TS2-echo). These results demonstrate the effectiveness of the proposed method applied to the E3Net and VoiceFilter-Lite models.

In the upper half of Table.2 and Table.3, the results for the three baseline models are shown. Our proposed AEC-PSE model applied to E3Net outperforms the recently proposed pGTCNN model for all metrics despite the lower computational cost. It also outperformed in terms of AEC performance at a computational cost similar to the AlignCruse-based AEC model. These results show that the proposed AEC-PSE model has improved performance compared to several baseline models.

4. CONCLUSION

The contributions of this paper are as follows.

- First, we propose to extract the feature from the far-end signal and microphone signal and to use the Cross-attention alignment network for the feature alignment of the far-end signal.
- Second, we propose to use x-vectors as the speaker embedding vector.
- Third, we propose a new causal model that can efficiently handle AEC, PSE, and joint AEC-PSE tasks. We applied a learnable encoder for the far-end signal so that the extracted features were used to align the far-end signal and estimate the echo intensity that would need to be removed. We also propose a method to mitigate TSOS by introducing bypass path and to apply multi-task learning.

We have applied the proposed model to typical PSE models E3Net and VoiceFilter-Lite to verify its effectiveness. In the future, we will further study neural networks that simultaneously remove echoes, interfering speech, background noise, and reverberation in more complex acoustic environments.

REFERENCES

[1] Tian Gao, Jun Dua, Li-Rong Dai and Chin-Hui Lee, "A Unified DNN Approach to Speaker-Dependent Simultaneous Speech Enhancement and Speech Separation

in Low SNR Environments", *Proceedings of International Conference on Speech Communication*, pp. 1-13, 2017.

[2] V. Nair and G.E. HinWon, "RecWified linear Units improve restricted Boltzmann Machines", *Proceedings of International Conference on on Machine Learning*, pp. 807-814, 2010.

[3] A. Krizhevsky, I. Sutskever and G.E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems*, Vol. 25, pp. 1-8, 2012.

[4] H. Zhao, S. Zarar, I. Tashev and C.H. Lee, "Convolutional-Recurrent Neural Networks for Speech Enhancement", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 2401-2405, 2018.

[5] K. Tan and D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement", *Proceedings of International Conference on Interspeech*, pp. 3229-3233, 2018.

[6] 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", *Proceedings of International Conference on Audio and Speech Processing*, pp. 1-9, 2020.

[7] C. Macartney and T. Weyde, "Improved Speech Enhancement with the Wave-U-Net", *Proceedings of International Conference on Audio and Speech Processing*, pp. 1-7, 2018.

[8] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-11, 2016.

[9] A. Pandey and D. Wang, "Densely Connected Neural Network with Dilated Convolutions for Real-Time Speech Enhancement in The Time Domain", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 6629-6633, 2020.

[10] Tom O'Malley, Arun Narayanan and Quan Wang, "A Universally-Deployable ASR Frontend for Joint Acoustic Echo Cancellation, Speech Enhancement and Voice Separation", *Proceedings of International Conference on Audio and Speech Processing*, pp. 1-9, 2022.

[11] Y. Luo and N. Mesgarani, "Conv-Tasnet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 27, No. 8, pp. 1256-1266, 2019.

[12] Y. Luo and J. Yu, "Music Source Separation with Band-Split RNN", *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1-7, 2023.

[13] J. Han, Y. Long, L. Burget and J. Cernock, "Dpccn: Densely-Connected Pyramid Complex Convolutional Network for Robust Speech Separation and Extraction", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 7292-7296, 2022.

[14] Xavier Anguera, Chuck Wooters and Javier Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 7, pp. 2011-2021, 2007.

[15] Yan-Hui Tu, Jun Du and Chin-Hui Lee, "An Efficient Encoder-Decoder Architecture with Top-Down Attention for Speech Separation", *Journal of Signal Processing Systems*, pp. 963-973, 2018.

[16] C. Li, J. Shi, W. Zhang, A. Shanmugam and X. Chang, "Espnet-SE: End-To-End Speech Enhancement and Separation Toolkit Designed for ASR Integaration", *Proceedings of International Conference on Audio and Speech Processing*, pp. 1-11, 2020.

[17] Ethan Manilow, Prem Seetharaman and Bryan Pardo, "The Northwestern University Source separAtion Library", *Proceedings of International Conference on Society for Music Information Retrieval*, pp. 297-305, 2018.

[18] Jinfang Zeng, Chao Li, Jiamei Huang and Wei Li, "Temporal Convolutional Network for Acoustic Echo Cancellation in Double-Talk Scenarios", *Proceedings of International Conference on* Acoustic *Signals Processing*, pp. 897-906, 2023.

[19] E. Indenbom, N.C. Ristea, A. Saabas, T. Parnamaa and J. Guzvin, "Deep Model with Built-in Self-Attention Alignment for Acoustic Echo Cancellation", *Proceedings of International Conference on Audio and Speech Processing*, pp. 1-10, 2022.

[20] R. Giri, S. Venkataramani, J.M. Valin, U. Isik and A. Krishnaswamy, "Personalized PercepNet: Real-Time, LowComplexity Target Voice Separation and Enhancement", *Proceedings of International Conference on Interspeech*, pp. 1124-1128, 2021.

[21] S.E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen and X. Huang, "Personalized Speech Enhancement: New Models and Comprehensive Evaluation", *Proceedings of International Conference on Audio and Speech Processing*, pp. 356-360, 2022.

[22] M. Thakker, S.E. Eskimez, T. Yoshioka and H. Wang, "Fast Real-Time Personalized Speech Enhancement: End-to-End Enhancement Network (E3Net) and Knowledge Distillation", *Proceedings of International Conference on Interspeech*, pp. 991-995, 2022.

[23] Q. Wang, I.L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika and A. Gruenstein, "VoiceFilter-Lite: Streaming Targeted Voice Separation for on Device Speech Recognition", *Proceedings of International Conference on Interspeech*, pp. 2677-2681, 2020.

[24] J.M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani and A. Krishnaswamy, "A Perceptually-Motivated Approach for LowComplexity, Real-Time Enhancement of Fullband Speech", *Proceedings of International Conference on Interspeech*, pp. 2482-2486, 2020.

[25] S. Zhang, Z. Wang, Y. Ju, Y. Fu, Y. Na, Q. Fu and L. Xie, "Personalized Acoustic Echo Cancellation for Full-Duplex Communications", *Proceedings of International Conference on* Acoustic *Signals Processing*, pp. 1-7, 2022.

[26] Sefi Emre Eskimez, Takuya Yoshioka, Alex Ju, Min Tang, Tanel Parnamaa and Huaming Wang, "Real-Time Joint Personalized Speech Enhancement and Acoustic Echo Cancellation", *Proceedings of International Conference on* Acoustic *Signals Processing*, pp. 1-9, 2023.

[27] D. Snyder, D. Garcia-Romero, D. Povey and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification", *Proceedings of International Conference on Interspeech*, pp. 999-1003, 2017.

[28] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S.E. Eskimez, M. Thakker, T. Yoshioka and H. Gamper, "ICASSP 2022 Deep Noise Suppression Challenge", *Proceedings of International Conference on Audio and Speech Processing*, pp. 9271-9275, 2022.

[29] J. Kearns, "LibriVox: Free Public Domain Audiobooks", *Reference Reviews*, Vol. 28, pp. 1-8, 2014.

[30] J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal and M. Ritter, "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events", *Proceedings of International Conference on Audio and Speech Processing*, pp. 776-780, 2017.

[31] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter and X. Serra, "Freesound Datasets: A Platform for the Creation of Open Audio Datasets", *Proceedings of International Conference on International Society for Music Information Retrieval*, pp. 486-493, 2017.

[32] V. Christophe, Y. Junichi and M. Kirsten, "CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit", *The Centre for Speech Technology Research*, pp. 1-8, 2016.

[33] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sørensen and R. Aichner, "ICASSP 2022 Acoustic Echo Cancellation Challenge", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 9107-9111, 2022.

[34] C.K. Reddy, V. Gopal and R. Cutler, "DNSMOS P. 835: A Nonintrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors", *Proceedings of International Conference on Audio and Speech Processing*, pp. 1-9, 2021.

[35] M. Purin, S. Sootla, M. Sponza, A. Saabas and R. Cutler, "AECMOS: A Speech Quality Assessment Metric for Echo Impairment", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 901-905, 2022.

[36] 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", *Proceedings of International Conference on Interspeech*, pp. 4748-4752, 2021.

[37] L. Ma, S. Yang, Y. Gong, X. Wang and Z. Wu, "Echofilter: End-to-End Neural Network for Acoustic Echo Cancellation", *Proceedings of International Conference on Audio and Speech Processing*, pp. 1-6, 2021.