JOINT AEC AND BEAMFORMING WITH DOUBLE-TALK DETECTION USING RNN-TRANSFORMER

Vinay Kothapally[*], *Yong Xu*[†], *Meng Yu*[†], *Shi-Xiong Zhang*[†], *Dong Yu*[†]

[*]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, TX, USA
[†]Tencent AI Lab, Bellevue, WA, USA
[*]vinay.kothapally@utdallas.edu, [†]{lucayongxu, raymondmyu, auszhang, dyu}@tencent.com

ABSTRACT

Acoustic echo cancellation (AEC) is a technique used in full-duplex communication systems to eliminate acoustic feedback of far-end speech. However, their performance degrades in naturalistic environments due to nonlinear distortions introduced by the speaker, as well as background noise, reverberation, and double-talk scenarios. To address nonlinear distortions and co-existing background noise, several deep neural network (DNN)-based joint AEC and denoising systems were developed. These systems are based on either purely "black-box" neural networks or "hybrid" systems that combine traditional AEC algorithms with neural networks. We propose an all-deep-learning framework that combines multi-channel AEC and our recently proposed self-attentive recurrent neural network (RNN) beamformer. We propose an all-deep-learning framework that combines multi-channel AEC and our recently proposed self-attentive recurrent neural network (RNN) beamformer. Furthermore, we propose a double-talk detection transformer (DTDT) module based on the multi-head attention transformer structure that computes attention over time by leveraging frame-wise double-talk predictions. Experiments show that our proposed method outperforms other approaches in terms of improving speech quality and speech recognition rate of an ASR system.

***Index Terms*—** acoustic echo cancellation, speech enhancement, deep learning, neural beamforming

1. INTRODUCTION

With an increasing demand for hands-free communication between speakers in two distant locations (far-end and near-end), effective communication necessitates high-quality audio transmission [1, 2]. Far-end speakers, on the other hand, tend to receive modified versions of their speech as feedback (far-end echo) due to acoustic coupling between the loudspeaker and the microphone locations at the near-end speaker, resulting in reduced speech intelligibility [3, 4]. To improve overall communication quality, an AEC system aims to remove far-end speech captured by the microphone at the near-end while preserving speech from the near-end speaker before transmission. Many AEC systems based on digital signal processing (DSP) have used linear and nonlinear adaptive filters to address this issue for more than two decades [5, 6, 7, 8, 9, 10]. Nonetheless, in practical scenarios involving nonlinear distortions caused by loudspeakers, the presence of reverberation and background noise at the near-end, and double-talk conditions, their performance in suppressing only far-end speech was insufficient.

Recent advances in deep learning have shown the potential to improve the performance of many speech processing systems. As a result, hybrid systems combining traditional adaptive filters and DNNs [11, 12, 13] have been proposed to address nonlinear distortions from loudspeakers by suppressing residual far-end speech from the adaptive filter output. For example, Speex and WebRTC [14, 15] have been combined with RNNs in [16]. Furthermore, multi-task networks were used to design AEC systems with the secondary task of detecting double-talk scenarios in order to avoid suppressing near-end speech in double-talk scenarios [17, 18, 19]. Later, advanced networks such as complex-valued DNNs [20, 21], Long Short Term Memory networks (LSTM), and multi-head self-attention [17, 22] were used to develop AEC systems to also compensate for the time lag between far-end speech and microphone captured signal alongside handling nonlinear distortions and double-talk. Early attempts at AEC for multi-channel speech systems included using single-channel AEC on individual microphones, followed by traditional beamforming techniques [23, 24]. Later, end-to-end DNN-based approaches were proposed for multi-microphone AEC systems [25].

In this work, we propose a two-stage joint AEC and beamformer with explicit deep learning based AEC and beamforming modules, as opposed to a "black-box" approach. The three contributions listed below contribute to the proposed system's overall performance: (i) we propose using a joint spatial covariance matrix computed using microphone signals and far-end speech as input features, which accounts for cross-correlation between far-end speech and multiple microphones, essential for designing an efficient multi-channel AEC system, (ii) we extend our recently proposed generalized spatio-temporal RNN beamformer (GRNNBF) [26] to a joint spatio-temporal RNN AEC beamformer (JRNN-AEC-BF) for handling AEC and beamforming simultaneously using original and AEC processed signals. We see that by doing so, JRNN-AEC-BF can learn a better beamforming solution from target speech and noise covariance matrices that accumulate the correlations between original and AEC processed signals, and (iii) we propose a double-talk detection transformer (DTDT) module based on the multi-head attention transformer structure [27], that computes attention over time while leveraging double-talk detection to suppress far-end residuals.

The remainder of the paper is organized as follows: The signal model for the joint AEC and beamformer task is introduced in Section 2, and the proposed system is described in Section 3. Section 4 describes the dataset and experimental setup. In Section 5, we report on the improvements achieved across various metrics. Finally, we conclude in Section 6.

2. SIGNAL MODEL

We consider the problem of enhancing near-end speech picked up by M-microphones in presence of reverberation, far-end echoes from the loudspeaker, and background noise. Let $x(t)$ and $s(t)$ represent the clean speech from far-end and near end speakers respec-

This work was done while V. Kothapally was an intern at Tencent.

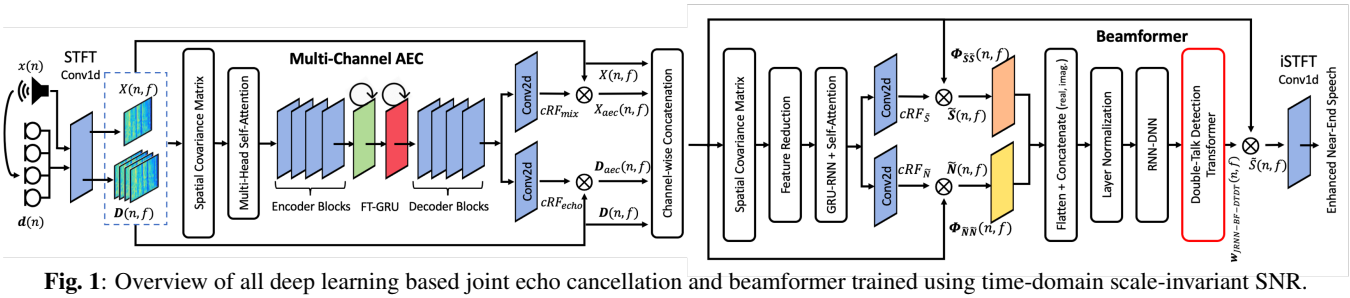

Fig. 1: Overview of all deep learning based joint echo cancellation and beamformer trained using time-domain scale-invariant SNR.

tively. The signals captured by an M-channel microphone array, $\mathbf{d}(t)$ (termed as "mixture") at time 't' can be represented as,

$$\mathbf{d}(t) = \mathbf{s}_r(t) + \tilde{\mathbf{x}}_r(t) + \mathbf{v}(t) \quad \in \mathbb{R}^{M\times 1} \tag{1}$$

$$\text{where, } \mathbf{s}_r(t) = \mathbf{h}_s(t) * s(t) \quad \text{and} \quad \tilde{\mathbf{x}}_r(t) = \mathbf{h}_x(t) * f_{\text{NL}}(x(t))$$

are the received reverberant copies of near-end and loudspeaker emitted nonlinearly distorted far-end speech components, f_{NL} is a function that mimics the loudspeaker's nonlinearities, $\mathbf{h}_s(t)$ and $\mathbf{h}_x(t)$ are M-channel room impulse responses (RIRs) from near-end speaker and the loudspeaker locations to the microphone array, '$*$' denotes the convolution, and $\mathbf{v}(t)$ represents the background noise.

$$\hat{s}_r(t) = \mathbf{\Psi}(\mathbf{y}(t)) = \mathbf{\Psi}([\mathbf{d}(t),\ x(t)]^T) \tag{2}$$

This study aims at designing a joint AEC and beamforming network ($\mathbf{\Psi}$) as a supervised speech enhancement system to suppress the far-end echo and background noise while preserving the embedded near-end speech using mixture and clean far-end signals. However, as stated in Eq.(2), the proposed network is limited to estimating reverberant near-end speech, $\hat{s}_r(t)$ and does not include dereverberation to estimate anechoic near-end speech, $\hat{s}(t)$.

3. PROPOSED SYSTEM OVERVIEW

This section describes proposed joint AEC and beamformer. As illustrated in Fig.1, the proposed system comprises of two stages: (i) a deep learning-based multi-channel AEC, and (ii) joint spatio-temporal RNN AEC beamformer (JRNN-AEC-BF) with double-talk detection transformer (DTDT) module trained using joint spatial covariance matrix as input features.

3.1. Joint Spatial Covariance Matrix

The proposed network is provided with stacked M-channel mixture and single-channel far-end signals, denoted as $\mathbf{y}(t)\in\mathbb{R}^{(M+1)\times T}$ where 'T' represents the number of samples. The audio samples are first transformed to frequency domain, $\mathbf{Y}(n,f)$ using one-dimensional convolution layers that employ a short-time fourier transform (STFT) operation. Here, 'n'$\in[0,N)$, and 'f'$\in[0,F)$ represents frame index and frequency bin. In general, multi-channel speech systems are trained using either stacked complex spectrum [25, 28] or log-power spectra (LPS) and interaural phase difference (IPD) features derived from mixture signals [26, 29]. Rather, we propose using the joint spatial covariance matrix $\mathbf{\Phi}_{\mathbf{y}}(n,f)\in\mathbb{C}^{(M+1)\times(M+1)}$ as input features. This accounts for cross-correlation between the far-end speech and the microphone(s), as well as inter-microphone phase delays, which are crucial in designing an efficient multi-channel AEC system.

$$\mathbf{Y}(n,f) = [\mathbf{D}(n,f),\ X(n,f)]^T;\ \bar{\mu}_y = \sum_{i=1}^{M+1} Y_i(n,f) \tag{3}$$

$$\mathbf{\Phi}_{\mathbf{y}}(n,f) = \big(\mathbf{Y}(n,f) - \bar{\mu}_y\big)\big(\mathbf{Y}(n,f) - \bar{\mu}_y\big)^H \tag{4}$$

We compute the joint spatial covariance matrix as shown in Eq.(3) & (4), where 'i' represents the channel number, '$\bar{\mu}_y$' represents the mean across all channels, and $(\cdot)^H$ represents Hermitian operation. We discard the upper half of the complex symmetrical matrix to reduce computational cost and memory usage. Since the spatial features include information about reverberation time, ambient noise, speakers, time delay(s), and far-end signal attenuation, we employ multi-head self-attention over time to dynamically emphasize relevant features which maximize the system's learning ability.

3.2. Joint AEC and Beamformer

3.2.1. Stage-I: Multi-Channel AEC

The first stage, multi-channel AEC is a deep convolutional recurrent neural network (DCRNN) [30] with two encoders, two decoders, and a frequency-time gated recurrent units (FT-GRU) with residual connections. Similar to [31, 28], FT-GRU comprises of two recurrent units with fully connected (FC) networks. The first GRU network scans all frequency bins to summarize spectral information from encoded features, U_{in}. Later, the output layer activations are reshaped and fed to the second GRU network which examines correlations over time producing U_{out}, see Eq.(5). Here, $(\cdot)^{\text{Tr}}$ represents transpose operation performed on time-frequency dimensions.

$$\text{FT-GRU}\begin{cases} Z_{out} = \big(U_{in} + \text{FC}\big(\text{GRU}\big(U_{in}[:,f,n]\big)\big)\big)^{\text{Tr}} \\ U_{out} = \big(Z_{out} + \text{FC}\big(\text{GRU}\big(Z_{out}[:,n,f]\big)\big)\big)^{\text{Tr}} \end{cases} \tag{5}$$

The decoder of the proposed multi-channel AEC system estimates $((2K+1)\times(2L+1))$ dimensional complex-valued ratio filters [32, 29] $\text{cRF}_{\text{mix}}(n,f)$ for mixture and $\text{cRF}_{\text{echo}}(n,f)$ for far-end signals, respectively. Eq.(6) demonstrates the computation of applying the estimated $\text{cRF}_{\text{mix}}(n,f)$ on time-frequency shifted version of mixture signals $\mathbf{D}(n,f)$ to produce far-end echo suppressed mixture signals $\mathbf{D}_{\text{aec}}(n,f)$.

$$\mathbf{D}_{\text{aec}}(n,f) = \sum_{\tau_1\in[-K,K],\tau_2\in[-L,L]} \text{cRF}_{\text{mix}}(n,f,\tau_1,\tau_2) * \mathbf{D}(n+\tau_1, f+\tau_2) \tag{6}$$

$$\mathbf{Y}_{\text{aec}}(n,f) = [\mathbf{D}_{\text{aec}}(n,f),\ X_{\text{aec}}(n,f)]^T \tag{7}$$

Similar computations are carried on far-end speech $X(n,f)$ to produce time-aligned signal $X_{\text{aec}}(n,f)$. Later, the processed signals $\mathbf{D}_{\text{aec}}(n,f)$, and $X_{\text{aec}}(n,f)$ are channel-wise concatenated to produce multi-channel AEC output, $\mathbf{Y}_{\text{aec}}(n,f)$, see Eq.(7)

3.2.2. Stage-II: Beamforming

In the second stage, we use our proposed joint spatio-temporal RNN AEC beamformer (JRNN-AEC-BF) to spatially filter co-existing far-end residuals and background-noise. This is adaptation to our previous work generalized spatio-temporal RNN beamformer (GRNN-BF) [26] which predicts frame-wise beamforming weights $\mathbf{w}_{\text{GRNN-BF}}(n,f)\in\mathbb{C}^M$ from frame-level target speech and noise covariance matrices estimated using complex spectrograms of only mixture signals. In the current work, proposed JRNN-AEC-BF uses mixture, far-end and the AEC processed signals to learn a better beamforming solution $\mathbf{w}_{\text{JRNN-AEC-BF}}(n,f)\in\mathbb{C}^{2\times(M+1)}$ from target speech and noise covariance matrices $\mathbf{\Phi}_{\widetilde{\mathbf{S}}\widetilde{\mathbf{S}}}(n,f)$ & $\mathbf{\Phi}_{\widetilde{\mathbf{N}}\widetilde{\mathbf{N}}}(n,f)$

estimated using channel-wise stacked mixture, far-end, and AEC processed complex spectrograms, $\widetilde{\mathbf{Y}}(n,f)$.

Similar to the precious stage, we extract joint spatial covariance matrix of the stacked inputs $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(n,f) \in \mathbb{C}^{2(M+1)\times 2(M+1)}$ to serve as features for the JRNN-AEC-BF, see Eq.(9). We discard the upper half and use one-dimensional convolution layers to project the features to lower dimension to reduces computational cost and memory usage. These low-dimensional features are then fed to multi-head self-attention to emphasize relevant features.

$$\widetilde{\mathbf{Y}}(n,f) = [\mathbf{Y}(n,f),\ \mathbf{Y}_{\text{aec}}(n,f)]^T \quad \bar{\mu}_{\widetilde{y}} = \sum_{i=1}^{2(M+1)} \widetilde{Y}_i(n,f) \quad (8)$$

$$\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(n,f) = \big(\big(\widetilde{\mathbf{Y}}(n,f) - \bar{\mu}_{\widetilde{y}}\big)\big(\widetilde{\mathbf{Y}}(n,f) - \bar{\mu}_{\widetilde{y}}\big)^H\big) \quad (9)$$

These extracted features are then passed through to one-dimensional convolution layers to estimate complex ratio filters [32] $\text{cRF}_{\widetilde{\mathbf{S}}}(n,f)$ and $\text{cRF}_{\widetilde{\mathbf{N}}}(n,f)$. As shown in Eq.(10), we employ $\text{cRF}_{\widetilde{\mathbf{S}}}(n,f)$ on stacked input spectrogram $\widetilde{\mathbf{Y}}(n,f)$ to estimate multi-channel target speech and noise signals, $\widetilde{\mathbf{S}}(n,f)$ and $\widetilde{\mathbf{N}}(n,f)$. To this end, we use Eq.(11) to compute frame-wise speech covariance matrix $\mathbf{\Phi}_{\widetilde{\mathbf{S}}\widetilde{\mathbf{S}}}(n,f)$ from multi-channel speech signals.

$$\widetilde{\mathbf{S}}(n,f) = \sum_{\tau_1\in[-K,K],\tau_2\in[-L,L]} \text{cRF}_{\text{S}}(n,f,\tau_1,\tau_2)\ *\ \widetilde{\mathbf{Y}}(n+\tau_1, f+\tau_2) \quad (10)$$

Similar to [26], we use layer normalization with learnable affine transforms to replace the the conventional mask normalization. Similar computations are carried out to estimate multi-channel noise signals $\widetilde{\mathbf{N}}(n,f)$ and frame-wise noise covariance matrix $\mathbf{\Phi}_{\widetilde{\mathbf{N}}\widetilde{\mathbf{N}}}(n,f)$. The real and imaginary parts of frame-wise speech and noise covariance matrices are concatenated and fed to a unified RNN-DNN ($\mathcal{F}_{\text{RD}}$) that predicts frame-level beamforming weights as,

$$\mathbf{\Phi}_{\widetilde{\mathbf{S}}\widetilde{\mathbf{S}}}(n,f) = \text{LayerNorm}\big(\widetilde{\mathbf{S}}(n,f)\widetilde{\mathbf{S}}(n,f)^H\big) \quad (11)$$

$$\mathbf{w}_{\text{JRNN-AEC-BF}}(n,f) = \mathcal{F}_{\text{RD}}\big(\big[\mathbf{\Phi}_{\widetilde{\mathbf{S}}\widetilde{\mathbf{S}}}(0{:}n,f), \mathbf{\Phi}_{\widetilde{\mathbf{N}}\widetilde{\mathbf{N}}}(0{:}n,f)\big]\big) \quad (12)$$

Finally, we employ the proposed double-talk detection transformer (DTDT) module, which attends to frames corresponding to near-end speech using double-talk detection and adjusts the computed weights to suppress far-end echo and co-existing background noise.

3.3. Double-Talk Detection Transformer (DTDT) module

A transformer network [27] is a collection of stacked sub-layers that include a multi-head self-attention (MHSA) module, a fully connected feed-forward (FF) network, residual connections, and layer normalization. As shown in Fig.1, the RNN-DNN is followed by a double-talk detection transformer (DTDT) module to further improve the spatio-temporal filtering capability of JRNN-AEC-BF. To avoid confusions, the transformer module in this paper refers to the RNN modified MHSA network. As shown in Eq. (13), the proposed DTDT computes global temporal correlation as,

$$\begin{aligned}
head_i &= \text{Attention}(Q_i, K_i, V_i) = softmax(Q_i K_i^T/\sqrt{d_k})V_i \\
\text{Mid} &= \text{LayerNorm}(V + \text{FF}(\text{Concat}(head_1, \ldots, head_j))) \\
\text{MHSA}(Q,K,V) &= \text{LayerNorm}(\text{Mid} + \text{FF}(\text{Mid})) \\
\mathbf{DTDT}(R) &= \sigma(\text{GRU}(R)) \times \text{MHSA}(R,R,R) \\
\mathbf{W}_{\text{JRNN-AEC-BF-DTDT}}(n,f) &= \mathbf{DTDT}\big(\mathbf{W}_{\text{JRNN-AEC-BF}}(n,f)\big)
\end{aligned} \quad (13)$$

where Q_i, K_i, V_i are the query/key/value transformations of its input (R) for attention heads indexed by i. 'd_k' stands for the hidden layer dimension of 'K_i'. To better fit for the specific double-talk scenario in the AEC problem, an additional gated recurrent unit (GRU) network coupled with sigmoid ($\sigma(\cdot)$) is trained to predict the frame-level double-talk detection. These prediction probabilities are then used to accordingly adjust the beamformer weights to further suppress time-frequency regions with consisting far-end speech.

$$\hat{S}_r(n,f) = \big(\mathbf{W}_{\text{JRNN-AEC-BF-DTDT}}(n,f)\big)^H \widetilde{\mathbf{Y}}(n,f) \quad (14)$$

Finally, the joint AEC and beamformer enhanced speech $\hat{S}_r(n,f)$ is obtained using the DTDT-attended beamformer weights, original mixture, far-end, and AEC processed signals.

4. DATASET AND EXPERIMENTAL SETUP

4.1. Dataset

We simulate multi-channel reverberant and noisy dataset using AISHELL-2 [33] and AEC-Challenge [34] corpus. We generate a total of 10k multi-channel RIRs with random room characteristics using image-source method. Each multi-channel RIR is a set consisting of RIRs from near-end speaker, loud-speaker, and background noise locations to 8-channel linear microphone array measuring 26 cm in length. The reverberation time (RT_{60}) ranges between [0,0.6s] across room configurations. We randomly select RIRs to simulate multi-channel AEC dataset. We use clean and nonlinear distorted versions of far-end speech from AEC-Challenge [34]. The nonlinear distortions include, but are not limited to: (i) clipping the maximum amplitude, (ii) using a sigmoidal function [35], and (iii) applying learned distortion functions. In addition, we include diffused noise with SNRs ranging from [0,40] dB and signal to echo ratio (SER) from [-10,10] dB. For '*Train*', '*Dev*', and '*Test*' of the dataset, a total of 90K utterances, 7.5K utterances, and 2K utterances are generated.

4.2. Experimental Setup

A 512-point STFT is employed with 32 ms Hann window and 16 ms step size to extract complex spectra for mixture and far-end signals. All systems in the study are trained on 4-second chunks with the Adam optimizer and a batch size of 12 to maximize the time-domain scale-invariant source-to-noise ratio (Si-SNR) [36] and minimize the frequency-domain mean square error (MSE), both of which are equally weighted. Initial learning rate is set to 1e-4 with a gradient norm clipped with max norm 10. All systems are designed to have $\sim$8.5M parameters and trained over 30 epochs. The estimated cRFs size in the proposed systems is empirically set to (3x1). In this study, we compare our proposed method to four baseline systems which include: (i) SpeexDSP [14], a purely signal-processing based AEC, (ii) FT-LSTM [28], a purely NN-based single-channel AEC adapted for multi-channel, (iii) GRNN-BF [26], a robust NN-based beamformer, (iv) hybrid models that combine AEC and traditional and NN-based beamformers.

4.3. Evaluation Metrics

The proposed system's performance is compared to other systems on the '*Test*' set using perceptual quality metrics such as PESQ and STOI, as well as objective metrics such as Si-SNR, signal-to-distortion ratio (SDR), and echo return loss enhancement (ERLE). Furthermore, a general-purpose mandarin speech recognition Tencent API [37] is used to test the ASR performance by computing word error rate (WER). As mentioned earlier, the current work only focuses on echo cancellation and denoising without dereverberation. Hence, the reverberant clean signal (near-end speech at the center of the array) is used as the reference signal for both training and evaluating the performance. To validate the improvement in speech intelligibility, PESQ and STOI metrics are computed only during double-talk periods. Similarly, ERLE is computed during periods where only far-end speech is active, providing an accurate measure of echo suppression.

Table 1: Experimental results for different joint AEC and spatial filtering networks across objective evaluation metrics.

Systems/Metrics	PESQ (↑)	STOI (↑)	SiSNR (↑)	SDR (↑)	ERLE (↑)	WER% (↓)
Reverberant clean reference	4.500	1.000	∞	∞	∞	2.190
Mixture (No Processing)	1.708	0.593	-4.275	-3.806	0.00	77.120
SpeexDSP [14]	1.935	0.637	-1.519	-0.590	3.652	44.994
FT-LSTM [28]	2.997	0.839	10.535	11.568	33.055	15.859
GRNN-BF [26]	2.765	0.798	9.530	10.589	34.940	23.805
JRNN-BF	2.872	0.822	10.268	11.233	34.395	17.393
SpeexDSP + JRNN-AEC-BF	2.811	0.810	7.212	10.849	34.100	20.465
FT-LSTM + MVDR	2.853	0.826	7.688	9.791	34.869	13.525
FT-LSTM + JRNN-AEC-BF	3.046	0.847	11.081	11.999	**37.420**	14.028
Proposed JRNN-AEC-BF-DTDT	**3.117**	**0.858**	**11.280**	**12.178**	36.620	**11.392**

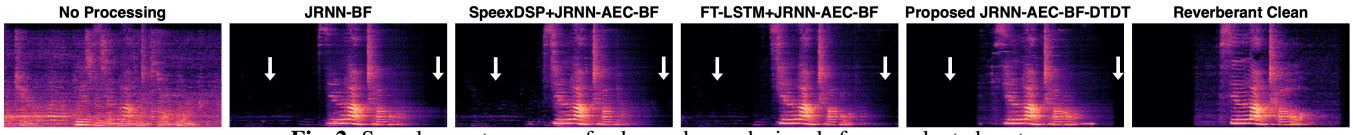

Fig. 2: Sample spectrograms of enhanced speech signals from evaluated systems

5. RESULTS AND DISCUSSIONS

We compare the performance of proposed system to other systems using quality scores, and word error rates on the '*Test*' set in Table.1.

["GRNN-BF vs. JRNN-BF" Beamformer]: We feed far-end signals to GRNN-BF in addition to the mixture signals in order to learn a beamforming solution for the mixture signals only. JRNN-BF is an adaptation GRNN-BF for AEC task, which learns beamforming weights for mixture and far-end signals. As a result, we see that the proposed JRNN-BF improves the performance of GRNN-BF. For example, an average PESQ of 2.76 vs. 2.87; WER: 23.80 vs. 17.93. Likewise, we also see 3%, 8%, and 6% relative improvements in STOI, Si-SNR, and SDR respectively. These findings suggest that estimating beamformer weights for far-end signal alongside mixture allows the network to learn a better beamforming solution.Therefore, further experiments in the study subsequently use JRNN-BF and its adaptations JRNN-AEC-BF for joint AEC and beamforming.

["Hyrbid vs. NN-based" Joint AEC beamformer]: For AEC tasks, we extend JRNN-BF to JRNN-AEC-BF, which employs on multi-channel AEC processed signals in addition to the original mixture and far-end signals. To design a hybrid joint AEC and beamformer system, we combine our proposed JRNN-AEC-BF with SpeexDSP [14], a widely used signal processing-based algorithm for AEC. Similarly, we create a second hybrid model by combining an adapted FT-LSTM for multi-channel applications with a traditional minimum variance distortionless response (MVDR) beamformer [38]. Finally, we compare them with a NN-based joint AEC and beamformer designed by combining FT-LSTM with the proposed JRNN-AEC-BF.

Table-1 shows that, while SpeexDSP has a marginal impact on overall speech quality, it has a greater impact on the performance of speech recognition systems, i.e., for an average PESQ improvement from 1.70 to 1.93, a corresponding improvement in WER from 77.12 to 44.99. Nonetheless, the performance of hybrid system outperforms SpeexDSP. The performance gains were not superior to our proposed JRNN-AEC-BF beamformer. We suspect that because of the severe non-linear distortions in the training We suspect that the linear adaptive filters in SpeexDSP do not converge due to the severe non-linear distortions in the training samples. This increases the range of uncertainty in nonlinear distortions, subsequently lowering the learning ability of JRNN-AEC-BF. The performance degradation can also be observed in in Fig.2. Likewise, while multi-channel adapted FT-LSTM performs well on its own, it does not perform well when combined with traditional MVDR on speech quality metrics. A probable reason for this is that traditional MVDR prioritizes a distortionless response over suppression to preserve the near-end speech. This can be observed from the improvements achieved in WER: 13.25 vs 20.46 and SiSNR:7.68 vs 10.53 over the hybrid system. However, the adapted FT-LSTM when combined with our proposed JRNN-AEC-BF outperforms both hybrid models. For example, an average PESQ of 3.05 vs. {2.81,2.85}; SiSNR: 11.08 vs. {7.21,7.68}' and ERLE: 37.42 vs {34.1,34.87} over hybrid models with the exception on WER which is still in a comparable range. The findings suggest that out proposed JRNN-AEC-BF optimizes well with NN-based AEC systems by including AEC processed signals within alongside mixture and far-end signals.

["Proposed JRNN-AEC-BF-DTDT vs FT-LSTM + JRNN-AEC-BF"]: Our proposed joint AEC and beamforming system differs from the "FT-LSTM + JRNN-AEC-BF" in the following ways: (i) we replace conventional LPS and IPD input features with spatial covariance matrix from Eq.(4) & (9), (ii) we use proposed DTDT module to adjust the estimated beamformer weights to further suppress the far-end via double-talk detection. The proposed system achieves better performance than "FT-LSTM + JRNN-AEC-BF" in terms of quality and speech recognition i.e., PESQ: 3.12 vs 3.04, WER: 11.39 vs 14.03. Likewise, we see {1.7,1.4}% relative improvements in SiSNR and SNR respectively. Fig. 2 also shows that the proposed system can enhance the spectrogram with less residual echo compared to other systems (highlighted with arrows). Although"FT-LSTM+JRNN-AEC-BF" achieves a bit higher ERLE than the proposed, 37.42 vs 36.62, we can conclude that the major contribution to this improvement comes from our proposed JRNN-AEC-BF when compared to "FT-LSTM," 37.42 vs 33.05.

6. CONCLUSION

To conclude, we present an all-deep learning strategy for joint AEC and beamforming with the following major contributions. First, we propose using spatial covariance matrices with multi-head self-attention to learn significant AEC features. Second, we propose JRNN-AEC-BF, a modification of our previous work GRNN-BF, which performs beamforming with mixture, far-end, and AEC processed signals. Finally, we propose DTDT module that predicts double-talk using RNN and adjusts attention weights to compensate for double-talk scenarios. Among systems evaluated, the proposed system achieves the highest objective scores and the lowest WER. Although the proposed system performs well in terms of recognition and echo suppression, we believe addressing dereverberation alongside AEC and beamforming can further improve performance.

7. REFERENCES

[1] Q. Wang and et al., "A frequency-domain nonlinear echo processing algorithm for high quality hands-free voice communication devices," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 10777–10796, 2021.

[2] Gerald Enzner and et al., "Acoustic echo control," in *Academic press library in signal processing*, vol. 4, pp. 807–877. 2014.

[3] J. Benesty and et al., "Advances in network and acoustic echo cancellation," 2001.

[4] E.Hänsler et al., *Acoustic echo and noise control: a practical approach*, vol. 40, John Wiley & Sons, 2005.

[5] C. Paleologu and et al., "An overview on optimized nlms algorithms for acoustic echo cancellation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–19, 2015.

[6] E. Ferrara, "Fast implementations of lms adaptive filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 474–475, 1980.

[7] Y. Park and et al., "Frequency domain acoustic echo suppression based on soft decision," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 53–56, 2008.

[8] G. Clark and et al., "Block implementation of adaptive digital filters," *IEEE Transactions on Circuits and Systems*, vol. 28, no. 6, pp. 584–592, 1981.

[9] J. Soo and et al., "Multidelay block frequency domain adaptive filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[10] A. Schwarz and et al., "Spectral feature-based nonlinear residual echo suppression," in *WASPAA*, 2013, pp. 1–4.

[11] J. M. Valin and et al., "Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet," in *ICASSP*, 2021, pp. 7133–7137.

[12] I. Amir and et al., "Nonlinear Acoustic Echo Cancellation with Deep Learning," in *Interspeech*, 2021.

[13] Z. Wang and et al., "Weighted recursive least square filter and neural network based residual echo suppression for the aec-challenge," in *ICASSP*, 2021, pp. 141–145.

[14] J. Valin, "Speex: A free codec for free speech," *ArXiv*, vol. abs/1602.08668, 2016.

[15] P. Srivastava and et al., "Performance evaluation of speex audio codec for wireless communication networks," in *International Conference on Wireless and Optical Communications Networks*, 2011, pp. 1–5.

[16] L. Ma and et al., "Acoustic echo cancellation by combining adaptive digital filter and recurrent neural network," *arXiv preprint arXiv:2005.09237*, 2020.

[17] L. Ma and et al., "Echofilter: End-to-end neural network for acoustic echo cancellation," *arXiv preprint arXiv:2105.14666*, 2021.

[18] X. Zhou and et al., "Residual acoustic echo suppression based on efficient multi-task convolutional neural network," *arXiv preprint arXiv:2009.13931*, 2020.

[19] C. Zhang and et al., "A robust and cascaded acoustic echo cancellation based on deep learning.," in *Interspeech*, 2020, pp. 3940–3944.

[20] M. M. Halimeh and et al., "Combining adaptive filtering and complex-valued deep postfiltering for acoustic echo cancellation," in *ICASSP*, 2021, pp. 121–125.

[21] R. Peng and et al., "Acoustic Echo Cancellation Using Deep Complex Neural Network with Nonlinear Magnitude Compression and Phase Information," in *Interspeech*, 2021, pp. 4768–4772.

[22] L. Ma and et al., "Multi-scale attention neural network for acoustic echo cancellation," *arXiv preprint arXiv:2106.00010*, 2021.

[23] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *ICASSP*, 1997, vol. 1, pp. 219–222 vol.1.

[24] W. Herbordt and et al., "Joint optimization of lcmv beamforming and acoustic echo cancellation," in *EUSIPCO*, 2004, pp. 2003–2006.

[25] H. Zhang and et al., "A Deep Learning Approach to Multi-Channel and Multi-Microphone Acoustic Echo Cancellation," in *Interspeech*, 2021, pp. 1139–1143.

[26] Y. Xu and et al., "Generalized spatio-temporal rnn beamformer for target speech separation," *Interspeech*, 2021.

[27] Ashish Vaswani, Noam Shazeer, and et al., "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[28] S. Zhang and et al., "F-T-LSTM Based Complex Network for Joint Acoustic Echo Cancellation and Speech Enhancement," in *Interspeech*, 2021, pp. 4758–4762.

[29] Zhuohuang Zhang, Yong Xu, and et al., "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," in *ICASSP*, 2021, pp. 6089–6093.

[30] Ke Tan and DeLiang Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Interspeech*, 2018, pp. 3229–3233.

[31] J. Li and et al., "LSTM time and frequency recurrence for automatic speech recognition," in *2015 IEEE Workshop on ASRU*, 2015, pp. 187–191.

[32] W. Mack and et al., "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2019.

[33] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[34] Kusha Sridhar, Ross Cutler, and et al., "ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *ICASSP*, 2021, pp. 151–155.

[35] Joachim Thiemann, Nobutaka Ito, and et al., "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA*, 2013.

[36] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[37] "Tencent ASR," https://ai.qq.com/product/aaiasr.shtml.

[38] Hakan Erdogan, John R Hershey, and et al., "Improved MVDR beamforming using single-channel mask prediction networks.," in *Interspeech*, 2016, pp. 1981–1985.