MULTI-TASK DEEP RESIDUAL ECHO SUPPRESSION WITH ECHO-AWARE LOSS

Shimin Zhang[1], *Ziteng Wang, Jiayao Sun*[1], *Yihui Fu*[1], *Biao Tian, Qiang Fu, Lei Xie*[1*]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science
Northwestern Polytechnical University, Xi'an, China
shmzhang@npu-aslp.org, lxie@nwpu.edu.cn

ABSTRACT

This paper introduces the NWPU Team's entry to the ICASSP 2022 AEC Challenge. We take a hybrid approach that cascades a linear AEC with a neural post-filter. The former is used to deal with the linear echo components while the latter suppresses the residual non-linear echo components. We use gated convolutional F-T-LSTM neural network (GFTNN) as the backbone and shape the post-filter by a multi-task learning (*MTL*) framework, where a voice activity detection (VAD) module is adopted as an auxiliary task along with echo suppression, with the aim to avoid over suppression that may cause speech distortion. Moreover, we adopt an *echo-aware* loss function, where the mean square error (MSE) loss can be optimized particularly for every time-frequency bin (TF-bin) according to the signal-to-echo ratio (SER), leading to further suppression on the echo. Extensive ablation study shows that the time delay estimation (TDE) module in neural post-filter leads to better perceptual quality, and an adaptive filter with better convergence will bring consistent performance gain for the post-filter. Besides, we find that using the linear echo as the input of our neural post-filter is a better choice than using the reference signal directly. In the ICASSP 2022 AEC-Challenge, our approach has ranked the 1st place on word accuracy (WAcc) (0.817) and the 3rd place on both mean opinion score (MOS) (4.502) and the final score (0.864).

Index Terms— Acoustic echo cancellation, noise suppression, multi-task learning

1. INTRODUCTION

Annoying acoustic echo arises when microphone and loudspeaker are coupled in a real-time communication (RTC) system such that the microphone picks up the desired near-end speech signal plus the unwanted loudspeaker signal. Acoustic echo has become one of the major sources of poor speech quality ratings in RTC [1]. As an indispensable module in RTC, *acoustic echo cancellation* (AEC) aims to remove such unpleasant acoustic echo from the near-end microphone signal while minimizing the distortion of the near-end speaker's speech. More challengingly, a complete AEC system may need to consider noise suppression [2] as well because noise is inevitable in actual acoustic environments. Moreover, AEC is also essential for downstream tasks such as speech recognition [3].

Digital signal processing (DSP) based linear echo cancellation has been adopted in RTC for a long time, which works by estimating near-end speech [4] or acoustic echo path [5] with an adaptive filter. Such algorithms usually give an estimate of the near-end speech and an estimate of the linear echo obtained by cancellation. A post-filter algorithm [6] is usually used to further remove the residual echo that

cannot be completely removed by the linear AEC algorithm. However, these methods are still not quite effective for echo suppression especially in the presence of noise interference.

Recent advances in deep learning have shown great potential in AEC due to the strong non-linear modeling ability of deep neural networks (DNN). Some neural AEC approaches tried to use a neural network directly without any conventional DSP module. Specifically, Zhang and Wang [2] formulated AEC as a supervised speech separation problem, where a bidirectional long-short term memory (LSTM) network was adopted to predict a magnitude mask of the microphone signal. West-hausen et al. [7] extended dual-signal transformation LSTM network (DTLN) [8] by adopting the far-end signal as additional information. Zhang et al. [9] extended deep complex convolution recurrent network (DCCRN) [10], which was originally designed for speech enhancement with superior performance, with frequency-time LSTM (F-T-LSTM) network [11] to better learn the relationship between frequency bands for effectively suppressing echo. In contrast, other methods effectively combine DSP with a neural network without losing the advantages of traditional signal processing in handling the linear component of mixed signal. As a typical approach, Wang et al. [12] cascaded an adaptive filter based linear echo canceller with a deep feedforward sequential memory network (DFSMN) [13] based post-filter. Peng et al. [14] adopted multiple filters for simultaneous linear echo cancellation and time delay estimation (TDE), cascaded further with a gated complex convolutional recurrent neural network (GCCRN) post-filter. The two approaches have shown superior performance in the recent AEC-Challenge series [1, 15].

Recently, processing of super-wideband and full-band speech signal has become more and more popular with the explosion of real-time communication and online collaboration. However, modeling more frequency bands and particularly high frequency components of speech signal is more challenging. The latest ICASSP 2022 AEC Challenge [16] has particularly focused on AEC for *full-band* speech. More challengingly, the performance of downstream speech recognition is also considered as a challenge metric. In other words, the AEC performance is measured by both human and machine auditory perception, i.e., human subjective listening score and machine speech recognition accuracy.

In this paper, we introduce our NWPU Team's entry to the ICASSP 2022 AEC Challenge. We submit a *hybrid* approach that combines a linear AEC with a neural post-filter. Specifically, inspired by our recent work [9], we propose a gated convolutional F-T-LSTM neural network (GFTNN) as the post-filter. More importantly, we shape our neural post-filter by a multi-task learning (*MTL*) framework, where a voice activity detection (VAD) module is adopted as an auxiliary task along with echo suppression. The MTL framework is designed to avoid over suppression and benefit speech recognition accuracy. Moreover, we adopt an *echo-aware*
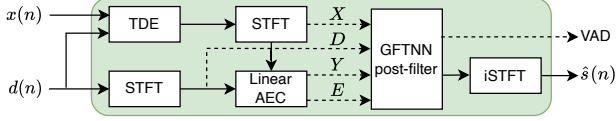
---

*: Corresponding author.

Fig. 1. Diagram of our proposed AEC system.

loss function, where the mean square error (MSE) loss can be optimized particularly for every time-frequency bin (TF-bin) according to the signal-to-echo ratio (SER), leading to further suppression of the echo. Through an extensive ablation study, we also find that the TDE module leads to better perceptual quality, and an adaptive filter with better convergence will bring consistent performance gain for the post-filter. Besides, using the linear echo as the input of our neural post-filter is a better choice than using the reference signal directly. According to the official results of the challenge, our approach has ranked the 1st place on word accuracy (WAcc) (0.817) and the 3rd place on both mean opinion score (MOS) (4.502) and the final score (0.864).

2. PROPOSED METHOD

2.1. Problem formulation

We illustrate the main architecture of the proposed AEC algorithm in Fig. 1. The microphone signal $d(n)$ consists of near-end speech $s(n)$, acoustic echo $z(n)$ and background noise $v(n)$:

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

where n refers to the time sample index. Here $z(n)$ is obtained by the far-end signal $x(n)$ convolved by echo path, with potential nonlinear distortions caused by loudspeakers. The AEC task aims to separate $s(n)$ apart from $d(n)$, on the premise that $x(n)$ is known. The error signal $e(n)$ and linear echo $y(n)$ are generated using $x(n)$ and $d(n)$ by adaptive filter. D, E, X and Y are the frequency domain representation of d, e, x and y, respectively. Note that the dotted lines in Fig. 1 are optional according to different configurations, which will be compared in Section 3. Specifically, the VAD module attached with the post-filter is used as the auxiliary task.

2.2. Split and synthesize

To confine our model with reasonable size and complexity, here we use 3-band finite impulse response (FIR) filter-bank with discrete cosine transform (DCT) modulation [17] to decompose the full-band (48k Hz) signal to subbands and only process the wide-band (16k Hz) signals, and the full-band signal finally is synthesized by average gain control.

As shown in Fig. 2(a), for full-band signal d_{full} and x_{full}, we use band-pass filters to obtain the wide-band signal d and x, where d_h denotes the remaining high-frequency bands of signal (8 to 16k Hz and 16 to 24k Hz). The average gain each frame for d_h is calculated as follows:

$$g(t) = \min\left(\frac{\sum_a^b \hat{S}(t,f)}{\sum_a^b D(t,f)}, \frac{\sum_c^d \hat{S}(t,f)}{\sum_c^d D(t,f)}\right) \tag{2}$$

where t is frame index and f denotes frequency bin index, $\{a = 11, b = 81\}$ and $\{c = 121, d = 161\}$ cover frequency range for 0.5 to 4k Hz and 6 to 8k Hz, respectively.

In the pipeline of the AEC system in Fig. 1, we implement TDE module using Generalized Cross Correlation with PHAse Transform (GCC-PHAT) [18] and the linear AEC using MDF [19] and

Table 1. Configuration of the VAD module. *-Dense means dense layer for the corresponding axis.

Layer Name	Input Size	Hyper Params	Output Size
F-Dense	$T \times C \times F$	(9, 16)	$T \times C \times 16$
Reshape	$T \times C \times 16$	-	$T \times 4C \times 4$
Maxpool1d	$T \times 4C \times 4$	kernel=4, stride=4	$T \times 4C \times 1$
Reshape	$T \times 4C \times 1$	-	$T \times C \times 4$
F-LSTM	$T \times C \times 4$	hidden size=C	$T \times C \times 4$
F-Dense	$T \times C \times 16$	(16,1)	$T \times C \times 1$
C-Dense	$T \times C \times 1$	(C,2)	$T \times 2$

wRLS [12]. Specifically, we make comparison on different adaptive filters which will be explained in detail in Section 3 later. The process of subband synthesis is as follows:

$$\hat{S}_{full}(t,f) = \text{SYN}\big(\hat{S}(t,f), g(t) \cdot D_h(t,f)\big) \tag{3}$$

where $\hat{S}, \hat{S}_{full}$ and D_h are the frequency domain representation of $\hat{s}, \hat{s}_{full}$ and d_h. SYN represents commonly used subband synthesis method in RTC scenarios [17].

2.3. GFTNN post-filter

Fig. 2(b), (c), (d) and (e) show the four sub-modules of the proposed GFTNN-based post-filter, namely GConv, TrGConv, VAD and FTLSTM. $*_r/*_i$ represents the real/imaginary part of a certain signal in frequency domain. The dotted box of the input feature indicates that a certain signal may not be used. In this paper, we explore three types of combinations, namely DX, EX and DEY.

Taking DEY as an example, the input feature $w \in \mathbb{R}^{3 \times n}$, where 3 denotes three signals – $d(n), e(n)$ stacks with $y(n)$. Performing short-time Fourier transform (STFT) on the input feature w, we obtain the complex spectra $W = W_r + jW_i$, where $W \in \mathbb{R}^{6 \times T \times F}$. T denotes the frame number and F denotes the frequency bins. We use the method in [14] to compress w and decompress them after the last layer of the GFTNN post-filter.

The GFTNN encoder consists of 4 GConv layers with same input and output channel for each Conv2d layer, excepting for the first GConv layer which needs to be selected according to the input signal (6 in the given example). 1×1 convolutional layers are used to connect shallow and deep feature representation, which proves to be beneficial according to [20]. The FTLSTM layers (FTLSTMs) are defined in [9]. The output tensor dimension of both the encoder and the FTLSTMs is $\mathbb{R}^{C \times T \times F}$, where C denotes the channel dimension.

The Real/Imag-Decoder consists of 4 TrGConv layers with same input and output channel for each Transpose-Conv2d layer, except that the output channel of the last TrGConv layer needs to be 1, which is used to estimate the real/imaginary part of the clean near-end spectra. $\oplus$ indicates the conv1×1 output concatenated with the previous TrGConv layer output.

Inspired by [21], we use the norm-layer implemented by LSTM as our VAD module. But different from [21], our structure recurs at channel axis, leading to a causal VAD module. The specific details of the VAD module can be found in Fig. 2(d) and Table 1, where $\otimes$ means the point-wise product of tensor $G \in \mathbb{R}^{T \times 1 \times 4 \times C}$ and tensor $H \in \mathbb{R}^{T \times 4 \times 4 \times C}$. The F-LSTM follows the details defined in our previous work [9].

2.4. Loss function

Our loss function is based on the power-law compressed phase-aware (PLCPA) loss [22] and the VAD information. As described
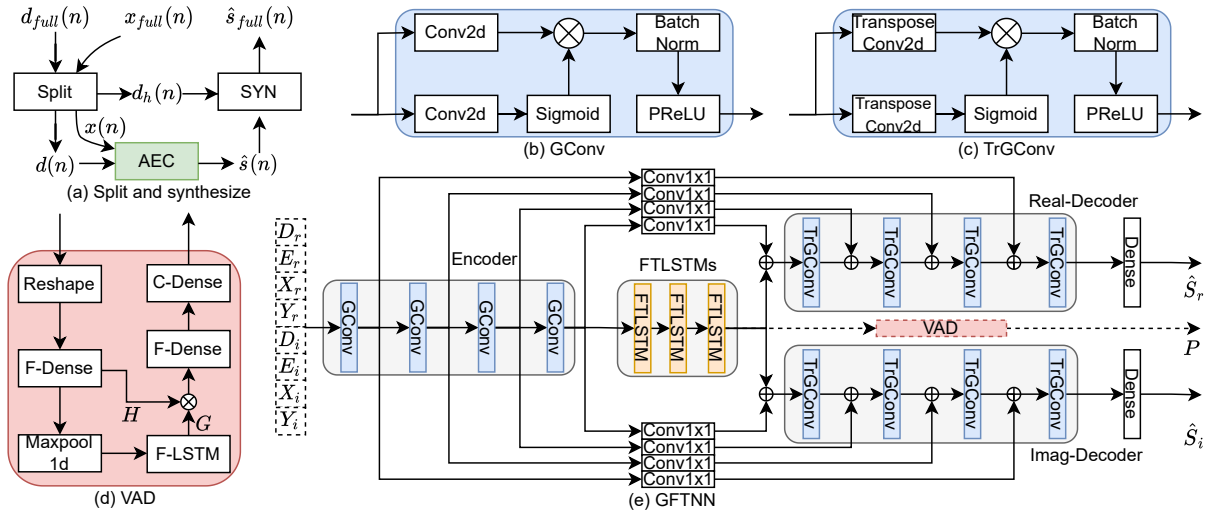
Fig. 2. Full-band signal processing and GFTNN post-filter. (b) GConv: Gated conv-2d. (c) TrGConv: Transpose gated conv-2d.

in [22], PLCPA is beneficial for both ASR accuracy and perceptual quality. It is mainly composed of two parts – amplitude loss $\mathcal{L}_{\text{mag}}$ and phase loss $\mathcal{L}_{\text{pha}}$, defined as follows:

$$\mathcal{L}_{\text{mag}}(t,f) = ||S(t,f)|^P - |\hat{S}(t,f)|^P|^2$$
$$\mathcal{L}_{\text{pha}}(t,f) = ||S(t,f)|^P e^{j\varphi(S(t,f))} - |\hat{S}(t,f)|^P e^{j\varphi(\hat{S}(t,f))}|^2. \quad (4)$$

The VAD loss $\mathcal{L}_{\text{vad}}$ is defined as

$$\mathcal{L}_{\text{vad}} = \text{CrossEntropy}\left(P, \bar{P}\right) \quad (5)$$

where $\bar{P} \in \mathbb{R}^{T \times 1}$ is the near-end speech activity label based on the short-term energy threshold and $P \in \mathbb{R}^{T \times 2}$ is the estimated state of the VAD module.

Echo-aware loss: Previous studies [23] have shown that the use of multiple losses, such as SI-SNR and MSE, are beneficial, where the weight of each loss is adjusted through a hyper-parameter. Inspired by this, in our approach, we adjust the weight of different TF-bins using the proportion of echo power. So we get the echo weighted coefficients defined as

$$\mathcal{W}_{\text{echo}}(t,f) = \frac{|Z(t,f)|^2}{|Z(t,f)|^2 + |S(t,f)|^2}. \quad (6)$$

Weighting the amplitude loss $\mathcal{L}_{\text{mag}}$, we get

$$\mathcal{L}_{\text{echo}} = \frac{1}{T}\frac{1}{F}\sum_T \sum_F \left[\mathcal{L}_{\text{mag}}(t,f)(1 + \mathcal{W}_{\text{echo}}(t,f)) + \mathcal{L}_{\text{pha}}(t,f)\right]. \quad (7)$$

When the echo component of a certain TF-bin is larger, the MSE weight of this TF-bin will also become larger without manual adjustment. The lower the SER, the more likely it is to suppress the echo; the higher the SER, the less the echoes are weighted.

To further improve the amount of echo suppression and ensure the perceptual quality, we make use of the VAD information as follows.

$$\mathcal{W}_{\text{vad}}(t) = \text{GS}(P_t) \cdot W_{\text{fix}}^{\text{T}}$$
$$\mathcal{L}_{\text{mask}} = \frac{1}{T}\frac{1}{F}\sum_T \sum_F ||S(t,f)|^P - |\hat{S}(t,f)|^P \cdot \mathcal{W}_{\text{vad}}(t)|^2, \quad (8)$$

we use the GS (Gumbel SoftMax) as a continuous, differentiable approximation to $\arg\max$ [24]. The fixed weight $W_{\text{fix}} = [0,1]^{1 \times 2}$ and $\mathcal{W}_{\text{vad}} \in \mathbb{R}^{T \times 1}$. $\mathcal{L}_{\text{mask}}$ can punish wrong judgment on VAD of a certain frame, thus reducing near-end over suppression.

The loss function we finally used in the submitted system is

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{echo}} + 0.2 \cdot \mathcal{L}_{\text{mask}} + 0.1 \cdot \mathcal{L}_{\text{vad}}. \quad (9)$$

3. EXPERIMENTS

3.1. Dataset

In our experiments, *train-clean-100* and *train-clean-360* from Librispeech [25], which has a sampling rate of 16k Hz, together with the speech data from DNS-Challenge [26], which has a sampling rate of both 16k and 48k Hz, are used as near-end signal and reference signal. For the noise signal, we use the noise data from DNS-Challenge, which has a sampling rate of both 16k and 48k Hz. For the echo signal, we use all the synthetic echo signals and real far-end single-talk recordings provided by the AEC-Challenge, which covers a variety of voice devices and echo signal delay. Furthermore, we also use the speech from Librispeech and DNS dataset to simulate echo data by convolving with simulated room impulse response (RIR). We use the HYB method described in [27] instead of the image method to generate 20,000 RIRs. The data generation scheme and the specific configuration such as room size of the RIRs are described in [9].

The training set has 860 h data in total, which contains 460 h data with simulated echo and 400 h data with real-recorded echo. Development and test set share the same generating method with training set, which contain 30 h and 15 h data, respectively. The source data among these three sets has no overlap.

3.2. Performance metrics

Echo suppression performance is evaluated in terms of echo return loss enhancement (ERLE) (defined in Eq. (10)) and wide-band perceptual evaluation of speech quality (WB-PESQ) [28] for single-talk (ST) periods and double-talk (DT) periods, respectively. The AEC-Challenge also provides mean opinion score (MOS) results based on the ITU-T Recommendations P.831, P.832 and P.808 [29]. The official ranking is based on the score which reflects both MOS and speech recognition WAcc [16].

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

3.3. Experimental setup

Window length and hop size are 20 ms and 10 ms, respectively, resulting in 30 ms overall algorithmic latency. We apply 320-point

Table 2. Echo suppression performance. DT: double talk, ST: single-talk, NE: near-end, FE: far-end. WB-PESQ used for DT and ST-NE scenarios and ERLE used for ST-FE scenario in both simulated test set (ST-FE) and ICASSP 2022 blind test set (Blind).

Method	TDE	Signal	Loss	DT						ST-NE		ST-FE			Blind	Data
SNR (in dB)				5			$+\infty$			5	$+\infty$	$+\infty$				
SER (in dB)				-5	5	15	-5	5	15	$+\infty$		-5	5	15		
Input	-	-	-	1.42	1.78	1.99	1.58	2.24	2.87	2.10	4.50	0				-
GCCRN	$\times$	DX	$\mathcal{L}_{\text{cmse}}$	1.83	2.25	2.37	2.15	2.79	3.18	2.48	3.55	30.22	30.17	22.47	21.76	
GCCRN	$\times$	EX-M	$\mathcal{L}_{\text{cmse}}$	2.06	2.43	2.58	2.37	2.96	3.36	2.66	3.71	36.03	35.15	28.61	27.25	
GFTNN	$\times$	EX-M	$\mathcal{L}_{\text{plcpa}}$	2.12	2.51	2.66	2.45	3.07	3.46	2.72	3.75	44.63	43.15	39.62	32.57	
GFTNN	$\times$	EX-M	$\mathcal{L}_{\text{echo}}$	2.20	2.59	2.74	2.57	3.22	3.61	2.81	3.77	82.38	76.46	67.30	58.63	460h
GFTNN	$\times$	DEY-M	$\mathcal{L}_{\text{echo}}$	2.27	2.66	2.83	2.62	3.24	3.65	2.91	4.14	96.69	97.52	90.37	80.45	
GFTNN	$\checkmark$	DEY-M	$\mathcal{L}_{\text{echo}}$	2.32	2.67	2.82	2.72	3.30	3.65	2.93	4.35	95.50	95.16	89.29	76.77	
GFTNN	$\checkmark$	DEY-W	$\mathcal{L}_{\text{echo}}$	2.38	2.74	2.85	2.74	3.32	3.70	2.99	4.35	78.24	71.33	64.25	62.32	
GFTNN			$\mathcal{L}_{\text{echo}}$	2.44	2.80	2.96	2.79	3.36	3.73	3.04	4.35	82.36	76.71	67.69	67.98	
GFTNN-VAD	$\checkmark$	DEY-W	$\mathcal{L}_{\text{final}}$	2.43	2.78	2.94	2.78	3.34	3.72	3.01	4.33	81.37	76.28	65.80	70.86	860h
GFTNN-VAD-L			$\mathcal{L}_{\text{final}}$	2.44	2.78	2.94	2.81	3.37	3.73	3.02	4.34	84.96	83.85	77.63	79.28	

Table 3. Subjective ratings in terms of MOS and WAcc for the blind test set of the AEC-challenge. The confidence interval is 0.02. DT-Echo DMOS: echo annoyance DMOS for DT scenario. DT-Other DMOS: other impairments DMOS of DT scenario [29].

Method	ST-NE MOS	ST-FE DMOS	DT-ECHO DMOS	DT-Other DMOS	WAcc	Score
Baseline [16]	4.152	4.563	4.122	3.563	0.659	0.752
GFTNN-VAD-L	**4.238**	**4.801**	**4.711**	**4.257**	**0.817**	**0.864**

STFT to each signal to produce the complex spectra. Chunk size of our generated data is set to 10 s. Hyper-parameter p is set to 0.5 for $\mathcal{L}_{\text{mag}}$ and $\mathcal{L}_{\text{pha}}$. We compare different models with various configurations, as shown in Table 2. Here $\mathcal{L}_{\text{cmse}}$ refers to the loss function used in [14] and $\mathcal{L}_{\text{plcpa}} = \mathcal{L}_{\text{mag}} + \mathcal{L}_{\text{pha}}$. DX means we directly feed the microphone signal and the near-end speech signal into the network, without using any DSP-based linear AEC. *-M/W means using MDF/wRLS as the adaptive filter respectively. All neural models are trained with the Adam optimizer [30] for 60 epochs with an initial learning rate of 1e-4, and the learning rate is halved if there is no loss decrease on development set for 2 epochs. The convolution kernels move with a stride of (1, 2). The number of output channels for each layer in encoder/decoder is 80 for GFTNN and GFTNN-VAD, and 128 for GFTNN-VAD-L. The whole size of the submitted model (GFTNN-VAD-L) is 4.7912M in parameters. Tested on Intel(R) Xeon(R) Platinum 8163 CPU@2.50GHz quad-core machine, the total real time factor (RTF) of our submitted model is 0.1706, which is composed of 0.1306 for the GFTNN post-filter and 0.04 for the DSP part, including sub-band decomposition, adaptive filter and synthesize. Some of the processed audio clips can be found in our demo page[1].

3.4. Results and analysis

From Table 2, we first see that the use of an adaptive filter (linear AEC) is beneficial – directly using a neural network for AEC leads to inferior performance (comparing DX with EX/DEY). We also notice that replacing $\mathcal{L}_{\text{plcpa}}$ with $\mathcal{L}_{\text{echo}}$ improves both PESQ and ERLE a lot, which proves that weighting the MSE by echo is very beneficial. Consistent with [31], replacing the input signals EX with DEY leads to improved PESQ and ERLE in each scenario. In real recordings of ST-NE scenario [16], the reference signal $x(n)$ may be completely irrelevant to microphone signal $d(n)$. This means that the performance of the neural AEC may be affected if the input signals contains X (such as DX and EX), because unlike the DSP

methods, it is difficult for a neural network to find consistent rules for such uncorrelated input combinations. This also explains why the input combination DEY has better performance than the other two combinations. The distortion in the subband decomposition and synthesis process makes the PESQ upper limit to be 4.35. When the SER $= +\infty$ and SNR $= +\infty$, which means the clean signal, the use of the TDE module in the GFTNN post-filter improves the perceptual quality of speech (PESQ $>= 4.3$). Comparing the two linear AEC algorithms (DEY-W vs. DEY-M) in our two-stage framework, wRLS leads to better performance than MDF due to its better convergence. Finally, the larger model GFTNN-VAD-L shows generally superior performance. We process the blind test clips using this model and submit to the challenge.

There are three observations we would like to explain further. First, replacing DEY-M with DEY-W results in almost 20 dB decline in ERLE despite of improvement on PESQ. In fact, there is no obvious difference in subjective listening performance when ERLE is greater than 60 dB. Second, the VAD module does not lead to better PESQ, but we find that many emotional clips in the blind test set are not over-suppressed because of the use of the VAD module. The better perceptual listening performance on these clips indicates that the network can effectively learn to distinguish speech frame from non-speech frame. Finally, we find that when $x(n)$ is uncorrelated with $d(n)$, the complex-valued neural network will cause distortion to the estimated $\hat{s}(n)$. Hence in this paper, we no longer use complex-valued structure.

The officially released results on the blind test set in Table 3 show that our method significantly outperforms the challenge baseline with absolute 0.158 *WAcc* gain and 0.112 *score* gain. In other words, our system ranked the 1st place in WAcc and the 3rd place in mean opinon score (MOS) and the final *score*.

4. CONCLUSIONS

This paper introduces our linear AEC + neural post-filter system submitted to the ICASSP 2022 AEC challenge. With the help of the specifically designed network structure GFTNN, multi-task learning with VAD and echo-aware loss function, our proposed AEC system can achieve better echo cancellation and noise suppression performance while ensuring that the near-end speech is not over suppressed. Our submitted system manages to rank the 3rd place in the challenge with good subjective quality (MOS) and speech recognition accuracy (WAcc). In future work, we will explore whether our echo-aware weighting loss can be transferred to related tasks like target speaker extraction (TSE), and explore more cascade schemes.

5. REFERENCES

[1] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *ICASSP*. IEEE, 2021, pp. 151–155.

[2] H. Zhang and D. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," *Training*, vol. 161, no. 2, pp. 322, 2018.

[3] T. O'Malley, A. Narayanan, Q. Wang, A. Park, J. Walker, and N. Howard, "A Conformer-based ASR frontend for joint acoustic echo cancellation, speech enhancement and speech separation," *arXiv preprint arXiv:2111.09935*, 2021.

[4] E. Ferrara, "Fast implementations of LMS adaptive filters," *ICASSP*, vol. 28, no. 4, pp. 474–475, 1980.

[5] G. Enzner and P. Vary, "Frequency-domain adaptive kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, no. 6, pp. 1140–1156, 2006.

[6] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*, vol. 40, John Wiley and Sons, 2005.

[7] N. Westhausen and B. Meyer, "Acoustic echo cancellation with the dual-signal transformation LSTM network," in *ICASSP*. IEEE, 2021, pp. 7138–7142.

[8] N. Westhausen and B. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," in *INTERSPEECH*. ISCA, 2020, pp. 2477–2481.

[9] 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 *INTERSPEECH*. ISCA, 2021, pp. 4758–4762.

[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 *INTERSPEECH*. ISCA, 2020, pp. 2472–2476.

[11] J. Li, A. Mohamed, G. Zweig, and Y. Gong, "LSTM time and frequency recurrence for automatic speech recognition," in *ASRU*, 2015, pp. 187–191.

[12] Z. Wang, Y. Na, Z. Liu, B. Tian, and Q. Fu, "Weighted recursive least square filter and neural network based residual echo suppression for the AEC-Challenge," in *ICASSP*. IEEE, 2021, pp. 141–145.

[13] S. Zhang, M. Lei, Z. Yan, and L. Dai, "Deep-FSMN for large vocabulary continuous speech recognition," in *ICASSP*. IEEE, 2018, pp. 5869–5873.

[14] R. Peng, L. Cheng, C. Zheng, and X. Li, "Acoustic echo cancellation using deep complex neural network with nonlinear magnitude compression and phase information," *INTERSPEECH*, pp. 4768–4772, 2021.

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

[16] 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.

[17] Fredric J. Harris, *Multirate Signal Processing for Communication Systems*, Prentice Hall PTR, USA, 2004.

[18] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[19] J. Soo and K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Trans. Signal Process.*, vol. 38, no. 2, pp. 373–376, 1990.

[20] J. Silva-Rodríguez, M. Dolz, M. Ferrer, A. Castelló, V. Naranjo, and G. Piñero, "Acoustic echo cancellation using residual U-Nets," *arXiv preprint arXiv:2109.09686*, 2021.

[21] S. Birnbaum, V. Kuleshov, Z. Enam, P. Koh, and S. Ermon, "Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations," *NeurIPS*, 2019.

[22] S. Eskimez, X. Wang, M. Tang, H. Yang, Z. Zhu, Z. Chen, H. Wang, and T. Yoshioka, "Human listening and live captioning: Multi-Task training for speech enhancement," *arXiv preprint arXiv:2106.02896*, 2021.

[23] Y. Fu, Y. Liu, J. Li, D. Luo, S. Lv, Y. Jv, and L. Xie, "Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation," *arXiv preprint arXiv:2111.06015*, 2021.

[24] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *ICLR*. 2017, OpenReview.net.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

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

[27] E. Bezzam, R. Scheibler, C. Cadoux, and T. Gisselbrecht, "A study on more realistic room simulation for far-field keyword spotting," in *APSIPA ASC*. IEEE, 2020, pp. 674–680.

[28] A. W Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*. IEEE, 2001, vol. 2, pp. 749–752.

[29] R. Cutler, B. Nadari, M. Loide, S. Sootla, and A. Saabas, "Crowdsourcing approach for subjective evaluation of echo impairment," in *ICASSP*. IEEE, 2021, pp. 406–410.

[30] D. Kingma and Jimmy B., "Adam: A method for stochastic optimization," in *ICLR*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[31] Z. Wang, Y. Na, B. Tian, and Q. Fu, "NN3A: Neural network supported acoustic echo cancellation, noise suppression and automatic gain control for real-time communications," *arXiv preprint arXiv:2110.08437*, 2021.