CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS FOR ACOUSTIC ECHO CANCELLATION

Fran Pastor-Naranjo*, Rocío del Amor*, Julio Silva-Rodríguez†, Miguel Ferrer‡, Gema Piñero‡, and Valery Naranjo*

*Instituto de Investigación e Innovación en Bioingeniería, I3B, Universitat Politècnica de València, Valencia, Spain

† Inst. Transport and Territory, Universitat Politècnica de València, Valencia, Spain

‡ ITEAM, Universitat Politècnica de València, Valencia, Spain

Email: {madeam2, vnaranjo}@i3b.upv.es

***Abstract*—This paper presents an acoustic echo canceller based on a conditional Generative Adversarial Network (cGAN) for single-talk and double-talk scenarios. cGANs have become a popular research topic in the audio processing area because of their ability to reproduce the finest details of audio signals. However, to the best of our knowledge, no previous works have used cGANs for Acoustic Echo Cancellation (AEC) in an end-to-end manner. The generator of the proposed cGAN framework is composed of a U-Net model able to synthesise the echo-free signal. The synthesised signal, conditioned by the estimated echo signal, is the input to the discriminator. The discriminator aims to refine the synthesised signals to convert them as realistic as possible. Experimental results have been carried out where the proposed cGAN has been compared to a GAN and a U-Net model in terms of echo return loss enhancement (ERLE) and perceptual evaluation of speech quality (PESQ) score for different values of Signal to Echo Ratio (SER). The cGAN outperforms the other two models for both ERLE and PESQ, and presents PESQ scores comparable to previous high-ranked echo cancellers of the AEC Challenge 2021.**

***Index Terms*—Acoustic echo cancellation, deep learning, U-Net, conditional GANs**

I. Introduction

The first Acoustic Echo Cancellation (AEC) systems were developed by AT&T Bell Labs [1] and were based on the model shown in Fig.1, where the AEC block was implemented as an adaptive filter. The aim of AEC systems is to suppress the undesired echo produced by the acoustic coupling between a loudspeaker and a microphone, usually driven by the same device. Traditionally, echo cancellation is accomplished by identifying the room impulse response (RIR) between the loudspeaker and the microphone denoted by $h(n)$ in Fig.1, and subtracting the estimated echo signal $x(n)*h(n)$ from the microphone signal $y(n)$ in order to obtain the clean speech $\hat{s}(n)$. In practice, the canceller must also deal with non-stationary scenarios, non-linear effects over the signals and possibly with ambient noise denoted by $v(n)$ in Fig.1. To combat these impairments, residual echo suppressors (RES) [2] and voice activity detectors for the far-end and near-end signals [3] are usually included in the AEC system design.

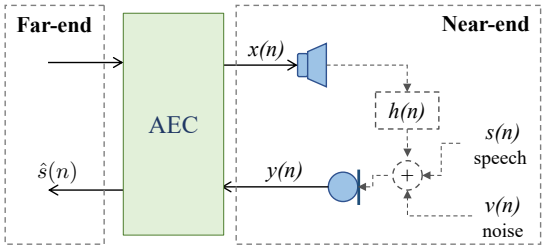


Fig. 1: Model of an acoustic echo cancellation system.

Recently, AEC systems based on artificial intelligence techniques have outperformed classical methods based on adaptive filters, specially when dealing with non-linear scenarios. In 2021 and 2022 Microsoft have been proposed three AEC Challenges [4] where competing AEC systems have been evaluated against the same blind dataset by means of a perceptual score [5]. Among the best solutions presented, on the one hand there were works that used neural networks (NN) in combination with adaptive cancellers to suppress the residual echo [6]–[8], and, on the other hand, there were complete end-to-end AEC solutions using deep learning models [9], [10]. Both types of models were very diverse, such as Gated Recurrent Units (GRU) [6], Deep Feedforward Sequential Memory Networks (Deep-FSMN) [7], Gate Complex Convolutional Recurrent Networks (GCCRN) [8] and Long Short Term Memory (LSTM) networks [9], [10].

Apart from the solutions presented to the AEC Challenges, other cancellers based on NN have been recently presented [11]–[14]. In [11], the authors propose a multiscale attention neural network for echo cancellation as an end-to-end implementation, which combines temporal convolutions to transform waveform into spectrograms and attention blocks to obtain features using LSTM units. In [12], a U-Net with multiple encoders is built to suppress the residual echo. Additionally, in [13], a very small network based on a densely-connected multidilated DenseNet (D3Net) is implemented to work in real time, eliminating the need of pooling thanks to

This work has been partially supported by grants RTI2018-098085-B-C41 (MCIU/AEI/FEDER) and PROMETEO/2019/109. Julio Silva-Rodríguez and Rocío del Amor work have also been supported by the Spanish Government under FPI [PRE2018-083443] and FPU Grant [FPU20/05263], respectively.

their building blocks. More recently, the usage of Generative Adversarial Networks (GANs) for echo cancellation has been studied [14]. However, in this case, the generator, based on a simple autoencoder, generates a TF mask used in a second step to resynthesize the enhanced signal. Therefore, the GAN developed in [14] is not able to map a noisy speech to a clean signal in an end-to-end way.

Nevertheless, there is no definite AEC solution yet and the problem remains open. Our contribution in this work is the implementation of an end-to-end AEC system based on a conditional Generative Adversarial Network (cGAN). The cGAN is an improved GAN version that has outperformed the state-of-the-art results for different tasks such as speech enhancement [15], font recognition and generation [16] and image generation [17], among others. However, to the best of our knowledge, this is the first time that a cGAN is used for acoustic echo cancellation. This sort of model is composed by two blocks: the generator, in charge of processing the signal recorded by the microphone $y(n)$ and synthesising it into a new signal without echo, $\hat{s}(n)$, and the discriminator, which ensures and improves the quality of the synthesised signal $\hat{s}(n)$. This last block determines if the input signal is real or synthesised, what prevents the generator to give unrealistic outputs, as will be detailed in the following section.

II. Methods

The methodological core of the proposed end-to-end AEC system is the cGAN shown in Fig.2. A detailed description of the different blocks and processes is given in the following.

A. Audio signal processing

In the context of the AEC system shown in Fig.1, the microphone signal $y(n)$ can be expressed as a mixture of the near-end speech $s(n)$ and the far-end speech $x(n)$ as:

$$y(n) = s(n) + f(x(n)) + v(n)\,, \tag{1}$$

where $v(n)$ denotes the ambient noise and $f(x(n)) = d(n)$ is the echo signal. Eq. (1) describes a generic double-talk scenario when $s(n) \neq 0$, or, alternatively a single-talk scenario when $s(n) = 0$. Regarding the echo $f(x(n))$, it is usually assumed to be a linear function such that $d(n) = x(n) * h(n)$. However, we will also consider non-linear effects on $x(n)$ as clipping or impairments due to time-varying RIRs.

As shown in Fig.2, the input to the cGAN are segments of 155ms long of the microphone signal, $y(n)$, and the far-end signal, $x(n)$ (highlighted with a red box at the bottom left of Fig.2). The first 135ms sample block corresponds to previous information, the following 10ms block conforms the frame of interest, while the last 10ms block is posterior information. Subsequently, the respective Short Time Fourier Transforms (STFT) of the 155ms segments are obtained using a Hanning window of $w = 318$ samples and 75% overlap. Then, normalised spectrograms $Y(k, m)$ and $X(k, m)$ of size 160×32 are obtained from $y(n)$ and $x(n)$, respectively. The modules of the complex spectrograms are used as input features to the proposed model. Note that during the inference stage, the normalization process is inverted, obtaining the predicted near-end speech spectrogram module, $|\hat{S}(k, m)|$. Then, the phase of $y(n)$ is used to calculate $\hat{s}(n)$ as the inverse STFT of $|\hat{S}(k, m)|e^{\angle j\, Y(k,m)}$. From every prediction, just the 10ms frame of interest out of the 155ms segment is selected to build the estimated near-end speech $\hat{s}(n)$. In the case of single-talk scenarios where $s(n)$ is not present, the output of the model would not have any speech content, although we also denote their spectrogram as $\hat{s}(n)$ for the sake of clarity. To simplify the notation, we will omit the module operator when referring to the spectrograms along the rest of the paper.

B. cGAN for echo cancellation

Given $\mathcal{T} = \{(S_1, X_1, Y_1), (S_2, X_2, Y_2), ..., (S_N, X_N, Y_N)\}$, consisting of N triplets of clean spectrogram (S), noisy spectrogram (Y) and far-end spectrogram (X), the problem of echo cancellation is to find a mapping $f(Y) : Y \longmapsto S$. Conforming to GAN's principle, the cGAN proposed has its generator (G) tasked for the echo-free mapping and a discriminator (D) to refine the synthesised signals. However, in the proposed framework, G and D receive some additional conditioning input information to control the output. Once presented Y together with the echo representation X (the conditional information), G produces the enhanced signal $\hat{S} = G(Y, X)$. Then, the discriminator (D) receives a pair of signals as input, $\{S \cup X\}$ and $\{\hat{S} \cup X\}$, and D learns to classify the pair (S, X) as real and ($\hat{S}$, X) as fake, while G tries to fool D such that D classifies ($\hat{S}$, X) as real.

*1) **cGAN Generator for signal synthesis**:* The generator is performing an image-to-image translation task. Usually, auto-encoder models are used for this problem, but, due to downsampling, a lot of information can be lost. Additionally, image information flow passes through all the layers, including the bottleneck. Thus, sometimes, many unwanted redundant features (inputs and outputs are sharing a lot of the same pixels) are exchanged. For this reason, we employ skip connections following the structure of a Residual U-Net [18]. The Residual U-Net is composed of two branches: encoder and decoder.

As shown in Fig.2, the spectrograms of the near-end microphone (Y) and the far-end speech (X) are used as inputs in a two-channel grayscale image of size $160 \times 32 \times 2$. The encoder branch consists of stacked convolutional blocks with residual connections, forcing the network to focus on temporally-close correlations in the input signal. These connections have improved deep learning models optimization, avoiding gradient vanishing problems in other U-Net applications. Regarding pooling operation between the convolutional blocks, it is performed only in the frequency-related dimension using filters of size 2×1. Due to the small size of the input time frames, a temporal pooling could soften the spectrogram and degrade the speech representation.

In the decoder branch, the convolutional blocks are deconvolutions that progressively recover the spatial dimension. The G network also features skip connections, connecting each encoding layer to its homologous decoding layer and

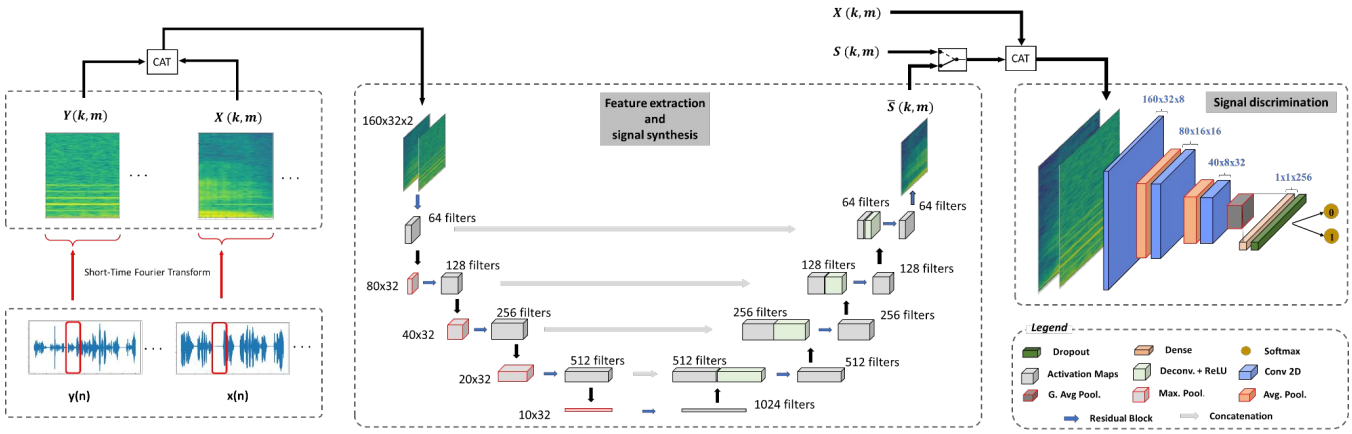


Fig. 2: Proposed framework to perform acoustic echo cancellation. The microphone (Y) and far-end (X) spectrograms are concatenated and fed into the generator (*Feature extraction and signal synthesis*). Subsequently, the predicted spectrogram ($\hat{S}$) is concatenated to the far-end spectrogram (X) and fed to the discriminator (*Signal discrimination*).

passing the fine-grained information of the input spectra to the decoder. In addition, they offer a better training behavior, as the gradients can flow deeper through the whole structure. Finally, a convolutional layer of size 1×1 reconstructs the estimated near-end speech spectrogram, $\hat{S}$.

*2) **cGAN Discriminator (synthesis vs real)**:* The discriminator is in charge of transmitting information to G of what is real and what is fake. In this way, Gcan slightly correct its output towards the realistic distribution, getting rid of the echo components as those are signaled to be fake. Therefore, D can be expressed as learning some sort of loss for G's output to look real. In this case, D is composed of a Convolutional Neural Network (CNN) classifier with two inputs, the predicted spectrogram by the generator ($\hat{S}$) and the conditional information (X). $D : \{\hat{S} \cup X\} \rightarrow Z$ maps the generated spectrograms $\hat{S}$ to an embedding vector Z, where the classification stage is addressed in a lower-dimensional space. At the end of the convolutional network, a softmax-activated dense layer is applied to address the classification (fake vs real).

*3) **cGAN optimization**:* The generator and discriminator are trained in an adversarial way where both networks play a mini-max game, and try to maximize their own utility function. The generator tries to fool the discriminator by producing samples which are very close to the samples from the training data, and the discriminator tries to be good at classifying real and fake data.

On the one hand, during the generator training stage, learning is driven by the error between the clean and predicted spectrograms and the loss provided by the discriminator every iteration. Let us denote the spectrogram module of the target signal as $S(k, m)$, where m denotes the frame index, $m = 1, \ldots, M$, and k is the frequency bin. The loss function that measures the error between the spectrograms is defined as the root mean square error (RMSE):

$$\mathcal{L}_{RMSE} = \sqrt{\frac{1}{M(w/2+1)} \sum_{m=1}^{M} \sum_{k=0}^{w/2} \left[\hat{S}(k,m) - S(k,m)\right]^2}. \tag{2}$$

The chosen discriminator loss function is the categorical cross-entropy and it is expected to predict the pair $\{\hat{S}, X\}$ as real:

$$\mathcal{L}_D = \sum_{m=1}^{M} D(\hat{S}, X) \cdot \log_{10} \hat{a}\,, \tag{3}$$

where $\hat{a}$ denotes the discriminator ground-truth. Thus, the objective function used in the back propagation is a weighted combination of both previous losses:

$$\mathcal{L}_G = \mathcal{L}_{RMSE} + \alpha \cdot \mathcal{L}_D\,. \tag{4}$$

On the other hand, during the discriminator training stage, learning process is just driven by the discriminator loss every second iteration. The loss function remains being categorical cross-entropy (3) but it is expected to predict the pair $\{\hat{S}, X\}$ as fake and $\{S, X\}$ as real.

III. Experiments and Results

A. Dataset

The chosen dataset to fit our model is the *synthetic dataset* released by Microsoft for the 2021 AEC Challenge [19]. It consists of 10,000 synthetic samples comprising single-talk, double-talk, near-end noise, far-end noise, and various nonlinear distortion situations. Each sample includes a far-end speech, an echo signal, a near-end microphone signal and the near-end speech that will be considered the ground truth S in (2)-(3). Further details on each generated sample are available in the github repository[1]. For the 10,000 synthetic samples, the SER was uniformly distributed between -10 and 9 dB in steps of 1 dB. For model evaluation the 400 first

[1]https://github.com/microsoft/AEC-Challenge.

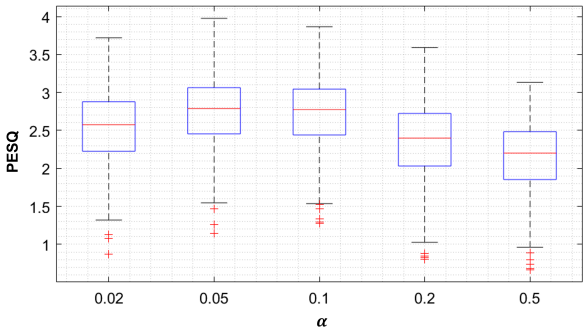


Fig. 3: Ablation study on discriminator loss. Hyperparameters study for α in (4) based on the PESQ score.

samples have been selected, as it was recommended in the Challenge. The rest of recordings, excluding those first 400, make up the training data.

B. Performance metrics

We have used the echo return loss enhancement (ERLE) and the perceptual evaluation of speech quality (PESQ) [20] as the performance metrics. The ERLE evaluates the echo reduction achieved during single-talk periods when the near-end speech is not present as

$$\text{ERLE} = 10\log_{10}\frac{E[y^2(n)]}{E[\hat{s}^2(n)]} \approx 10\log_{10}\frac{\sum_n y^2(n)}{\sum_n \hat{s}^2(n)}, \quad (5)$$

where the statistical expectation operation $E[\cdot]$ is estimated by the mean square value over the whole speech duration. The PESQ is an objective measure of the quality of the estimated near-end speech $\hat{s}(n)$ compared to that of the original speech $s(n)$, thus it will be used for double-talk scenarios. The PESQ scores range from -0.5 to 4.5, and a higher score indicates a better speech quality.

C. Ablation experiments

In order to optimize the cGAN parameters, we have carried out several experiments. Using the training setting, we have cross-validated different values of $\alpha = \{0.02, 0.05, 0.1, 0.5\}$ in (4) and the corresponding PESQ scores of their synthesised speech signals $\hat{s}(n)$ have been obtained and depicted in Fig.3.

These results show that the inclusion of the discriminator loss term in the whole training improves the performance of the synthesised speech. Nevertheless, using a too large slope once the performance is satisfied can lead to a worsening of the results. Thus, we have selected $\alpha = 0.05$ due to its good PESQ performance in Fig.3. The rest of the parameters' values used to train our model are: 130 epochs with batch size of 128, step $\mu = 5 \cdot 10^{-4}$ for the generator and step $\mu = 5 \cdot 10^{-5}$ for the discriminator. Both steps suffer an exponential declination from epoch 80.

D. Experimental Results

After training our model with the mentioned dataset and the optimized parameters, we have evaluated our methodology in the test set by means of the PESQ and ERLE metrics.

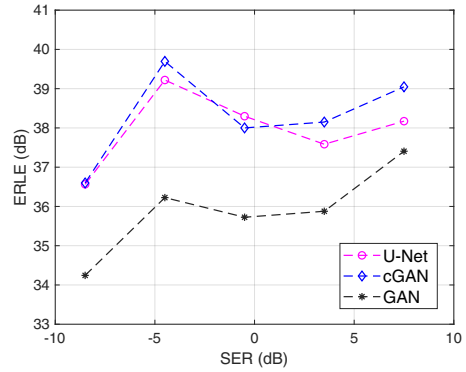
TABLE I: PESQ and ERLE scores for different scenarios. All: Whole testset; FN: Far-end Noisy; FnN: Far-end not Noisy; NN: Near-end Noisy; NnN: Near-end not Noisy; N+FnL: Both Noisy + Far-end not Linear.

Metrics	Models	Testing parameters					
		All	**FN**	**FnN**	**NN**	**NnN**	**N+FnL**
PESQ	GAN	2.68	2.66	2.7	2.6	2.75	2.58
	U-Net	2.72	2.71	2.74	2.65	2.79	2.63
	cGAN	**2.76**	**2.74**	**2.78**	**2.69**	**2.82**	**2.66**
ERLE	GAN	34.24	33.73	34.78	33.62	34.82	33.43
	U-Net	35.94	35.57	36.32	35.36	36.47	35.46
	cGAN	**36.31**	**36.11**	**36.51**	**35.85**	**36.73**	**36.28**

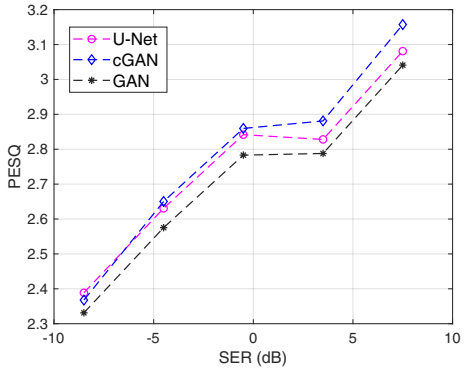
Additionally, to assess the performance improvement of the proposed cGAN, we have compared the results with two baseline models: a Residual U-Net and a GAN. The core structure is the same for all the networks, maintaining the number of filters and layers of the generator. In the case of the Residual U-Net, we omit the discriminator block so it only consists of a generator. In contrast, in the GAN, the discriminator input is a one-channel image, so the far-end spectrogram, X, is not used as the condition. The ERLE and PESQ values for the three models are shown in Table I for different scenarios and for the whole range of SER values between -10 dB and 9 dB. Table I shows that the best performance in both metrics is achieved by the cGAN for every particular scenario.

As a further discussion on the results shown in Table I, we can compare the PESQ scores obtained by our models to those reported by [8], [9], [21], whose AEC solutions obtained the fourth, third and second position, respectively, in the "Acoustic Echo Cancellation Challenge - ICASSP 2021" [4]. It was not possible to compare our model with the best one [6] because they do not measure their performance with an implementable metric. First, the authors in [8] use 100 utterances randomly chosen from the *synthetic dataset* of Microsoft as their test set, reporting a PESQ of 2.61, which is lower than the PESQ values obtained by our models in the "All" column of Table I. Second, Westhausen et al. [9] use an alternative test set also provided by Microsoft (called *test set* in [19]), which presents similar characteristics to the *synthetic dataset*. Their PESQ scores are 2.53 for near-end noisy speech (comparable to the PESQ values in the "NN" column of Table I), 2.73 for far-end noisy speech (comparable to the values in the "FN" column), and 2.43 for both noisy signals (comparable to the values in the "N+FnL" column). Third, in [21] the authors use the first 500 clips from *synthetic dataset* to evaluate their model, reaching a PESQ score of 2.07. It can be appreciated that the PESQ achieved by our cGAN framework outperforms the previous AEC solutions for all conditions.

Finally, we have studied the influence of the signal-to-echo ratio in the performance of our model gathering the PESQ and ERLE values in uniform sets according to their SER: low SER in the range of $[-10, -7]$ dB, medium-low in the range of $[-6, -3]$ dB, medium SER in the range of $[-2, 1]$ dB, medium-high in the range of $[2, 5]$ dB and high SER in the



(a)



(b)

Fig. 4: Average ERLE and PESQ values for different SER intervals of the test set.

range of $[6, 9]$ dB. Fig.4 shows the average ERLE (top) and the average PESQ (bottom) of the synthesised signals for every set, where the SER value in Fig.4 is taken as the center point of the interval. It can be appreciated that the cGAN values for both PESQ and ERLE metrics are higher than the other models for most of the SER intervals. Nevertheless, the PESQ and ERLE gains obtained by the cGAN with respect to the U-Net is higher in medium to high values of SER, whereas for low to low-medium SER values the discriminator cannot improve the signal obtained at the generator output. Regarding the GAN behaviour, the U-Net and cGAN clearly outperform the GAN for all SER inervals. It is remarkable that the GAN performance is worse than that of the U-Net, what shows the importance of building a robust conditional discriminator.

IV. Conclusions

In this paper, we present an AEC system based on a conditional Generative Adversarial Network (cGAN). For the model training and testing, we have used a synthetic dataset provided by Microsoft. We have evaluated the performance of this architecture both in single-talk and double-talk scenarios by means of ERLE and PESQ metrics, respectively. We have also compared the results of our proposal with U-Net and GAN methods, and state-of-the-art works showing the benefits of our model.

References

[1] M. M. Sondhi, "An adaptive echo canceller," *The Bell System Technical Journal*, vol. 46, no. 3, pp. 497–511, 1967.

[2] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Proc.*, vol. 64, no. 1, pp. 21–32, 1998.

[3] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Proc. Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[4] M. Research, "Acoustic echo cancellation challenge - ICASSP 2021 [Online]," Available at https://www.microsoft.com/en-us/research/academic-program/acoustic-echo-cancellation-challenge-icassp-2021/, accessed on March 4th, 2022 2022.

[5] B. Naderi and R. Cutler, "An open source implementation of ITU-T recommendation P.808 with validation," *arXiv preprint arXiv:2005.08138*, 2020. [Online]. Available: http://arxiv.org/abs/2005.08138

[6] J. Valin, S. Tenneti, K. Helwani, U. Isik, and A. Krishnaswamy, "Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet," in *Proc. of the 2021 IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2021, pp. 7133–7137.

[7] 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 *Proc. of the 2021 IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2021, pp. 141–145.

[8] R. Peng, L. Cheng, C. Zheng, and X. Li, "ICASSP 2021 acoustic echo cancellation challenge: Integrated adaptive echo cancellation with time alignment and deep learning-based residual echo plus noise suppression," in *Proc. of the 2021 IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2021, pp. 146–150.

[9] N. Westhausen and B. Meyer, "Acoustic echo cancellation with the dual-signal transformation lstm network," in *Proc. of the 2021 IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2021, pp. 7138–7142.

[10] H. Zhang and D. Wang, "A deep learning approach to multi-channel and multi-microphone acoustic echo cancellation," in *Proc. of Interspeech 2021*, 2021, pp. 1139–1143.

[11] L. Ma, S. Yang, Y. Gong, and Z. Wu, "Multi-scale attention neural network for acoustic echo cancellation," *arXiv:2106.00010*, 2021.

[12] E. Kim, J.-J. Jeon, and H. Seo, "U-convolution based residual echo suppression with multiple encoders," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Proc. (ICASSP)*. IEEE, 2021, pp. 925–929.

[13] K. Watcharasupat, T. Nguyen, W.-S. Gan, S. Zhao, and B. Ma, "End-to-end complex-valued multidilated convolutional neural network for joint acoustic echo cancellation and noise suppression," *arXiv preprint arXiv:2110.00745*, 2021.

[14] Y. Zhang, C. Deng, S. Ma, Y. Sha, H. Song, and X. Li, "Generative adversarial network based acoustic echo cancellation." in *INTERSPEECH*, 2020, pp. 3945–3949.

[15] S. Routray and Q. Mao, "Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network," *Computer Speech & Language*, vol. 71, p. 101270, 2022.

[16] R. Liu, X. Wang, H. Lu, Z. Wu, Q. Fan, S. Li, and X. Jin, "Sccgan: style and characters inpainting based on cgan," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 3–12, 2021.

[17] X. Cao, "Image-to-image translation with application to biometrics," Master's thesis, Schulich School of Engineering, 2022.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241.

[19] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "Icassp 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Proc. (ICASSP)*, 2021, pp. 151–155.

[20] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862*, 2001.

[21] 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 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Proc. (ICASSP)*. IEEE, 2021, pp. 141–145.