IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 June 2024, accepted 12 August 2024, date of publication 16 August 2024, date of current version 27 August 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3445279*

RESEARCH ARTICLE

# Novel TransQT Neural Network: A Deep Learning Framework for Acoustic Echo Cancellation in Noisy Double-Talk Scenario

**V. SONI ISHWARYA AND MOHANAPRASAD KOTHANDARAMAN**
School of Electronics Engineering (SENSE), VIT, Chennai, Tamil Nadu 600127, India

Corresponding author: Mohanaprasad Kothandaraman (kmohanaprasad@vit.ac.in)

**ABSTRACT** Acoustic echo is a persistent issue in telecommunication that degrades the quality of speech and breaks down communication either entirely or for a period of time; therefore, acoustic echo cancellation (AEC) systems were developed. The demand for AEC has significantly risen after the global pandemic 2020 as the speaker and the listener communicate in unpredictable environments such as home environments where echo and noise significantly disrupt communication. Numerous AEC solutions have been proposed, including adaptive filters and deep learning techniques. However, their effectiveness is notably lowered during double-talk scenarios, where both nearend and farend speakers talk simultaneously, as well as in noisy environments. This paper proposes a novel transQT neural network (TNN), an end-to-end neural network that leverages the constant Q transform (CQT) and transformer-inspired self-attention module to eliminate the echo and noise in double-talk noisy scenarios. Additionally, it utilizes the smooth L1 loss function to enable efficient training and enhance the overall performance of the proposed model. In the proposed TNN, the CQT is used as the front end to convert the signal from time domain to time-frequency domain. The primary aim of CQT is to improve speech quality as it aligns more closely with the human auditory system due to its use of a logarithmic frequency scale. The attention module has been incorporated among the layers of the proposed models to focus on double-talk and noisy parts of speech. It aids the AEC model by making it easier to separate the clean target signal from the parts affected by double-talk and noise. The smooth L1 loss is employed to ensure smooth training and stable and efficient convergence. It is also less sensitive to variability in data, therefore reducing large errors and overall loss. An experimental implementation was conducted for both causal and non-causal scenarios. The proposed TNN model demonstrated superior performance in terms of speech quality, as measured by the perceptual evaluation of speech quality (PESQ) and it also showed a significant reduction of echo, quantified by echo return loss enhancement (ERLE). The performance was further evaluated using the correlation coefficient, which indicates the relationship between the clean and the echo signal.

**INDEX TERMS** Transformers, self-attention, constant Q transform, acoustic echo cancellation, convolutional recurrent neural network, deep learning.

## I. INTRODUCTION

Acoustic echo cancellation (AEC) has always been employed in hands-free communication. However, with the widespread adoption of remote communication due to the COVID-19 pandemic, there is a need to enhance the performance and efficiency of AEC systems, ensuring uninterrupted and clear communication between two or more people. The primary function of AEC is to cancel the acoustic echo, a type of noise signal that occurs when the farend signal reflects off a surface, gets captured by the nearend microphone, and is subsequently transmitted back to the

The associate editor coordinating the review of this manuscript and approving it for publication was Li Zhang.

farend speaker [1]. This echo phenomenon often results in the farend speaker hearing their own voice after a delay, thereby hindering communication. AEC aims to cancel this echo, thereby facilitating clearer communication. The AEC system is essential in various applications, including Voice over Internet Protocol (VoIP) services, smart speakers and virtual assistants like Amazon Echo and Google Home, video conferencing systems, hearing aids, and gaming. The conventional methods for echo cancellation involve adaptive filters, echo barriers, and echo suppressors [2], with adaptive filters being particularly prominent and recognized for their efficiency, stability and broad scope for improvement. The adaptive filters aim to replicate the echo by finding the room impulse response. The estimated echo signal is subtracted from the mixed microphone signal. Figure 1 shows the block diagram of an adaptive filter based acoustic echo cancellation system. In this system, the mixed signal $\mathbf{z}(n)$ is represented as

$$\mathbf{z}(n) = \mathbf{s}(n) + \mathbf{q}(n) + \mathbf{r}(n). \tag{1}$$

where, $\mathbf{s}(n)$ is the nearend signal, $\mathbf{q}(n)$ is the echo signal which is obtained by convolving farend signal $\mathbf{p}(n)$ and room impulse response $\mathbf{w}$ and $\mathbf{r}(n)$ is the added noise. Figure 1 either represents the singletalk scenario where $\mathbf{s}(n) = 0$, the double-talk scenario where $\mathbf{s}(n) \neq 0$, or the noisy double-talk scenario where $\mathbf{s}(n) \neq 0$ and $\mathbf{r}(n) \neq 0$.

$$\hat{\mathbf{s}}(n) = \mathbf{z}(n) - \hat{\mathbf{q}}(n). \tag{2}$$

where, $\hat{\mathbf{s}}(n)$ is the clean target nearend signal, $\hat{\mathbf{q}}(n)$ is the estimated echo signal. The noise r(n) is later removed by using post filters [14], [15]. In real-world scenarios, the AEC encounters challenges such as double-talk, where the farend and nearend speaker are simultaneously active, echo path change, low convergence speed due to high order FIR filter, and nonlinearities [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], leading to a performance decline. A lot of adaptive filter algorithms were devised to solve these issues, like subband filter algorithms [3], [4], [5], affine projection algorithms [6], [7], [8], wavelet-based algorithms [9], [10], Kalman filters [11], variable step size algorithms [12], [13], post filters [14], [15], volterra filters [16], [17], and, kernelized adaptive filters [18], [19]. Despite the development of numerous algorithms to address these challenges, adaptive filters demand careful selection and tuning of parameters such as step size, filter length, and regularization terms, making their performance highly sensitive to these settings. They rely on simple mathematical models, which may not capture complex signal patterns, and are focused specifically on echo cancellation, necessitating separate systems for tasks like noise suppression and speech enhancement. Therefore, a better substitute is needed. As deep learning can solve complicated problems, learn and extract features, handle multiple audio processing tasks at once, and provide a single solution for echo cancellation, noise suppression, and speech enhancement, it has become increasingly popular in recent years.

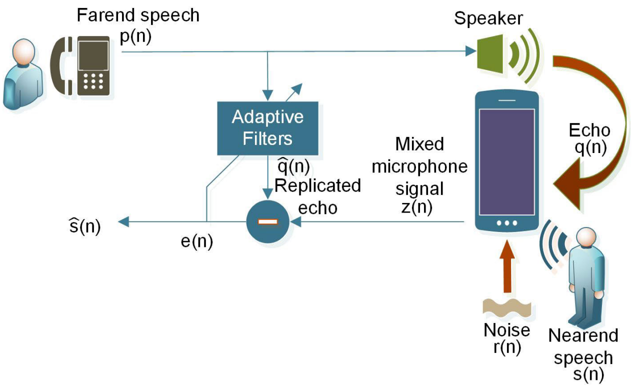


**FIGURE 1.** **Block diagram of traditional method.**

## II. RELATED WORK

In recent years, many deep learning models [21], [22], [23], [24], [25], [26], [27] have addressed the earlier challenges and performed better than traditional methods. The models employ three primary strategies: first, treating the acoustic echo problem as a speech separation issue, where a mask segregates the target signal. Secondly, it can be addressed by modeling the echo path similar to adaptive filters. Thirdly, a combined approach is adopted, integrating both deep learning and adaptive filters, referred to as the cascade model, designed to handle both linear and nonlinear echo scenarios. Predominantly, deep learning models approach the AEC problem as speech separation problem where the deep learning model is trained to predict a ratio mask [20] which is subsequently applied to the mixed signal to obtain the separated target signal. Numerous deep learning models have been developed, including recurrent neural network (RNN) [21], long short term memory (LSTM) [22], gated recurrent unit (GRU) [23] which, leverage temporal features to separate the target signal and, convolution recurrent neural network (CRNN) [24], [25], [26], [27] which employs both spatial and temporal features for the separation task. Despite outperforming the conventional AEC in terms of performance, these models are not without drawbacks. It's enormous and intricate; it is challenging to isolate the clean signal from the mixed signal when noise and double-talk are present, and it is difficult to achieve good speech quality.

Until now, the conventional approach to transforming time to time-frequency (T-F) domain has involved using the short-time fourier transform (STFT). Given the nonlinear nature of the human auditory system, it is essential to utilize a transform that better corresponds to this characteristic, as opposed to the STFT, which has a linear frequency representation. Therefore, STFT can be replaced by constant Q transform. CQT was initially presented by Brown [28], which features a frequency resolution that adapts based on the center frequencies of the windows assigned to each bin. Notably, the center frequencies of the frequency bins are distributed in a geometric, rather than linear fashion, similar to the human auditory system. Numerous works have

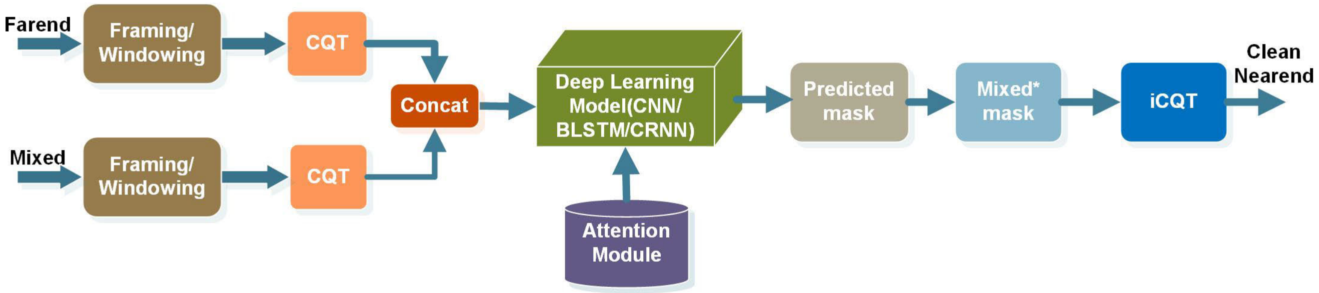


**FIGURE 2.** Block diagram of TransQT neural network.

used CQT as the front end in various audio processing applications [29], [30], [31], [32], [36].

The challenge of dealing with double-talk and noisy environments persists, hindering the existing model's ability to distinguish between clean speech and mixed signals. To address this issue, an attention module inspired by transformers [33] is considered. This module efficiently identifies patterns associated with double-talk and noise, directing the model's focus toward these elements and facilitating the separation of speech signals [34]. It also captures global dependencies so that information from all parts is used to predict the mask. This idea was inspired by the transformer model [33]. Several models, such as those referenced in [39] and [40], employ transformer-based attention networks. These models integrate CNN, LSTM, GRU, and attention layers to separate echo and noise from the noisy signal. However, a significant drawback of these models is their complexity. Additionally, they use the STFT as the front-end processing, which is inefficient for non-linear frequency signals like speech, and can suffer from spectral leakage. These models rely on mean square error (MSE) for loss calculation, which is sensitive to outliers as the square function is used.

Therefore, due to the drawbacks of using STFT and the challenges faced during noisy double-talk scenarios, a novel transQT neural network (TNN) is proposed, which replaces STFT by CQT and addresses the issues faced during the double-talk scenario by adding attention modules. The main modifications made from existing models include replacing STFT with CQT, as speech exhibits non-linear characteristics. Additionally, Smooth L1 loss is used instead of MSE loss in this paper, balancing the advantages of both L1 and L2 losses. This approach maintains sensitivity to small errors (like MSE) while offering robustness to large errors (like MAE). Finally, a simpler model is developed to effectively and simultaneously remove echo and noise. To ensure real-time processing of audio signals in applications like Voice over Internet Protocol (VoIP) and live broadcasting, the AEC system must be causal. This means that the system's output relies exclusively on past and current inputs, without requiring future data. In this paper, the CQT and self-attention module are integrated into three models (CNN, LSTM, and CRNN). Both causal and non-causal systems are implemented, and their results and performance are compared and analyzed against existing models.

The remainder of the section is organized as follows. Section III introduces the proposed transQT neural network (TNN) models, experimental setup and data creation are shown in section IV, results are discussed in section V and finally, the paper concludes in section VI.

## III. PROPOSED METHOD

In this section, the novel transQT neural networks (TNN) in which three deep learning models (CNN, BLSTM [22], CRNN [24]), that are based on constant Q transform and self-attention module is proposed. Figure 2 shows the block diagram of the proposed TNN architecture. Figure 4 shows the expanded TNN's deep learning block which explains CNN, BLSTM and CRNN individually. The TNN model is explained below in the following subsections.

### A. FRONT-END TRANSFORM

For preprocessing, first framing and windowing are done to separate the long speech into short frames with overlapping windows. The proposed model opts for CQT over the conventional STFT to convert the time series signal to a frequency domain. This transform is used due to the limitation of STFT, which exhibits linearity with respect to frequency, a mismatch with the nonlinear characteristics of the human auditory system.

The CQT is a frequency transform that provides a logarithmically spaced frequency axis. CQT of a signal z(n) can be represented as

$$Z^{\mathrm{CQT}}(k) = \frac{1}{N(k)} \sum_{n=0}^{N(k)-1} z(n)\, w(n,k)\, e^{-j\left(\frac{2\pi}{N(k)} Qn\right)} \tag{3}$$

where, window function $w(n)$ is hanning window which has the identical shape for $k_{th}$ frequency component. The window length $N(k)$ varies with respect to $f_k$. With its extremely long window length for lower frequency areas, it offers high frequency resolution and, as a result, aids in the effective capture of low noise and speech [29].

$$Q = \frac{f_k}{\Delta f_k} = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{1/B} - 1} \tag{4}$$

where, $Q$ is the quality factor which is the ratio of the center frequency $f_k$ to the bandwidth of each window $\Delta f_k$, $f_k$ is the center frequency of the $k_{th}$ bin in the transform and $B$ is the number of frequency bins per octave [28].

The primary reason for using CQT to extract features is

- it is less sensitive to pitch and pitch-related features; therefore, gender doesn't affect the performance of the system [31], [32],
- it is represented in logarithmic frequency scale therefore, it is considered to be nonlinear as the frequency bin varies logarithmically when the frequency varies [28], [31], [32],
- it provides high resolution at low frequencies [36],
- it exhibits greater resilience to noise due to its utilization of a more extended window size, making it less susceptible to the influence of isolated noise peaks [30], [31].

The CQT for the farend signal p(n) and mixed signal z(n) is obtained, from which the magnitude of both the signals are calculated. These magnitudes are concatenated and given as input to the deep learning models.

### B. NEURAL NETWORK MODELS

The proposed TNN uses three deep learning models. The first is the CNN model, which uses only 1D convolution layers without fully connected (FC) layers. This model represents a novel approach of using only the convolution layer as both input and output, which is absent in prior works. The significant advantage of this model is that it can capture the spatial and spectral properties of the audio signals and identify different patterns in the speech signal. A non-causal convolution layer can be easily converted to a causal by padding extra zeros in the beginning. For kernel size m, a padding length m-1 is added. The second model is the BLSTM model, which has bidirectional LSTM layers that are a forward and a backward LSTM and a fully connected layer, as proposed by Zhang et al. [22]. Its primary strength is leveraging information from both preceding and subsequent audio frames, enabling it to learn complex temporal dependencies. For causal and non-causal BLSTMs, the forward pass is the same as given in equation 5 and 7, but only the backward pass changes, as shown in equations 6 and 8. To convert the non-causal BLSTM layer to causal, the backward LSTM is applied manually to ensure causality, processing up to the current time step $t$ and flipping the sequence for each time step.

Non-causal:

$$\overrightarrow{h}_t = LSTM_{forward}(x_t, \overrightarrow{h}_{t-1}) \tag{5}$$

$$\overleftarrow{h}_t = LSTM_{backward}(x_t, \overleftarrow{h}_{t+1}) \tag{6}$$

Causal:

$$\overrightarrow{h}_t = LSTM_{forward}(x_t, \overrightarrow{h}_{t-1}) \tag{7}$$

$$\overleftarrow{h}_t = LSTM_{backward}(x_t, \overleftarrow{h}_{t-1}) \tag{8}$$

The linear layer is also made sure to process only the current input for causal system at any given point in time t.

The final and last model combines the CNN and BLSTM layers to form a CRNN architecture, adopted from [24] model. This model's advantage lies in capturing the signals' global temporal and local spatial features together.

An ideal ratio mask (IRM) [20] is used as the training target. The mask-based approach is chosen as it can adapt to varying noise conditions, making the AEC system more robust in real-world scenarios where noise levels and types can fluctuate. These three models are trained to predict an ideal ratio mask, which is subsequently applied to the mixed signal through element-wise multiplication to obtain the clean target signal.

$$IRM = \left(\frac{S(t,f)^2}{Z(t,f)^2}\right)^{\frac{1}{2}} \tag{9}$$

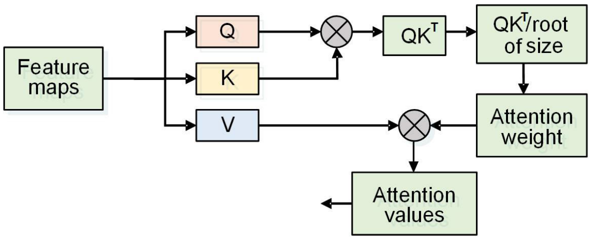


**FIGURE 3. Block diagram of self attention module.**

where, $\mathbf{S}(t,f)$ is the magnitude of the clean nearend target signal and $\mathbf{Z}(t,f)$ is the magnitude of the mixed signal. Here, $(t,f)$ signifies the specific time-frequency bin. The models undergo training utilizing the rectified linear unit (ReLU) and sigmoid activation function, given that the value of the IRM varies between 0 and 1. Batch normalization, employing a batch size of 32, is applied during both training and validation. Given the depth of the models and the increased number of neurons, a dropout of 20% is implemented in the hidden layers.

### C. SELF ATTENTION

Self-attention modules based on the Transformers [33] are added between the hidden layers. Self-attention is also called scaled dot product attention, where the query, key, and values are formed into a matrix Q,K, and V, respectively [31].

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

where, $\mathbf{d_k}$ is the dimension of the queries and keys. Dot product attention is used as it is faster and space efficient [33]. The block diagram of the self-attention module is shown in Figure 3. Here, weights are assigned to different parts of the signal based on their relevance to the task. The inclusion of these modules directs attention to specific segments of the audio signal, focusing particularly on areas associated with double-talk and noisy regions in this context. Attention mechanisms facilitate parallel computation across various segments of the input signal, resulting in faster training and

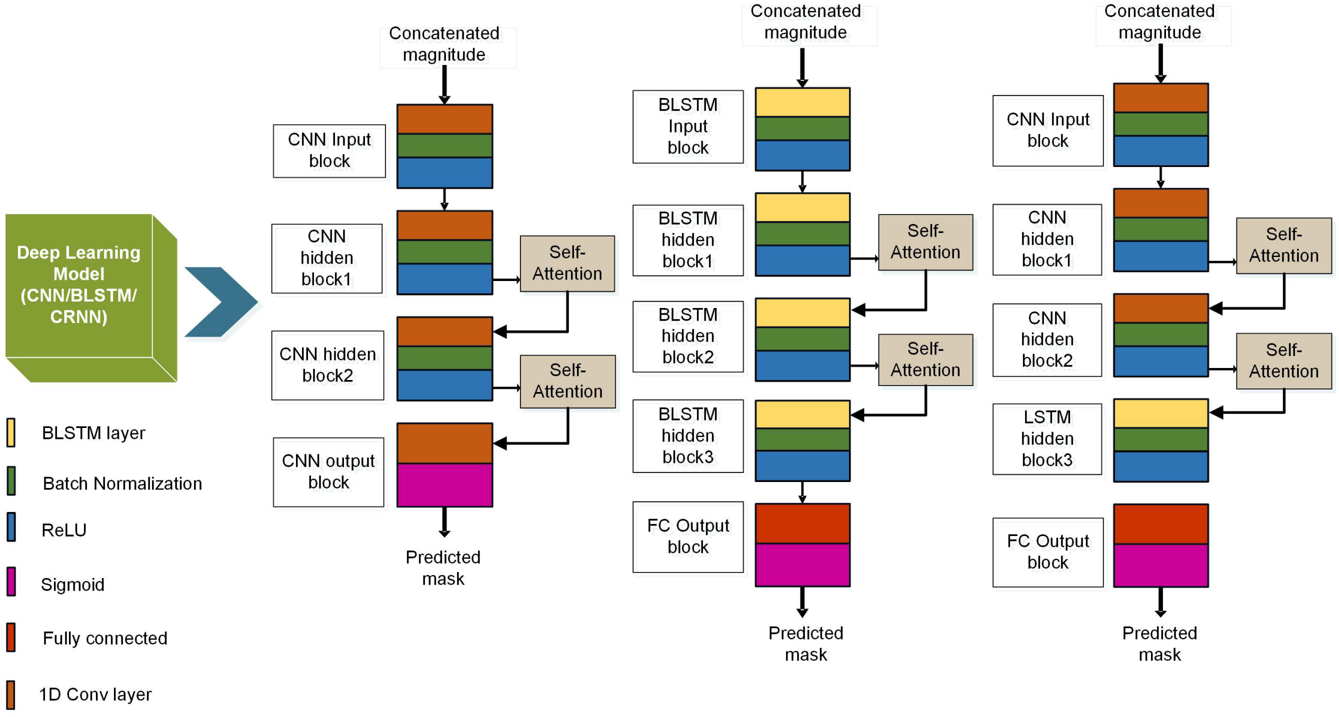


**FIGURE 4.** Expanded block diagram of the deep learning module (CNN(left), BLSTM(middle), CRNN(right)).

**TABLE 1.** Hyper parameters and size of the models.

| Model | Hyper Parameters | Model Size (MB) |
|---|---|---|
| CNN+STFT+Attention | 4054079 | 33.86 |
| CNN+CQT | 8390592 | 69.29 |
| CNN+CQT+Attention | 7964064 | 80.05 |
| BLSTM+STFT+Attention | 6586509 | 59.67 |
| BLSTM+CQT | 19521896 | 124.32 |
| BLSTM+CQT+Attention | 18079496 | 120.3 |
| CRNN+STFT+Attention | 6081068 | 57.8 |
| CRNN+CQT | 23181216 | 133.96 |
| CRNN+CQT+Attention | 13100704 | 111.73 |

inference times. Causal attention can be implemented by applying a mask to the $QK^T$ as given in equation 11, where the elements of the main diagonal are set to a very large negative value (e.g., negative infinity) so that after applying softmax, the values of the future frame becomes zero.

$$Attention_{causal}(Q, K, V) = softmax\left(\frac{Mask(QK^T)}{\sqrt{d_k}}\right)V \quad (11)$$

where,

$$Mask(R)(i,j) = \begin{cases} R(i,j), & i \leq j \\ -\infty, & otherwise \end{cases} \quad (12)$$

where, $R = QK^T$.

## IV. EXPERIMENTAL SETUP

In this section, the data creation process is elucidated, followed by the evaluation of the structure of each model, and the evaluation metrics are explained. This work implements and benchmarks three deep-learning models encompassing CNN, BLSTM, and CRNN, where CQT is used as the front-end transform step to extract features. Additionally, a transformer-based self-attention module is incorporated into each model, resulting in six different and distinct models. The results demonstrate that the proposed frameworks significantly enhance echo cancellation and achieve a tremendous improvement in the quality of speech.

### A. DATASET

The proposed method is evaluated based on two databases: the customised TIMIT database [22] and the AEC-Challenge dataset [38]. The TIMIT dataset contains recordings of 630 speakers, each speaking ten sentences of different lengths. The speech waveforms are sampled at 16kHz sampling frequency. In this experiment, the speech signals are used to create four subdivided data, the farend, echo, nearend and mixed. The farend and nearend are taken as male-female pair, male-male pair, and female-female pair directly from the TIMIT dataset. The speech signals from the single speaker are combined to form 9 second long farend speech signals. The room impulse response (RIR) is generated using the image method [37] where the room dimension is [5,4,6]m, the source position is [2,3.5,2]m, the receiver position is [2,1.5,2]m, with a reverberation time of 0.7s. Then, the echo signals are created by convolving the farend signals and the room impulse response. The nearend is created by adding silence before and after the speech signals and is made into 9s long sentences. The mixed signal

**TABLE 2.** Comparison of results in double-talk and noisy double-talk scenario using TIMIT dataset.

| Sl.No | Model | | double-talk | | | | Noisy double-talk | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | ERLE (dB) | PESQ | Correlation Coefficient | | ERLE (dB) | PESQ | Correlation Coefficient | |
| | | | | | Input | Output | | | Input | Output |
| 1 | CNN | STFT | 43.15 | 2.79 | 0.99712 | 0.0015 | 39.09 | 2.55 | 0.97687 | 0.003 |
| | | STFT and attention | 45.82 | 2.95 | 0.99712 | 6.50E-04 | 43.78 | 2.84 | 0.97687 | 0.0009 |
| | | CQT | 44.98 | 3.07 | 0.99712 | 0.00057 | 42.18 | 2.84 | 0.97687 | 0.00097 |
| | | CQT and attention(TNN) | 48.00 | 3.40 | 0.99712 | 0.0003 | 46.35 | 3.20 | 0.97687 | 0.0007 |
| 2 | BLSTM | STFT [22] | 44.45 | 2.56 | 0.99712 | 0.0035 | 36.16 | 2.48 | 0.97687 | 0.00438 |
| | | STFT and attention | 46.85 | 3.17 | 0.99712 | 0.0007 | 41.78 | 2.76 | 0.97687 | 0.001 |
| | | CQT | 45.12 | 2.84 | 0.99712 | 0.00091 | 41.38 | 2.79 | 0.97687 | 0.0017 |
| | | CQT and attention(TNN) | 47.93 | 3.23 | 0.99712 | 0.0006 | 45.88 | 3.16 | 0.97687 | 0.0008 |
| 3 | CRNN | STFT [24] | 46.55 | 2.63 | 0.99712 | 0.00157 | 39.29 | 2.77 | 0.97687 | 0.00242 |
| | | STFT and attention | 48.03 | 3.27 | 0.99712 | 0.0004 | 46.72 | 3.05 | 0.97687 | 0.0005 |
| | | CQT | 48.23 | 3.13 | 0.99712 | 0.00035 | 43.14 | 2.97 | 0.97687 | 0.00068 |
| | | CQT and attention(TNN) | **49.56** | **3.44** | **0.99712** | **0.0001** | **48.76** | **3.25** | **0.97687** | **0.0004** |

is created by adding the echo, nearend signals and, white noise. Altogether, there are 3000 speech files for training, 340 speech files for validation and 30 speech files for testing.

The other dataset is the AEC-Challenge dataset [38], an open-source database offered by Microsoft. It contains about 10,000 synthetic speech signals of length 10s each, including farend speech, nearend speech, microphone speech signal, and echo signal. The dataset includes different RIRs, varying signal to echo ratios (SER), which measure the level of echo present in the signal, and varying nearend to farend ratios (NER), which measure the relative strength of the nearend signal to the farend signal. The SER in the dataset is between $-10$dB and 10dB. Various real world noises are incorporated into both farend and nearend speech signal. Out of the 10,000 speech files, 9,000 speech data are taken for training, 900 are taken for validation, and 100 files are used for testing.

### B. TRAINING FRAMEWORK SETUP

There are two sets of experiments being implemented in this paper. One employs STFT+attention, and the other is CQT+attention. In the initial set of experiments, the long sequence is segmented into 20ms frames, as this duration can effectively capture detailed features in both the frequency and time domains. These frames have a 50% overlap, which facilitates smoother transitions between frames and offers a balanced trade-off between computational complexity and model performance. Then the time domain signal is transformed to the frequency domain using STFT with the help of a Hanning window of size 320, thus finding its magnitude and phase. In the second experiment, CQT is performed using a Hanning window with a hop size of 256 samples for both the farend and mixed signal. The octave bins are 72, and the bins are 504 again leading to finding the magnitude and phase of farend and mixed signal. The magnitudes are concatenated in both experiments and given to the deep learning models. There are three deep learning models. Firstly, the CNN model has one input layer with 1008 neurons, two hidden 1DCNN layers with 1008 neurons and one output 1DCNN layer with 504. The BLSTM(bidirectional LSTM) model has an input layer of 1008 neurons, four BLSTM hidden layers with 600 neurons and one FC output layer with 504 neurons. The CRNN model has one input layer with 1008 neurons, two 1DCNN hidden layers with 504 neurons, two BLSTM layers with 600 neurons, and one FC output layer with 504 neurons. All three models are embedded with a self-attention mechanism. The network is trained with 32 utterances per mini-batch, and the Adam optimizer is used with a learning rate of 0.0003. After 20 epochs the learning rate changes to 0.01. Smooth L1 loss function is used as Loss function.

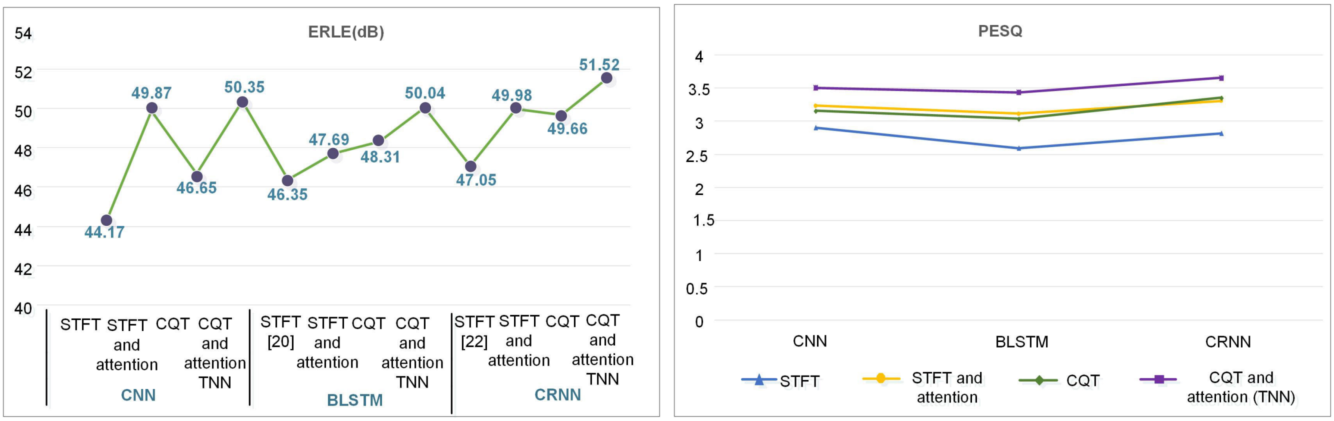
### C. LOSS FUNCTION

A loss function that combines the characteristics of L2 loss, also known as mean squared error (MSE), and L1 loss, also referred to as mean absolute error (MAE), is called Smooth L1 Loss. MAE is more robust to outliers due to its linear scaling with error, but it can result in unstable gradients. On the other hand, MSE is more sensitive to outliers because it squares the error, which leads to larger gradients. Smooth L1 loss addresses these challenges by blending the advantages of both, minimizing the effect of outliers while ensuring stable gradients during the training process. When the error is minor, it behaves like MSE; when the error is huge, it behaves like MAE [41]. The loss is defined as

$$loss = \begin{cases} 0.5(s_i - \hat{s}_i)^2/beta, & if\,|s_i - \hat{s}_i| < beta \\ (|s_i - \hat{s}_i| - 0.5) * beta, & otherwise \end{cases} \tag{13}$$

where, $s_i$ is the clean nearend speech and $\hat{s}_i$ is predicted nearend speech and beta is set to 1 for this paper. The major advantage of using smooth L1 loss is it transitions from quadratic to linear behaviour when the errors are larger [41]. Therefore, it is

- less sensitive to outliers
- differentiable at all points. The differentiability is essential for efficient training using gradient descent algorithm.



(a) ERLE Comparison

(b) PESQ Comparison

**FIGURE 5.** ERLE and PESQ comparison for singletalk scenario using STFT, CQT, and attention for various deep learning models.

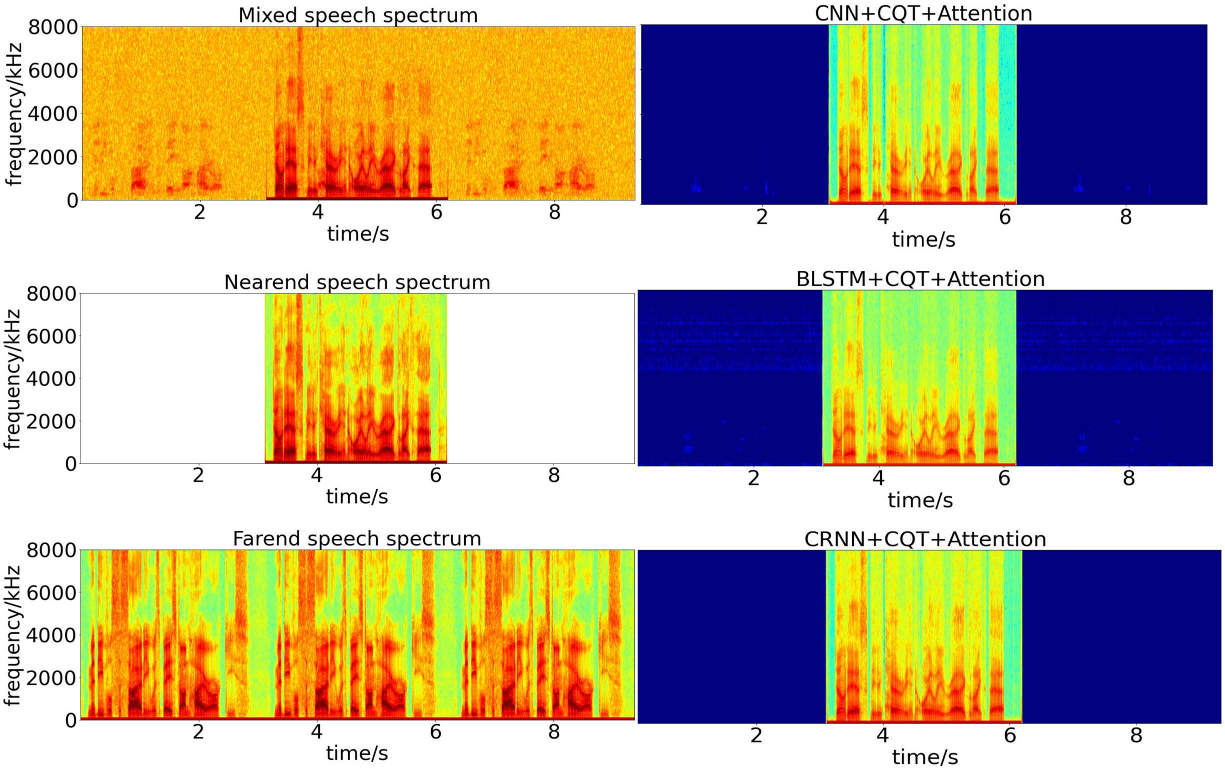


**FIGURE 6.** Spectrogram of the proposed TNN model.

- leads to smoother optimization which prevents issues related to exploding and vanishing gradients
- leads to faster convergence.

### D. PERFORMANCE METRICS

Three parameters are used to measure the performance of the echo cancellation model. One is echo return loss enhancement (ERLE) [22], [24], which measures the echo reduction between the mixed and the clean signal.

$$ERLE = 10log_{10}\frac{E[\mathbf{z}^2(n)]}{E[\hat{\mathbf{s}}^2(n)]}, \tag{14}$$

where, $\mathbf{z}(n)$ is the mixed microphone signal and $\hat{\mathbf{s}}(n)$ is the calculated near-end speech signal and E[·] denotes the statistical expectation operation. To measure the quality of speech, perpetual evaluation of speech quality (PESQ) [22], [24] is used. The range of PESQ is –0.5 to 4.5 [35]. A larger PESQ also indicates better speech quality.

The input correlation, represented by the correlation between the mixed and echo signals, and the output correlation, denoted by the correlation between the nearend clean signal and echo, are also determined [9]. This gives the relation between the clean speech and echo and how similar they are. The correlation coefficient formula

**TABLE 3. Comparison of results in noisy double-talk scenario using AEC challenge dataset.**

| Sl.No | Causal (C)/Non-causal (NC) | Model | SER = -5dB and Input correlation = 0.91 | | SER = 0dB and Input correlation = 0.87 | | SER = 5dB and Input correlation = 0.83 | |
|---|---|---|---|---|---|---|---|---|
| | | | PESQ | Output correlation | PESQ | Output correlation | PESQ | Output correlation |
| 1 | NC | T-F complex mask network [27] | 1.92 | 0.0785 | 2.24 | 0.0612 | 2.36 | 0.0470 |
| 2 | C | CNN+Transformer [39] | 1.89 | 0.0621 | 2.31 | 0.0599 | 2.41 | 0.0290 |
| 3 | C | CRNN+Transformer [40] | 1.99 | 0.061 | 2.39 | 0.0520 | 2.44 | 0.0333 |
| 4 | NC | Proposed TNN (CNN+CQT+attention) | 2.29 | 0.0599 | 2.33 | 0.0320 | 2.48 | 0.0156 |
| 5 | NC | Proposed TNN (BLSTM+CQT+attention) | 2.24 | 0.051 | 2.37 | 0.0210 | 2.45 | 0.0290 |
| 6 | NC | Proposed TNN (CRNN+CQT+attention) | **2.32** | **0.0472** | **2.51** | **0.0150** | **2.79** | **0.0093** |
| 7 | C | Proposed TNN (CNN+CQT+attention) | 1.94 | 0.0745 | 2.22 | 0.0637 | 2.38 | 0.0442 |
| 8 | C | Proposed TNN (BLSTM+CQT+attention) | 1.99 | 0.0798 | 2.18 | 0.0699 | 2.26 | 0.064 |
| 9 | C | Proposed TNN (CRNN+CQT+attention) | **2.20** | **0.0712** | **2.42** | **0.035** | **2.53** | **0.0139** |

is given as

$$r_{ab} = \frac{covariance(a, b)}{\sqrt{covariance(a, a)covariance(a, b)}}, \tag{15}$$

where, $r_{ab}$ is the correlation coefficient of two signals a and b.

## V. RESULTS AND DISCUSSION

The results for this paper are based on the test dataset present in both the TIMIT[22] and AEC-Challenge[38] datasets, which are from different speakers but generated the same as that of training and validation datasets. The outcomes of these models are subsequently compared. Three different deep learning models (CNN, BLSTM, CRNN) are implemented with three different modifications (STFT+atttention, CQT, CQT+attention), and the results are compared with the existing models and among themselves using TIMIT dataset. Here, the CNN model is a new model that has only 4 layers of 1D convolution layer. The first implementation is adding self-attention to the existing models which uses STFT; secondly, instead of STFT, CQT is used to convert the time signal to frequency, and thirdly, attention is included along with the CQT. Experiments were conducted, and performance evaluations were tabulated for singletalk, double-talk, and noisy double-talk scenarios. The proposed TNN model ((CNN/BLSTM/CRNN)+CQT+attention) is compared with existing models using the AEC challenge dataset for both causal and noncausal systems. The model size, results and advantages of the proposed model over the existing models are discussed below.

### A. COMPARISON OF MODEL SIZE

Table 1 outlines the hyperparameters and model size of the proposed TNN, revealing it to be a medium-sized model that occupies minimal space during execution. Despite its moderate size, the TNN delivers excellent performance compared to larger, more complex models. This is attributed to the precise front-end transform CQT and the effective smooth L1 loss function. The preprocessing step ensures accurate feature extraction, reducing the need for additional layers, while the loss function captures every detail and is robust to outliers, enhancing the model's efficiency. The CNN model, though smaller than the BLSTM model, yields more ERLE and has better PESQ than the BLSTM model, indicating the effectiveness of the CNN model.

### B. COMPARISON OF TIMIT TEST DATASET RESULTS

Figure 5 and Table 2 illustrate the ERLE, PESQ, input, and output correlation coefficients for the TIMIT test dataset. These metrics are compared and thoroughly examined in this section. According to the evaluation metrics, the proposed TNN architecture with CQT and attention performs better than the existing architectures, which have STFT as frontend, with no attention module. The CRNN with CQT+attention was able to produce an ERLE of 51.52dB and a PESQ of 3.65 for singletalk scenario. The results of the proposed models are not only assessed with the existing models but also compared with each other. Among the three proposed TNN models, the CRNN model implemented with CQT and attention achieves the highest echo cancellation and speech quality over all the other models in all scenarios, as seen in Figure 6. CRNN uses the advantage of CNN's ability to capture spatial patterns and LSTM's capability to capture temporal dependencies.

### C. COMPARISON OF AEC TEST DATASET RESULTS

For AEC challenge dataset, 3 existing models [27], [39], [40] are compared with the proposed TNN model (both causal and noncausal) using PESQ and correlation coefficients. Three conditions where, SER is $-5$dB, 0dB and, 5dB are taken for analysis of the system. As we can see the proposed TNN model excels the existing models both with respect to quality of speech and correlation coefficient in all three SER conditions. Initially, a noncausal system is designed, which is then transformed into a causal (real-time) system by ensuring it relies solely on past and present inputs. The performance of both the noncausal and causal systems is subsequently analyzed and assessed.

It is noted that the results of the noncausal systems are superior to those of the causal systems. This is because causal systems do not take future values into account, leading to a decline in performance. Table 3 shows that the performance of CNN and BLSTM is nearly identical at low SER levels. However, at higher SER levels, the CNN model surpasses BLSTM. This advantage is due to the CNN's ability to effectively capture spatial features and patterns, and the use of smooth L1 loss, which works well with CNN [41]. BLSTM excels in capturing long-term dependencies in sequential data, due to its bidirectional nature, which allows it to consider both past and future context. This can be particularly beneficial in cases where audio frame sequences exhibit complex temporal patterns. Combining CNN and BLSTM layers harnesses the strengths of both, resulting in even better performance. The proposed TNN achieves an average PESQ of 2.38.

The use of CQT for the front end contributes majorly to the clarity of speech. CQT being more resilient to noise, unaffected by pitch-related features, with a logarithmic scale, and providing timbral speech quality contributes to the excellent performance, as shown in Table 3. Typically, the self-attention module focuses on specific signal segments; in this context, it directs attention to regions characterized by double-talk and noise. Consequently, this targeted attention leads to an enhancement in the speech quality of the pristine nearend signal. The benefit of employing a hybrid loss is that it facilitates smooth model training, thereby boosting the system's overall performance in producing clean speech. It effectively preserves data details, particularly in noisy datasets, making it ideally suited for this neural network. Incorporating CQT and attention in the proposed model results in enhanced ERLE, PESQ, and output correlation coefficients compared to models utilizing STFT without attention. This reaffirms the importance of the Q transform and self-attention in improving model performance.

## VI. CONCLUSION

AEC is crucial in communication. As the human auditory system is nonlinear, the use of STFT cannot bring out the fullness of the AEC models, and during double-talk scenarios, special focus is not given to the existing models. The loss function significantly impacts the overall performance of the model, indicating the need for a hybrid loss function. Therefore, in this paper, a novel TransQT neural network (TNN) based AEC is presented in which STFT is replaced by CQT to mimic the human auditory processing and the integration of an attention module inspired by transformers to focus on noisy double-talk scenarios with a smooth L1 loss function to ensure smooth and effective training. First, a noncausal system is designed and implemented. This noncausal system is then converted to a causal system and compared with existing causal transformer-based networks. It uses two databases, the TIMIT and AEC datasets, to train and validate the proposed system. The proposed TNN-AEC shows promising results with respect to ERLE, PESQ and correlation coefficient. The utilization of CQT allows the TNN-AEC to effectively capture both low and high frequency components with improved resolution compared to STFT and also exhibits greater resilience to noise. By introducing the attention module, the double-talk and noisy patterns in the speech gain focus, thereby increasing the overall performance of the AEC system during noisy double-talk scenarios. The loss function is crucial for effective training and minimizing overall loss, thus enhancing the performance of the proposed model. The TNN model outperforms existing models in both echo and noise cancellation. Future work will focus on developing multichannel and personalized AEC.

## REFERENCES

[1] A. Deb, A. Kar, and M. Chandra, "A technical review on adaptive algorithms for acoustic echo cancellation," in *Proc. Int. Conf. Commun. Signal Process.*, Apr. 2014, pp. 041–045, doi: 10.1109/ICCSP.2014.6949795.

[2] J. Benesty, T. Gãnsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. Berlin, Heidelberg: Springer, 2001, doi: 10.1007/978-3-662-04437-7.

[3] G. Guo, Y. Yu, R. C. d. Lamare, Z. Zheng, L. Lu, and Q. Cai, "Proximal normalized subband adaptive filtering for acoustic echo cancellation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, no. 1, pp. 2174–2188, Aug. 2021, doi: 10.1109/TASLP.2021.3087951.

[4] Y. Yu, Z. Huang, H. He, Y. Zakharov, and R. C. de Lamare, "Sparsity-aware robust normalized subband adaptive filtering algorithms with alternating optimization of parameters," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 9, pp. 3934–3938, Sep. 2022, doi: 10.1109/TCSII.2022.3171672.

[5] H. Zhao, Y. Gao, and Y. Zhu, "Robust subband adaptive filter algorithms-based mixture correntropy and application to acoustic echo cancellation," *IEEE/ACM Trans. Audio Speech Language Process*, vol. 31, no. 1, pp. 1223–1233, Jul. 2023, doi: 10.1109/TASLP.2023.3250845.

[6] J. J. Jeong, "A robust affine projection algorithm against impulsive noise," *IEEE Signal Process. Lett.*, vol. 27, pp. 1530–1534, 2020, doi: 10.1109/LSP.2020.3018652.

[7] S. M. Boopalan, S. Alagala, and A. Ramalingam, "A memory sparse proportionate affine projection algorithm for echo cancellation: Analysis and simulations," *Arabian J. Sci. Eng.*, vol. 47, no. 3, pp. 3367–3381, Mar. 2022, doi: 10.1007/s13369-021-06219-w.

[8] V. I. Djigan, "Simplified fast affine projection algorithm in echo cancellation tasks," in *Proc. Wave Electron. Appl. Inf. Telecommun. Syst. (WECONF)*, May 2023, pp. 1–5, doi: 10.1109/WECONF57201.2023.10147942.

[9] K. Mohanaprasad and P. Arulmozhivarman, "Wavelet based ICA using maximisation of non-Gaussianity for acoustic echo cancellation during double talk situation," *Appl. Acoust.*, vol. 97, pp. 37–45, Oct. 2015, doi: 10.1016/j.apacoust.2015.04.004.

[10] J. Raghuwanshi, A. Mishra, and N. Singh, "The wavelet transform-domain adaptive filter for nonlinear acoustic echo cancellation," *Multimedia Tools Appl.*, vol. 79, nos. 35–36, pp. 25853–25871, Sep. 2020, doi: 10.1007/s11042-020-09218-5.

[11] H. Qian, F. Wang, and X. Liu, "Acoustic echo cancellation algorithm based on Kalman filtering of skewed observation noise," *IEEE Sensors J.*, vol. 24, no. 5, pp. 6626–6633, Mar. 2024, doi: 10.1109/JSEN.2024.3351879.

[12] Z. Yan, F. Yang, and J. Yang, "Optimum step-size control for a variable step-size stereo acoustic echo canceller in the frequency domain," *Speech Commun.*, vol. 124, pp. 21–27, Nov. 2020, doi: 10.1016/j.specom.2020.08.004.

[13] M. Salah, M. Dessouky, and B. Abdelhamid, "Design and implementation of an improved variable step-size NLMS-based algorithm for acoustic noise cancellation," *Circuits, Syst., Signal Process.*, vol. 41, no. 1, pp. 551–578, Jan. 2022, doi: 10.1007/s00034-021-01796-5.

[14] F. Ykhlef and H. Ykhlef, "A post-filter for acoustic echo cancellation in frequency domain," in *Proc. 2nd World Conf. Complex Syst. (WCCS)*, Nov. 2014, pp. 446–450, doi: 10.1109/ICoCS.2014.7060938.

[15] K. Chen, P.-Y. Xu, J. Lu, and B.-L. Xu, "An improved post-filter of acoustic echo canceller based on subband implementation," *Appl. Acoust.*, vol. 70, no. 6, pp. 886–893, Jun. 2009, doi: 10.1016/j.apacoust.2008.10.004.

[16] W. Lu and L. Zhang, "Collaborative block-delay Volterra filters for nonlinear acoustic echo cancellation," *Appl. Acoust.*, vol. 156, pp. 83–91, Dec. 2019, doi: 10.1016/j.apacoust.2019.06.024.

[17] C. Contan, B. S. Kirei, and M. D. Topa, "Modified NLMF adaptation of Volterra filters used for nonlinear acoustic echo cancellation," *Signal Process.*, vol. 93, no. 5, pp. 1152–1161, May 2013, doi: 10.1016/j.sigpro.2012.11.017.

[18] S. Burra and A. Kar, "Adaptive kernelized subfilter nonlinear AEC algorithm," in *Proc. Adv. Commun. Technol. Signal Process. (ACTS)*, Dec. 2021, pp. 1–4, doi: 10.1109/ACTS53447.2021.9708184.

[19] S. Sankar, A. Kar, S. Burra, M. N. S. Swamy, and V. Mladenovic, "Nonlinear acoustic echo cancellation with kernelized adaptive filters," *Appl. Acoust.*, vol. 166, Sep. 2020, Art. no. 107329, doi: 10.1016/j.apacoust.2020.107329.

[20] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018, doi: 10.1109/TASLP.2018.2842159.

[21] L. Pfeifenberger and F. Pernkopf, "Nonlinear residual echo suppression using a recurrent neural network," in *Proc. Interspeech*, Oct. 2020, pp. 3950–3954, doi: 10.21437/interspeech.2020-1473.

[22] H. Zhang and D. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," in *Proc. Interspeech*, Sep. 2018, pp. 1–26, doi: 10.21437/interspeech.2018-1484.

[23] A. Fazel, M. El-Khamy, and J. Lee, "Deep multitask acoustic echo cancellation," in *Proc. Interspeech*, 2019, pp. 4250–4254, doi: 10.21437/Interspeech.2019-2908.

[24] H. Zhang, K. Tan, and D. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions," in *Proc. Interspeech*, Sep. 2019, pp. 4255–4259, doi: 10.21437/interspeech.2019-2651.

[25] Y. Zhang, C. Deng, S. Ma, Y. Sha, and H. Song, "Deep multi-task network for delay estimation and echo cancellation," 2020, *arXiv:2011.02109*.

[26] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," in *Proc. Interspeech*, Oct. 2020, pp. 2477–2481, doi: 10.21437/interspeech.2020-2631.

[27] N. Sun, H. Liu, L. Gan, Y. Zhao, Z. Luo, and Y. Zhou, "Time-frequency complex mask network for echo cancellation and noise suppression," in *Proc. IEEE 24th Int. Conf. High Perform. Comput. Commun., 8th Int. Conf. Data Sci. Systems, 20th Int. Conf. Smart City, 8th Int. Conf. Dependability Sensor, Cloud Big Data Syst. Appl. (HPCC/DSS/SmartCity/DependSys)*, Dec. 2022, pp. 2275–2279.

[28] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.

[29] A. T. Patil, K. Khoria, and H. A. Patil, "Voice liveness detection using constant-Q transform-based features," in *Proc. 30th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2022, pp. 110–114.

[30] Z. Shi, H. Lin, L. Liu, R. Liu, and J. Han, "Is CQT more suitable for monaural speech separation than STFT? An empirical study," 2019, *arXiv:1902.00631*.

[31] R. Hemavathi and R. Kumaraswamy, "A study on unsupervised monaural reverberant speech separation," *Int. J. Speech Technol.*, vol. 23, no. 2, pp. 451–457, Jun. 2020, doi: 10.1007/s10772-020-09706-x.

[32] N. Holighaus, M. Dörfler, G. A. Velasco, and T. Grill, "A framework for invertible, real-time constant-Q transforms," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 4, pp. 775–785, Apr. 2013, doi: 10.1109/TASL.2012.2234114.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[34] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2021, pp. 21–25, doi: 10.1109/ICASSP39728.2021.9413901.

[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1109, pp. 749–752, doi: 10.1109/ICASSP.2001.941023.

[36] L. Xu, Z. Wei, S. F. A. Zaidi, B. Ren, and J. Yang, "Speech enhancement based on nonnegative matrix factorization in constant-Q frequency domain," *Appl. Acoust.*, vol. 174, Mar. 2021, Art. no. 107732, doi: 10.1016/j.apacoust.2020.107732.

[37] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979, doi: 10.1121/1.382599.

[38] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sorensen, and R. Aichner, "ICASSP 2022 acoustic echo cancellation challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9107–9111, doi: 10.1109/icassp43922.2022.9747215.

[39] X. Sun, C. Cao, Q. Li, L. Wang, and F. Xiang, "Explore relative and context information with transformer for joint acoustic echo cancellation and speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9117–9121.

[40] Y. Zhang, X. Xu, and W. Tu, "Improving acoustic echo cancellation by exploring speech and echo affinity with multi-head attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 401–405.

[41] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

**V. SONI ISHWARYA** was born in 1987. She received the Bachelor of Engineering degree from the Department of Electronics and Communication Engineering, Karunya University, Coimbatore, Tamil Nadu, in 2009, and the Master of Engineering degree in embedded system technologies from Anna University, Chennai, Tamil Nadu, in 2011. She is currently pursuing the Ph.D. degree with the School of Electronics Engineering, VIT, Chennai. Her research interests include signal processing, acoustic speech processing, and deep learning.

**MOHANAPRASAD KOTHANDARAMAN** was born in 1981. He received the Bachelor of Engineering degree from the University of Madras, Chennai, the Master of Engineering degree from Anna University, Chennai, in 2006, and the Ph.D. degree in speech signal processing from VIT, Vellore, India, in 2016. He completed the Postdoctoral Research Fellowship from UTAR Malaysia. He is currently an Associate Professor Senior with the School of Electronics Engineering, VIT, Chennai. He has more than 40 international journals and international conference publications. His research interests include signal processing, acoustic speech processing, deep learning, and natural language processing.