Speech Enhancement for Cochlear Implant Recipients using Deep Complex Convolution Transformer with Frequency Transformation

Nursadul Mamun, John H. L. Hansen
CRSS: Center for Robust Speech Systems; Cochlear Implant Processing Laboratory (CILab)
Department of Electrical and Computer Engineering, University of Texas at Dallas, USA
(nursadul.mamun, john.hansen)@utdallas.edu

***Abstract*—The presence of background noise or competing talkers is one of the main communication challenges for cochlear implant (CI) users in speech understanding in naturalistic spaces. These external factors distort the time-frequency (T-F) content including magnitude spectrum and phase of speech signals. While most existing speech enhancement (SE) solutions focus solely on enhancing the magnitude response, recent research highlights the importance of phase in perceptual speech quality. Motivated by multi-task machine learning, this study proposes a deep complex convolution transformer network (DCCTN) for complex spectral mapping, which simultaneously enhances the magnitude and phase responses of speech. The proposed network leverages a complex-valued U-Net structure with a transformer within the bottleneck layer to capture sufficient low-level detail of contextual information in the T-F domain. To capture the harmonic correlation in speech, DCCTN incorporates a frequency transformation block in the encoder structure of the U-Net architecture. The DCCTN learns a complex transformation matrix to accurately recover speech in the T-F domain from a noisy input spectrogram. Experimental results demonstrate that the proposed DCCTN outperforms existing model solutions such as the convolutional recurrent network (CRN), deep complex convolutional recurrent network (DCCRN), and gated convolutional recurrent network (GCRN) in terms of objective speech intelligibility and quality, both for seen and unseen noise conditions. To evaluate the effectiveness of the proposed SE solution, a formal listener evaluation involving four CI recipients was conducted. Results indicate a significant improvement in speech intelligibility performance for CI recipients in noisy environments. Additionally, DCCTN demonstrates the capability to suppress highly non-stationary noise without introducing musical artifacts commonly observed in conventional SE methods.**

***Index Terms*—Speech Enhancement, Complex-valued Network, Frequency Transformation Block, Transformer, Deep Neural Network, U-Net.**

I. Introduction

COCHLEAR implants (CI) provide a valuable solution for individuals with severe to profound sensorineural hearing loss, but speech perception quality can still be impacted by factors such as background noise, distortions, and reverberation [1], [2]. The presence of noise can lead to fatigue, strain, and other adverse effects when individuals listen to speech in noisy environments for extended periods. Speech enhancement (SE) techniques aim to mitigate the impact of background noise on speech signals, thereby improving speech perception.

More recently, network-based SE can be categorized as supervised or unsupervised. Unsupervised SE methods, such as traditional Wiener filtering [3] and model-based approaches [4], rely on estimating speech production properties of speech and statistical characteristics of noise. These methods demonstrate good performance when accurate knowledge estimation of these characteristics is possible. However, their effectiveness degrades in non-stationary noisy environments or when the accurate speech/environment properties are difficult to estimate. To address these challenges, various supervised SE approaches have been developed [5]–[8]. Text-directed SE [9] and deep learning-based SE systems have made a breakthrough in the speech community, especially in dealing with non-stationary audio environments [10]–[12]. These methods employ deep learning techniques and leverage large amounts of labeled data to train neural networks. Among them, supervised SE networks learn to discriminate between speech and noise by exploiting the information contained in the training data. By incorporating supervised SE techniques, researchers aim to enhance the performance of CI in challenging listening environments. These advancements can significantly improve speech perception for individuals with CI, offering better communication and reducing the negative effects of noise interference.

Over the last few decades, monaural SE or single-channel SE techniques have been extensively studied and have shown tremendous success, particularly in low Signal-to-Noise Ratio (SNR) environments, surpassing traditional methods. However, deep neural network (DNN) approaches, while successful in noise-independent SE, have limitations in generalizing speaker characteristics [13]. Convolutional Neural Networks (CNNs) were originally designed to capture local information in images but have found wide utility in analyzing local patterns of input signals through local connections [14], [15]. While CNNs excel at representation, they struggle with modeling explicit long-range dependencies due to the locality of convolution operations. In contrast, recurrent neural network (RNN)-based networks such as long short-term memory (LSTM) and gated recurrent unit (GRU) are commonly used for modeling long-term sequences with order information. However, they lack parallel processing capabilities, leading to high computational complexity. To address these challenges, researchers have proposed incorporating LSTM layers between the encoder and decoder to extract high-level features and enlarge receptive fields. Despite these efforts, contextual information in speech is often overlooked, impacting denoising performance. Combining the strengths of CNNs and RNNs, researchers have been exploring hybrid models that leverage both CNNs and RNNs to capture both local and long-range dependencies. Tan et al. [16] introduced a convolutional recurrent neural network (CRN) as an encoder-decoder network for SE. In 2019, an extended CRN was proposed by introducing a gated convolutional recurrent network (GCRN), achieving improved SE results [17]. Strake et al. [18] argued that the internal relations and local structures of CNN feature maps in CRN are compromised due to data reshaping among different CRN components. To address this, they utilized convolutional LSTM for SE, replacing fully connected mappings in LSTM with convolutional mappings. These models aim to address the limitations of CNNs in modeling explicit long-range dependencies while preserving the strengths of both CNNs and RNNs.

In contrast, transformers have achieved remarkable success in natural language processing tasks by effectively incorporating long-range dependency structures while operating efficiently for parallel processing tasks [19], [20]. Unlike traditional CNN-based methods, transformers excel at modeling the global context and demonstrate superior transferability for downstream tasks through large-scale pre-training. Self-attention, commonly used as an intra-attention module for learning task-independent sequence representations, allows the model to focus on relevant features in different parts of the input spectrogram while disregarding less salient information. In addition to self-attention, transformers incorporate other components such as multi-head attention, feed-forward neural networks, and residual connections, which generate hidden representations for input data and improve the quality and robustness of enhanced speech signals produced by the model. It is noted that the full potential of transformers in audio-visual speech enhancement research has yet to be explored.

Despite impressive improvement in short-time Fourier transform (STFT) magnitude, most DNN models in SE combine it with the noisy phase to resynthesize the time-frequency (T-F) waveform. Some studies have claimed that phase is unimportant for SE [21]. However, the importance of phase enhancement has been recognized, leading to proposed methods that estimate both magnitude and phase. Early efforts by [22]–[24] incorporated phase information into magnitude processing. [16], [25]–[28] proposed different networks to reconstruct the complex spectrogram of clean speech. Tan and Wang [17] extended the CRN model to GCRN, incorporating a gated linear unit (GLU) block to control information flow. While these approaches have considered phase for SE, their processing is learned under a real-valued network, and therefore still removes some portions of speech while suppressing the background noise.

To overcome this, complex-valued SE networks represent a cutting-edge approach in the field, departing from traditional real-valued methods by incorporating both magnitude and phase information in speech signals [5], [28]–[30]. This paradigm enhances the representation of complex audio characteristics, leading to improved performance in challenging acoustic environments. By exploiting the interdependence between real and imaginary components, these networks effectively mitigate noise and interference, showing promise for advancing speech enhancement technology [31]–[33]. For instance, [34] introduced a deep complex CRN (DCCRN) that combines both CNN and RNN to emulate complex-valued targets, demonstrating performance advantages with objective and subjective metrics. The DCCRN is further extended to DCCRN+ [35] by exploring sub-band processing, leading to a faster noise suppressor. While these methods improve speech quality, they still introduce distortion by removing parts of the speech while suppressing background noise, thereby introducing added processing distortion in the enhanced speech.

The substantial success achieved by these networks has underscored the importance of exploring and developing various network architectures to further enhance system performance in speech-related tasks. The correlation that exists among harmonics plays a significant role in speech quality [36]–[38]. However, common noise reduction algorithms suppress some of these harmonics present in the T-F spectrogram of the original noisy signal and result in added processing artifacts in the outputs [36]. The regeneration of distorted speech frequency can be used to restore harmonic characteristics of speech. Krawczyk and Germann [39] suggested that phase correlation between harmonics can be used for phase reconstruction. This motivated a phase reconstruction method based on harmonic enhancement [37]. A recent study [38] also investigated the correlation among harmonics in the T-F spectrogram and proposed a phase-and-harmonics-aware SE network [38]. Although most of these SE methods are designed for normal-hearing (NH) individuals, none of them have explored the complex T-F residual content delivery for CI users. Motivated by these results, a recent study by Mamun and Hansen [40] achieved significant success in SE by formant restoration. The objective results suggested that CI recipients could experience significant benefits in improving speech perception from formant restoration. Motivated by this, the current study introduces a deep complex convolution transformer network (DCCTN) with a frequency transformation module for SE for CI users. The proposed network consists of a complex-valued U-Net style CNN as a backbone, frequency transformation layers (FTL) in the encoder layer, a two-layer transformer network in the bottleneck layer, and convolutional blocks in the skip connection.

Our main contributions in this study to the proposed network are four-fold. (1) propose a fully complex-valued deep complex convolution transformer network, DCCTN, that uses a complex audio transformer and complex frequency transformation network (2) complex FTL in the encoder to leverage correlation among harmonics to capture global correlations over frequency for more effective T-F representations (3) a complex audio transformer within the bottleneck layer of the network. This transformer offers several advantages: the self-attention mechanism captures long-range relationships in speech by focusing on key input sequence features; its parallel processing capabilities outperform RNNs by enabling simultaneous use of multiple processing units for expedited computation [41]; accurately captures both local and global contexts, and the multi-head attention mechanism improves

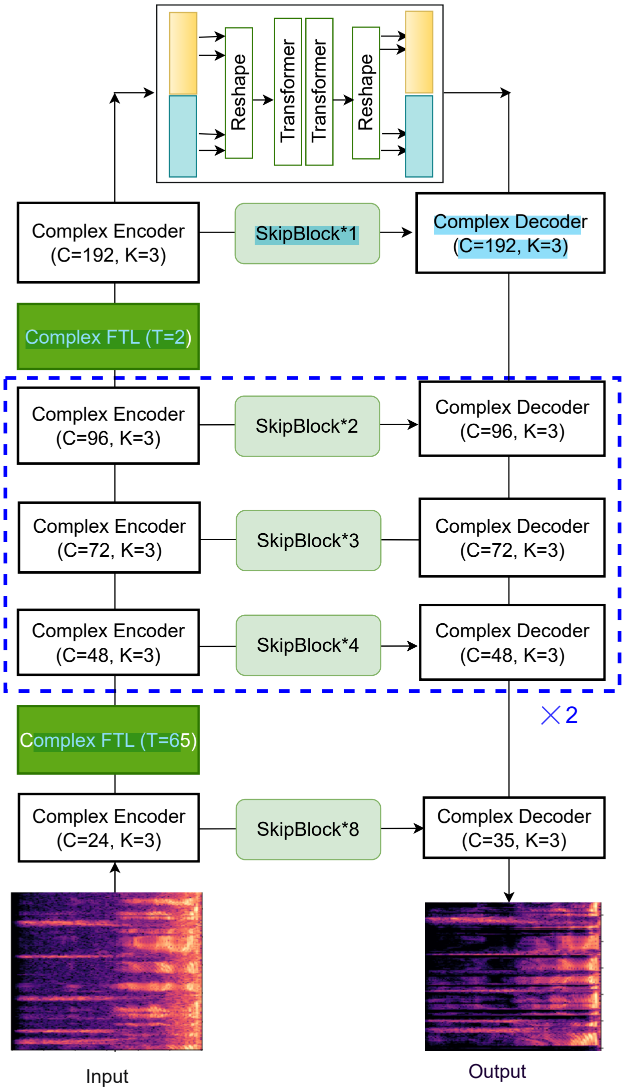
Reshape
Transformer
Transformer
Reshape
Complex Encoder (C=192, K=3)
SkipBlock*1
Complex Decoder (C=192, K=3)
Complex FTL (T=2)
Complex Encoder (C=96, K=3)
SkipBlock*2
Complex Decoder (C=96, K=3)
Complex Encoder (C=72, K=3)
SkipBlock*3
Complex Decoder (C=72, K=3)
Complex Encoder (C=48, K=3)
SkipBlock*4
Complex Decoder (C=48, K=3)
×2
Complex FTL (T=65)
Complex Encoder (C=24, K=3)
SkipBlock*8
Complex Decoder (C=35, K=3)
Input
Output

Fig. 1. Basic block diagram of DCCTN with the complex frequency transformed module. C, K, and T represent the number of channels, kernels, and input feature dimensions of FTL. L in "SkipBlocks*L" represents the number of SkipBlocks in each skip connection.

robustness and generalization for unseen data (4) to bridge the semantic gap between encoder and decoder feature maps, we substitute the skip connection in the U-Net with our proposed complex-valued convolutional blocks. This ensures that the network shares spectral information to enhance learning capabilities and regenerate missing frequency components in the distorted signal using minimal available information. Note that, the proposed network employs complex-valued convolution in all processing components to ensure effective phase reconstruction alongside magnitude reconstruction.

This paper is organized as follows: Sec. II provides a detailed implementation of our proposed DCCTN. The experimental setup and evaluation results are described in Sec. III, and IV, respectively, followed by discussion in Sec. V. Finally, we conclude this study in Sec. VI.

II. Methodology

The proposed network, DCCTN, is formulated on the concept of reconstructing lost harmonics in recovered speech signals. It addresses the challenge of degraded speech signals by employing a complex audio transformer-based network, which is specifically designed to handle complex-valued representations of audio signals, incorporating both magnitude and phase information. By leveraging this architecture, DCCTN aims to restore missing harmonics in speech signals, thereby enhancing quality and hopefully benefiting intelligibility. The overall architecture of DCCTN is discussed in this section.

A. *Overall architecture*

The goal of DCCTN is to parse degraded speech and recover high-quality signal content to improve quality, and hopefully help intelligibility. To achieve this goal, the proposed DCCTN architecture (see Fig. 1) is composed of four main processing components: (1) a fully convolutional complex-valued encoder-decoder network (Cplx-UNet); (2) a complex-valued audio transformer in the bottleneck layer, which helps the network effectively model long-range dependencies that convolutional operations cannot capture; (3) complex-valued frequency transformation modules, and (4) a complex-valued convolutional block within the skip connections between the encoder and decoder.

Each encoder/decoder block in the network is constructed upon complex-valued convolution layers to ensure successive enhancement of both magnitude and phase. The encoder network is designed using a series of encoder blocks followed by an FTL before and after the encoder layers. This ensures that the decoder uses all spectral and temporal features learned by the encoder. Several complex convolutional blocks (SkipBlock) in the skip connection reduce the semantic gap between features from the encoder and decoder blocks and thus guide the decoder to more effectively reconstruct the enhanced output.

The transformer modules in the bottleneck layer help the DCCTN to model explicit long-range dependencies that convolution operations cannot capture. This is essential for modeling long-range sequences such as speech. Furthermore, the transformer modules allow the network to operate in a parallel framework.

B. *Complex-Valued Encoder-Decoder Layer*

The complex-valued encoder-decoder architecture utilizes complex convolution to improve quality of reconstructed speech signals, distinguishing it from real-valued networks. This section focuses on the algorithmic formulation and its application.

The complex-valued encoder layer within the proposed network consists of three main components: complex convolution, complex batch normalization, and complex nonlinear activation functions. Complex convolution is employed in the U-Net architecture to enhance both the magnitude and phase components of the T-F representation of noisy speech. Conventional convolution layers operate by sliding a kernel matrix over the input matrix and performing point-wise multiplication. However, in the complex-valued generalization of a convolution layer, both input and kernel matrices are complex-valued. Despite this distinction, the convolution operation remains unchanged. The complex-valued convolution layer performs the same point-wise multiplication as the conventional convolution layer, but the output is now a complex-valued matrix, incorporating both magnitude and phase information. This additional information allows the complex-valued convolution layer to capture enhanced detailed features and patterns in the input data, making it a powerful tool in complex-valued networks. Similar to the encoder layer, the complex-valued decoder layer employs a complex-transpose convolution instead of a complex convolution. This layer aims to reconstruct a clean speech signal by applying complex-transpose convolution operations. By utilizing complex-transpose convolution, the decoder layer efficiently utilizes information captured by the encoder layer to generate the final output.

The algorithmic formulation of the complex convolution involves operating on the input complex variable $X = X_r + jX_i$ and the network's complex kernel $W = W_r + jW_i$. The complex convolution operation combines real and imaginary components of X and W to generate a complex-valued output,

$$Z = W * X = (W_r + jW_i) * (X_r + jX_i) \tag{1}$$

$$Z = (W_r * X_r - W_i * X_i) + j(W_r * X_i + W_i * X_r) \tag{2}$$

This output incorporates information from both magnitude and phase, enabling the complex-valued encoder-decoder architecture to reconstruct cleaner speech signals.

C. Complex-Valued Audio Transformer

The transformer, a machine learning module, employs a self-attention mechanism that enables a selective focus on distinct components of the input signal at each network layer. This study incorporates a complex-valued audio transformer within the bottleneck layer of the DCCTN model instead of a pure transformer for SE as shown in Fig. 2(a). This decision enhances the DCCTN model's ability to capture long-range dependencies in the input features. Within the complex transformer network, complex convolutions operate on the real and imaginary components of feature maps and weights, performing real-valued convolutions. Consider, X be the complex-valued feature map from the encoder, where X_r and X_i denote the real and imaginary components, respectively and T(.) represents the output of a real-valued transformer for each input. The output of the transformer, denoted as Y_T is presented as,

$$Y_T = (T(X_r) - T(X_i)) + j(T(X_r) + T(X_i)) \tag{3}$$

where T(.) represents the output of a real-valued transformer for each input.

Fig. 2(b) illustrates the architecture of this transformer module which consists of positional encoding, multiple layers of self-attention, and feed-forward neural networks (FFN). However, in this study, positional encoding was omitted as it was found ineffective in capturing acoustic sequences [42]. Self-attention serves as an intra-attention module, facilitating the learning of task-independent sequence representations. This allows the model to attend to relevant features in the input spectrogram while disregarding irrelevant information. Subsequently, feed-forward layers process the attended features, generating a hidden representation for the input.

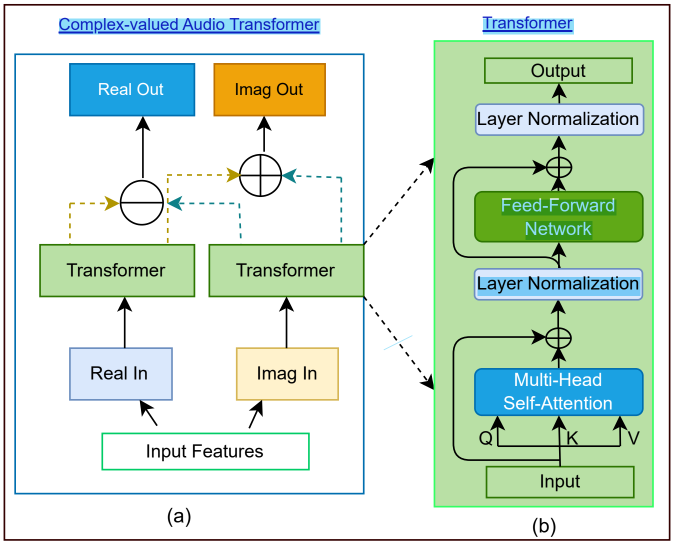

Fig. 2. Architecture of the complex-valued audio transformer used in the bottleneck layer of DCCTN model. (a) Complex-valued audio transformer (b) Real-valued audio transformer

The attention module operates on queries, keys, and values, which are obtained by linear transformations of the input sequences. Queries determine the elements in the sequence to focus on, keys determine the similarity between elements, and values scale the importance of each element. These matrices are processed in parallel, with each parallel attention mechanism referred to as a 'head'. In a typical transformer model, the input representation undergoes several fully connected layers to produce the matrices Q, K, and V:

$$Q = W_q * X; \quad K = W_k * X; \quad V = W_v * X;$$

Here, W_q, W_k, and W_v represent weight matrices of the linear transformation, which are learned during training, and X denotes the input signal representation. The resulting matrices Q, K, and V are utilized in the attention mechanism, allowing the model to learn the relationships between elements in the input sequence relevant to the given task and adapt to different input data types.

The feed-forward network in the transformer is applied independently to each position in the input sequence. Specifically, it consists of a GRU with a ReLU activation function followed by a linear transformation layer, $T(.)$. The output of the feed-forward network at each position is then passed on to subsequent layers of the Transformer.

The output of the attention module is represented as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

$$O_1 = LayerNorm(X + Attention(Q, K, V)) \tag{5}$$

$$FFN(O_1) = T(ReLu(GRU(O_1))) \tag{6}$$

$$O_2 = LayerNorm(O_1 + FFN(O_1)) \tag{7}$$

The output of the attention module is then concatenated with the input sequence, X and passed to the layer normalization to provide the output O_1. FFN module is then processed to introduce nonlinearity into the model, enabling it to capture complex relationships between different positions in the input speech sequence. This allows the Transformer to learn sophisticated patterns in language and perform tasks such as translation and summarization. The output of the FFN layer is then concatenated with the previous output (O_1) and passed through the layer normalization to obtain the final output.

Therefore, the complex-valued audio transformer, integrated into the DCCTN model's bottleneck layer, enhances its ability to capture long-range dependencies from input features. By leveraging self-attention and omitting positional encoding, the audio transformer selectively attends to pertinent information and generates a robust hidden representation. This empowers the model to improve SE performance and effectively process complex audio data.

D. Complex-valued Frequency Transformation

When a speech signal undergoes degradation due to noise or other distortions, it can affect harmonics and lead to a decrease in speech quality. To improve signal quality, it is possible to preserve or reconstruct these harmonics. However, conventional CNN kernels are designed to work in the spatial domain with images or matrices as input, using a sliding window technique to perform convolution operations. When applied to a T-F spectrogram, which represents speech in the frequency domain, these spatial-domain kernels are not well-suited to capture harmonics as they are distributed across the frequency axis of the T-F spectrogram and not localized in the spatial domain.

To capture these harmonics effectively, specialized T-F CNN kernels such as gamma-tone filter-bank kernels or wavelet convolutional kernels are required. Previous studies have demonstrated that utilizing the attention module can effectively capture harmonic components of speech and restore the enhanced signal, as well as reconstruct missing frequencies in a band-limited signal [38]. Although, most of the existing networks employed attention modules across the frequency axis with real-valued networks operating only on the magnitude response, [8], [43] employed attention modules on both real and image spectrograms.

In this study, the FTL is extended to address the complex domain by proposing a complex-valued FTL that attends to features in frequency while maintaining the interdependence between real and imaginary components of the complex-valued feature map. To utilize the inter-channel relationship of features, we apply the attention module to the incoming feature maps, point-wise multiply it with the input features, and output the result. Next, we apply the trainable frequency transformation matrix (FTM) to the feature maps at time step t to ensure global frequency correlation along the frequency axis. Finally, we concatenate the output of FTM module with the input features, using a CNN layer to ensure global and local frequency correlation among harmonics.

As shown in Fig. 3, the FTL comprises three stacked CNN layers: a fully connected layer, a CNN layer for FTM, and a CNN layer for concatenation. The set of complex-valued feature maps is extracted from the encoder's stacked CNN layers, consisting of a sequence of F frequency vectors with C channels and T frames in total. The input feature vector is defined as,

$$U \in R^{T\times F\times C} \tag{8}$$

The trainable FTM is applied to the feature map slice at each point in time. If $W_{FTM} \in R^{F\times F}$ denotes the trainable FTM and $U(t_0) \in R^{F\times C}$ denotes the feature slice at time step t_0, the transformation feature slice at time step t_0 can be represented by the following equation:

$$U^{tr}(t_0) = W_{FTM}.U(t_0) \tag{9}$$

where $t_0 \in 0, 1, ..., T - 1$.

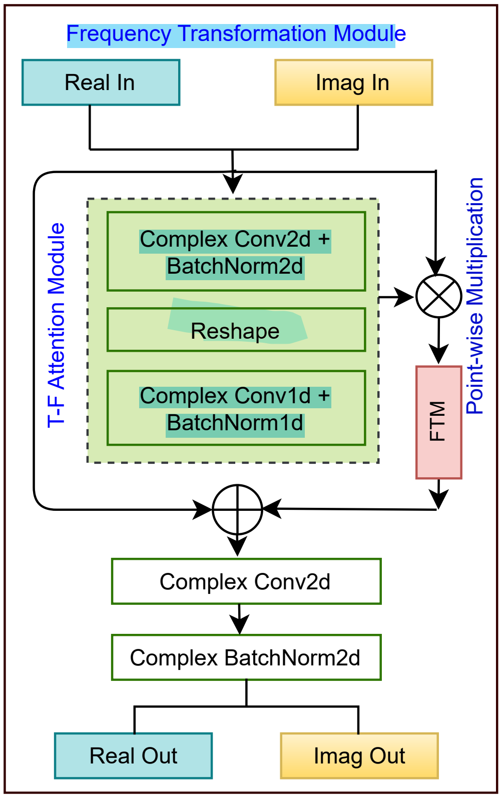
Frequency Transformation Module
Real In
Imag In
T-F Attention Module
Complex Conv2d + BatchNorm2d
Reshape
Complex Conv1d + BatchNorm1d
Point-wise Multiplication
FTM
Complex Conv2d
Complex BatchNorm2d
Real Out
Imag Out

Fig. 3. Architecture of the complex-valued frequency transformation module used in the encoder of DCCTN model.

E. Complex-valued SkipBlocks

A recent study has found that in the U-Net architecture, there is a potential issue with the exchange of features between the encoder and decoder due to a probabilistic semantic gap [44]. Specifically, the first layer of the encoder extracts low-level local spectral and temporal features, and as the layer number increases, it extracts higher-level features. Meanwhile, the final layer of the decoder receives highly processed features from its previous layer and low-level features from the first layer of the encoder through a skip connection. The incompatibility between these two feature sets could potentially limit the learning ability of the network. To address this issue, our study proposes adding convolution layers within the skip connection to directly transform the encoder features into a more intuitive form for the decoder, thus compensating for this incompatibility. Despite the importance of skip connections in developing robust networks, the identified semantic gap can hinder speech synthesis quality. However, by incorporating these proposed convolution blocks within the skip connection, our study aims to improve the network's ability to learn and share spectral information, as has been demonstrated in the success of image segmentation and speech dereverberation. We introduce a series of 'SkipBlocks' along each skip connection path within the architecture. Each 'SkipBlock' is composed of a complex convolution layer, a normalization layer, and is activated by a complex ReLU function. Crucially, the quantity of SkipBlocks deployed is inversely proportional to the depth of the corresponding encoder layer (as depicted in Fig. 1). Therefore, a skip connection linked to the encoder's final layer will have just a single SkipBlock, whereas one connected to the first layer of the encoder will comprise as many as eight SkipBlocks, ensuring a tailored approach to the varying levels of feature abstraction across the network.

F. Loss function

Most state-of-art networks use either a time domain or frequency domain loss function to optimize the machine learning models which may not be fully correlated with the perceptual quality of the reconstructed signal. Therefore, this study optimizes the proposed network by calculating both time and frequency domain losses of the real and imaginary components. We adopt a frame-level auxiliary loss, STFT-based auxiliary loss along with a scale-invariant signal-to-distortion ratio (SISDR) [45] loss to minimize mean square error between the network prediction, $\hat{y_a}$ and corresponding clean spectrogram, y_a.

$$L_{MCFT}(\hat{y_a}, y_a) = L_{SISDR}(\hat{y_a}, y_a) + \alpha \cdot L_{Freq}(\hat{y_a}, y_a) \quad (10)$$

where α is the weight (chosen $\alpha = 50$ in this study to be compatible with SISDR loss) and the time domain SISDR loss, $L_{SISDR}(\hat{y_a}, y_a)$ is defined as

$$L_{SISDR}(\hat{y_a}, y_a) = 10 \cdot \log_{10} \frac{\|y_a\|^2}{\|y_a - \hat{y_a}\|^2} \quad (11)$$

and frequency domain loss, L_{Freq} is a combination of spectral convergence loss, L_{SC} and logarithmic value of STFT magnitude loss, L_{Mag} and can be written as

$$L_{Freq}(\hat{y_a}, y_a) = L_{SC}(\hat{y_a}, y_a) + L_{Mag}(\hat{y_a}, y_a) \quad (12)$$

$$L_{SC}(\hat{y_a}, y_a) = \frac{\| |STFT(y_a)| - |STFT(\hat{y_a})| \|_F}{\| |STFT(y_a)| \|_F} \quad (13)$$

$$L_{Mag}(\hat{y_a}, y_a) = \| \log |STFT(y_a)| - \log |STFT(\hat{y_a})| \|_1 \quad (14)$$

where $\|.\|_F$ and $\|.\|_1$ denote the Frobenius and L1 normalization, respectively and $|STFT(.)|$ denotes the magnitudes of the spectrogram.

III. Experimental Setup

A. Speech Database

In our experiments, we evaluate the proposed SE model on the TIMIT database [46]. TIMIT dataset is comprised of 6300 speech utterances of American English speakers, each phonetically transcribed. The duration of each sentence ranges from 3 to 5 seconds. The training set consists of a subset of the TIMIT database consisting of 950 utterances from 50 speakers. These sentences were altered by adding eight distinct noise sources from the AURORA dataset, using five different SNRs: -10, -5, 0, 5, and 10 dB. For utterances from the training set, we hold out 150 randomly selected utterances to create a validation test set. The environmental noise conditions included samples from various sources such as airports, babble, cars, exhibitions, train stations, city streets, speech-shaped noise (SSN), and white Gaussian noise.

For testing the model, a second subset of 50 samples was utilized. These utterances were mixed with one of 3 seen noise types (babble, car, and SSN) and two unseen noise types (restaurant and train), using 7 different SNR levels: -7.5, -5, -2.5, 0, 2.5, 5, and 10 dB (note: 4 of 7 SNRs are the same as train set). 'Seen' noise refers to the noise type that was encountered by the model during training, whereas 'unseen' noise denotes noise types that were entirely unknown to the trained model. Therefore, the training set includes 38,000 noisy-clean pairs, while the test set contains 1750 pairs. All speech samples and noises were resampled at a rate of 16 kHz.

B. Subjective Listener Evaluation

1) Stimuli and Subject Demographics: Three post-lingually and one prelingually deafened CI user (two males and two females) participated in this study. The age of the CI subjects at the time of the test varies from 50 to 75 years, with a mean age of 64.5 years. Implant use ranged from 5 to 14 years, with a mean of 6.2 years. Table I presents the demographic information of the CI participants. All were native English speakers and were implanted with the Nucleus cochlear implant system by Cochlear Corporation. Subjects were paid for their participation in this study.

The test stimuli used in this evaluation were taken from the TIMIT corpus. Each sentence consists of 3-5 keywords uttered by multiple speakers. The root-mean-square value of all sentences was equalized to approximately 65 dB. All stimuli were sampled at 16 kHz. Two types of noise, babble and car noise were used as maskers at 0 db and 5 dB SNR to corrupt the sentences to simulate noisy speech. This database was developed to evaluate the speech perception abilities of CI users in a noisy environment. The speech corpus includes 12 lists, each containing 5 phonetically balanced sentences recorded from six speakers.

This article has been accepted for publication in IEEE/ACM Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2024.3366760

TABLE I
DEMOGRAPHIC INFORMATION OF CI RECIPIENTS FOR SUBJECTIVE EXPERIMENTS.

		S1	S2	S3	S4
		Right/Left	Right/Left	Right/Left	Right/Left
Implant Specific	Implant type	CI512	CI512/ CI422	CI512/ CI422	CI422/ CI24RE
	Uni/Bilateral	B	B	B	B
	Speech Processor	CP1000	CP1110	CP910/920	CP1110/ CP1110
Processor Specific	Device Exp. (years)	13.5/13 y	5 y	11.5/8 y	6 y
	Active electrodes	22/21	20	22	22
	Stim. Rate (Hz)	900	900	900	900
	'n-maxima'	8	8	8	8
Demographics	Age (years)	75	66	56	61
	Linguistic Exp.	Post-lingual	Post-lingual	Pre-lingual	Post-lingual

2) *Experiment I: Speech Intelligibility Assessment:* The test was performed using CCi-cloud [47], an online research platform designed by UTD-CRSS-CILab, hosting a MATLAB GUI with a test dataset running in the backend. CI users performed the test fitted with their daily clinical processor. Recordings from the TIMIT database were used to create the experiment dataset. The test consists of 60 randomly selected samples throughout the experiment. The experiment began with a short training phase to listen to a set of five clean stimuli to become familiar with the testing procedure. After the training phase, a speech token was presented to the listener. Participants were asked to listen to 60 samples in different conditions. Each subject participated in a total of 12 test conditions (2 noisy types * 2 SNR levels * 3 processing conditions). The noisy and enhanced samples were chosen to be discrete from each other. Participants were allowed to listen to the test token only twice. Participants were asked to type what they had heard in a designated box within the MATLAB GUI interface. The total number of speech samples was 60 (2 noisy types * 2 SNR levels * 3 processing conditions * 5 samples/conditions) throughout the experiment. The presentation order of the enhanced, noisy speech, and SNR level was randomized throughout the session. The average testing time for the experiment was one hour.

3) *Experiment II: Speech Quality Assessment:* The test set for this experiment is similar to Experiment I. Each test consists of four speech samples: clean speech as a reference sample, noisy speech, and two enhanced speech processed by a current state-of-the-art (SOA) network and our proposed network. Each subject participated in a total of 12 test conditions (2 noisy types * 2 SNR levels * 3 processing conditions). The total number of test sets was 40 (2 noisy types * 2 SNR levels * 10 samples/conditions) and speech samples were 160 (40 samples/networks * 4 conditions) throughout the experiment. Participants were asked to listen to a clean speech token in each test set as a reference sample. After that, three test audio samples (one original distorted and two enhanced using two networks) were randomly presented. Participants were allowed to listen to each speech token as many times as they wanted. Participants were asked to select the best sample (among three test samples), closer to the reference clean sample. Moreover, they were asked to perceptually rate each of the three test samples on a scale from 1 to 5, with 1 being the poor quality and 5 being the highest quality. In each set, the clean and processed samples were chosen to be unique in this experiment. The average testing time for the experiment was 30 minutes.

C. *Evaluation metrics*

To evaluate the performance of the proposed network in terms of speech quality and intelligibility, we utilize several objective intelligibility and quality metrics. Specifically, we calculate the short-time objective intelligibility (STOI) [48] and Spectrogram Orthogonal Polynomial Measure (SOPM) [49] for intelligibility assessment. For measuring speech quality, we employ the Perceptual Evaluation of Speech Quality (PESQ) [50] and SISDR [45] metrics. Additionally, we evaluate speech distortion using the Log-spectral distance (LSD) [51] and Itakura-Saito (IS) [52] metrics.

In all objective metrics (except LSD and IS), a higher value indicates better performance, suggesting that the enhancement system effectively reduces distortion while preserving the quality of the target speech. The PESQ scores typically range between -0.5 and 4.5, with higher values denoting improved speech quality. Both SOPM and STOI map objective scores to the range of [0, 1], where higher values indicate enhanced intelligibility.

On the other hand, the SISDR, LSD, and IS scores are unbounded. Smaller values (LSD and IS measure) indicate better similarity or less distortion between the signals. IS values between 0 and 0.5 reflect waveform coding level distortion, while values between 1.5 and 5.0 represent greater additive noise distortion.

By considering these various metrics, we can comprehensively assess the performance of the proposed network and gain insights into its ability to enhance speech quality and intelligibility effectively.

D. *Comparison Systems*

This study compared the proposed network with four SOA algorithms: CRN, DCCRN, GCRN, and CFTNet. In order to validate the proposed network, we utilized the same training and testing dataset, maintaining consistent test conditions during our evaluation.

CRN [16] combines CNNs and RNNs, with convolutional layers capturing local spatial or temporal patterns and recurrent layers modeling long-term dependencies and sequential information in the input data.

DCCRN [34] is designed to process complex-valued sequential or time-dependent data, utilizing complex-valued operations to model complex relationships within the data effectively. It incorporates convolutional layers for spatial or temporal pattern learning and recurrent connections for capturing temporal dependencies. As both DCCRN and DCCTN are designed based on CRN, this study has used DCCRN as a baseline network in this study.

GCRN [17] incorporates CNNs and RNNs with gated mechanisms, specifically leveraging gated units to model sequential or temporal dependencies in the data. These gated mechanisms allow the network to selectively update and retain relevant information over time, making GCRNs highly effective in processing sequential data with long-term dependencies and temporal patterns.

CFTNet [40] utilizes a frequency transformation module to capture distorted frequency components in speech. It also incorporates skip connections to promote gradient flow and mitigate the vanishing gradient problem in DNNs, enabling the network to learn residual mappings and emphasize differences between input and desired output.

E. Network Architecture

The DCCTN employs components to estimate non-linear mappings from a noisy speech T-F spectrum to a clean speech spectrum. The process begins by computing the STFT of the speech signal, using a frame size of 16 ms and an overlap of 8 ms.

The network architecture consists of eight layers of encoder-decoder pairs, incorporating two FTLs and two transformers in the bottleneck layer. The skip connections in the architecture also utilize convolutional layers. The encoder layers employ convolution layers with a specified kernel size and stride, while the decoder layers use the same parameters except for the transposed convolution.

To ensure harmonic correlation along the frequency axis, an FTL is included after the input layer and before the bottleneck layer in the encoder. The parameters of the FTL are selected based on the parameters in the corresponding encoder layer. This approach promotes consistency in both magnitude and phase advancements by leveraging the complex spectrogram, utilizing complex-valued convolutions, and employing complex-valued LSTM layers.

The network undergoes training for 100 epochs using an Adam optimizer with an initial learning rate of 0.0003 and a batch size of 16. The objective function combines the SISDR and STFT loss to minimize the mean square error (MSE) between the network's predictions and the corresponding clean spectrogram. The STFT loss calculates the spectral convergence and spectral magnitude losses in the STFT domain, while SISDR accounts for channel variations, interference, and artifacts in the time domain signal. The total number of parameters of the DCCTN model is 10.1M and Multiply-Accumulate operations (MACs) is 130M.

In summary, the DCCTN incorporates complex-valued convolution, complex-valued frequency transformation block, and complex-valued transformer blocks to estimate the non-linear mapping from a noisy T-F spectrum to a clean speech spectrum. The network architecture, training process, and objective function are carefully designed to ensure accurate magnitude and phase representation, addressing the challenges posed by noisy speech signals.

IV. RESULTS

This section provides an assessment of performance of the proposed SE technique. This assessment involves analyzing results from subjective listening tests and objective measurements for speech intelligibility and quality. The obtained scores are compared with those from existing networks (e.g., CRN, DCCRN, GCRN, and CFTNet) for seen and unseen noises and SNRs.

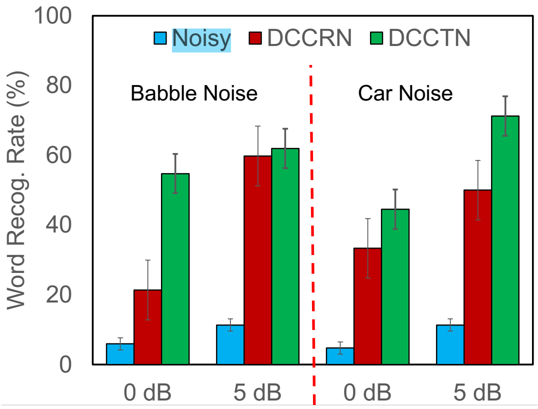

Fig. 4. Mean word recognition rate in babble and car noise condition for CI recipients

A. Experiment I: Subjective Speech Enhancement Performance based on Speech Intelligibility

The speech intelligibility of CI recipients is evaluated using word recognition rate (WRR) measured from test samples. Fig. 4 shows mean WRR results for CI recipients in babble and car noise conditions. Error bars indicate ±1 standard deviation in CI performance. Overall, both proposed and baseline approaches enhance speech intelligibility for CI recipients across all SNR levels and noise conditions. The proposed DCCTN approach consistently outperforms the baseline network regardless of SNR level or noise category. Under babble noise at 0 dB SNR, mean subject WRR scores improved from 5.95% to 21.39% and 54.76% with DCCRN and DCCTN processing, respectively. At 5 dB, mean WRR scores improved from 11.36% to 59.78% and 61.96% with DCCRN and DCCTN, respectively. In the presence of car noise, DCCTN showed improved performance over DCCRN by 33.7% and 42.5% at 0 and 5 dB SNR, respectively.

To investigate improvement over SNR and noise cases, an analysis of variance (ANOVA) is conducted on listener performance, using a significance value of 0.05. The ANOVA results for car noise at 0 and 5 dB SNR are as follows: [F(2, 11) = 34.95, $p < 0.0006$] and [F(2, 11) = 84.45, $p < 0.0000015$], respectively. Similarly, for babble noise, the ANOVA results are [F(2, 11) = 15.79, $p < 0.001$] and [F(2, 11) = 13.48, $p < 0.0019$], respectively. These results indicate a significant difference across processing conditions.

A posthoc analysis reveals a noteworthy disparity in scores between DCCRN and DCCTN, particularly evident in babble noise at 0 dB SNR and car noise at 5 dB SNR, as evidenced by ANOVA results [F(1, 7) = 13.45, $p < 0.01$] and [F(1, 7) = 26.14, $p < 0.002$], respectively. However, the observed difference is statistically insignificant for car noise at 0 dB SNR [F(1, 7) = 3.89, $p < 0.096$].

B. Experiment II: Subjective Speech Enhancement Performance for Speech Quality

To investigate speech quality preference, we extend our study to perform pair preference tests between original noisy speech, DCCRN, and DCCTN-processed speech. The listener rating for processed speech was on a scale of 1-5 (5 being

TABLE II
MEAN OBJECTIVE SCORES FOR THE DCCTN FOR BABBLE, CAR, AND SPEECH-SHAPED NOISE.SCORES WERE CALCULATED FOR STOI(0 ∼ 1), SOPM(0 ∼ 1), PESQ(-0.5 ∼ 4.5), SISDR(UNBOUNDED).

SNR	Method	STOI			SOPM			PESQ			SISDR		
		Babble	Car	SSN	Babble	Car	SSN	Babble	Car	SSN	Babble	Car	SSN
-5 dB	Noisy	0.53	0.55	0.61	0.58	0.59	0.61	1.18	1.14	1.09	-5.02	-4.97	-4.95
	DCCRN [34]	0.78	0.78	0.79	0.88	0.88	0.89	1.59	1.56	1.32	3.65	3.68	3.51
	DCCTN	**0.83**	**0.84**	**0.82**	**0.93**	**0.92**	**0.92**	**1.86**	**1.91**	**1.73**	**5.43**	**5.55**	**5.06**
0 dB	Noisy	0.68	0.68	0.72	0.75	0.77	0.78	1.25	1.22	1.17	-0.01	0.03	-0.01
	DCCRN [34]	0.84	0.84	0.85	0.92	0.93	0.93	1.87	1.86	1.48	6.78	6.80	6.64
	DCCTN	**0.90**	**0.89**	**0.89**	**0.96**	**0.96**	**0.96**	**2.36**	**2.39**	**2.14**	**8.99**	**9.05**	**8.74**
5 dB	Noisy	0.81	0.82	0.82	0.89	0.90	0.90	1.42	1.37	1.36	4.97	5.01	4.99
	DCCRN [34]	0.90	0.90	0.90	0.96	0.97	0.97	2.15	2.18	1.75	10.16	10.31	10.12
	DCCTN	**0.94**	**0.94**	**0.94**	**0.98**	**0.98**	**0.98**	**2.80**	**2.83**	**2.67**	**12.38**	**12.43**	**12.47**
10 dB	Noisy	0.90	0.91	0.90	0.96	0.96	0.96	1.72	1.64	1.63	9.99	10.00	10.00
	DCCRN [34]	0.94	0.94	0.94	0.98	0.98	0.98	2.43	2.44	2.07	13.85	13.88	13.94
	DCCTN	**0.96**	**0.96**	**0.97**	**0.99**	**0.99**	**0.99**	**3.09**	**3.12**	**3.15**	**15.58**	**15.56**	**15.94**

the highest quality) and presented in Fig. 5. In general, the proposed DCCTN solution was the highly preferred network over baseline in every condition, as well as the highest quality score over baseline. Overall mean quality for DCCTN was 97.9%, vs 76.8% for DCCRN, whereas 28.12% for original noisy speech. Moreover, DCCTN-processed speech achieved the highest +27.5% improvement in speech quality over DCCRN-processed for car noise conditions.

ANOVA shows that improvement in speech quality with proposed and baseline networks vs. original noisy speech is statistically significant across all SNRs and noise types. ANOVA results for babble noise at 0 and 5 dB SNR is [F(2, 11) = 63.2, $p < 0.000005$] and [F(2, 11) = 33.28, $p < 0.00007$], respectively, indicating a significant difference across processing conditions. Similarly, ANOVA results for car noise at 0 and 5 dB SNR are [F(2, 11) = 125.6, $p < 0.00003$] and [F(2, 11) = 50.83, $p < 0.00001$], respectively.

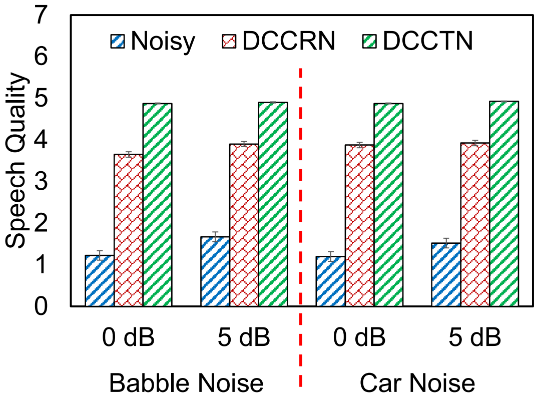

Fig. 5. Mean speech quality assessment in babble and car noise condition for CI recipients.

C. Experiment III: Speech Enhancement Performance Assessment using Objective Measures

1) DCCTN Model Evaluation: Table II presents objective scores obtained with the proposed DCCTN and baseline DCCRN method, assessing the improvement in speech intelligibility, quality, and distortion. Scores were calculated using four different objective metrics: STOI, SOPM, PESQ, and SISDR.

The evaluation involved calculating objective scores for both unprocessed noisy signals and enhanced signals processed by DCCRN and DCCTN. Three different noise types were considered: babble, car, and SSN, across a -5 to +10 dB range of SNRs. Each objective score represents the average speech intelligibility or quality obtained from 50 utterances. In general, objective scores for enhanced speech were better than for original noisy speech. The DCCTN network showed better performance over the DCCRN network for all noises in terms of all metrics. The relative improvement in PESQ was particularly noticeable for higher SNRs, whereas the trend was the opposite for other metrics. Moreover, results indicated that the proposed network performs better for babble noise compared to car and SSN noise. Specifically, DCCTN achieves +18.9%, +22.4%, and +31% improvement in PESQ over the DCCRN for babble, car, and SSN at -5 dB, respectively.

2) Comparison with Existing networks: Cochlear implants rely on a reduced T-F signal representation based on electrical stimulation of the auditory nerve to transmit sound information. In noisy environments, neural processing of the auditory system can be disrupted, affecting the CI user's ability to effectively decode speech signal content in the brain. Background noise can mask important speech cues, making it difficult for CI users to perceive and understand speech. As SNR decreases, the impact of noise interference increases, potentially leading to decreased speech intelligibility.

This study aimed to assess the impact of noisy and enhanced speech on CI recipients by categorizing the data into three groups based on SNR: 'High' (SNRs 5-10 dB), 'Medium' (SNRs 0-5 dB), and 'Low' (SNRs less than 0 dB). The 'High' group represents a relatively high SNR range, with the speech signal being generally clear and suitable for CI recipients. Here, CI recipients can expect good speech intelligibility and quality. The 'Medium' group represents a more challenging environmental condition with moderate SNRs. The speech signal in this range has moderate noise, which depending on noise type, can still be beneficial. The 'Low' group represents extremely noisy situations for CI recipients, with SNRs falling below 0 dB. In this range, speech intelligibility is significantly compromised, making it very challenging for CI recipients to understand speech content. It should be noted that while

TABLE III
MEAN OBJECTIVE SCORES FOR DIFFERENT NETWORKS FOR THREE SNR GROUPS. 'HIGH', 'MED.', 'LOW', AND 'AVG.' REPRESENT HIGH, MEDIUM, LOW GROUPS AND AVERAGE, RESPECTIVELY. THE OBJECTIVE SCORES ARE PRESENTED FOR BOTH SEEN AND UNSEEN NOISES.

Condition Method		STOI				SOPM				PESQ				SISDR			
		SNR Range				SNR Range				SNR Range				SNR Range			
		High	Med.	Low	Avg.	High	Med.	Low	Avg.	High	Med.	Low	Avg.	High	Med.	Low	Avg.
Seen	Noisy	0.86	0.73	0.59	0.73	0.93	0.80	0.64	0.79	1.52	1.25	1.15	1.31	7.49	1.26	-3.76	1.66
	CRN [16]	0.87	0.77	0.65	0.76	0.92	0.88	0.80	0.87	1.55	1.31	1.17	1.34	6.00	3.30	0.16	3.15
	DCCRN [34]	0.92	0.86	0.80	0.86	0.97	0.94	0.89	0.93	2.17	1.80	1.54	1.84	12.04	7.61	4.36	8.00
	GCRN [17]	0.87	0.8	0.71	0.79	0.94	0.91	0.85	0.90	1.76	1.46	1.28	1.50	7.66	5.17	1.92	4.92
	CFTNet [40]	0.93	0.88	0.82	0.88	0.91	0.91	0.90	0.91	2.70	2.20	1.80	2.23	10.42	7.33	4.84	7.53
	DCCTN	**0.95**	**0.91**	**0.85**	**0.90**	**0.98**	**0.96**	**0.93**	**0.96**	**2.94**	**2.41**	**1.94**	**2.43**	**14.04**	**9.82**	**6.23**	**10.03**
Unseen	Noisy	0.87	0.74	0.61	0.74	0.93	0.79	0.61	0.78	1.54	1.27	1.17	1.33	7.50	1.25	-3.75	1.67
	CRN [16]	0.86	0.74	0.61	0.74	0.91	0.84	0.70	0.82	1.56	1.33	1.20	1.36	5.76	2.20	-2.19	1.92
	DCCRN [34]	0.88	0.77	0.64	0.76	0.94	0.83	0.66	0.81	1.72	1.36	1.21	1.43	8.78	2.56	-2.67	2.89
	GCRN [17]	0.87	0.79	0.69	0.78	0.94	0.89	0.78	0.87	1.74	1.44	1.27	1.48	7.55	4.41	0.22	4.06
	CFTNet [40]	0.90	0.81	0.70	0.80	0.89	0.85	0.78	0.84	2.18	1.60	1.32	1.70	8.82	4.18	-0.23	4.26
	DCCTN	**0.92**	**0.83**	**0.71**	**0.82**	**0.97**	**0.92**	**0.78**	**0.89**	**2.35**	**1.65**	**1.32**	**1.77**	**12.38**	**6.45**	**0.64**	**6.49**

listeners with normal hearing may also experience some speech intelligibility loss in all three, CI recipients are greatly impacted across all SNR ranges due to the reduced T-F content delivery of their implants (e.g., ∼ 10% of what normal hearing subject experience).

To address these challenges, SE techniques are necessary and can provide benefits for CI recipients. State-of-the-art SE methods have shown benefits for CI users in the 'High' zone, where speech content is relatively clean. However, these methods tend to introduce processing distortion in the 'Medium', and 'Low' ranges, making it more difficult for CI users to perceive and understand speech. By categorizing noisy speech into these three ranges, this study has aimed to highlight the varying impact of different SNR levels and noise types along with the relative improvement provided by the proposed network in terms of speech intelligibility and quality for CI recipients.

To analyze performance of the proposed network in different ranges, objective scores were also calculated in both seen and unseen noisy conditions and presented in Table III. Performance was evaluated using two speech intelligibility metrics (STOI and SOPM), one speech quality metric (PESQ), and two speech distortion metrics (SISDR and LSD). The 'High', 'Medium', and 'Low' ranges represent mean objective scores for SNRs of -5 and -2.5 dB, 0 and 2.5 dB, and 5 and 10 dB, respectively. Each score for a particular group represents the average for a specific number of speech samples. For seen noise conditions, this represents the average for 300 speech samples (50 samples * 3 noises * 2 SNRs), while for unseen noise conditions, it represents the average score for 200 speech samples (50 samples * 2 noises * 2 SNRs). The scores are presented for the proposed DCCTN algorithm and four baseline networks (e.g., CRN, DCCRN, GCRN, and CFTNet). DCCTN outperforms all prior methods.

The proposed DCCTN algorithm offers several clear advantages over existing algorithms, as demonstrated by objective scores and performance evaluation. Objective scores for each network show improvement compared to original unprocessed speech, regardless of the group and noise type. However, relative improvement is higher for seen versus unseen noise conditions. The results indicate that the proposed algorithm consistently outperforms all baseline networks in every condition. It excels in both speech intelligibility and distortion reduction, making it a superior choice across the three SNR ranges. Unlike other algorithms, the proposed DCCTN algorithm shows relatively high improvement in speech intelligibility and distortion reduction in the challenging 'Red' zone, where speech intelligibility is significantly compromised. Additionally, the network achieves competitive speech quality scores even in unseen noise conditions, along with maintaining high speech intelligibility. This advantage is important as it demonstrates the algorithm's potential usability in naturalistic environments for CI recipients.

TABLE IV
COMPARISON OF DIFFERENT ALGORITHMS IN OBJECTIVE INTELLIGIBILITY, QUALITY, AND DISTORTION SCORE(ARROWS INDICATE DIRECTIONS FOR IMPROVEMENT)

Method	STOI↑	SOPM↑	PESQ↑	SISDR↑	LSD↓	IS↓
Noisy	0.70	0.75	1.29	0.35	7.00	1.31
CRN [16]	0.72	0.82	1.31	1.70	8.87	3.78
DCCRN [34]	0.80	0.86	1.62	4.90	5.24	1.01
GCRN [17]	0.76	0.86	1.45	3.59	5.45	1.63
CFTNet [40]	0.83	0.86	1.92	5.23	4.71	2.44
DCCTN	**0.85**	**0.91**	**2.07**	**7.44**	**3.38**	**0.67**

The mean objective scores for seven SNRs (-7.5, -5, -2.5, 0, 2.5, 5, and 10 dB) and five noises (three seen and two unseen) are provided in Table IV. DCCTN demonstrates significant performance improvements compared to noisy speech and baseline networks. Specifically, it achieves +21.4% improvement in STOI over the unprocessed signal, and relative improvements of +11.8%, +6.5%, and +2.4% in STOI over the CRN, GCRN, and DCCRN networks, respectively. Furthermore, it achieves a +7.8% improvement in PESQ and a remarkable +42.3% improvement in SISDR compared to our initial investigation network, CFTNet. These findings highlight the effectiveness of the proposed DCCTN algorithm in enhancing speech intelligibility, speech quality, and reducing speech distortion, especially for CI listeners.

TABLE V
ABLATION STUDY OF THE PROPOSED DCCTN. 'CPLX' REPRESENTS THE COMPLEX-VALUED NETWORK, 'CTN' REPRESENTS THE CONVOLUTIONAL TRANSFORMER NETWORK, AND 'N' REPRESENTS THE NUMBER OF FTL IN THE NETWORK.

Method	STOI↑	SOPM↑	PESQ↑	SISDR↑	LSD↓	IS↓
Noisy	0.69	0.75	1.29	0.35	7.00	1.31
CRN [16]	0.72	0.82	1.31	1.70	8.87	1.31
CRN+SkipConvNet	0.72	0.78	1.35	2.14	6.50	1.32
Cplx-CRN+SkipConv+FTL(n=2)	0.84	0.91	2.02	6.90	3.38	2.44
Cplx-CRN+SkipConv+FTL(n=4)	0.83	0.86	1.92	5.23	4.71	2.44
Cplx-CTN+SkipConv+FTL(n=2) (DCCTN)	**0.85**	**0.91**	**2.07**	**7.44**	**3.38**	**0.67**
DCCTN W/O SkipConvNet	0.84	0.91	2.07	7.30	3.66	2.30
DCCTN Real-valued	0.75	0.82	1.35	2.14	5.96	4.50

To illustrate the contributions of each block, we conducted a series of training and testing steps for the proposed model, presenting the mean objective scores in Table V. The inclusion of convolutional blocks in the skip connection led to improvements in speech quality and distortion, as indicated by enhanced PESQ, SISDR, and LSD scores while maintaining consistent speech intelligibility. Notably, the model exhibited significant performance gains when implemented in a complex domain, underscoring the efficacy of a complex-valued network compared to a real-valued one.

The convolution layers in the skip connection, CRNSkipConv (CRN+SkipConvNet), showcased a substantial +25.9% relative improvement in SISDR compared to CRN. Introducing complex-valued FTL (FTL only in the first and last layer) in the encoder demonstrated notable +16.7% and +51.1% relative enhancements in STOI and PESQ, respectively, compared to the real-valued CRNSkipConv (CRN+SkipConvNet). Notably, the model's performance exhibited diminishing returns with an increase in FTL layers, with the optimum FTL being 2 for the proposed network.

Further improvements were observed when the proposed DCCTN incorporated a transformer in the bottleneck layer. Specifically, DCCTN achieved relative improvements of +2.4%, +3%, and +72.4% in STOI, PESQ, and IS scores over Cplx-CRN+SkipConv+FTL. The absence of SkipConvNet in DCCTN had a noticeable impact on the objective scores. To evaluate the contribution of a complex-valued DCCTN over a real-valued network, we implemented a similar real-valued network as DCCTN, revealing that DCCTN outperformed its real-valued counterpart, highlighting the advantages of a complex network.

V. DISCUSSION

Cochlear implants provide electrical stimulation to the auditory nerve through an implanted array of electrodes. Each CI recipient will have their own CI MAP configuration, and therefore their unique residual hearing and individual auditory perception differences. An electrodogram is a two-dimensional time vs. electrode/channel representation of the auditory CI stimulation response. The electrode/channel is related to frequency over the auditory space [53]. Alternatively, linear predictive coefficients (LPC) provide information concerning spectral envelopes. Analyzing electrodogram and LPC responses provides insights into cochlear implant stimulation patterns.

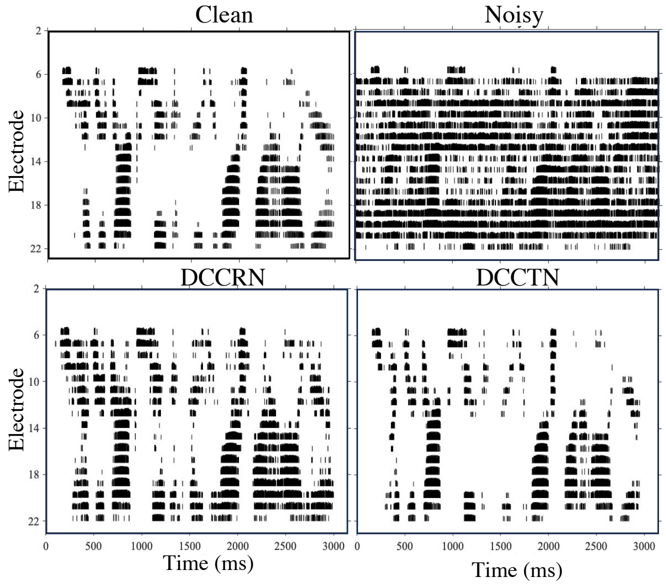

Fig. 6. CI electrode stimulation response (electrodogram) in babble noise at SNR 0 dB for (a) clean, (b) noisy, (c) DCCRN, and (d) DCCTN processed speech

A. Effect of Proposed Network on CI Electrodogram responses

The impact of background noise on speech perception for CI users is significant. Speech enhancement can help to mitigate this interference. To assess impact of the proposed algorithm on electrodograms, Fig. 6 illustrates electrodograms of processed signals. The original clean signal is corrupted with babble noise at an SNR of 0 dB. The noisy signal is then subjected to enhancement by both the baseline and proposed networks. A CI Advanced Combined Encoder (ACE) signal processing strategy [54] is subsequently utilized to simulate the received signal for RF pulse generation and generate the corresponding electrodograms. A standard CI parameter setting for biphasic electric RF pulse stimuli generation with 22 electrodes is employed.

To compare CI device output, this study presents electrodograms of the clean, noisy, baseline processed, and finally proposed network processed signal. Results demonstrate that the proposed DCCTN network effectively attenuates noise while preserving harmonic speech structure in the electrodogram. Alternatively, the baseline DCCRN processed signal exhibits residual noise, either retained or introduced, in the electrodogram, along with processing artifacts that ultimately decrease speech intelligibility for cochlear implant recipients.

B. Effect of DCCTN on Formant Restoration

In challenging listening environments, background noise and distortion degrade formant information, contributing to a

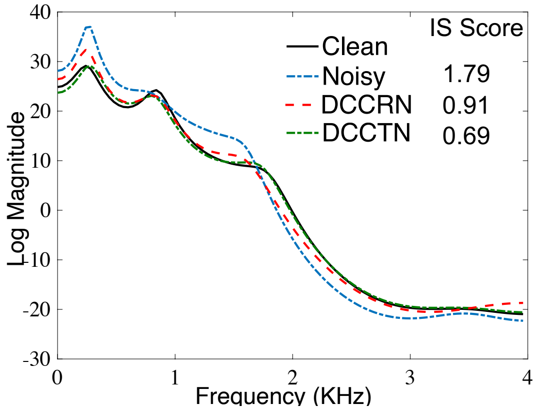

Fig. 7. Effects of the proposed network in harmonic restoration

low perception of phonetic sounds. SE can restore formant frequencies to improve the overall quality and naturalness of speech. Fig. 7 depicts formant analysis of speech signals under different conditions: clean, noisy, baseline DCCRN, and proposed DCCTN processed. Analysis reveals that LPC parameters for speech processed with DCCTN overlap closely with clean speech, indicating successful restoration of the formant frequencies. In contrast, the LPC parameters for DCCRN speech process align more closely with the noisy speech, albeit with lower magnitudes.

Overall, formant restoration through speech enhancement algorithms is crucial for improving speech perception and quality, especially in adverse listening conditions. This LPC analysis provides further evidence of the effectiveness of the DCCTN algorithm in recovering formant structure in speech.

VI. Conclusion

A machine learning-based SE method, entitled DCCTN, was proposed and validated to improve speech perception in naturalistic environments. The method employs a complex-valued U-Net architecture with a transformer integrated into the bottleneck layer. Additionally, a frequency transformation module was incorporated to successfully reconstruct harmonic components from the distorted speech. DCCTN is aimed at enhancing both magnitude and phase response of speech, contributing to improved speech quality. To validate the effectiveness of the proposed network, objective and subjective evaluations were conducted. The human listener evaluation revealed significant improvements achieved by DCCTN compared to the baseline network DCCRN, with a +40% increase in speech intelligibility and a +31% improvement in speech quality. Objective measures related to speech intelligibility, quality, and distortion also demonstrated similar performance. Furthermore, a pair-wise preference test was conducted, which indicated that DCCTN method was preferred by listeners regardless of SNR and noise conditions. These results highlight the potential of the proposed network to significantly enhance speech intelligibility for CI recipients, particularly in challenging environmental situations. Additionally, DCCTN holds promise as a preprocessor for future speech technology challenges such as speaker identification and speech recognition algorithms tailored to CI users. Overall, DCCTN provides a valuable contribution to the field, offering a potential solution to enhance speech perception and quality for CI listeners in various real-world scenarios.

Acknowledgment

The authors would like to thank all NH and CI subjects who voluntarily participated in the experiments. This work was supported by Grant No. R01 DC016839-02 from the National Institute on Deafness and Other Communication Disorders (NIDCD), National Institutes of Health (NIH).

References

[1] A. Natarajan, J. H.L. Hansen, K. H. Arehart, and J. Rossi-Katz, “An auditory-masking-threshold-based noise suppression algorithm gmmse-amt for listeners with sensorineural hearing loss,” *EURASIP Journal on Advances in Signal Proc.*, vol. 2005, pp. 1–16, 2005.

[2] F.-G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, “Cochlear implants: system design, integration, and evaluation,” *IEEE Reviews in Biomedical Engineering*, vol. 1, pp. 115–142, 2008.

[3] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[4] J. H.L. Hansen and M. A. Clements, “Constrained iterative speech enhancement with application to speech recognition,” *IEEE Trans. on Signal Proc.*, vol. 39, no. 4, pp. 795–805, 1991.

[5] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[6] N. Mamun, S. Majumder, and K. Akter, “A self-supervised convolutional neural network approach for speech enhancement,” in *2021 5th International Conference on Electrical Engineering and Information Communication Technology*, 2021, pp. 1–5.

[7] S. Zhao, T. H. Nguyen, and B. Ma, “Monaural speech enhancement with complex convolutional block attention module and joint t-f losses,” in *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, 2021, pp. 6648–6652.

[8] V.-A. Nguyen, A. H. Nguyen, and A. W. Khong, “Tunet: A block-online bandwidth extension model based on transformers and self-supervised pretraining,” in *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, 2022, pp. 161–165.

[9] B. L. Pellom and J. H.L. Hansen, “Text-directed speech enhancement using phoneme classification and feature map constrained vector quantization,” in *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 2, 1996, pp. 645–648.

[10] T. Vuong, Y. Xia, and R. M. Stern, “A modulation-domain loss for neural-network-based real-time speech enhancement,” in *IEEE ICASSP Inter. Conf. on Acoustics, Speech and Signal Proc.*, 2021, pp. 6643–6647.

[11] M. Sun, X. Zhang, T. F. Zheng *et al.*, “Unseen noise estimation using a separable deep autoencoder for speech enhancement,” *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 24, no. 1, pp. 93–104, 2015.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio,” *New Era for Robust Speech Recognition: Exploiting Deep Learning*, pp. 165–186, 2017.

[13] J. Chen and D. Wang, “Long short-term memory for speaker generalization in supervised speech separation,” *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[14] S. R. Park and J. Lee, “A fully convolutional neural network for speech enhancement,” *ISCA Interspeech*, pp. 1993–1997, 2016.

[15] N. Mamun, S. Khorram, and J. H.L. Hansen, “Convolutional neural network-based speech enhancement for cochlear implant recipients,” in *ISCA Interspeech*, 2019, pp. 4265–4269.

[16] K. Tan and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement.” in *ISCA Interspeech*, 2018, pp. 3229–3233.

[17] K. Tan and D. Wang, “Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement,” *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 28, pp. 380–390, 2019.

[18] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, “Fully convolutional recurrent networks for speech enhancement,” in *IEEE ICASSP Inter. Conf. on Acoustics, Speech and Signal Proc.*, 2020, pp. 6674–6678.

[19] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[20] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.* IEEE, 2022, pp. 6857–6861.

[21] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 30, no. 4, pp. 679–681, 1982.

[22] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 2, pp. 126–137, 1999.

[23] N. Zheng and X.-L. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 27, pp. 63–76, 2018.

[24] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, 2015, pp. 708–712.

[25] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *IEEE Asia-Pacific Signal and Info. Proc. Assoc. Annual Summit and Conf. (APSIPA ASC)*, 2017, pp. 006–012.

[26] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, 2019, pp. 5756–5760.

[27] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[28] A. Pandey and D. Wang, "Exploring deep complex networks for complex spectrogram enhancement," in *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, 2019, pp. 6885–6889.

[29] J. Lee and H.-G. Kang, "A joint learning algorithm for complex-valued tf masks in deep learning-based single-channel speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1098–1108, 2019.

[30] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.

[31] Y. Sun, L. Yang, H. Zhu, and J. Hao, "Funnel deep complex u-net for phase-aware speech enhancement." in *Interspeech*, 2021, pp. 161–165.

[32] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, 2022, pp. 7857–7861.

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

[34] 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," *ISCA Interspeech*, pp. 2472–2476, 2020.

[35] S. Lv, Y. Hu, S. Zhang, and L. Xie, "Dccrn+: Channel-wise subband dccrn with snr estimation for speech enhancement," *ISCA Interspeech*, pp. 1–5, 2021.

[36] C. Plapous, C. Marro, and P. Scalart, "Speech enhancement using harmonic regeneration," in *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 1, 2005, pp. 157–160.

[37] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura, and Y. Yamashita, "Single-channel speech enhancement with phase reconstruction based on phase distortion averaging," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 26, no. 9, pp. 1559–1569, 2018.

[38] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *AAAI Conf. on Artificial Intell.*, vol. 34, no. 05, 2020, pp. 9458–9465.

[39] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[40] N. Mamun and J. H.L. Hansen, "CFTNet: Complex-valued frequency transformation network for speech enhancement," *ISCA Interspeech*, vol. 2023, pp. 809–813, 2023.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 1–11, 2017.

[42] K. Ramesh, C. Xing, W. Wang, D. Wang, and X. Chen, "Vset: A multimodal transformer for visual speech enhancement," in *IEEE ICASSP Inter. Conf. on Acoustics, Speech and Signal Proc.*, 2021, pp. 6658–6662.

[43] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, 2022, pp. 7847–7851.

[44] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[45] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" IEEE, 2019, pp. 626–630.

[46] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[47] H. Younis and J. H.L. Hansen, "CCi-CLOUD: A framework for community-based remote cochlear implant user experiments based on the CCi-MOBILE research platform," *The Journal of the Acoustical Society of America*, vol. 152, no. 4, pp. 141–141, 2022.

[48] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[49] N. Mamun, M. S. Zilany, J. H.L. Hansen, and E. E. Davies-Venn, "An intrusive method for estimating speech intelligibility from noisy and distorted signals," *Journal of Acoustic Society of America*, vol. 150, no. 3, pp. 1762–1778, 2021.

[50] 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," vol. 2, 2001, pp. 749–752.

[51] J. S. Erkelens, M. R. J. Bastiaans, and J. J. M. Kalker, "Perceptual evaluation of speech quality based on log-spectral distortion," *IEEE ICASSP Inter. Conf. on Acoustics, Speech, and Signal Proc.*, pp. 661–664, 1994.

[52] F. Itakura, "Analysis/synthesis of telephony speech based on the maximum likelihood method," *Reports on 6th Int. Cong. Acoust., 1968*, 1968.

[53] N. Mamun, R. Ghosh, and J. H.L. Hansen, "Quantifying cochlear implant users' ability for speaker identification using CI auditory stimuli," in *ISCA Interspeech*, 2019, pp. 3118–3122.

[54] M. W. Skinner, L. K. Holden, L. A. Whitford, K. L. Plant, C. Psarros, and T. A. Holden, "Speech recognition with the nucleus 24 SPEAK, ACE, and CIS speech coding strategies in newly implanted adults," *Ear and Hearing*, vol. 23, no. 3, pp. 207–223, 2002.