Multi-Channel Speech Enhancement using a Minimum Variance Distortionless Response Beamformer based on Graph Convolutional Network

Nguyen Huu Binh, Duong Van Hai, Bui Tien Dat, Hoang Ngoc Chau and Nguyen Quoc Cuong*
School of Electrical and Electronic Engineering, Hanoi University of Science and Technology
Hanoi, Vietnam
*Corresponding author

***Abstract*—The Minimum Variance Distortionless Response (MVDR) beamforming algorithm is frequently utilized to extract speech and noise from noisy signals captured from multiple microphones. A frequency-time mask should be employed to compute the Power Spectral Density (PSD) matrices of the noise and the speech signal of interest to obtain the optimal weights for the beamformer. Deep Neural Networks (DNNs) are widely used for estimating time-frequency masks. This paper adopts a novel method using Graph Convolutional Networks (GCNs) to learn spatial correlations among the different channels. GCNs are integrated into the embedding space of a U-Net architecture to estimate a Complex Ideal Ratio Mask (cIRM). We use the cIRM in an MVDR beamformer to further improve the enhancement system. We simulate room acoustics data to experiment extensively with our approach using different types of the microphone array. Results indicate the superiority of our approach when compared to current state-of-the-art methods. The metrics obtained by the proposed method are significantly improved, except the Scale-Invariant Source-to-Distortion Ratio (SI-SDR) score. The Perceptual Evaluation of Speech Quality (PESQ) score shows a noticeable improvement over the baseline models (i.e., 2.207 vs. 2.104 and 2.076). Our implementation of the proposed method can be found in the following link: *https://github.com/3i-hust-asr/gnn-mvdr-final*.**

***Keywords*—Multi-channel speech enhancement; graph convolutional networks; minimum variance distortionless response beamformer; complex ideal ratio mask**

I. Introduction

Speech enhancement is a subject studied and applied in many applications, e.g., automatic speech recognition, teleconference, or aided hearing [1-4]. There are two algorithm categories of speech enhancement: single-channel algorithms (using a single microphone) [5-8] and multi-channel algorithms (using multi microphones) [9], [10]. The performance of multi-channel algorithms is generally better than that of single-channel algorithms because they use not only statistical information related to signals but also more spatial information [11].

Speech enhancement using microphone array beamforming is a type of multi-channel approach. The basis of these techniques is to enhance signals from desired directions (signals of interest) and attenuate signals from uninterested directions (noise signals). One of the beamforming algorithms used in speech enhancement is the MVDR beamforming [12-14]. There are two conventional approaches to determine MVDR filter weights for noise reduction and/or dereverberation.

The first approach is to estimate the characteristic vector of Acoustic Transfer Functions (ATF) from the speech source to the microphone array based on a priori assumptions such as the position of the desired signal source, the microphone array's configuration, and room acoustics. In a real environment, the performance of this approach is reduced since the effect of multi-paths [15], [16].

The second approach does not involve any such a priori assumptions. Instead of using an ATF vector, a Relative Transfer Function (RTF) vector is estimated based on data collected from the microphone array. The (RTF) vector is defined as the (ATF) vector normalized to a reference microphone of the microphone array. The (RTF) vector estimate is calculated from (PSD) matrices of noise and desired signal. A time-frequency mask is used to estimate the matrices. There are some techniques to create the mask [17], [18].

Recently DNNs have been widely used for speech-related task for better robustness and performance [10], [19-23]. GCNs are considered a generalization of Convolutional Neural Networks (CNNs) [24]. In [23], the GCNs are used to learn spatial features and incorporate them with a U-net to estimate a cIRM. The cIRM is used directly to estimate clean speech based on spectral information obtained from multi-channel. Some experiments in [23] show that speech enhancement using GCNs and U-Net has results that outperform the prior state-of-the-art approach.

This paper adopted the idea of using GCNs and U-Net architecture of [23] for a speech enhancement system with two contributions. Firstly, instead of using the number of nodes of GCNs in [23] as the number of microphones, we increase the node number of GCNs. It helps GCNs learn spatial features more precisely. Therefore a cIRM can be better estimated by incorporating GCNs in the U-Net architecture. Secondly, we use an MVDR beamformer based on the obtained cIRM to estimate the clean speech rather than the attention layer as in [23].

These works are implemented and tested on the dataset provided from ConferencingSpeech2021 Challenge [25]. The results demonstrate that the combination between MVDR beamforming and GCNs improves the performance of the speech enhancement system. The metrics obtained by the proposed method are significantly improved, except for the SI-SDR score. The PESQ score shows a noticeable improvement over the baseline models (i.e., 2.207 vs. 2.104 and 2.076).

The rest of the paper is organized as follows. In Section II, a brief basis of speech enhancement, MVDR beamformer, GCNs, as well as evaluation metrics is described. In Section III, we review some related works. In Section IV, the proposed speech enhancement approach is detailed. In Section V, we explain some experimental setups and analyze the results of the proposed approach. Finally, the conclusions are presented in Section VI.

II. Preliminary

A. Speech Enhancement

The noisy speech signal can be represented in the Short-Time Fourier Transform (STFT) domain:

$$X(t,f) = S(t,f) + N(t,f) \tag{1}$$

where $X(t,f)$ represents the complex-valued time-frequency (t,f) bin of noisy speech, $S(t,f)$ denotes the reverberated signal received at the microphone and $N(t,f)$ indicates the interference noise at time frame t and frequency bin f with $t = 0, ..., T-1$ and $f = 0, ..., F-1$. T and F are the number of frames and frequency bins, respectively. The neural beamformer here focuses on the task of suppressing noise. The objective is to remove the interference noise $N(t,f)$ and retrieve the speech signal $S(t,f)$.

Deep learning-based speech enhancement approaches are usually designed in a supervised manner. Based on how to obtain the target, the applied techniques can be classified into *mapping-based* or *masking-based* methods. In *mapping-based* approaches, the goal is to approximate a non-linear function from the noisy speech into the desired speech through a learning process. Meanwhile, the most popular methods recently used are *masking-based*, where the target is masks computed between desired and noisy speech.

The masking-based approaches try to approximate a non-linear function from an observed noisy speech spectrum $X(t,f)$ to a Time-Frequency $(T-F)$ mask $M(t,f)$ through the learning/training process. The commonly used masks in recent researches include: binary-based mask [17] and ratio-based mask [18].

The binary-based mask usually indicates the Ideal Binary Mask (IBM). Each entry of the $T-F$ mask is set to 1 when the local Signal-to-Noise Ratio (SNR) is greater than a predefined threshold value R (indicates that speech is dominated over noise), or 0 if otherwise (indicates that noise is dominated over speech). In particular,

$$M_{\text{IBM}}(t,f) = \begin{cases} 1, & \text{if } SNR(t,f) > R. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

here $SNR(t,f)$ indicates the SNR at the frame index t and the frequency bin f within the $T-F$ mask.

Typical ratio-based mask commonly refers to Ideal Ratio Mask (IRM), where each entry of the $T-F$ mask is set by the soft ratio of the reverberated speech over the observed noisy signal, that is:

$$M_{\text{IRM}}(t,f) = \frac{|S(t,f)|^{\alpha}}{|S(t,f)|^{\alpha} + |N(t,f)|^{\alpha}} \tag{3}$$

here $|S(t,f)|$ indicates the magnitudes of reverberated speech, $|N(t,f)|$ denotes the noise in the $T-F$ domain, and α is a factor over the magnitudes, which is to scale the value of each entry of the mask or change the dynamic ranges of the features. From Eq. 2 and 3, we could deduce that IRM-based methods could provide an enhanced signal with less distortion, while it may possibly lead to much computation [26].

Williamson et al. [27] further improved this approach, called cIRM. The complex ratio mask (CRM), demonstrated to be more effective than the ideal ratio mask (IRM) [28-30]. Given the complex spectrum of noisy speech, $X(t,f)$, we get the spectrum of reverberated speech, $S(t,f)$, that is:

$$S(t,f) = X(t,f) \odot M_{\text{cIRM}}(t,f) \tag{4}$$

where $\odot$ is the element-wise multiplication. Note that, $X(t,f)$, $S(t,f)$ and $M_{\text{cIRM}}(t,f)$ are complex-valued matrices.

Given the observed input noisy signals $X(t,f)$ from the $T-F$ domain and the target mask $M(t,f)$, the deep neural networks are optimized by the Mask Approximation (MA) objective function, which minimizes the Mean Squared Error (MSE) loss between the estimated and the target mask. On the other hand, recently, more approaches have been starting to employ Signal Approximation (SA) objective functions [31-33]. This objective aims to minimize the MSE loss between the estimated and the target reverberated speech spectrum. Besides, another approach minimizes MSE between the estimated reverberated signal and the target one in the time-domain by additionally applying inverse STFT. Furthermore, the conclusions in [31], [34] show that mixing the objectives (i.e., MA and SA) could lead to further improvement in both the magnitude and the spectral domains.

B. MVDR Beamformer

The separated speech can be obtained as

$$\hat{S}_{MVDR}(t,f) = \mathbf{h}^{\mathbf{H}}(f) X(t,f) \tag{5}$$

here $\mathbf{h}(f) \in \mathbb{C}^M$ denotes the weights of MVDR beamformer at frequency index f, M denotes the number of channels and $(\cdot)^{\mathbf{H}}$ indicates Hermitian operation. The main target of the MVDR beamformer is to suppress the interference noise while keeping the desired signal undistorted as much as possible, that is:

$$\begin{aligned} \mathbf{h}(f) = \underset{\mathbf{h}}{\text{argmin}}\ & \mathbf{h}^{\mathbf{H}}(f) \mathbf{\Phi}_N(f) \mathbf{h}(f) \\ \text{s.t.}\quad & \mathbf{h}^{\mathbf{H}}(f) \mathbf{v}(f) = 1 \end{aligned} \tag{6}$$

Here $\mathbf{\Phi}_N(f)$ is the PSD matrix of the noise, and $\mathbf{v}(f) \in \mathbb{C}^M$ represents the steering vector to the target source.

Different approaches could be adopted to find the optimal weights of the MVDR beamformer. To reduce the computation in the beamforming block, we employ the MVDR solution of Souden et al. [12]:

$$\mathbf{h} = \frac{(\mathbf{\Phi}_N(f))^{-1} \mathbf{\Phi}_S(f)}{trace((\mathbf{\Phi}_N(f))^{-1} \mathbf{\Phi}_S(f))} \mathbf{u} \tag{7}$$

where $\mathbf{\Phi}_S(f) \in \mathbb{C}^{M\times M}$ is the PSD matrix of speech, $\mathbf{\Phi}_N(f) \in \mathbb{C}^{M\times M}$ indicates the PSD matrix for noise. $\mathbf{u} \in \mathbb{R}^M$ is a one-hot vector representing a reference microphone.

Two masks, $M_S(t,f)$ and $M_N(t,f)$, will be used to estimate the desired PSD matrices:

$$\mathbf{\Phi}_S(f) = \sum_{t=1}^{T} (X(t,f) \odot M_S(t,f))(X(t,f) \odot M_S(t,f))^{\mathbf{H}} \tag{8}$$

$$\mathbf{\Phi}_N(f) = \sum_{t=1}^{T} (X(t,f) \odot M_N(t,f))(X(t,f) \odot M_N(t,f))^{\mathbf{H}} \tag{9}$$

where $\odot$ is the element-wise multiplication.

C. Evaluation Metrics

The standard metric to measure the performance of Automatic Speech Recognition (ASR) systems is Word Error Rate (WER). However, other objective metrics are also employed to evaluate the performance of the front-end techniques, such as denoising. The typical metrics include Short-Time Objective Intelligibility (STOI) [35], Extended Short-Time Objective Intelligibility (ESTOI) [36], SI-SDR [37], and PESQ [38]. It is worth looking at [39] for more detailed definitions and explanations of the objective metrics.

D. Graph Neural Networks

1) Definition: Graph Neural Network (GNN) [24] is a new type of deep neural network designed to work with graph data. Graphs provide a much more flexible way to process and aggregate information. GNN allows for generalizing DNN operations to graph-structured processing [40], [41]. By aggregating information from neighboring nodes, GNN models encode structural-relational information into the representation, which then is applied in a wide range of tasks, including biochemical structure discovery [42], [43], computer vision [44], and recommendation systems [45].

A specific variance of GNN is the convolutional GNN (so-called GCN) which is similar to CNN [46] with the basis of shared weights through training. There are two approaches for building GCN, Spectral GCN and Spatial GCN [47-49]. Spectral GCN, infrequently used nowadays, is based on the Eigen-decomposition of graph Laplacian. Spatial GCN defines convolution operations that work directly on a graph through the nodes and edges and aggregate spatial information between neighboring nodes and edges. Therefore, Spatial GCN is less computational and complex and can generalize better than spectral GCN. Recently, new convolutional GNN structures [47] have drastically leveraged the performance of GNN by employing various techniques, including normalization [46], attention [50], and activation [51].

2) Graph Convolutional Network: Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of nodes v_i of the graph and $\mathcal{E}$ represents the edges of the graph between two nodes (v_i, v_j). The GCN applies non-linear transformation on the input $\mathcal{X} \in \mathbb{R}^{|\mathcal{V}|\times N}$, where $|\mathcal{V}|$ is the number of nodes and N is node feature size. In particular, GCN can be mathematically represented as follows:

$$\mathcal{H}^{(l)} = g\big(\mathcal{D}^{-1/2}\mathcal{A}\mathcal{D}^{-1/2}\mathcal{H}^{(l-1)}\mathcal{W}^{(l-1)}\big) \tag{10}$$

where $\mathcal{D} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is the diagonal matrix, $\mathcal{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is the adjacency matrix, $\mathcal{H}^{(l)} \in \mathbb{R}^{|\mathcal{V}|\times K}$ is the l^{th} GCN layer with K hidden features, $\mathcal{H}^{(0)} = \mathcal{X}$, $\mathcal{W}^{(l-1)}$ is the trainable parameters at the $l-1^{th}$ layer, and g is a non-linear activation function.

III. Related Work

A. LSTM-Based Speech Enhancement

For the recognition of sequence-based data, context information is essential. Certain straightforward approaches for processing context-dependent data have been adopted, such as concatenating several consecutive features to construct long-context input features [52]. Moreover, Recurrent Neural Networks (RNNs), especially the Long Short-Term Memories (LSTMs), have been experimented with to be able to capture the information of the long sequence [53-56].

ConferencingSpeech 2021 Challenge[1] [25] adopted LSTM to suppress noise in distorted input signal. In particular, the multi-channel noisy speech is converted into frequency-domain by STFT transformation.

The STFT features were stacked with "cosIPD" features [57], a smoother version of Inter-channel Phase Difference (IPD), to obtain the input features, and then used to train the model. A 3-layer real-valued LSTM is used to capture the temporal information of input features. The output of the LSTM model is treated as the cIRM then a real-valued fully connection layer is added to map the output into real and imaginary components of the mask, respectively.

The cIRM mask was multiplied with the first microphone channel of STFT features of noisy speech to filter out the noise. The model was trained with SA objective functions, minimizing MSE between the estimated reverberant signal and the target signal in time-domain. Code and samples can be found in this repository[2].

B. GCN-Based Speech Enhancement

Recently proposed solutions are introduced to tackle the problem of speech enhancement by employing DNN models with spatial post-filtering techniques such as the filter-and-sum beamformer [23], [58-60]. Tzirakis et al. [23] proposed a novel approach by treating each audio channel (microphone) as a node of a graph structure.

The well-known U-Net architecture was incorporated to learn representations for the inputs, especially in speech improvement problems [61], [62]. GCN was used in the embedding space of a U-Net architecture to learn spatial correlations between the different nodes (or channels/microphones).

This approach utilizes both real and imaginary parts of the complex features in the STFT domain. Complex spectrograms from each channel are fed into the encoder part of the architecture. The higher level features obtained after the Encoder part are used to construct the multi-channel GCN. After that, the Decoder produces the estimated cIRM for a reference microphone with the same dimension as the input. Finally, cIRM for noisy STFT features is computed, which is used to estimate the desired clean speech.

[1] https://tea-lab.qq.com/conferencingspeech-2021/

[2] https://github.com/ConferencingSpeech/ConferencingSpeech2021

IV. Proposed Multi-Channel Processing Method

A. Proposed System

This paper proposes a novel pipeline for multi-channel speech enhancement tasks by incorporating GNN with a neural MVDR beamformer. GNN is successfully applied in a wide range of tasks with structural data, including computer vision [44] and speech processing [23]. At the same time, MVDR demonstrates its superior performance over the Filter-and-Sum technique [63]. However, to the best of our knowledge, no prior work focuses on integrating both these techniques to tackle the speech enhancement problem.

The overall proposed model is schematically depicted as Fig. 1. The whole pipeline consists of three main processes:

1) Signal transformation includes signal conversion from the time domain to the STFT domain (STFT block) and signal inversion from the STFT domain back to the time domain (iSTFT block);
2) Mask estimation for both clean speech and noise (Mask Estimator block);
3) Applying MVDR for noise suppression (MVDR block). However, only the parameters in the Mask Estimator block are trainable (details illustrated in Algorithm 1).

The two signal transformation blocks are trivial, while the mask estimation and the MVDR blocks are more advanced and described below.

1) Mask Estimator Block: Firstly, we employ the well-known U-Net architecture for mask estimation blocks. The u-Net model has been shown very successfully in many computer vision tasks and used in some recent approaches in speech processing [64]. In addition, U-Net architecture comprises an encoder/decoder part with an embedding layer. The encoder/decoder parts in U-Net are set to be cascaded CNN layers, while the GCN is used as a core embedding layer. We adopt this idea from the currently published approach of [23].

Secondly, our proposed U-Net model is unique and different from [23] because we use other graph construction methods for the embedding layer. In [23], the input channels, denotes as M, is preserved as GCN nodes, so that, GCN has only M nodes ($M = \{2, 4\}$ as reported in [23]). With such a few nodes, we realize that GCN's ability to capture spatial information is restricted.

From that deduction, we propose a novel graph construction method such that GCN's nodes are now set equal to the number of channels (kernels) of the last CNN layer in the encoder part of U-Net. In detail, suppose that the STFT features with the shape of $(M \times T \times 2F)$ dimensional feeds to the encoder. The latter then represents with the dimension of $(H \times T' \times F')$, where H is the number of kernels of the last CNN layer in the encoder, while T', F' are the reduced size of T and F after all layers of CNN in the encoder, respectively. As a result, the number of nodes in GCN is H, allowing us to choose an appropriate number of nodes during the training process. For more detail about the graph construction process, see Section IV-C.

Finally, outputs of the mask estimation process are two masks for clean speech and noise, denote as *mask_speech* and *mask_noise* in Fig. 1, respectively.

2) MVDR Beamformer: We utilize a neural MVDR beamformer as posterior filtering to leverage the enhancement performance. The MVDR uses these aforementioned estimated masks from the previous step to compute beamforming weights on-the-fly. The detailed computation is introduced in section II-B. The process of MVDR beamformer is integrated with mask estimation and trained as a unified model so that, making the enhancement system more robust.

B. System Procedure

The overall procedure of this approach in Fig. 1 is shown in Algorithm 1.

First, the multi-channel speech signals are transformed into the time-frequency domain with STFT transformation. The components of the complex STFT features are stacked together to create a new feature with a two-channel of size $(T \times F \times 2)$, where T denotes the number of frames, and F indicates the number of frequency bins, in total.

Considering M channels, the input features will have the shape of $(M \times 2 \times T \times F)$ dimensional, in which each entry is a real value. After that, these STFT input features are fed to the Encoder of Mask Estimator, which produces more complex and high-level representations. The feature is reshaped into $(M \times T \times 2F)$ dimensional to fit the requirement of our proposed method. The Encoder then produces representations with the dimension of $(H \times T' \times F')$, where H is a number of filters of the last CNN layer in the encoder. At the same time, T' and F' are the reduced size of T and F after the entire layers of CNN in the encoder, respectively. Next, the GCN is used as a core embedding layer between the encoder and decoder parts. The representations produced from the Encoder are utilized in constructing a graph with H nodes, which captures the spatial information by aggregating the information of its nodes and edges. The GCN construction process is described in detail in Section IV-C.

The output of GCN layers is forwarded through the decoder part, which converts the hidden features to the original dimension. The decoder outputs could be treated as two cIRM masks for clean and noisy speech, then used to compute the PSD matrices and MVDR weights. Estimated STFT features of reverberant speech $\hat{S}$ are computed by applying MVDR processing as in Section II-B. Finally, the inverse STFT transformation is applied to obtain the estimated reverberant speech in the time domain.

C. Graph Construction

We adopt a recently published approach that captures multi-channel signal information with graph structure [23]. The graph structure is first constructed using the hidden feature representations obtained after the Encoder. We construct an undirected graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of nodes v_i of the graph. For example, with hidden features of shape $(H \times T' \times F')$ from the previous step, a graph with H nodes with a feature size is $N = T'F'$ will be constructed by flattening function. $\mathcal{E}$ represents the edges of the graph

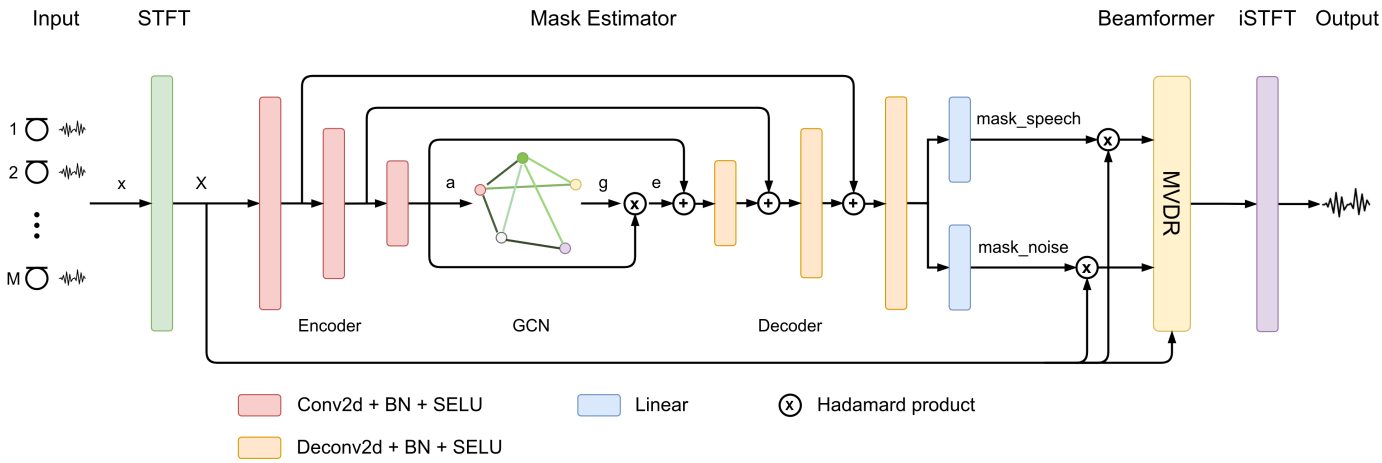

Fig. 1. Proposed MVDR System using GNN-Based U-Net Architecture as mask Estimator. The Estimated Masks are then used to Compute the Beamformer Weights and Applied to a Noisy Signal to Provide the Enhanced Signal.

between two nodes (v_i, v_j). Then, an adjacency matrix of the graph, $\mathcal{A} \in \mathbb{R}^{H \times H}$, is computed.

We also employ a learnable adjacency matrix, where each entry of $\mathcal{A}$ is treated as a weighted obtaining from edge $(v_i, v_j) \in \mathcal{E}$. Intuitively, each entry represents a similarity between two nodes in the graph. In our approach, the learnable weights, $w_{ij}, \{i, j \in \mathcal{V}\}$, of the adjacency matrix $\mathcal{A}$ are optimized during the training process. For two nodes v_i and v_j, their representations will be concatenated, $\mathbf{f}_{v_i}, \mathbf{f}_{v_j} \in \mathbb{R}^N$ as $[\mathbf{f}_{v_i} \| \mathbf{f}_{v_j}]$ and then passed through a non-linear layer $F([\mathbf{f}_{v_i} \| \mathbf{f}_{v_j}])$. The node degree matrix $\mathcal{D}$ is a diagonal matrix, where $\mathcal{D}_{ii} = \sum_j \mathcal{A}_{ij}$.

The graph constructed $\mathcal{G}$ provides an efficient way to capture the structured information from its nodes (e.g., microphones). We use the GCN to produce high-level abstraction for the hidden node representations by learning aggregated features for each node w.r.t its neighbors. The mathematical detail of the GCN layer can be seen in Section II-D2.

D. Loss Functions

In the training process of the proposed network, we adopt loss computations in different forms. We use a loss function with magnitude features and raw signals in the time domain. More specifically, these losses are:

$$\mathcal{L}_{mag} = \left\| |\hat{S}| - |S| \right\|_1 \tag{11}$$

$$\mathcal{L}_{raw} = \left\| \hat{s} - s \right\|_1 \tag{12}$$

$$\mathcal{L} = \mathcal{L}_{mag} + \mathcal{L}_{raw} \tag{13}$$

where $\| \cdot \|_1$ indicates the L1 norm, $|S|$ indicates the magnitude spectrogram of the complex spectrogram S, s indicates reverberant signal, and $\hat{}$ sign indicates the corresponding predicted entities.

V. Experimental Evaluation

A. Dataset

With some missing information about the dataset configurations and the authors did not publish the implementation of their proposed model in [23], we decided to utilize the dataset provided from ConferencingSpeech 2021 Challenge [25]. The simulation set was provided for all participants to develop the enhancement systems and estimate the objective scores. To focus on the development of algorithms, the authors designed the challenge with the *close* training condition. In other words, only the provided list of open-source clean speech and noise datasets could be used in the training process.

1) Training Set: Clean training speech set signals are chosen from three open source speech databases: AISHELL-1[3] [65], AISHELL-3[4] [66], and Librispeech [67]. The speech utterances with SNR higher than 15 dB are selected for training. The total duration of the clean training example is around 550 hours. The noise set is selected from MUSAN [68] and AudioSet [69]. The total duration is around 120 hours.

The imaging method is used to simulate Room Impulse Response (RIR) for three types of microphone arrays: (i) a linear microphone array with uniformly distributed 8 microphones, (ii) a circular microphone array, and (iii) a linear microphone array with non-uniformly distributed 8 microphones. The room size ranged from $3 \times 3 \times 3$ m^3 to $8 \times 8 \times 3$ m^3, and provided RIR set contains more than 2500 rooms.

The microphone array is randomly placed in the room with a height ranging from 1.0 to 1.5 m. The sound source, including speech and noise, comes from any possible position in the room with a height ranging from 1.2 to 1.9 m. The angle between two sources is wider than 20°. The distance between the source and microphone array are ranged from 0.5 to 5.0 m. The total number of RIR is more than 10000 for each microphone array. The simulated SNR ranges from 0 to 30 dB, and the duration of each clip is 6 seconds.

[3]https://www.openslr.org/33/
[4]http://www.openslr.org/93/

Algorithm 1 Summary the Process of the Proposed Model for Multi-Channel Speech Enhancement

Require: M-channel noisy speech $x \in \mathbb{R}^{M \times L}$ in time domain; reference microphone vector $\mathbf{u}$.

Ensure: Enhanced speech $\hat{s} \in \mathbb{R}^{L}$ in time domain.

$X = STFT(x)$ ▷ $X \in \mathbb{R}^{M \times 2 \times T \times F}$

$X = reshape(X, [M, T, 2 * F])$ ▷ $X \in \mathbb{R}^{M \times T \times 2F}$

skips $= [\ ]$

$a = copy(X)$ ▷ $a \in \mathbb{R}^{M \times T \times 2F}$

for *enc_layer* **in** *Encoder* **do** ▷ UNet Encoder

    $a =$ *enc_layer*(a)

    skips.append(a)

end for

$\mathcal{A} =$ *construct_adj(a)* ▷ $a \in \mathbb{R}^{H \times T' \times F'}, \mathcal{A} \in \mathbb{R}^{H \times H}$, see Section IV-C

$g = GCN(\mathcal{A}, a)$ ▷ $g \in \mathbb{R}^{H \times T' \times F'}$, Applying GCN, see Equation 10

$e = a \otimes g$ ▷ $e \in \mathbb{R}^{H \times T' \times F'}$, $\otimes$ Hadamard product

for *dec_layer* **in** *Decoder* **do** ▷ UNet Decoder

    skip $=$ *skips.pop*(-1)

    $e = e +$ *skip*

    $e =$ *dec_layer*(e)

end for

mask_speech $=$ *linear_speech*(e) ▷ $e \in \mathbb{R}^{M \times T \times 2F}$, *mask_speech* $\in \mathbb{R}^{M \times T \times 2F}$

mask_noise $=$ *linear_noise*(e) ▷ $e \in \mathbb{R}^{M \times T \times 2F}$, *mask_noise* $\in \mathbb{R}^{M \times T \times 2F}$

$\Phi_S =$ *get_psd_matrix*(*mask_speech*, X) ▷ $\Phi_S \in \mathbb{C}^{F \times M \times M}$, see Equation 8

$\Phi_N =$ *get_psd_matrix*(*mask_noise*, X) ▷ $\Phi_N \in \mathbb{C}^{F \times M \times M}$, see Equation 9

$\mathbf{h} =$ *get_mvdr_weights*$(\Phi_S, \Phi_N, \mathbf{u})$ ▷ $\mathbf{h} \in \mathbb{C}^{F \times M}$, see Equations 7

$\hat{S} = \mathbf{h}^{\mathbf{H}} X$ ▷ $\hat{S} \in \mathbb{R}^{T \times F \times 2}$, Applying MVDR, see Equation 5

$\hat{s} =$ *iSTFT*$(\hat{S})$ ▷ Enhanced speech $\hat{s} \in \mathbb{R}^{L}$

2) Development Set: The development set is categorized into three parts: Simulation clips, Semi-real recordings, and Real recordings. In this experiment, we only experiment with simulated audio with a single microphone array scenario (there also exists another task using multiple microphone arrays). 1588 clips are simulated for three types of the microphone array. 1624 clean speech selected from AISHELL-1, AISHELL-3, and 800 noise clips selected from MUSAN are used to simulate these sets. The simulated SNR ranges from 0 to 30 dB, and the duration of clips is 6 seconds.

B. Experimental Setup

For a convenient comparison with other approaches, we set up the training configurations as follows. The AdamW [70] optimization algorithm is adopted to optimize the proposed models with a fixed learning rate of 10^{-4} and a mini-batch of size 16. The number of microphones in the experiments is set to $M = 8$. The complex features are the STFT computed with a window of length 1024, the window's type is set to Hanning, and an overlap size of 512.

Our proposed model (MVDR-GCN) uses the optimized configuration. Each block in the Encoder (Decoder) part of U-net architecture comprises one CNN layer, followed by batch normalization and a SELU activation function. Each Encoder block's kernels of CNN layers are set to $\{64, 128, 128, 128, 32\}$, respectively. The blocks in the Decoder have the same configurations as the Encoder but in reverse order. All the kernel sizes of CNN layers are 3×3, and the stride is 2×2, with no padding.

For the embedding layer of U-Net, GCN, a bottle-neck layer is used with the hidden size of 64, then two GCN layers are integrated with hidden units as same as the dimension of the bottle-neck layer. In order to generate two masks for speech and noise, after the Decoder part, two Linear layers are added with an input size equal to the hidden size of the last CNN layer in the Decoder, while the output size is set to the same as feature size of STFT features.

The LSTM-based baseline model (Section III-A) is set up as same as the model in [25]. The model is composed of three layers of RNN with 512 hidden units. The input features are

TABLE I. DETAILED ENHANCEMENT RESULTS OF BASELINES AND OUR PROPOSED SYSTEM ON THE DEVELOPMENT SIMULATION SET

MA	Model	PESQ	STOI	E-STOI	SI-SDR
Linear	Noisy	1.551	0.807	0.716	4.673
	Baseline (LSTM)	2.091	0.867	**0.779**	**13.065**
	Baseline (GNN)	2.088	0.853	0.772	12.001
	Proposed System (MVDR-GCN)	**2.101**	**0.861**	0.773	10.996
Circular	Noisy	1.558	0.807	0.716	4.633
	Baseline (LSTM)	2.129	0.872	0.782	**13.118**
	Baseline (GNN)	2.077	0.852	0.771	12.007
	Proposed System (MVDR-GCN)	**2.260**	**0.886**	**0.804**	12.270
Non-uniform	Noisy	1.543	0.804	0.712	4.536
	Baseline (LSTM)	2.091	0.867	0.777	**13.042**
	Baseline (GNN)	2.061	0.851	0.768	11.769
	Proposed System (MVDR-GCN)	**2.261**	**0.889**	**0.806**	12.333

TABLE II. OVERALL RESULT OF BASELINES AND OUR PROPOSED SYSTEM ON THE DEVELOPMENT SIMULATION SET

Model	PESQ	STOI	E-STOI	SI-SDR
Noisy	1.551	0.806	0.715	4.614
Baseline (LSTM)	2.104	0.869	0.758	**13.075**
Baseline (GNN)	2.076	0.852	0.770	11.926
Proposed System (MVDR-GCN)	**2.207**	**0.879**	**0.794**	11.866

STFT stacking up with cosine of IPD features.

The GNN-based baseline model (Section III-B) has the same configurations as our proposed system, which includes input STFT features and parameters of layers in Encoder/Decoder blocks, except that number of kernels in each CNN layer, are slightly different from our proposed model ($\{64, 128, 256, 128, 32\}$). Note that this is the optimized configuration in [23].

Finally, a noisy scenario is obtained by directly computing the metrics with noisy data, simulated with SNR ranging from 0 to 10, and scaling from 0.2 to 0.9.

C. Enhancement Results and Evaluation

Our proposed approach's results are compared with Tzirakis et al. [23], a novel GCN-based multi-channel enhancement model. The detailed enhancement results are presented in Table I.

For overall comparison, the result of scenarios is averaged and reported as in Table II. Combining the MVDR algorithm showed an improvement in scores. The metrics obtained by the MVDR method are significantly improved, as expected, except for the SI-SDR score. The PESQ score achieved by the MVDR system shows a noticeable improvement over the baseline or GCN-based model (i.e., PESQ 2.207 vs. 2.104 and 2.076).

However, in the linear array scenario, our proposed system obtains worse scores than others, except for the PESQ metric, because of the distribution of the microphone in arrays. In circular and non-uniform scenarios, the position of microphones is various to capture more spatial information. The masks are more accurately estimated, and the model can achieve decent overall metrics. Conversely, the microphones are placed equidistant for the linear array, the information may be symmetric, and the mask estimator model may receive less information than others and get worse scores.

VI. CONCLUSION

In this approach, we propose a new method of using a graph neural network to exploit the spatial correlations among the different channels in the speech enhancement task. We use the U-Net architecture with the encoder, which tries to produce higher-level representations for each channel. After that, the GCN is constructed using these hidden features. GCN is used to learn spatial features by propagating and aggregating information in the graph. Then the features are fed to the decoder to reconstruct into the original forms of each channel. By integrating the GCN-based U-Net into the MVDR system, the experimental results validate our approach's effectiveness when compared with recent state-of-the-art approaches.

ACKNOWLEDGMENT

This research is funded by the Hanoi University of Science and Technology under grant number T2021-PC-003.

REFERENCES

[1] M. R. Bai, J.-G. Ih, and J. Benesty, *Acoustic array systems: theory, implementation, and application*. John Wiley & Sons, 2013.

[2] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5751–5755.

[3] S. A. Nossier, M. Rizk, N. D. Moussa, and S. el Shehaby, "Enhanced smart hearing aid using deep neural networks," *Alexandria Engineering Journal*, vol. 58, no. 2, pp. 539–550, 2019.

[4] C. R. Kumar and M. P. Chitra, "Implementation of modified wiener filtering in frequency domain in speech enhancement," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 2, 2022. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2022.0130251

[5] Y. Li and S. Kang, "Deep neural network-based linear predictive parameter estimations for speech enhancement," *IET Signal Processing*, vol. 11, no. 4, pp. 469–476, 2017.

[6] B. M. Mahmmod, S. H. Abdulhussian, S. A. R. Al-Haddad, W. A. Jassim *et al.*, "Low-distortion mmse speech enhancement estimator based on laplacian prior," *IEEE Access*, vol. 5, pp. 9866–9881, 2017.

[7] B. M. Mahmmod, T. Baker, F. Al-Obeidat, S. H. Abdulhussain, W. A. Jassim *et al.*, "Speech enhancement algorithm based on super-gaussian modeling and orthogonal polynomials," *IEEE Access*, vol. 7, pp. 103 485–103 504, 2019.

[8] N. Jamal, N. Fuad, and M. Shaabani, "A hybrid approach for single channel speech enhancement using deep neural network and harmonic regeneration noise reduction," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0111033

[9] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[10] S. Chakrabarty and E. A. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[11] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[12] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2009.

[13] R. Ali, T. Van Waterschoot, and M. Moonen, "Integration of a priori and estimated constraints into an mvdr beamformer for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2288–2300, 2019.

[14] J. Malek, Z. Koldovskỳ, and M. Bohac, "Block-online multi-channel speech enhancement using deep neural network-supported relative transfer function estimates," *IET Signal Processing*, vol. 14, no. 3, pp. 124–133, 2020.

[15] J. Benesty, M. M. Sondhi, Y. Huang *et al.*, *Springer handbook of speech processing*. Springer, 2008, vol. 1.

[16] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Eigenvector-based speech mask estimation for multi-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2162–2172, 2019.

[17] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[18] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[19] A. A. Alnuaim, M. Zakariah, A. Alhadlaq, C. Shashidhar, W. A. Hatamleh, H. Tarazi, P. K. Shukla, and R. Ratna, "Human-computer interaction with detection of speaker emotions using convolution neural networks," *Computational Intelligence and Neuroscience*, vol. 2022, 2022. [Online]. Available: https://doi.org/10.1155/2022/6005446

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

[21] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 531–535.

[22] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense u-net for multichannel speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 836–840.

[23] P. Tzirakis, A. Kumar, and J. Donley, "Multi-channel speech enhancement using graph neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3415–3419.

[24] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[25] W. Rao, Y. Fu, Y. Hu, X. Xu, Y. Jv, J. Han, Z. Jiang, L. Xie, Y. Wang, S. Watanabe *et al.*, "Interspeech 2021 conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing," *arXiv preprint arXiv:2104.00960*, 2021.

[26] E. M. Grais, G. Roma, A. J. Simpson, and M. Plumbley, "Combining mask estimates for single channel audio source separation using deep neural networks," *Interspeech2016 Proceedings*, 2016.

[27] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[28] Y. Xu, C. Weng, L. Hui, J. Liu, M. Yu, D. Su, and D. Yu, "Joint Training of Complex Ratio Mask Based Beamformer and Acoustic Model for Noise Robust Asr," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6745–6749, iSSN: 2379-190X.

[29] 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, Mar. 2016, conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[30] D. S. Williamson and D. Wang, "Time-Frequency Masking in the Complex Domain for Speech Dereverberation and Denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017, conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[31] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 577–581.

[32] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1562–1566.

[33] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[34] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4390–4394.

[35] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice-Hall, 1988.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm

for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[37] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[38] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[39] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[40] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[41] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[42] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*. PMLR, 2017, pp. 1263–1272.

[43] F. Wan, L. Hong, A. Xiao, T. Jiang, and J. Zeng, "Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions," *Bioinformatics*, vol. 35, no. 1, pp. 104–111, 2019.

[44] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing, "Rethinking knowledge graph propagation for zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 487–11 496.

[45] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 974–983.

[46] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[47] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

[48] M. Balcilar, G. Renton, P. Héroux, B. Gauzere, S. Adam, and P. Honeine, "Bridging the gap between spectral and spatial domains in graph neural networks," *arXiv preprint arXiv:2003.11702*, 2020.

[49] A. Senthilkumar, M. Gupte, and S. S, "Dynamic spatial-temporal graph model for disease prediction," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, 2022. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2022.01306112

[50] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[51] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *International conference on machine learning*. PMLR, 2019, pp. 6861–6871.

[52] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[53] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of selected topics in signal processing*, vol. 4, no. 5, pp. 867–881, 2010.

[54] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[55] S. Gupta, R. S. Shukla, R. K. Shukla, and R. Verma, "Deep learning bidirectional lstm based detection of prolongation and repetition in stuttered speech using weighted mfcc," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0110941

[56] M. A. A. Al-Rababah, A. Al-Marghilani, and A. A. Hamarshi, "Automatic detection technique for speech recognition based on neural networks inter-disciplinary," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 3, 2018. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2018.090326

[57] C. Deng, H. Song, Y. Zhang, Y. Sha, and X. Li, "Dnn-based mask estimation integrating spectral and spatial features for robust beamforming," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4647–4651.

[58] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 3246–3250, iSSN: 2379-190X.

[59] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 286–290, iSSN: 2379-190X.

[60] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 5210–5214, iSSN: 2379-190X.

[61] R. Giri, U. Isik, and A. Krishnaswamy, "Attention Wave-U-Net for Speech Enhancement," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2019, pp. 249–253, iSSN: 1947-1629.

[62] C. Macartney and T. Weyde, "Improved Speech Enhancement with the Wave-U-Net," Nov. 2018, arXiv:1811.11307.

[63] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[64] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[65] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[66] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[67] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[68] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[69] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[70] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.