Fusing Bone-Conduction and Air-Conduction Sensors for Complex-Domain Speech Enhancement

Heming Wang, *Graduate Student Member, IEEE*, Xueliang Zhang, *Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Speech enhancement aims to improve the listening quality and intelligibility of noisy speech in adverse environments. It proves to be challenging to perform speech enhancement in very low signal-to-noise ratio (SNR) conditions. Conventional speech enhancement utilizes air-conduction (AC) microphones, which are sensitive to background noise but capable of capturing full-band signals. On the other hand, bone-conduction (BC) sensors are unaffected by acoustic noise, but recorded speech has limited bandwidth. This study proposes an attention-based fusion method to combine the strengths of AC and BC signals and perform complex spectral mapping for speech enhancement. Experiments on the EMSB dataset demonstrate that the proposed approach effectively leverages the advantages of AC and BC sensors, and outperforms a recent time-domain baseline in all conditions. We also show that the sensor fusion method is superior to single-sensor counterparts, especially in low SNR conditions. As the amount of BC data is very limited, we additionally propose a semi-supervised technique to utilize both parallelly and unparallelly recorded AC and BC speech signals. With additional AC speech from the AISHELL-1 dataset, we achieve similar performance to supervised learning with only 50% parallel data.**

***Index Terms*—Speech enhancement, air-conduction, bone-conduction, attention-based fusion, complex spectral mapping.**

I. Introduction

NOISE interference degrades the quality and intelligibility of speech signals in real-world environments. Speech enhancement aims to remove or reduce the background noise of a given speech signal. The recent introduction of deep learning has led to dramatic advances in this field, and deep neural networks (DNNs) effectively suppress background noise for untrained speakers and noise types [11], [24], [42]. However, speech enhancement in non-stationary noises at very low SNRs remains challenging, as noise dominates the acoustic signal making it difficult to recover clean speech.

Conventional speech enhancement operates on speech recorded by air-conduction (AC) sensors or microphones. AC microphones can capture full-band speech, but are susceptible to background noise. Bone-conduction (BC) sensors directly convert articulation-induced vibrations on the human skull to electric signals [33]. As a result, BC signals are not subject to background interference transmitted in air. On the other hand, BC speech has a limited bandwidth as high-frequency components are attenuated or lost due to the nature of bone conduction, resulting in muffled sound.

In the speech telecommunication scenario where AC and BC signals are both available at the speaker end, how to leverage AC and BC recordings for speech processing before transmitting the processed result to the remote listener end becomes a significant research issue. In early efforts, BC signals are used to extract auxiliary speech information in noisy conditions, e.g., voice activity detection [54], SNR estimation [32] and pitch extraction [27]. Later, researchers attempt to extend the bandwidth of BC signals to improve speech quality. These methods can be categorized into three groups: equalization, analysis and synthesis, and DNN-based. Simulating BC signals by passing AC signals through a low-pass filter, Shimamura and Tamiya [31] proposed an equalization method that estimates the inverse of such transformation. Specifically, they derive a linear-phase filter by first calculating the ratio of long-term discrete Fourier transform of AC and BC speech spectra, and then taking the inverse and applying it to BC speech to recover the AC counterpart. Kondo et al. [17] improve the equalization method by estimating the filter in a frame-by-frame fashion. Although the proposed equalization method improves speech quality, the performance is sensitive to filter length and order and expected to degrade for unknown speakers. In addition, this approach mainly considers the magnitude ratio, and the phase is kept the same as that of the input signal, so perfect speech reconstruction is impossible in the ideal case. Analysis and synthesis models assume the excitation signals are the same for both AC and BC signals. The task is then to obtain the envelope feature for AC signals. Past work uses features like linear predictive coding (LPC) [38], mel-frequency cepstral coefficient (MFCC) [34], and linear spectral frequency (LSF) [12] to predict the spectral envelope of AC signals, and then perform speech synthesis. This

Manuscript received 29 January 2022; revised 10 June 2022 and 25 August 2022; accepted 5 September 2022. Date of publication 26 September 2022; date of current version 6 October 2022. This work was supported in part by NIDCD under Grant R01 DC012048 and in part by Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tom Bäckström. *(Corresponding author: Heming Wang.)*

Heming Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: wang.11401@osu.edu).

Xueliang Zhang is with the Department of Computer Science, Inner Mongolia University, Hohhot 010021, China (e-mail: cszxl@imu.edu.cn).

DeLiang Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2022.3209943

approach has several limitations. First, the assumption about the excitation does not always hold in real applications, causing distorted speech reconstruction. Second, excitation signals are hard to model as they are highly non-stationary. Recently DNN based methods are introduced to perform bandwidth extension on BC signals. Shan et al. [30] proposed a speaker-dependent approach to extend the bandwidth of BC speech. An encoder-decoder based network is employed to reconstruct the magnitude of AC speech, and magnitude-based features of spectral magnitude, MFCC and LPC are concatenated as the training input. Given the spectra of BC speech, Zheng et al. [51] introduce attention-based bidirectional long short-term memory (LSTM) to reconstruct the magnitude spectrogram of the corresponding AC speech. A structural similarity metric and a spectral distance metric are employed to guide optimization. Nguyen and Unoki [22] also employ bidirectional LSTM to recover AC speech. It predicts the LSFs of the corresponding full-band speech given the LSFs of BC speech, and then performs inverse filtering with the filter derived from the predicted LSFs to restore AC speech. Zheng et al. [52] use the vocoder WaveNet [23] to perform bandwidth extension for BC spectrograms, and attempt to reconstruct the full-band waveform from the bandwidth-limited BC magnitude spectrogram. Hussain et al. [14] proposed a hierarchical extreme learning machine to extend the bandwidth of BC spectrogram, which improves the automatic speech recognition accuracy with a limited amount of training data. Despite DNN-based methods showing improved performance, it remains challenging to recover high-resolution speech from BC speech alone. One reason is that the bandwidth of BC speech is usually limited to 1-2 kHz depending on sensor position [4], [15], [21], which makes it very difficult to perform bandwidth extension to 8 kHz or 16 kHz with high quality. As the majority of a spectrogram is missing, the extended spectrogram suffers from the over-smoothing issue [29]. The other reason is that low-intensity, wide-band sounds such as /f/ and /s/ are poorly captured by BC sensors as they induce weak, narrowband vibrations [26], making them especially hard to reconstruct via bandwidth extension.

Earbud devices like Apple Airpods have become popular consumer electronics, and they include both AC and BC sensors. For a typical bone-conduction earbud, the BC sensor is placed on the pinna and the AC sensor serves as a close-talk microphone, making it easier to obtain parallelly recorded AC and BC speech. A recent study by Yu et al. [47] proposed a DNN-based method that regards BC sensors as another modality. They investigate ensemble learning methods to integrate the two types of signal and employ a fully convolutional network (FCN) to perform time-domain speech enhancement, demonstrating the efficacy of combining AC and BC signals in speech enhancement.

In a preliminary study [44], we proposed to leverage AC-BC signals by performing attention based fusion and employing a convolutional recurrent network (CRN) [36] and to perform speech enhancement in the complex domain. The attention mechanism is first introduced in [41] and has produced superior performance for sequence-to-sequence modeling. Since then, it has been widely employed in tasks like automatic speech recognition [25], natural language processing [5] and computer vision [9]. The core idea of attention is to generate a context vector that "attends to" subsets of a sequence through weights that highlight salient features and suppress irrelevant information. This also allows the network to model the long-term dependencies. Recent speech enhancement studies [8], [24] also report significant performance gain by incorporating attention modules. Experiments show that the proposed attention based AC-BC fusion offers an advantage over conventional speech enhancement. In this study, we extend the preliminary work in two main aspects. First, we improve the design of attention-based fusion by concatenating the original feature maps and attention-mapped features. Second, considering the limited availability of parallel AC and BC speech data, we propose a novel semi-supervised framework that trains with both parallel and unparallel AC and BC speech. Our semi-supervised method outperforms its full-supervised counterpart.

The rest of the paper is organized as follows. In Section II, we formulate AC-BC fused speech enhancement. Section III describes our proposed network and pipeline. We describe the semi-supervised AC-BC enhancement framework in Section IV. Section V presents datasets and experimental results. Finally, Section V-B concludes the paper.

II. Problem Formulation

We propose to utilize both AC and BC sensors to perform speech enhancement. It is assumed that we simultaneously collect a noise-insensitive signal y_{BC} from the BC sensor and a noisy speech signal y from the AC sensor, which is composed of background noise n and clean speech s,

$$y[k] = s[k] + n[k], \tag{1}$$

where k denotes the sample index of a waveform signal. Applying short-time Fourier transform (STFT) to the signals we have,

$$Y[t, f] = S[t, f] + N[t, f], \tag{2}$$

where Y, S and N are the corresponding STFTs of y, s and n. Symbols t, f index time frame and frequency bin, respectively. The STFTs can be written in terms of real and imaginary parts,

$$\begin{aligned} Y_r[t, f] + iY_i[t, f] = (S_r[t, f] + N_r[t, f]) \\ + i(S_i[t, f] + N_i[t, f]). \end{aligned} \tag{3}$$

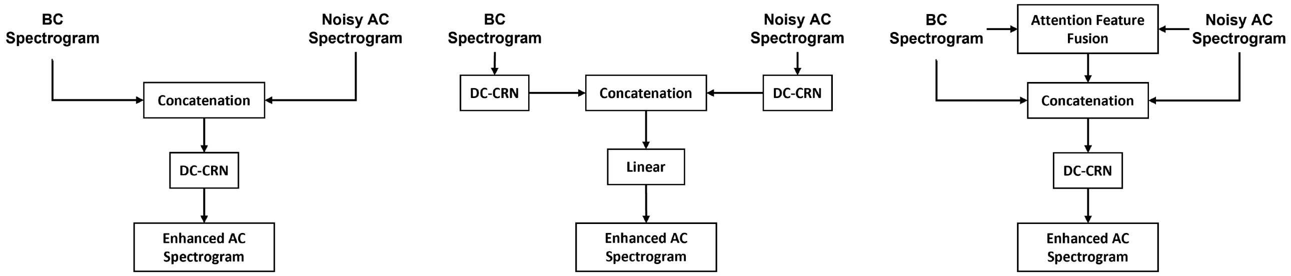
The subscripts r and i denote real and imaginary numbers, respectively, and i the imaginary unit. Using the proposed complex-domain enhancement model g, whose parameters are denoted as θ, our goal is to recover the clean speech S using the signals collected from both Y and Y_{BC}. The task is defined as,

$$\hat{S}[t, f] = g\left(\theta, Y[t, f], Y_{BC}[t, f]\right), \tag{4}$$

where $\hat{S}[t, f]$ is the enhanced speech in the complex domain.

III. Attention-Based Sensor Fusion for Complex Speech Enhancement

We propose an attention-based method to fuse AC and BC signals and perform complex spectral mapping for speech enhancement. The proposed strategy is illustrated in Fig. 1(c). Two other fusion strategies, namely early-fusion and late-fusion as depicted in Figs. 1(a) and 1(b), are also investigated for comparison. In the following subsections, we describe the components



(a) Early-Fusion (b) Late-Fusion (c) Attention-based Fusion

Fig. 1. Diagrams showing different fusion strategies, where both BC and noisy AC spectra are utilized to produce an enhanced AC complex spectrogram.

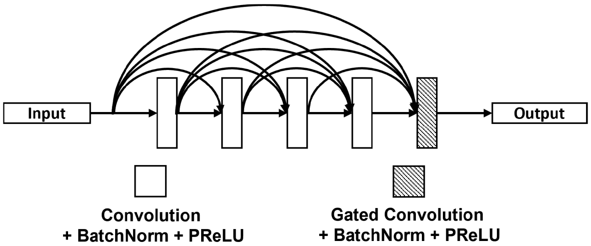


Fig. 2. Diagram of a DC block. The first four layers are standard 2D convolutions, and the last one utilizes gated convolutions.

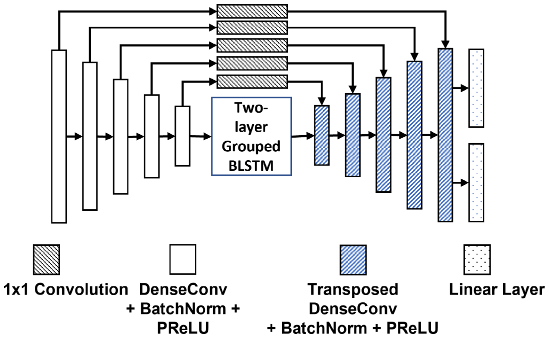

Two-layer Grouped BLSTM
1x1 Convolution
DenseConv + BatchNorm + PReLU
Transposed DenseConv + BatchNorm + PReLU
Linear Layer

Fig. 3. Diagram of the DC-CRN that performs complex spectral mapping for speech enhancement.

of the proposed system and present fusion strategies and the training objective.

A. Densely Connected Block

Motivated by the success of the densely connected (DC) network [13], [37], [50], we incorporate densely connected blocks into our network to replace standard convolution layers, as illustrated in Fig. 2. These studies suggest a DC network outperforms the same architecture without dense connections. In a DC block, one convolutional operation is split into multiple convolution layers, each with fewer channels, and all layers have direct connections to subsequent layers. This design encourages the reusage of feature maps while also addressing the gradient vanishing issue. We use DC blocks to replace standard convolutions in our network. Specifically, a DC block consists of five convolutional layers, and the first four are 2-D convolutions with the number of output channels set to 8. Each convolution is followed by a batch normalization and a parametric rectified linear unit (PReLU) activation [10]. The final layer accepts outputs from all preceding layers and performs a gated convolution [36]. The gated convolution is employed to facilitate the feature fusion across convolution channels. The kernel size for each convolution layer is $(1, 4)$ along the time and frequency axis, respectively. The dense block with gated convolutions can be formulated as,

$$x_{cat} = Concat(x_1, x_2, x_3, x_4) \tag{5}$$

$$x = conv1(x_{cat}) \odot (\sigma\, conv2(x_{cat})), \tag{6}$$

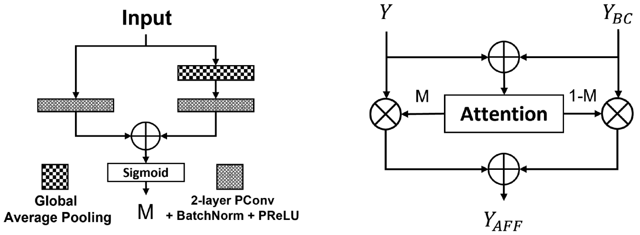
where x_l denotes the output at convolution layer $l\,(l = 1, 2, 3, 4)$, and x is the dense block output. Symbol $\odot$ represents element-wise multiplication, and σ denotes the sigmoidal activation function. $Concate()$ is the concatenation operation of the feature vectors, and we use two distinct convolutions $conv1$ and $conv2$ to perform gated convolutions on the concatenated feature x_{cat}.

B. Dc-Crn

We use the densely connected CRN (DC-CRN) as the primary component to perform complex spectral mapping based speech enhancement, and illustrate its details in Fig. 3. The network architecture is based on CRN [36], [37], which builds on the convolutional encoder-decoder structure and a recurrent neural network (RNN) bottleneck to model temporal dependencies. Such an architecture effectively captures the local and global contexts of a given input. We concatenate the real and imaginary parts of the complex spectrogram and feed the DC-CRN with 3-D feature maps. The CRN encoder is a convolutional neural network (CNN) downsampler that uses standard convolutions to reduce the feature dimension along the frequency axis, and the decoder mirrors the encoder architecture to restore the feature dimension with transposed convolutions. In DC-CRN, each convolutional layer within the CRN encoder and decoder is replaced by a DC block as described in Section III-A. The encoder comprises 7 DC blocks, and the number of output convolutional channels is set to be 16, 32, 64, 128, 256, respectively. These blocks and channels are mirrored for the decoder. The major difference with [37] is that we employ pointwise convolutions as skip connections to connect the encoder to the decoder in order to make our DC-CRN model lightweight and memory efficient. Table I lists the efficiency gain by adopting these modifications. For memory consumption, we measure the GPU memory usage by passing a batch of 8 utterances.For the bottleneck RNN, we employ a two-layer grouped bidirectional long short-term memory (BLSTM)

TABLE I
EFFICIENCY GAIN OF THE MODIFIED DC-CRN. M DENOTES MILLIONS AND G REPRESENTS GIGABYTES

	# of parameters	GPU memory used
Original DC-CRN	6.43 M	4.97 G
Modified DC-CRN	5.84 M	4.53 G



(a) Attention Module

(b) Attention-based Feature Fusion

Fig. 4. Illustration of attention-based feature fusion. (a) process of calculating the attention score M, and (b) process of using M to perform soft selection and feature concatenation. Symbol $\bigotimes$ represents element-wise multiplication, and $\bigoplus$ summation.

module [6], [36], which reduces the computational complexity while maintaining enhancement performance. Specifically, to reduce inter-layer calculations, we divide the feature maps into four disjoint groups. To model the intra-group relationship, we perform a representation rearrangement and a layer normalization after each LSTM layer. Finally, the output of the CNN decoder is halved and then reshaped into one-dimensional features. Each half passes through a linear layer to produce real and imaginary spectrogram estimates (see Fig. 3). One thing worth noting is that we can easily convert our model to the causal version by switching BLSTM to uni-directional LSTM.

C. Attention-Based Fusion

Different from the attention based methods that focus on a single modality, we regard AC and BC complex spectrograms as different modalities, and employ attention-based modality fusion techniques similar to [3], [49] to fully exploit cross-modal and single-modal features. The attention-based fusion of AC and BC feature maps is illustrated in Fig. 4. First, we implement a channel attention module in multiple scales. To make attention calculations efficient, we only consider local and global contexts. The local context is calculated by applying a two-layer pointwise convolution followed by a batch normalization and a PReLU activation. The global context is acquired similarly, except that we employ a global average pooling before the convolution operation. We aggregate context information and then calculate the attention score M using a sigmoidal activation. Note during the attention calculation, the global context vector has a smaller shape compared with the local context vector, so we expand the vector such that they have the compatible shape before summation. Then, we perform element-wise addition on two input features and assign weights M and $1 - M$ to each feature map to produce an attention-fused feature (AFF). Finally, as shown in Fig. 1(c), we concatenate the AC and BC complex spectrograms with the attention-fused feature as the input to the DC-CRN model. That is,

$$Y_{AFF}[t,f] = MY[t,f] + (1-M)Y_{BC}[t,f] \tag{7}$$

$$Y_{feat}[t,f] = Concat(Y[t,f], Y_{BC}[t,f], Y_{AFF}[t,f]). \tag{8}$$

We investigate two other fusion strategies, early-fusion (EF) and late-fusion (LF) [18], which are depicted in Fig. 1(a) and 1(b). Early-fusion concatenates AC and BC signals before feeding them to the DC-CRN. For the late-fusion strategy, AC and BC signals are fed to separate DC-CRN models, and we merge the outputs of the two models using a linear layer.

D. Training Objective

We define the training objective in the complex domain. Recent studies [45], [46], [48] have demonstrated that including a magnitude loss in complex spectral mapping is beneficial, reflecting the relative importance of magnitude over phase. Based on this observation, we construct the loss function by calculating the mean absolute error (MAE) for the real and imaginary parts, plus the MAE of magnitudes. With the total number of time frames and frequency bins denoted as T and F respectively, the loss is defined as,

$$L_{RI-Mag}(S,\hat{S}) = L_{RI} + L_{Mag} \tag{9}$$

$$\begin{aligned} L_{RI} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}(&|\hat{S}_r[t,f] - S_r[t,f]| \\ &+ |\hat{S}_i[t,f] - S_i[t,f]|) \end{aligned} \tag{10}$$

$$L_{Mag} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}||\hat{S}[t,f]| - |S[t,f]||. \tag{11}$$

IV. SEMI-SUPERVISED LEARNING FOR AC-BC FUSION

The vast majority of existing speech corpora are recorded with AC microphones. The availability of BC speech is limited, and parallelly recorded AC and BC data is even scarcer. This brings difficulties to the application of our sensor fusion method for speech enhancement. To address this issue, we propose a semi-supervised method for AC-BC fusion. Semi-supervised learning is a kind of weakly-supervised learning where both paired and unpaired data are utilized to facilitate training [1], [39]. In this study, we regard parallel AC and BC speech as paired data, and AC speech provides the 'label' of its corresponding BC signal. For unpaired data, the 'label' of a given BC speech signal is unavailable. Our proposed framework is based on the Cycle-consistent Generative Adversarial Network (CycleGAN) [53], which is shown to be effective for tasks with unpaired data, like image-to-image translation [53], image segmentation [20], and voice conversion [7]. This framework enables us to train with unpaired speech data, and improves the enhancement performance when paired data is limited.

A. Cyclegan

CycleGAN [53] is a GAN architecture extension and it is typically applied when there is a lack of paired training data.

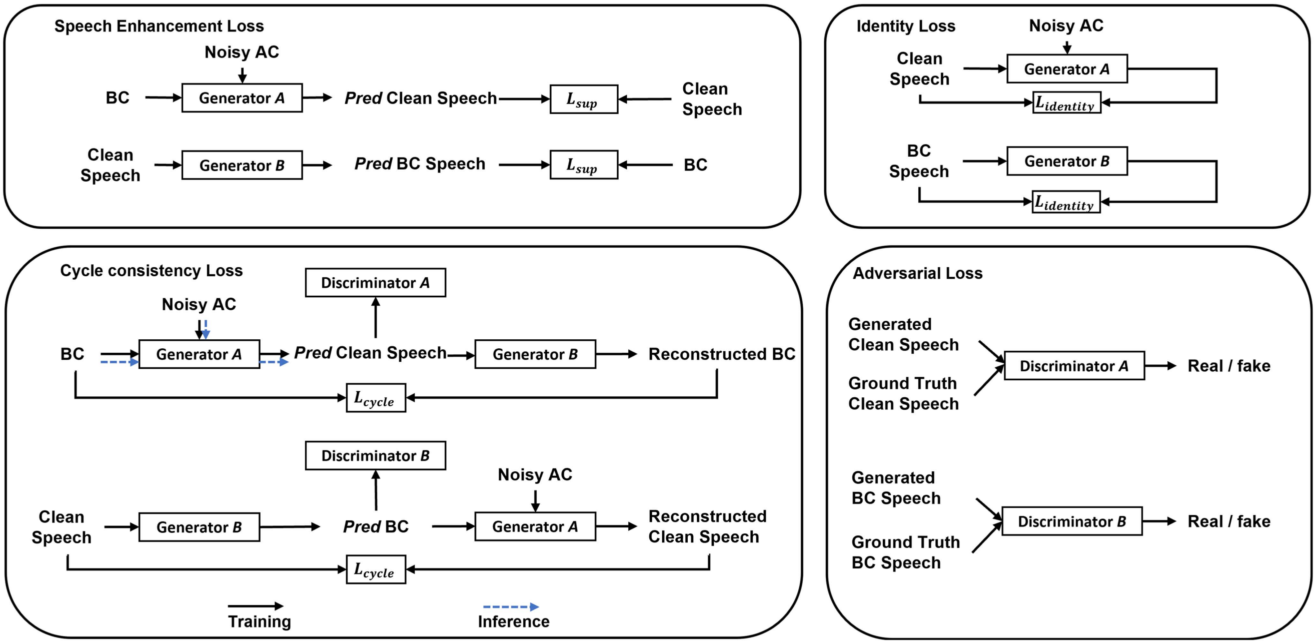


Fig. 5. Schematic of the CycleGAN-based semi-supervised framework. The proposed model contains two generators and two discriminators, which are trained in a competitive manner. The solid arrow denotes the training process, and the dashed arrow represents the pipeline of inference. *Pred* stands for predicted, and the subscript *sup* denotes supervised.

There are four modules in CycleGAN, two conditional generators and two discriminators. The generators are employed to learn a bidirectional mapping between two domains. The first generator takes input from the first domain, and produces output to the second domain. Meanwhile, the second generator learns the reverse mapping. By applying two generators sequentially, we map the input to its original domain, i.e., recover the original input. The discriminators are designed to determine whether the generated output is real or fake. Adversarial training is performed such that generators and discriminators compete with each other, and generators aim to produce outputs realistic enough to trick discriminators. This model is capable of generating plausible predictions even if there is limited paired data.

B. Model Description

Our semi-supervised AC-BC fusion speech enhancement model is illustrated in Fig. 5, and it contains two CNN-based discriminators and two generators that build on the proposed DC-CRN model. During training, we adopt the attention based fusion DC-CRN model as Generator A, which takes as input both noisy speech and BC speech and predicts clean speech. Generator B is the DC-CRN that converts clean speech to its BC counterpart. Discriminator A determines whether a given input is an authentic clean signal, and Discriminator B is trained to discriminate whether a given signal belongs to BC speech or not. Unlike image data, speech signals are of variable lengths, so we construct a 7-layer CNN with adaptive pooling as our discriminator, which converts variable-sized features into vectors of fixed dimension. Each CNN layer in the discriminator is followed by a batch normalization and a PReLU activation. The number of convolution channels in each layer is set to 32, 64, 128, 256, 512, 256, 1, sequentially. During interference, we feed Generator A with BC speech and noisy speech to produce a clean speech estimate.

C. Training Objective

The training objective for the semi-supervised framework is composed of two parts, supervised loss and semi-supervised loss. Both paired and unpaired data are involved in the loss calculation. We denote the paired data with the superscript P and the unpaired data with the superscript U. For instance, the clean speech that has no parallel BC counterpart is denoted as S^U, and the corresponding noisy speech as Y^U.

For supervised speech enhancement loss L_{sup}, we employ the complex-domain loss function defined in Section III-D to measure the complex spectrogram difference of the generated speech and its corresponding ground truth. It consists of L^A_{sup} and L^B_{sup}, which optimizes Generator A and B respectively. The supervised enhancement loss is defined as,

$$L_{sup} = L^A_{sup} + L^B_{sup} \tag{12}$$

$$L^A_{sup} = L_{RI-Mag}(G_A(Y^P, Y^P_{BC}), S^P) \tag{13}$$

$$L^B_{sup} = L_{RI-Mag}(G_B(S^P), Y^P_{BC}). \tag{14}$$

The semi-supervised loss consists of three components, an adversarial loss, a cycle consistency loss and an identity loss. Instead of the cross-entropy loss in regular GANs, we employ the least square loss [19] as the adversarial loss to stabilize adversarial training. It has been shown that this loss minimizes the Pearson χ^2 divergence. We define the adversarial loss as,

$$L_D = L_D^A + L_D^B \tag{15}$$

$$L_D^A = \frac{1}{2}\,\mathbb{E}_{S\sim p_S}\left[(D_A(S)-1)^2\right] + \frac{1}{2}\,\mathbb{E}_{Y,Y_{BC}\sim p_{Y,Y_{BC}}}\left[(D_A(G_A(Y,Y_{BC}))^2\right] \tag{16}$$

$$L_D^B = \frac{1}{2}\,\mathbb{E}_{Y_{BC}\sim p_{Y_{BC}}}\left[(D_B(Y_{BC})-1)^2\right] + \frac{1}{2}\,\mathbb{E}_{S\sim p_S}\left[D_B(G_B(S))^2\right] \tag{17}$$

$$L_G = \frac{1}{2}\,\mathbb{E}_{Y,Y_{BC}\sim p_{Y,Y_{BC}}}\left[(D_A(G_A(Y,Y_{BC}))-1)^2\right] + \frac{1}{2}\,\mathbb{E}_{S\sim p_S}\left[(D_B(G_B(S))-1)^2\right], \tag{18}$$

where $X \sim p_X$ represents a random variable X drawn from the probability distribution p_S, and $X, Y \sim p_{X,Y}$ random variables X and Y from the joint probability distribution $p_{X,Y}$. $\mathbb{E}$ is the expectation operator. Superscripts A and B indicate discriminator A and B, respectively. The discriminators seek to classify real speech as 1 and generated speech as 0, whereas the generators intend to deceive the discriminators and identify the label of generated speech to be 1. Note that superscripts U and P are absent in the above equation, as this loss term applies to both paired and unpaired data.

To exploit unparallel speech data, we use a cycle consistency loss. Applying two generators sequentially, we obtain a reconstructed complex spectrogram that corresponds to the original input. Again, we measure the complex spectrogram difference using L_{RI-Mag},

$$L_{cycle} = L_{RI-Mag}\left(G_B(G_A(Y^P, Y_{BC}^P)), Y_{BC}^P\right) + L_{RI-Mag}\left(G_A(Y^U, G_B(S^U)), S^U\right). \tag{19}$$

An identity loss is added to regularize adversarial training for which, if given a target speech signal, the generator should output the same speech [53], i.e.,

$$L_{identity} = \mathbb{E}_{S,Y\sim p_{S,Y}}[G_A(S,Y)-S)] + \mathbb{E}_{Y_{BC}\sim p_{Y_{BC}}}[G_B(Y_{BC})-Y_{BC})]. \tag{20}$$

The purpose of this loss term is to preserve the feature correlations between the input and output [53]. Without the identity loss, the generators produce complex spectrograms reasonable enough to deceive the discriminators, but might deviate from the ground truth, as both mappings are equally valid under the adversarial loss and the cycle consistency loss.

Finally, the total loss of our training objective combines all loss terms,

$$L_{total} = L_D + L_G + \alpha L_{cycle} + \beta L_{identity} + \gamma L_{sup}, \tag{21}$$

where α, β, γ control the relative importance of their respective loss terms, and we set $\alpha = 5.0$, $\beta = 2.0$, $\gamma = 5.0$ based on the performance on a validation set.

V. Experiments

A. Datasets and Evaluation Metrics

We perform supervised experiments on the Elevoc Simultaneously-recorded Microphone/Bone-sensor (ESMB) speech corpus,[1] which is a Chinese corpus consisting of 128 hours of speech uttered by 131 male and 156 female speakers. Speech is recorded using a pair of Elevoc Clear earbuds, and each earbud contains a ST25ba BC sensor near the entry of the ear canal to gather skull vibrations during articulation and an AC sensor outside the ear that acts as a close-talk microphone. During the recording, every speaker reads Chinese prompts for around 20 minutes, producing 16 kHz stereo speech data, for which each channel corresponds to one earbud. We use the same noise set for training and validation, which is generated by randomly selecting 20000 files from the DNS challenge dataset.[2] For each utterance, we generate a noisy speech signal by mixing an AC signal with a noise segment cut to the same length from the noise set at an SNR level uniformly sampled from the range $\{-5, -4, -3, -2, -1, 0\}$ dB. We set aside two male and two female speakers for validation and evaluate on two male and two female speakers that are not included in training and validation sets. The remainder of the corpus constitutes the training set. For evaluation, we select four challenging noises: babble and cafeteria from an Auditec CD,[3] and factory and engine from the NOISEX92 dataset [40]. Each test utterance is mixed with these four noises at three SNR levels -5, 0 and 5 dB.

For semi-supervised experiments, paired AC and BC speech are extracted from the ESMB corpus, and we employ the AISHELL-1 dataset [2] as the source for unpaired data. AISHELL-1 is a Chinese Mandarin speech corpus that consists of around 120000 utterances with a total duration of about 178 hours. Four hundred speakers participated in the recording, which was conducted in a quiet indoor environment using a high-fidelity microphone and then downsampled to 16 kHz. The validation and test settings are the same as in supervised experiments. A similar procedure to supervised experiments is used to generate noisy mixtures for both AISHELL and ESMB.

We use two standard metrics to assess enhancement performance, short-time objective intelligibility (STOI) [35] and perceptual evaluation of speech quality (PESQ) [28]. STOI has a typical value range from 0 to 1, which can be typically interpreted as percent correct. PESQ ranges from -0.5 to 4.5. Higher values indicate better performance for both metrics.

B. Experimental Setup

For all experiments, we resample recordings to the sampling rate of 8 kHz. During training and validation, we discard for each recording silent portions whose energy is 60 dB below the peak power reference. A window length of 32 ms with 50% overlap between adjacent frames is used in calculating STFTs, which correspond to 129-dimensional spectra. We apply mean-variance normalization (MVN) to each noisy utterance, and the

[1] [Online]. Available: https://github.com/elevoctech/ESMB-corpus

[2] [Online]. Available: https://github.com/microsoft/DNS-Challenge

[3] [Online]. Available: http://www.auditec.com

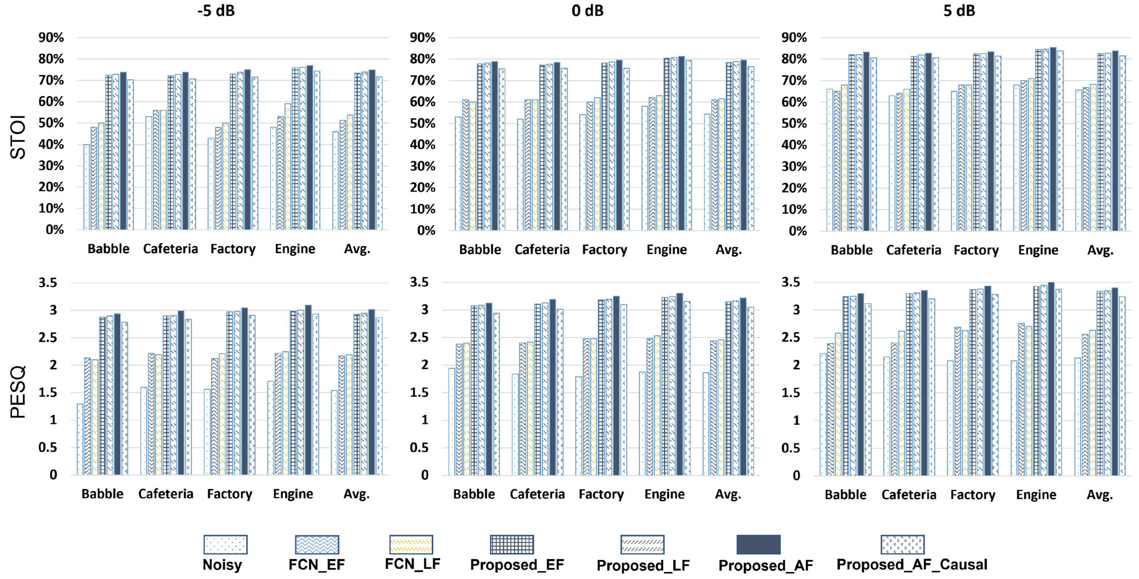


Fig. 6. Enhancement performance of the FCN baseline and the proposed method using different fusion strategies in terms of STOI and PESQ on the ESMB corpus at different SNRs.

corresponding clean utterance is scaled accordingly. Each BC utterance passes through an eighth-order Butterworth low-pass filter, and is then normalized using MVN. This low-pass filtering serves two purposes. First, there is still residual energy in the upperband of BC spectrograms, which is not helpful for speech enhancement. We find that removing the upperband energy slightly improves enhancement performance. Second, it enforces the same cutoff frequency of all BC utterances, which improves the generalization of the trained model to devices with different cutoff frequencies.

For the fully-supervised model, we use the Adam optimizer [16] and train with the batch size of 16 utterances for 30 epochs. The initial learning rate is set to 0.0006, and is halved if the validation loss has not improved for three consecutive epochs. We also employ a gradient clipping with a maximum value of 5.0 to avoid gradient explosion.

For the semi-supervised model, both generators and discriminators are optimized using the Adam optimizer. The learning rate for the generators is set to 4e-4, and for the discriminators to 2e-4. We train the CycleGAN in an alternating fashion, i.e., when the generators are optimized, the parameters of the discriminators are fixed, and vice versa. To balance the adversarial training, we optimize the discriminators less frequently, and update their parameters every 5 iterations. Furthermore, we set the batch size to 8 utterances and train for 120000 iterations. For the first 10% of the iterations, we only train with paired data using L_{sup} to initialize, and the learning rate is fixed to 0.0004. For the rest of the training, we use L_{total} and the learning rate is linearly decayed from 0.0004 to 0.0001.

VI. Results and Analyses

A. Supervised Experiments

Fig. 6 plots the enhancement performance of AC-BC sensor fusion approaches on the ESMB dataset. We present the results of our proposed method and the baseline FCN [47], and compare different fusion strategies. Subscripts AF, EF and LF denote the proposed attention-based fusion, early-fusion and late-fusion strategies, respectively. We also provide a causal version of the proposed DC-CRN for a fairer comparison with FCN. For the causal implementation, we use unidirectional LSTM instead of BLSTM, and only keep the local context computation in the attention module to avoid global average pooling. As shown in the figure, our complex-domain DC-CRN outperforms the time-domain baseline FCN [47] in all conditions. Especially at -5 dB SNR, our attention-based fusion achieves 21.1% higher STOI, and PESQ is improved by 0.83 compared with the best FCN fusion. In terms of fusion strategies, the proposed attention-based fusion shows a consistent improvement over early fusion and late fusion. For instance, at the SNR of -5 dB, on average the attention-based fusion has 1.0% STOI and 0.08 PESQ advantage over the late fusion. Furthermore, for both FCN and DC-CRN, late-fusion performs slightly better than early fusion (see also [47]). However, requiring separate DNNs for two types of sensor signal, late-fusion tends to be computationally heavier and may not be preferable in real applications.

Additionally, we compare sensor fusion with single-sensor counterparts in Fig. 7. Specifically, we feed DNNs with only AC or BC signals, and compare them with the AC-BC fusion. From the figure, we observe that the networks that employ AC-BC fusion always outperform conventional speech enhancement that only utilizes AC signals. Especially at -5 dB SNR, sensor fusion substantially boosts the enhancement performance. For example, STOI is improved by 11.6% and PESQ by 0.65 for the proposed DC-CRN. Incorporating BC signals becomes less beneficial as SNR rises. This is to be expected, as noise interference is not that severe in these conditions, and the noise insensitivity of BC signals is less useful. At 5 dB SNR, STOI is merely 1.7% higher, and PESQ is improved by 0.10 for DC-CRN.

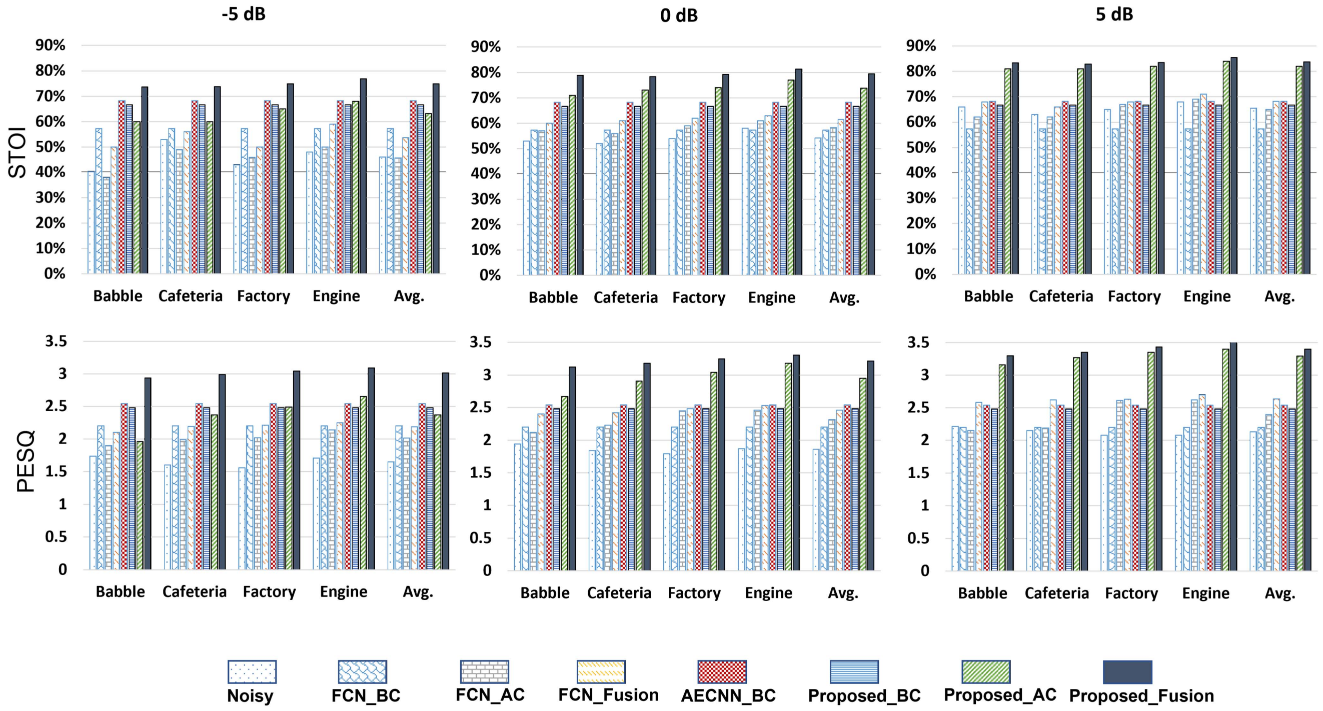


Fig. 7. Enhancement performance of single-sensor versus sensor-fusion methods.

TABLE II
ENHANCEMENT PERFORMANCE OF FULLY-SUPERVISED AND SEMI-SUPERVISED LEARNING MODELS USING DIFFERENT PROPORTIONS OF PAIRED DATA AT -5 DB SNR

	Fully-supervised		Semi-supervised	
paired portion	STOI (%)	PESQ	STOI (%)	PESQ
1%	57.6	2.27	66.2	2.65
2%	61.2	2.40	69.1	2.74
5%	64.2	2.55	70.7	2.79
10%	67.6	2.71	72.6	2.86
20%	70.0	2.83	73.9	2.99
50%	73.0	2.96	74.8	3.02
100%	74.8	3.01	74.9	3.03

TABLE III
ABLATION STUDY OF THE PROPOSED NETWORK AT -5 DB SNR

	STOI (%)	PESQ
Proposed_AF	74.8	3.01
– DC blocks (i)	69.5	2.72
– gated convolution (ii)	72.6	2.82
– pointwise convolution skip connections (iii)	74.1	2.94
attention-based fusion with addition (iv)	68.4	2.73

We also provide the results of employing BC signals only, which essentially amounts to bandwidth extension. An advanced bandwidth extension baseline (AECNN_BC) [43] for comparison. Due to the nature of BC signals, it performs the same in all noisy conditions. Compared to sensor fusion, the enhancement performance is worse, but the gap is relatively small in lower SNR conditions. It is worth noting that, at -5 dB SNR, speech enhancement with only BC signals yields on average better results than with only AC signals.

B. Semi-Supervised Experiments

Table II reports the results of training with different portions of paired data of the ESMB corpus for supervised and semi-supervised learning, where we present average evaluation results of four test noises at -5 dB SNR. We train both the fully-supervised model and the CycleGAN model using 1%, 2%, 5%, 10%, 20%, 50% and 100% paired data, and the semi-supervised model additionally exploits unpaired AC data from the AISHELL corpus.

Compared to fully-supervised baselines, semi-supervised learning has a clear advantage on different paired portions, suggesting we have effectively benefited from unpaired data. Especially when training with only 1% of paired data, the semi-supervised approach considerably boosts the enhancement performance, improving STOI by 8.6% and PESQ by 0.38. As the paired portions rise, the improvement becomes smaller as expected. Using 50% paired data, we are able to match the performance of the full-supervised baseline using the complete ESMB corpus. This shows that the proposed semi-supervised technique can improve the enhancement performance when paired data is limited.

C. Ablation Study

An ablation study is conducted to investigate the effects of different components within the proposed model, and the results are given in Table III. We use the attention-based fusion of our DC-CRN as the baseline and compare several variants at -5 dB

SNR: (i) replacing DC blocks with standard convolutions; (ii) replacing the gated convolutions within DC blocks with standard convolutions; (iii) replacing pointwise convolution-based skip connections with concatenation-based skip connections. (iv) employing addition instead of concatenation when performing attention-based fusion. As shown in the table, these variants all underperform the proposed design. Among these factors, dense connectivity plays a significant role in enhancement performance, as removing DC blocks degrades STOI by 5.3% and PESQ by 0.29. Gated convolutions are beneficial for merging cross-channel features, and removing them from DC blocks results in 2.2% and 0.19 drop in STOI and PESQ, respectively. Furthermore, pointwise skip connections are an efficient way to boost feature fusion compared to simple concatenations, as it improves the performance without introducing many extra parameters. Lastly, performing attention-based fusion using addition leads to a significant performance drop. This is expected as using concatenation can leverage both cross-modal and single-modal features, whereas addition only utilizes cross-modal features.

VII. Conclusion

In this study, we have proposed a novel attention-based approach for fusing AC and BC sensor signals for complex-domain speech enhancement. To restore clean speech in adverse environments, we take advantage of the full bandwidth of AC microphones and the noise insensitivity of BC sensors. Systematic evaluations show that our approach substantially boosts the enhancement performance compared with conventional monaural speech enhancement that only utilizes AC microphones, especially in very low SNR conditions. Furthermore, our DC-CRN model significantly outperforms a recent time-domain baseline in all conditions. Additionally, as the availability of parallelly recorded AC and BC speech is limited, we have proposed a semi-supervised CycleGAN-based framework to utilize AC and BC speech data in unrelated recordings. We have demonstrated that this framework achieves similar performance with only 50% paired data compared to the fully supervised counterpart. For future work, we plan to reduce the DC-CRN model complexity and improve inference efficiency so that the proposed algorithm can be deployed on mobile devices.

Acknowledgment

This work began when the first author was an intern at Elevoc Co. Ltd. The authors would like to thank Yongjie Yan for his assistance in organizing the ESMB corpus.

References

[1] C. Baur, S. Albarqouni, and N. Navab, "Semi-supervised deep learning for fully convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, 2017, pp. 311–319.

[2] 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 *Proc. 20th Conf. Oriental Chapter Int. Coordinating Committee Speech Databases Speech I/O Syst. Assessment*, 2017, pp. 1–5.

[3] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3560–3569.

[4] E. Erzin, "Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1316–1324, Sep. 2009.

[5] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4291–4308, Oct. 2021.

[6] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T.-Y. Liu, "Efficient sequence learning with group recurrent networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol., Volume 1 (Long Papers)*, 2018, pp. 799–808.

[7] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2506–2510.

[8] R. Giri, U. Isik, and A. Krishnaswamy, "Attention Wave-U-uet for speech enhancement," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 249–253.

[9] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, pp. 1–38, 2022.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[11] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2482–2486.

[12] B. Huang, Y. Gong, J. Sun, and Y. Shen, "A wearable bone-conducted speech enhancement system for strong background noises," in *Proc. 18th Int. Conf. Electron. Packag. Technol.*, 2017, pp. 1682–1684.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[14] T. Hussain, Y. Tsao, S. M. Siniscalchi, J.-C. Wang, H.-M. Wang, and W.-H. Liao, "Bone-conducted speech enhancement using hierarchical extreme learning machine," in *Proc. Increasing Naturalness Flexibility Spoken Dialogue Interact.*, 2021, pp. 153–162.

[15] T. Ito, C. Röösli, C. J. Kim, J. H. Sim, A. M. Huber, and R. Probst, "Bone conduction thresholds and skull vibration measured on the teeth during stimulation at different sites on the human head," *Audiol. Neurotol.*, vol. 16, no. 1, pp. 12–22, 2011.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[17] K. Kondo, T. Fujita, and K. Nakagawa, "On equalization of bone conducted speech for improved speech quality," in *Proc. IEEE Int. Symp. Signal Process. Informat. Technol.*, 2006, pp. 426–431.

[18] J. Liu, T. Li, P. Xie, S. Du, F. Teng, and X. Yang, "Urban Big Data fusion based on deep learning: An overview," *Inf. Fusion*, vol. 53, pp. 123–133, 2020.

[19] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2794–2802.

[20] A. K. Mondal, A. Agarwal, J. Dolz, and C. Desrosiers, "Revisiting CycleGAN for semi-supervised segmentation," 2019, *arXiv:1908.11569*.

[21] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2003, pp. V– 708.

[22] H. Q. Nguyen and M. Unoki, "Improvement in bone-conducted speech restoration using linear prediction and long short-term memory model," *J. Signal Process.*, vol. 24, pp. 175–178, 2020.

[23] A. V. D. Oord et al., "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[24] A. Pandey and D. L. Wang, "Dense CNN with self-attention for time-domain speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1270–1279, 2021.

[25] D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khudanpur, "A time-restricted self-attention layer for ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5874–5878.

[26] N. Prasad and T. K. Kumar, "Bandwidth extension of speech signals: A comprehensive review," *Int. J. Intell. Syst. Appl.*, vol. 8, no. 2, pp. 45–52, 2016.

[27] M. S. Rahman and T. Shimamura, "Pitch characteristics of bone conducted speech," in *Proc. IEEE 18th Eur. Signal Process. Conf.*, 2010, pp. 795–799.

[28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.

[29] J. Sautter, F. Faubel, M. Buck, and G. Schmidt, "Discriminative training of deep regression networks for artificial bandwidth extension," in *Proc. IEEE 16th Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 540–544.

[30] D. Shan, X. Zhang, C. Zhang, and L. Li, "A novel encoder-decoder model via NS-LSTM used for bone-conducted speech enhancement," *IEEE Access*, vol. 6, pp. 62638–62644, 2018.

[31] T. Shimamura and T. Tamiya, "A reconstruction filter for bone-conducted speech," in *Proc. IEEE 48th Midwest Symp. Circuits Syst.*, 2005, pp. 1847–1850.

[32] H. S. Shin, T. Fingscheidt, and H.-G. Kang, "A priori SNR estimation using air- and bone-conduction microphones," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 11, pp. 2015–2025, Nov. 2015.

[33] H. S. Shin, H.-G. Kang, and T. Fingscheidt, "Survey of speech enhancement supported by a bone conduction microphone," in *Proc. IEEE ITG Conf. Speech Commun.*, 2012, pp. 1–4.

[34] P. Singh, M. K. Mukul, and R. Prasad, "Bone conducted speech signal enhancement using LPC and MFCC," in *Proc. Int. Conf. Intell. Hum. Comput. Interact.*, 2018, pp. 148–158.

[35] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, Sep. 2011.

[36] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 380–390, 2020.

[37] K. Tan, X. Zhang, and D. L. Wang, "Deep learning based real-time speech enhancement for dual-microphone mobile phones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1853–1863, 2021.

[38] T. V. Thang, K. Kimura, M. Unoki, and M. Akagi, "A study on restoration of bone-conducted speech with MTF-based and LP-based models," *J. Signal Process.*, vol. 10, pp. 407–417, 2006.

[39] J. E. Van E. and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[40] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.

[41] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

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

[43] H. Wang and D. L. Wang, "Towards robust speech super-resolution," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2058–2066, 2021.

[44] H. Wang and D. L. Wang, "Attention-based fusion for bone-conducted and air-conducted speech enhancement in the complex domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7757–7761.

[45] Z.-Q. Wang, P. Wang, and D. L. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[46] Z.-Q. Wang, G. Wichern, and J. L. Roux, "On the compensation between magnitude and phase in speech separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 2018–2022, 2021.

[47] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, "Time-domain multi-modal bone/air conducted speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 1035–1039, 2020.

[48] J. Zhang, M. D. Plumbley, and W. Wang, "Weighted magnitude-phase loss for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5794–5798.

[49] P. Zhang et al., "A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data," *Remote Sens.*, vol. 12, no. 22, 2020, Art. no. 3764.

[50] Y. Zhao and D. L. Wang, "Noisy-reverberant speech enhancement using DenseUNet with time-frequency attention," in *Proc. Interspeech*, 2020, pp. 3261–3265.

[51] C. Zheng, T. Cao, J. Yang, X. Zhang, and M. Sun, "Spectra restoration of bone-conducted speech via attention-based contextual information and spectro-temporal structure constraint," *IEICE Trans. Fundamentals Electron. Commun. Comput. Sci.*, vol. 102, pp. 2001–2007, 2019.

[52] C. Zheng, J. Yang, X. Zhang, T. Cao, M. Sun, and L. Zheng, "Bandwidth extension WaveNet for bone-conducted speech enhancement," in *Proc. 7th Conf. Sound Music Technol.*, 2020, pp. 3–14.

[53] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[54] M. Zhu, H. Ji, F. Luo, and W. Chen, "A robust speech enhancement scheme on the basis of bone-conductive microphones," in *Proc. 3rd Int. Workshop Signal Des. Its Appl. Commun.*, 2007, pp. 353–355.

Heming Wang (Graduate Student Member, IEEE) received the bachelor's degree in physics in 2016, and the M.S. degree in applied mathematics in 2018 from the University of Waterloo, Waterloo, ON, Canada. He is currently working toward the Ph.D. degree with the Ohio State University, Columbus, OH, USA. His research interests include speech enhancement, speech super-resolution, and deep learning.

Xueliang Zhang (Member, IEEE) photograph and biography not available at the time of publication.

DeLiang Wang (Fellow, IEEE) photograph and biography not available at the time of publication.