A Two-Stage Approach to Quality Restoration of Bone-Conducted Speech

Changtao Li, Feiran Yang, *Member, IEEE*, and Jun Yang, *Senior Member, IEEE*

***Abstract*—Bone-conducted speech is not susceptible to background noise but suffers from poor speech quality and intelligibility due to the limited bandwidth. This paper proposes a two-stage approach to restore the quality of bone-conducted speech, namely, bandwidth extension and speech vocoder. In the first stage, a deep neural network is trained to learn mappings from a low-resolution representation of the bone-conducted speech, i.e., log Mel-scale spectrogram, to that of the air-conducted speech, which extends the bandwidth of the bone-conducted speech. In the second stage, a speech vocoder is employed to transform the extended log Mel-scale spectrogram of the bone-conducted speech back to time-domain waveforms. Due to the many-to-many correspondence between the air-conducted and bone-conducted speech, supervised learning may not be the best training protocol for the bone-conducted/air-conducted feature mapping. We thus propose to leverage adversarial training to further improve the bandwidth extension performance in the first stage. The two stages are decoupled and can be trained independently. The vocoder is trained on a large multi-speaker dataset and can generalize well to unknown speakers. Also, the vocoder can help to remedy the spectral artifacts introduced in the bandwidth extension stage. Objective and subjective evaluations on ESMB dataset show that the proposed two-stage system substantially outperforms existing bone-conducted speech enhancement systems.**

***Index Terms*—Bone conduction, vocoder, adversarial training, speech enhancement, bandwidth extension.**

I. Introduction

AIR-CONDUCTION sensors capture the desired full-band speech signals as well as the background noise simultaneously. In contrast, bone-conduction sensors convert articulation-induced vibrations on human tissue into electrical signals, which is not subject to the background noise [1]. However, bone-conducted speech has two major limitations [1]. First, human body tissue where the bone-conducted speech is transmitted acts like a low-pass filter, which attenuates the speech signals significantly. Accordingly, high-frequency components of the bone-conducted speech recorded by a vibration sensor are missing. Second, vibration sensors cannot record unvoiced sounds that are only audible due to turbulent airflows and cause very small vibrations in the body of the speaker. Hence, bone-conducted speech suffers from poor sound quality and intelligibility, which limits its application in speech communications.

Bone-conducted speech has been utilized in two different ways. In [1], [2], [3], [4], [5], [6], [7], [8], bone-conducted speech is fused with noisy air-conducted speech and plays an auxiliary role in the task at hand. In [1], phone-dependent filtering techniques are adopted to equalize the bone-conducted speech, which is then fused with the air-conducted speech to estimate the spectra amplitude and phase of the clean speech based on the minimum mean square error criterion. In [2], the equalized bone-conducted speech is used to estimate the *a priori* signal-to-noise ratio (SNR) required in the Wiener filter, and the desired speech signal is estimated as a weighted sum of the equalized bone-conducted speech and the enhanced air-conducted speech. In [3], the bandwidth of bone-conducted speech is extended by incorporating mutual information of noisy air-conducted speech captured by the outer ear microphone. In [4], an involution network is used to predict the mask of clean speech component using both noisy air-conducted speech and synchronized bone-conducted speech. In [5], two ensemble-learning-based fusion strategies, i.e., early fusion and late fusion, are presented to integrate the bone-conducted and air-conducted signals for time-domain speech enhancement based on a fully convolutional network (FCN). In [6], an attention-based approach is proposed to fuse the bone-conducted and air-conducted signals, and the resulting attention-fused feature is concatenated with the original bone-conducted and air-conducted complex spectrograms as input to a convolutional recurrent network (CRN) for speech enhancement. Also, the bone-conducted and air-conducted signals are combined as multi-modal data to improve the performance of automatic speech recognition (ASR) system in adverse acoustic environments [7], [8].

In very low SNR conditions, air-conducted speech may become completely unintelligible, and hence many efforts have been made to restore bone-conducted speech in the absence of air-conducted speech. This task is challenging because of the limited frequency range and the nonlinear distortion of the bone-conducted speech signal. Both bone-conducted and air-conducted speech signals can be represented by the source-filter model. It is assumed that bone-conducted and air-conducted

Manuscript received 20 April 2023; revised 19 September 2023 and 13 November 2023; accepted 25 November 2023. Date of publication 30 November 2023; date of current version 14 December 2023. This work was supported in part by National Natural Science Foundation of China under Grant 62171438, in part by Beijing Natural Science Foundation under Grand L223032 and in part by IACAS Frontier Exploration Project under Grant QYTS202111. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhizheng Wu. *(Corresponding author: Feiran Yang.)*

Changtao Li and Jun Yang are with the Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: lichangtao@mail.ioa.ac.cn; jyang@mail.ioa.ac.cn).

Feiran Yang is with the State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: feiran@mail.ioa.ac.cn).

Digital Object Identifier 10.1109/TASLP.2023.3337988

speech originate from the same excitation source, and their difference only lies in the filter, i.e., the transfer function of the speech path. In [9], [10], a fixed equalization filter is designed based on the long-term or short-term spectra of the bone-conducted and air-conducted speech, which is then applied to the bone-conducted speech. However, the equalization filter is not fixed but depends on the specific speaker, the measurement position and the pronounced syllables, and hence speech artifacts such as musical noise and echoes are common in the equalization method [11]. To address this issue, a more accurate frame-based equalization filter is designed using the linear prediction (LP) analysis of speech signal [12], [13], [14], [15]. LP residues are related to the source information that are the same for both bone-conducted and air-conducted speech, and LP coefficients are related to the filter, i.e., the spectral characteristics. The equalization filter can be derived from the LP coefficients of the bone-conducted and air-conducted speech. Several methods have been proposed to predict the LP coefficients of air-conducted speech from that of bone-conducted speech signal [12], [13], [14]. In [12], an equivalent representation of LP coefficients, i.e., the line spectral frequencies (LSF), are adopted as suitable prediction parameters, and a recurrent neural network (RNN) is trained for mapping LSF of bone-conducted speech to that of air-conducted speech. To deal with the potential over-training problem of simple RNNs, a Gaussian mixture model (GMM) is thus employed to enhance LSF prediction of air-conducted speech [13]. Also, a long short-term memory (LSTM) network is used to predict the LSF of air-conducted speech given the LSF of bone-conducted speech [14]. In [15], both the LSF and excitation signals of the bone-conducted speech are enhanced using the phone-dependent GMMs, and it is shown that phone-dependent mappings exhibit significant performance improvements over phone-independent mappings. However, the assumption on the excitation signals may not hold in the presence of bone-conduction channel noise and physical noise, and this approach often leads to distorted speech reconstruction [16].

Deep neural networks (DNNs) are also utilized for bone-conducted speech enhancement after their widespread application in acoustic signal processing [17]. In [18], [19], DNN-based models are used to predict the magnitude spectrogram of air-conducted speech from that of bone-conducted speech. The enhanced magnitude spectrogram is combined with the phase information of bone-conducted speech to restore bone-conducted speech. In [18], a denoising auto-encoder is used to extend the magnitude spectrogram of bone-conducted speech. In [19], a speaker-dependent bone-conducted speech enhancement system is presented, where an attention-based bidirectional LSTM is utilized to map the magnitude spectrogram of bone-conducted speech to that of air-conducted speech. In [20], a bandwidth extension WaveNet is employed to reconstruct full-band waveform from the narrow-band bone-conducted magnitude spectrogram. A time-domain dual-path transformer-based network is adopted for speaker-dependent bone-conducted speech enhancement in [21], where a transformer-based model maps the bone-conducted speech waveform directly to the air-conducted speech waveform. However, most DNN-based models are speaker-dependent, and the generalization capability to unknown speakers is limited.

This paper proposes a two-stage framework for restoring bone-conducted speech, which consists of bandwidth extension and vocoder. Bandwidth extension in our context refers to the restoration of high-frequency information missing in bone-conducted speech, rather than upsampling [20], [22]. Moreover, due to the nonlinear distortion of the low-frequency part of the bone-conducted speech, our bandwidth extension also includes the mapping of the low-frequency part, which is different from [22]. In contrast to discriminative models used in [3], [20], we employ a conditional generative model in compact latent feature space to extend the bandwidth of bone-conducted speech. The latent speech feature used in the first stage is log Mel-scale spectrogram that complies more with human hearing and possesses low dimensionality. We adopt an improved version of UNet [23] as the backbone of bandwidth extension. The UNet backbone can be trained based on supervised learning by utilizing paired bone- and air-conducted speech. However, we have found that the relationship between the bone- and air-conducted speech is not a bijection, but rather a many-to-many correspondence. Adversarial training has been proved to outperform supervised learning under such a circumstance, which is therefore employed for bandwidth extension. We train a discriminator to distinguish the outputs of bone-conducted/air-conducted speech feature mapping from air-conducted speech features, which makes the bandwidth extension a conditional generative model and improves the overall restoration performance. In the second stage, we transform extended time-frequency features of the bone-conducted speech to the time-domain waveforms using a DiffWave vocoder [24]. Such a vocoder can produce enhanced speech with high quality and generalizes well to unseen speakers. Also, either the first stage or second stage of the proposed framework can be used independently to restore bone-conducted speech. However, extensive experiments consistently indicate that single-stage systems often introduce noticeable speech artifacts. The combination of two stages, on the other hand, leads to a significant improvement in the quality of the restored speech.

II. Methodology

In this section, we describe the proposed two-stage approach to quality restoration of bone-conducted speech in details. An overview of the system is presented in Section II-A, followed by the description of the bandwidth extension and speech vocoder in Sections II-B and II-C, respectively.

A. Overall Structure

Inspired by current latent-variable models [25], [26], we propose a two-stage approach to restore the quality of the bone-conducted speech blindly. The block diagram of the proposed bone-conducted speech enhancement system is illustrated in Fig. 1, consisting of bandwidth extension and vocoder. In the first stage, we perform bandwidth extension in a compact latent feature space to compensate the missing high frequency component of the bone-conducted speech. Small network size and low complexity can be achieved by utilizing the low-dimensional feature

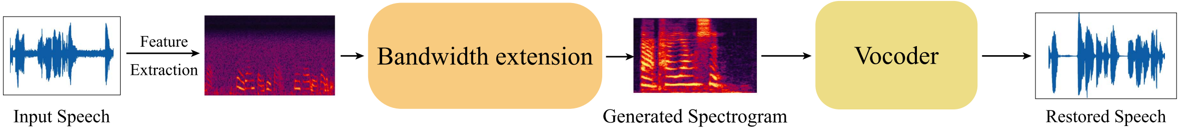


Fig. 1. Proposed two-stage approach to bone-conducted speech enhancement.

space of the input speech, and we thus choose the log Mel-scale spectrogram as the speech feature. Compared to the short time Fourier transform (STFT) commonly used in time-frequency domain models, the log Mel-scale spectrogram feature is known to be more aligned with human hearing [27], [28], [29]. Other learned embedding feature spaces, such as SoundStream [30], could also be considered, but are not within the scope of this paper. In the second stage, we transform the enhanced log Mel-scale spectrogram to the time-domain waveforms using the vocoder. Many efficient and lightweight DNN-based vocoder are developed in recent years, which should be beneficial for our system.

The two stages of the proposed algorithm, i.e., bandwidth extension and vocoder, are decoupled, and can be trained separately. Many telecommunication systems follow a similar two-stage setup where speech features are transferred and speech waveforms are synthesized by a vocoder at the receiver end. Our system can be easily incorporated into such telecommunication systems.

B. Bandwidth Extension of Bone-Conducted Speech

To extend the bandwidth, we train a DNN that implements a mapping from the feature of bone-conducted speech to that of the corresponding air-conducted speech. Both the bone-conducted and air-conducted speech are utilized to train the network, while only the former is used to perform the bandwidth and improve the speech quality.

We simultaneously collect the bone-conducted speech signal $x_{bc}(n)$ and the air-conducted speech signal $x_{ac}(n)$ to construct the training dataset for the DNN, where n denotes the discrete time index. Applying the STFT to the time-aligned signals $x_{ac}(n)$ and $x_{bc}(n)$, their time-frequency representations are given by

$$X_{ac}(k,l) = \sum_{m=0}^{N-1} x_{ac}(lR+m)w(m)e^{-j2\pi km/N} \quad (1)$$

$$X_{bc}(k,l) = \sum_{m=0}^{N-1} x_{bc}(lR+m)w(m)e^{-j2\pi km/N} \quad (2)$$

where k and l denote the frequency bin index and the frame index, respectively, N is the STFT length, R is the frame shift, and $w(\cdot)$ is an analysis window.

Compared with the air-conducted speech, the frequency components of the bone-conducted speech, e.g., more than 2 kHz, are highly attenuated. Also, the low frequency components of the bone-conducted speech are distorted. One solution to restoration of the bone-conducted speech is to directly predict the magnitude spectrogram of the air-conducted speech $|X_{ac}(k,l)|$ from that of the bone-conducted speech $|X_{bc}(k,l)|$ using a DNN $\mathcal{V}$ [19]

$$|X_{ac}(k,l)| \leftarrow \mathcal{V}(|X_{bc}(k,l)|). \quad (3)$$

In this paper, however, we choose a low-resolution representation of the speech signal, i.e., the log Mel-scale spectrogram, for bandwidth extension, which is shown to preserve faithful information. The magnitude spectrum $|X_{ac}(k,l)|$ and $|X_{bc}(k,l)|$ are transformed into Mel scale by applying a Mel filter bank $M(p,k)$ [23]

$$S_{ac}(p,l) = \ln\left(\sum_{k=0}^{N-1} M(p,k)|X_{ac}(k,l)|\right) \quad (4)$$

$$S_{bc}(p,l) = \ln\left(\sum_{k=0}^{N-1} M(p,k)|X_{bc}(k,l)|\right) \quad (5)$$

where $p = 0, 1, \ldots, P-1$, P is the number of Mel filter banks and $P \ll N$. For simplicity, we omit p and l latter.

The goal of the bandwidth extension is to approximate the log Mel-scale spectrogram of air-conducted speech S_{ac}. The task is defined as

$$S_{ac} \leftarrow \mathcal{U}_{\theta}(S_{bc}) \quad (6)$$

where $\mathcal{U}$ is the backbone network for bandwidth extension, whose parameters are denoted as θ. In this paper, we set the number of Mel filter banks to 128 ($P = 128$), resulting in log Mel-scale spectrograms $S_{ac}, S_{bc} \in \mathbb{R}^{128\times L}$ where L denotes the length of the log Mel-scale spectrogram.

UNet has been commonly employed in image segmentation and speech enhancement [31], and is utilized here as the backbone of the bandwidth extension. Fig. 2 presents the details of the proposed UNet architecture, which is similar to that in [23]. The global structure of UNet resembles an encoder-decoder arrangement, where an information bottleneck is imposed at the middle of the neural network [31]. The encoder of original UNet consists multiple layers of convolution neural networks (CNNs) along with max-pooling operations to down-sample the speech feature, and the original decoder consists of the same number of transposed CNN blocks for up-sampling. The encoder and decoder are skip-connected, which can aggregate both global and local features of the input data. We here modify the original UNet structure to better capture the global information of speech features. The improved version of UNet retains the original macro-structure but replaces the original CNN blocks with residual connected CNNs adopted from [32]. To down-sample the speech feature, we use an additional CNN layer with a kernel size of 1×1 and a stride of 2. For up-sampling the speech feature, we utilize linear interpolation at the decoder stage. We present one example of feature maps for different blocks when using log Mel-scale spectrograms $S_{bc} \in \mathbb{R}^{128\times L}$ as input in Fig. 2. Note

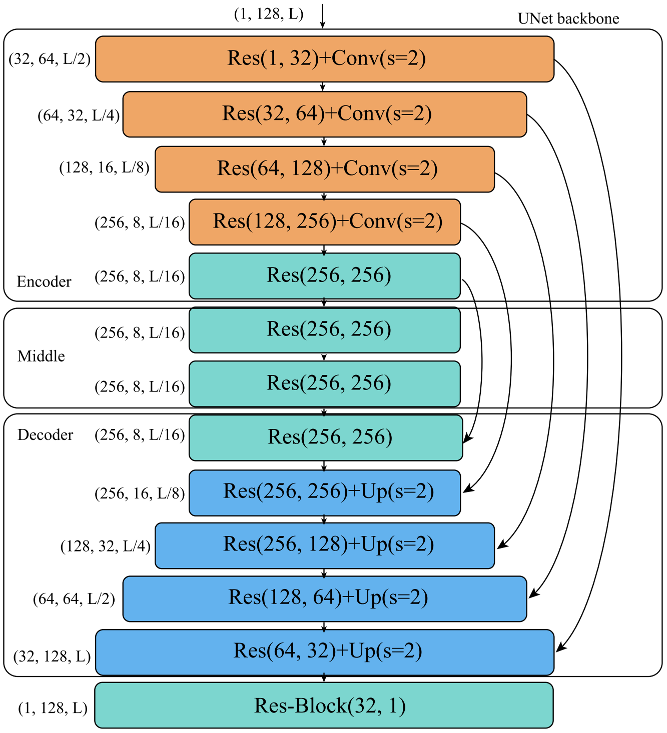

(1, 128, L)
UNet backbone
(32, 64, L/2) Res(1, 32)+Conv(s=2)
(64, 32, L/4) Res(32, 64)+Conv(s=2)
(128, 16, L/8) Res(64, 128)+Conv(s=2)
(256, 8, L/16) Res(128, 256)+Conv(s=2)
Encoder
(256, 8, L/16) Res(256, 256)
Middle
(256, 8, L/16) Res(256, 256)
(256, 8, L/16) Res(256, 256)
Decoder
(256, 8, L/16) Res(256, 256)
(256, 16, L/8) Res(256, 256)+Up(s=2)
(128, 32, L/4) Res(256, 128)+Up(s=2)
(64, 64, L/2) Res(128, 64)+Up(s=2)
(32, 128, L) Res(64, 32)+Up(s=2)
(1, 128, L) Res-Block(32, 1)

Fig. 2. The structure of the UNet backbone.

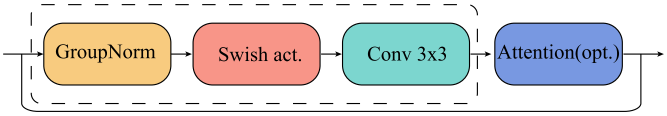


Fig. 3. Block diagram of the basic building structure in the UNet backbone.

that an extra channels dimension is added to the log Mel-scale spectrograms S_{bc}, and the input becomes $S_{in\_bc} \in \mathbb{R}^{1\times 128\times L}$.

The basic building structure of the residual-connected block is shown in Fig. 3, where the neural network layers enclosed in the dashed line are repeated twice. To better model the long-range dependency of speech features [33], we have employed a self-attention mechanism. Additionally, we adopt GroupNorm layers and Swish activation instead of BatchNorm and ReLU activation [34], [35]. Statistical features of speech signals vary significantly across diverse speakers, and hence GroupNorm is expected to outperform BatchNorm in this situation [34]. Due to the quadratic memory and computational complexities of existing self-attention modules [35], it is impractical to use self-attention in all the blocks. We use self-attention only in the middle four blocks of the UNet backbone to balance the system performance and the computational complexity. Because only CNNs and attention mechanisms are utilized, the UNet backbone can be a causal system with a proper padding and feature mask [29], [33].

We investigate two training approaches for the UNet backbone, i.e., supervised learning and adversarial training.

1) Supervised Learning: Since we have obtained paired air-conducted/bone-conducted speech features in (4) and (5), the mapping from log Mel-scale spectrogram of bone-conducted speech S_{bc} to that of the air-conducted speech S_{ac} can be realized via a supervised learning method as shown in Fig. 4. In the training phase, we adopt the l_1 loss between the output of backbone $\mathcal{U}$ and the log Mel-scale spectrogram of air-conducted speech as the objective function

$$L(S_{ac}, \mathcal{U}(S_{bc})) = \|S_{ac} - \mathcal{U}(S_{bc})\|_1 \tag{7}$$

where $\|\cdot\|_1$ denotes the l_1 norm. During inference, the converged feature mapping $\mathcal{U}$ is employed to predict the air-conducted speech features S_{ac} using only the bone-conducted speech features S_{bc}.

2) Adversarial Training: Under supervised learning framework, we assume only one correct answer for the input bone-conducted speech feature, i.e., one-to-one correspondence [36]. However, it should be mentioned that the relationship between the bone-conducted speech and the air-conducted speech is a many-to-many correspondence. Supposing that we place microphones of different types at different positions around the speaker, each microphone will record different air-conducted speech. All different versions of air-conducted speech can be the proper bandwidth extension target of the bone-conducted speech. Consequently, supervised learning may be not the best training approach for the UNet backbone [37], and we leverage a discriminator to implement adversarial training of the bandwidth extension system $\mathcal{U}$. To validate our many-to-many correspondence claim, we conduct a recording experiment. Two vibration sensors are placed on cranial vertex and mastoid of the speaker for bone-conducted speech recording. One Neumann U87 microphone and one Android mobile phone are placed at different places near the speaker for air-conducted speech recording. The power spectrum of recorded signals are presented in Fig. 5. Clear distinctions can be observed when comparing Fig. 5(a) with Fig. 5(b), as well as Fig. 5(c) with Fig. 5(d), which confirms our claim of many-to-many correspondence.

Fig. 6 illustrates the adversarial training procedure of the bandwidth extension system. Different from supervised learning, adversarial training introduces a discriminator D and two extra losses. The UNet generator $\mathcal{U}$ is in dynamic competition with the discriminator D. By alternatively training the generator $\mathcal{U}$ and the discriminator D, the performance of feature mapping can be further enhanced. During the inference phase, the discriminator is discarded, and the generator is used for mapping the bone-conducted speech feature S_{bc} to the air-conducted counterpart S_{ac}.

The discriminator D punishes the generator $\mathcal{U}$ by differentiating the enhanced log Mel-scale spectrogram $\mathcal{U}(S_{bc})$ from the true log Mel-scale spectrogram of air-conducted speech S_{ac}. Therefore, alongside the existing l_1 loss $L(S_{ac}, \mathcal{U}(S_{bc}))$, extra adversarial loss and feature matching loss are imposed to the bone-conducted/air-conducted feature mapping network $\mathcal{U}$. The adversarial loss imposed on $\mathcal{U}$ by the discriminator D is

$$L^{adv}_{\mathcal{U}}(S_{bc}) = -D(\mathcal{U}(S_{bc})). \tag{8}$$

We also consider deep feature matching loss originally applied to speech denoising and speech synthesis [27], [39], [40], [41], [42], which catches noticeable differences between the generated feature $\hat{S}_{ac}$ and the original air-conducted speech feature

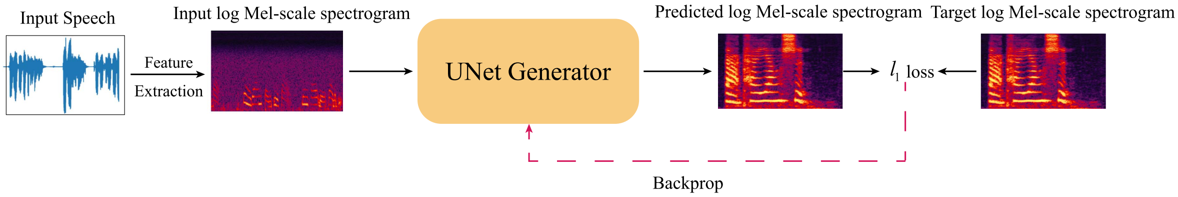

Input Speech
Feature
Extraction
Input log Mel-scale spectrogram
UNet Generator
Predicted log Mel-scale spectrogram
l_1 loss
Target log Mel-scale spectrogram
Backprop

Fig. 4. Training process of the UNet backbone by supervised learning.

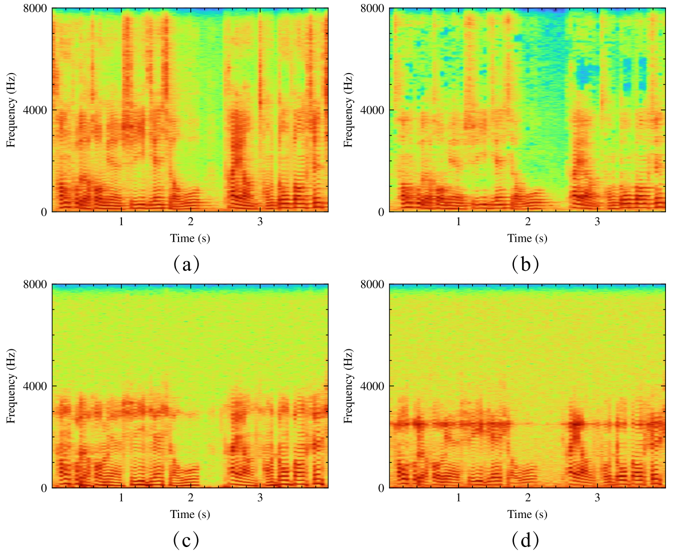


Fig. 5. Magnitude spectrograms of signals recorded by (a) Neumann U87 microphone, (b) Android mobile phone, (c) vibration sensor placed on cranial vertex and (d) mastoid.

S_{ac}, thus enhancing the power of variable discrimination. Mode collapse can be prevented by deep feature matching loss. The generator is forced to match the reference content, which prevents it from cheating by producing monotonous examples. The deep feature matching loss can be formulated as

$$L_{\mathcal{U}}^{FM}(S_{ac}, S_{bc}; D) = \sum_{i=1}^{T} \frac{1}{N_i} \left\| D^{(i)}(\mathcal{U}(S_{bc})) - D^{(i)}(S_{ac}) \right\|_1 \tag{9}$$

where T represents the number of intermediate layers of D, $D^{(i)}(\centerdot)$ is the output of the ith intermediate layer, and N_i is node number of the ith intermediate layer of discriminator D. The complete objective function imposed on the bone-conducted/air-conducted feature mapping is

$$L_{\mathcal{U}} = L(S_{ac}, \mathcal{U}(S_{bc})) + L_{\mathcal{U}}^{adv}(S_{bc}) + L_{\mathcal{U}}^{FM}(S_{ac}, S_{bc}; D) \tag{10}$$

The objective loss for the discriminator D is Hinge loss

$$\begin{aligned} L_D(S_{ac}, S_{bc}) = &\max[1 - D(S_{ac}), 0] \\ &+ \max[1 + D(\mathcal{U}(S_{bc})), 0] \end{aligned} \tag{11}$$

Fig. 7 presents the structure of the discriminator D, which resembles the STFT discriminator in SoundStream [30]. The discriminator D contains 6 residual-connected convolutional blocks, which provides a large receptive field and maintains the training stability. Two convolution layers are utilized in the residual-connected block, and strided convolution is used in each residual-connected block for down-sampling. Our discriminator D employs no global pooling layer but a window-based output [38], which means that the discriminator D needs differentiate each patch of the input log Mel-scale spectrogram. In other words, the output of the discriminator D is a two-dimensional matrix, and each number of the output logits represents the probability for the corresponding patch of the input log Mel-scale spectrogram to be generated or true. As observed in image translation and speech synthesis tasks, such a setup leads to a better performance [38], [39].

C. Speech Vocoder

In the first stage, we have obtained enhanced log Mel-scale spectrogram of the bone-conducted speech $\hat{S}_{ac}(p, l)$. However, such log Mel-scale spectrogram is the intermediate representation and cannot be inversed into speech waveform directly. In the second stage, we hence employ a vocoder to synthesize the time-domain waveform $\hat{x}_{ac}(n)$ from the enhanced speech feature $\hat{S}_{ac}(p, l)$.

A vocoder is used to generate speech waveforms from the temporal-frequency representations, which is an essential part of speech processing [43], [44]. Recent years have seen a rise in popularity of neural vocoders that use generative models [27], [30], [39], [45]. A neural vocoder can convert the low-dimensional spectral representations of an audio signal to the high-quality waveforms efficiently. Therefore, the utilization of a vocoder may eliminate speech artifacts commonly seen in speech enhancement models. Moreover, a speech vocoder trained on a multi-speaker speech dataset generalizes well to unknown speakers [27]. Thanks to the rapid advances of DNN, many low-complexity and high-performance speech Vocoders such as MelGAN [27], HiFi-GAN [39] and SoundStream [30] have been developed. We here employ a DiffWave vocoder based on diffusion model to generate high fidelity speech waveforms. It should be mentioned that other Vocoders can be adopted, which is beyond the scope of this paper.

Diffusion model is a multi-latent-variable generation model, which consists of a diffusion process and a denoising process. In the diffusion process, the data point x_0 is gradually transformed into Gaussian noise x_I by adding Gaussian noise step-by-step following a preset noise schedule $\{\beta_1, \beta_2, \ldots, \beta_I\}$. The original data distribution is transformed to a Gaussian distribution by repeated application of a Markov diffusion kernel $q(x_{i+1}|x_i)$ for I steps, where $x_i, i = 1, \ldots, I-1$ is the perturbed data with

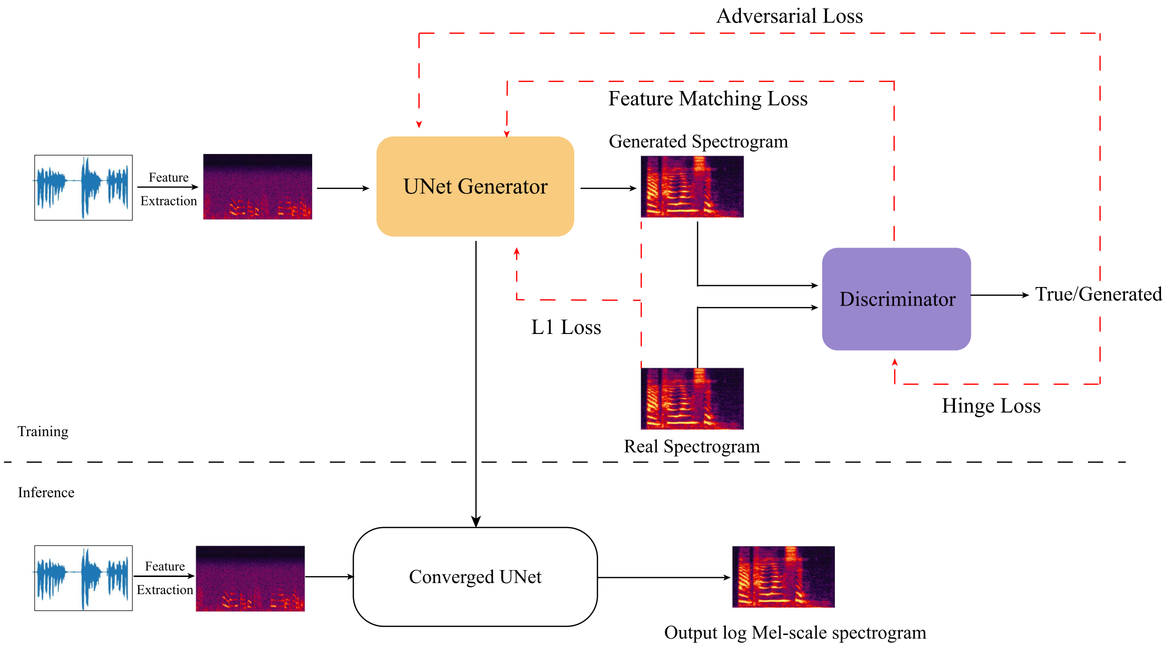


Fig. 6. Bone-conducted/air-conducted feature mapping in adversarial training.

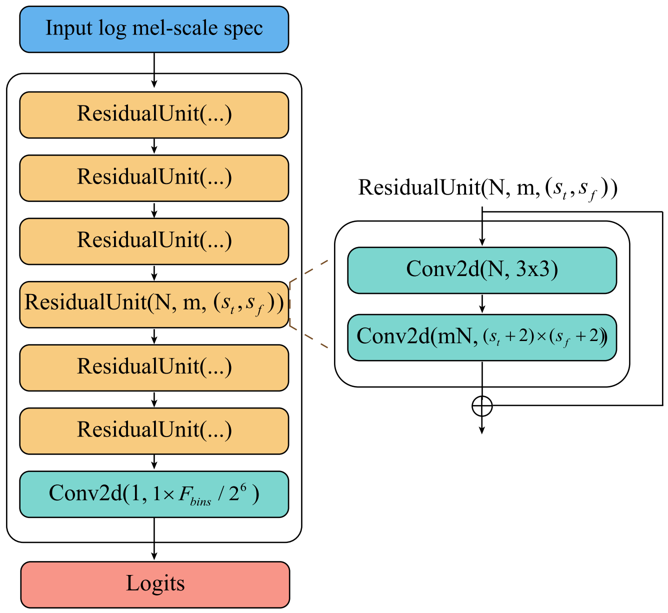

Input log mel-scale spec
ResidualUnit(...)
ResidualUnit(...)
ResidualUnit(...)
ResidualUnit(N, m, (s_t, s_f))
ResidualUnit(...)
ResidualUnit(...)
Conv2d(1, $1\times F_{bins}/2^6$)
Logits
ResidualUnit(N, m, (s_t, s_f))
Conv2d(N, 3x3)
Conv2d(mN, $(s_t+2)\times(s_f+2)$)

Fig. 7. Architecture of Discriminator D. Conv2d$(n, (s_t+2)\times(s_f+2))$ is two-dimensional convolution with output channels n, kernel size $(s_t+2)\times(s_f+2)$, and convolution stride $s_t \times s_f$.

varying levels of Gaussian noise. In the denoising process, a neural network ϵ_θ (θ being the learnable parameters) is trained to predict the added Gaussian noise of every diffusion step so that the intermediate variables can be denoised. By conducting denoising step-by-step, we can generate the original data from Gaussian distribution. Specifically, DiffWave uses WaveNet that is locally conditioned on log Mel-scale spectrogram as ϵ_θ of the aforementioned Diffusion Model [24], [29]. The log Mel-scale spectrogram can be transformed into high-quality speech by iteratively applying the WaveNet denoiser to Gaussian noise. One limitation of DiffWave vocoder is that the speech generation process is often time-consuming [46].

The DiffWave vocoder can be trained on any multi-speaker speech dataset. We initially chose air-conducted speech clips in the Elevoc Simultaneously-recorded Microphone/Bone-sensor (ESMB) dataset[1] as a suitable corpus for training the vocoder. However, we found that the vocoder trained with air-conducted speech clips requires a long convergence time, which may be caused by the low speech quality and limited quantity of air-conducted speech in ESMB dataset. Therefore, we have trained the DiffWave vocoder on the AISHELL-3 dataset [47], which contains enough high-quality speech from native Chinese speakers. Despite the mismatch between the AISHELL-3 dataset and the bone-conducted/air-conducted speech dataset, the DiffWave vocoder trained on the former is capable of producing high-quality speech from the enhanced features $\mathcal{U}(S_{bc})$ as shown latter.

III. Experimental Setup

This section introduces speech datasets, evaluation metrics as well as the training details of the proposed model.

A. Air-Conducted/bone-Conducted Speech Data

We use ESMB speech corpus to evaluate the performance of the proposed system. The ESMB dataset comprises Chinese speech corpus of 128 hours, which are uttered by 131 male speaker and 156 female speakers. The recording device for the ESMB dataset is Elevoc Clear earbuds that consists of one ST25ba bone-conduction vibration sensor and one air-conduction sensor. The bone-conducted vibration sensor located near the entry of the ear canal is utilized to gather skull vibrations

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

during articulation, and the air-conducted sensor outside the ear acts as the close-talk microphone. The sampling rate is 16 kHz. Each speaker reads Chinses prompts for 20 minutes. We set aside three female and three male speakers for evaluation. One tenth of the training data is randomly chosen as validation data to adjust the hyperparameters. After obtaining the optimal hyperparameters, we train the bone-conducted/air-conducted feature mapping system on the whole training data and evaluate the converged system on the evaluation data.

B. Multi-Speaker Speech Dataset

AISHELL-3 dataset [47] is utilized to train the DiffWave vocoder. AISHELL-3 is a multi-speaker Mandarin speech corpus that contains roughly 85 hours of emotion-neutral recordings spoken by 218 native mandarin speakers. All utterances in AISHELL-3 are recorded in a quiet indoor environment using high-fidelity microphones (44.1 kHz, 16-bit depth). We resample the speech data to 16 kHz for the training of our vocoder.

C. Training Details

For all experiments, utterances are divided into small 4-second clips to form mini-batch. For the calculation of the log Mel-scale spectrograms of air-conducted/bone-conducted speech, we choose a window length of 64 ms ($N = 1024$) and a window shift length of 16 ms ($R = 256$). For bone-conducted/air-conducted feature mapping system, we set the mini-batch size as 16. When being trained with supervised learning method, the bone-conducted/air-conducted feature mapping system is optimized with AdamW optimizer [48] at the learning rate of 0.0001 for 500k steps on our training data. We adopt a cosine learning rate schedule with 50k warmup steps. Notably, when being trained with supervised learning method, the bone-conducted/air-conducted feature mapping becomes unstable and may diverge if a larger learning rate is used.

When employing adversarial training method, the generator $\mathcal{U}$ and the discriminator D are alternatively optimized with Adam optimizer [49] at the learning rate of 0.0001. Specifically, for a batch of data, we first fix the parameters of the generator $\mathcal{U}$ and only update the parameters of the discriminator D, and vice versa. The batch size is 16 and the generator $\mathcal{U}$ is adversarial trained for 500k steps. After training, the generator $\mathcal{U}$ is kept for mapping the log Mel-scale spectrogram of bone-conducted speech S_{bc} into that of air-conducted speech S_{ac}, while the discriminator is deserted.

For DiffWave vocoder, we follow the setting of DiffWave Base in [24]. The only difference is that we use the 128-dimensional log Mel-scale spectrogram instead of the original 80-dimensional one as the input of DiffWave vocoder. The DiffWave vocoder is trained for 1M steps on AISHELL-3 dataset with a mini-batch size of 8.

D. Evaluation Metrics

To evaluate the proposed two-stage restoration system, we perform two kinds of subjective evaluations. For speech quality judgement, we conduct a paired comparison test [50] between different restoration systems. In paired comparison test, different restoration systems restore the same bone-conducted speech. Pairs of restored speech clips generated by different systems are presented to the test subject in a randomized order, and the test subject has to indicate the one with better quality. We randomly choose six bone-conducted speech clips (one for each speaker) as the evaluation data for paired comparison test. It should be noted that the paired comparison test is a relative preference method and cannot directly provide a measurable level of speech quality improvement. For intelligibility judgement, we conduct an absolute category rating (ACR) experiment [51]. We randomly choose 30 bone-conducted speech clips for test (five for each individual speaker). Subjects give each speech file a single opinion on the five-point ACR scale. The final results of this subjective test are expressed in terms of mean opinion scores (MOS). In the training phase of the ACR experiment, we use air-conducted speech (excellent, 5) and bone-conducted speech (bad, 1) to equalize the subjective range of intelligibility ratings of all listeners. Twenty Chinese listeners with no hearing impairment participated in the subjective tests.

For development and hyperparameter selection, we choose two objective metrics: short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) [52]. STOI typically has a value range from 0 to 1, which can be considered as correct percentage. PESQ ranges from -0.5 to 4.5. Higher values of STOI and PESQ imply a better enhancement performance.

IV. Experiment Results

A. Comparison With Existing Restoration Models

We set four baselines to demonstrate the effectiveness of the proposed two-stage model. We follow the code implementation[2] and train DPT-EGNet on ESMB corpus. We implement DC-CRN [6] and train DCCRN-BC following [6]. It is worth noting that the training and evaluation speech signals for DCCRN-BC have a sampling rate of 16 kHz, instead of 8 kHz used in [6]. We implement two other baselines to validate the necessity of combining bandwidth extension and vocoder. Instead of restoring the log Mel-scale spectrogram of bone-conducted speech, we employ the proposed UNet architecture combined with adversarial training to restore the magnitude spectrogram of bone-conducted speech. The restored magnitude spectrogram is combined with the phase of bone-conducted speech to generate restored speech. This restoration system is referred to as S1. Additionally, we train a restoration DiffWave vocoder (S2) that converts the log Mel-scale spectrogram of the oracle bone-conducted speech into the corresponding air-conducted speech waveforms. S1 and S2 are trained using the same training strategies as the first and second stage models, respectively.

Table I presents the MOS of different restoration systems. The 95% confidence intervals under t-distribution are also presented. As seen, the proposed two-stage system outperforms all the baselines in terms of MOS, which demonstrates the effectiveness of

[2] [Online]. Available: https://github.com/echoaimaomao/Codes-and-Demos-for-DPT-EGNet

TABLE I
COMPARISION WITH EXISTING MODELS (MOS)

Systems	MOS	CI_{95}
Proposed	4.43	[4.22, 4.65]
S1	3.66	[3.41, 3.91]
S2	2.29	[2.08, 2.51]
DCCRN-BC [6]	2.29	[2.08, 2.50]
DPT-EGNet [21]	1.87	[1.67, 2.08]

TABLE II
RESULTS OF PAIRED COMPARISON TEST

Frequency	Ranking				
	Proposed	S1	S2	DCCRN-BC [6]	DPT-EGNet [21]
34	1	2	4	3	5
20	1	2	5	3	4
6	1	2	3	5	4
23	1	2	3	4	5
7	1	3	4	2	5
5	1	4	3	2	5
1	1	2	3	4	5
2	1	2	4	4	4
2	1	2	5	4	3
3	2	1	5	3	4
1	2	1	3	5	3
2	2	1	4	3	5
1	1	2.5	2.5	4.5	4.5
2	1	3	2	5	4
2	1	4	2	3	5
1	1	3	5	2	4
6	2	1	3	4	5
1	3	1	2	4	4
1	2	2	2	5	5

TABLE III
RANK SUMS OF DIFFERENT RESTORATION SYSTEMS

Systems	Rank sum
Proposed	135
S1	251.5
S2	450.5
DCCRN-BC [6]	403.5
DPT-EGNet [21]	559.5

the proposed approach in terms of speech intelligibility. Table II shows the results of pairwise comparison between five systems in two presentation orders of 120 blocks. We use Friedman's rank sum test [50] to analyze the results. For each presentation the five systems are scored, i.e., a system gets a score 1, 2, 3, 4 or 5, when it is preferred four times, three times, twice, once or not at all, respectively. If there are tied values, we assign to each tied value the average of the ranks that would have been assigned without ties. The rank scores are summed for each system as presented in Table III. The Friedman's test static [50] for our paired comparison test is 374.27. This result should be compared to a chi-square distribution with 4 degrees of freedom. As the Friedman's test statistic is greater than the 5% critical level (i.e., $374.27 > 9.49$) of a chi-square distribution with 4 degrees of freedom, we conclude that a significant difference exists among the five compared systems. The rank sums in Friedman's analysis provide a scale on which to position the different systems. Large difference on this scale indicates a significant variation among the systems being compared. As seen, the proposed two-stage model achieves a better bone-conducted speech enhancement performance than baselines in terms of speech quality.

TABLE IV
COMPARISION WITH DENOISING MODELS AT A SNR OF -10 DB (MOS)

Systems	MOS	CI_{95}
Proposed System	4.43	[4.22, 4.65]
DCCRN-AF [6]	2.32	[2.12, 2.51]
WaveNet [53]	1.78	[1.58, 1.98]

It is worth noting that S1 and S2 achieve good restoration performance. However, as reported from test subjects, the single-stage system S1 results in annoying speech artifacts. Additionally, although S2 produces speech with good speech quality, it is prone to generating speech that is characterized by blurriness and lack of clarity. A speech vocoder alone is unable to fully recover the missing information of bone-conducted speech. Bone-conducted speech enhancement requires the generation of missing high-frequency content, which poses a significant challenge. The combination of bandwidth extension and speech vocoder can yield better restoration performance. Hence, for challenging speech restoration tasks, the two-stage approach may be a superior choice compared to single-stage models. Some speech samples are presented in the website. [3]

B. Comparison With Speech Denoising

We compare the proposed approach with two speech denoising systems, i.e., WaveNet [53] and DCCRN-AF [6] in very low SNR environment. WaveNet solely incorporates noisy speech as its input, whereas DCCRN-AF is a multimodal model that takes in both noisy speech and the bone-conducted speech. We generate noisy training speech by mixing the air-conducted speech of our training dataset with noise data from three corpus: DNS challenge,[4] Nonespeech115 [54] and NOISEX92 [55]. The SNR level is uniformly sampled from [-5, 5] dB. We construct the evaluation dataset by using clean air-conducted speech and noise data from CHIME-3 [56]. The street noise, café noise, pedestrian noise, and bus noise, which are utilized to construct the testing dataset, are not present in the training dataset. Particularly, evaluation samples are synthetically mixed for SNR = -10 dB. The training dataset consists of 39,258 pairs of noisy speech and corresponding bone-conducted speech, while the testing dataset comprises 961 pairs of noisy speech and corresponding bone-conducted speech. We adhere to the training methodology outlined in [53] while training WaveNet. We train and evaluate DCCRN-AF using speech samples with a sampling rate of 16 kHz, instead of the 8 kHz sampling rate in [6]. Apart from this difference, the training methodology of DCCRN-AF remains consistent with the description in [6].

Table IV shows the MOS of different restored speech and the 95% confidence intervals under t-distribution. As observed, the proposed two-stage system achieves promising enhancement performance, outperforming speech denoising systems in very low SNR environment. Listeners report that the denoising model WaveNet, which only utilizes noisy speech as input, shows limited effectiveness in attenuating noise in very low

[3] [Online]. Available: https://ioa-audio.github.io/2023/09/15/two-stage-demo/

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

TABLE V
PESQ Comparison Between Different Two-Stage Systems

	Original UNet	CRN	UNet w/o attention	UNet
DiffWave	2.10	2.06	2.16	2.16
MelGAN	1.86	1.81	1.94	1.94

TABLE VI
STOI Comparison Between Different Two-Stage Systems

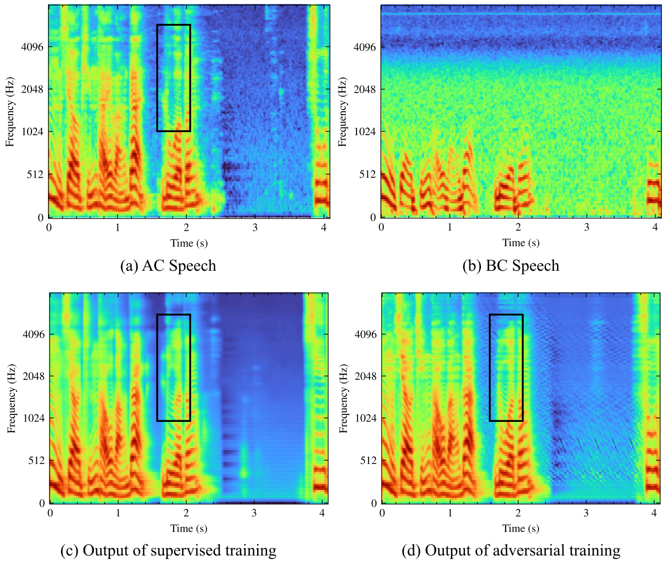
	Original UNet	CRN	UNet w/o attention	UNet
DiffWave	0.73	0.72	0.77	0.78
MelGAN	0.71	0.72	0.76	0.75

SNR scenarios. The enhanced speech generated by WaveNet still retains a noticeable presence of residual noise. The fusion model, DCCRN-AF, exhibits commendable noise reduction capabilities. However, DCCRN-AF may introduce some speech distortions and imperfections. Furthermore, by comparing Tables I and IV, it can be observed that DCCRN-AF does not show significant advantages over DCCRN-BC at a SNR of -10 dB, which demonstrates the merits of bone-conducted speech for telecommunication in very low SNR scenarios. We also conducted informal evaluations under different SNR conditions using the WV-MOS [57] metric (not shown here), which shows that the fusion model DCCRN-AF and the denoising model WaveNet outperform the proposed system for SNR $>$ 0 dB. We hence stress that the bone-conducted speech restoration system is preferred in very low SNR conditions.

C. Comparison With Other Two-Stage Models

For the two-stage approach, it is feasible to employ distinct network architectures as the backbone network for bandwidth extension and utilize different vocoders in the second stage. To validate the effectiveness of the proposed UNet network and DiffWave vocoder, we have implemented various bone-conducted speech enhancement systems within the two-stage framework. Specifically, we have employed the CRN network [58] and the original UNet [31] as replacements for our improved UNet. Furthermore, to investigate the role of self-attention mechanism in our improved UNet network, we have also trained another improved UNet, which does not incorporate any attention mechanism. The aforementioned models are trained using the same adversarial training strategy and training data. We utilize a GAN-based vocoder called MelGAN [27] to replace DiffWave vocoder in our bone-conducted speech enhancement system. The MelGAN vocoder is trained using Aishell-3 dataset following the description in [27]. The PESQ and STOI scores for the different two-stage models are respectively presented in Tables V and VI.

For reference, the PESQ and STOI for bone-conducted speech are 1.32 and 0.60, respectively. Consequently, all two-stage systems have substantially improved the speech quality and intelligibility of bone-conducted speech. As seen, it is evident that the original UNet network exhibits advantages over the CRN network. The encoder-decoder architecture and skip connections of UNet effectively capture both long-term and short-term dependencies in data, which confirms its suitability as a backbone network for various tasks. The improved version of UNet that we utilize has yielded superior results compared to the original UNet. This verifies the effectiveness of the residual connection-based building blocks we employed in our framework. It is worth noting that the self-attention mechanism does not yield significant performance improvements in our study. The self-attention mechanism tends to demonstrate its advantages more prominently in the context of modeling long sequences. However, in our UNet network, the feature maps in the intermediate layers often have relatively low dimensions. Consequently, the utilization of self-attention mechanism in these intermediate layers may not fully exploit the potential benefits.



(a) AC Speech
(b) BC Speech
(c) Output of supervised training
(d) Output of adversarial training

Fig. 8. Log Mel-scale spectrograms of different speech signals.

The DiffWave vocoder demonstrates a more notable advantage over MelGAN in terms of the PESQ. However, the difference between the two vocoders in terms of STOI, a metric that is more relevant to bandwidth extension [59], is not significant. Considering that DiffWave requires longer inference time and more computational resources, the MelGAN vocoder also has practical value.

D. The Merits of Adversarial Training

We conduct an ablation study to validate the merits of adversarial training in bone-conducted/air-conducted feature mapping. Fig. 8 shows the log Mel-scale spectrograms of speech enhanced by bone-conducted/air-conducted feature mapping using various training protocols. For comparison, the log Mel-scale spectrograms of air-conducted and bone-conducted speech are also presented. As observed from Fig. 8, adversarial training improves the ability of the bone-conducted/air-conducted feature mapping to generate log Mel-scale spectrograms with crisp and distinct high-frequency content, compared to supervised learning. This is consistent with the experimental results shown in [30], [40], [42], where additional discriminators were used to improve speech enhancement performance. Bone-conducted

TABLE VII
PERFORMANCE COMPARISON BETWEEN FEATURE MAPPING TRAINED WITH DIFFERENT METHODS (PESQ, STOI)

Systems	PESQ	STOI
Adversarial training w/ l_1-loss	2.16	0.78
Supervised learning	2.10	0.75
Bone-conducted speech	1.32	0.60

speech enhancement should be treated as a generative task, suggesting that adversarial training may be a promising approach for tackling this challenge.

We also investigate how adversarial learning influences the quality and intelligibility of the enhanced speech. Table VII reports the quality and intelligibility of the speech enhanced by bone-conducted/air-conducted feature mapping under different training protocols in terms of PESQ and STOI. Both training protocols significantly enhance the PESQ and STOI of bone-conducted speech. Notice from Table VII that adversarial training further improves the PESQ and STOI of the enhanced speech, which shows that adversarial training can help produce better log Mel-scale spectrogram.

However, the difference between adversarial training and supervised learning is not very significant in subjective listening tests. Though adversarial training can lead to slightly better objective metrics such as PESQ and STOI, it does not significantly improve the subjective perception of the synthesized speech. This may be because the speech vocoder at the second stage plays an important role in determining the perceptual quality of the enhanced speech. In our experiments, it is observed that the speech vocoder can even help to remove the spectral artifacts produced in the bandwidth extension stage. Nevertheless, our experiments have shown that the feature mapping trained with supervised learning may not converge under certain learning rates, whereas the feature mapping trained with adversarial training with the same learning rate still converges. Hence, adversarial training can improve the robustness of the feature mapping to hyperparameter settings, which is highly appreciated.

E. Comparison Between Different Speech Features

At this point, we investigate the impact of different speech features on bone-conducted speech restoration. Specifically, we will explore the utilization of linear predictive coding (LPC) and line spectral pairs (LSP) as alternatives to the log Mel-scale spectrogram in our two-stage system. We employ the open-source software spafe[5] to extract 20-dimensional LPC and LSP features from the speech signals. For the calculation of LPC, we choose a window length of 64 ms and a window shift length of 16 ms. Due to the lower dimensionality of LPC and LSP features, we have appropriately reduced the number of layers in the UNet used for feature mapping. The encoder and decoder of the UNet, which maps the speech log Mel-scale spectrogram, consist of 5 residual connections. In contrast, the encoder and decoder of UNet used for mapping LPC and LSP only consists of 3 layers. Apart from appropriately reducing the number of layers in the feature mapping UNet and utilizing LPC and LSP features, no further modifications have been made to the two-stage model. After training, we obtain four distinct models. To distinguish them from the proposed two-stage model, we denote these four models as UNet-LPC, UNet-LSP, DiffWave-LPC, and DiffWave-LSP. The PESQ and STOI of two-stage models with different speech features are presented in Table VIII. As seen, the two-stage systems utilizing LPC and LSP features fail to generate high-quality restored speech.

TABLE VIII
TWO-STAGE SYSTEMS USING DIFFERENT SPEECH FEATURES (PESQ, STOI)

	log Mel-scale spectrogram	LPC	LSP
PESQ	2.16	1.18	1.14
STOI	0.78	0.60	0.42

TABLE IX
QUALITY AND INTELLIGIBILITY OF VOCODED SIGNALS PRODUCED BY VARIOUS VOCODERS (PESQ, STOI)

	DiffWave	DiffWave-LPC	DiffWave-LSP
PESQ	3.02	1.30	1.16
STOI	0.91	0.72	0.49

We found that the reason behind the failure of these two-stage models lies in their vocoder component. The DiffWave vocoder struggles to effectively convert LPC and LSP features into speech waveforms. We directly input features extracted from air-conducted speech into the vocoder, and we present the PESQ and STOI of the vocoder-generated speech in Table IX. As seen, DiffWave Vocoder using log Mel-scale spectrogram can generate speech with a PESQ of 3.02 and a STOI of 0.91 when the real log Mel-scale spectrogram is fed directly. Therefore, DiffWave Vocoder using log Mel-scale spectrogram does not limit the performance of the two-stage approach, and the improvement of bone-conducted/air-conducted feature mapping can further enhance the performance of the overall system. However, both DiffWave-LPC and DiffWave-LSP fail to generate high-quality speech even when provided with speech features of air-conducted speech as input. Therefore, LPC and LSP may not be the optimal feature for our two-stage system.

F. Ablation Study About Multi-Speaker Dataset

Finally, we investigate the impact of the training dataset used for the vocoder on the overall performance of the two-stage system. As previously mentioned, the low-quality air-conducted speech data in the ESMB dataset poses challenges in training the DiffWave vocoder. We employed 4 Nvidia RTX2080Ti GPUs to accelerate the training of the vocoder. When training the model with air-conducted speech from the ESMB dataset, it requires 15 days to converge. In contrast, when utilizing the Aishell-3 dataset, DiffWave achieves convergence in 5 days. We present the PESQ and STOI of two-stage models with DiffWave trained using different datasets in Tables X and XI, respectively. We have found that training the vocoder using the ESMB dataset proves to be more beneficial for the overall performance of the two-stage system. However, it is worth noting

[5] [Online]. Available: https://spafe.readthedocs.io/en/latest/

TABLE X
PESQ OF DIFFERENT COMBINATIONS OF DIFFWAVE VOCODERS AND SPEECH FEATURES

Features	DiffWave (Aishell-3)	DiffWave (ESMB)
Enhanced spectrogram	2.16	2.22
Groundtruth spectrogram	3.02	3.31

TABLE XI
STOI OF DIFFERENT COMBINATIONS OF DIFFWAVE VOCODERS AND SPEECH FEATURES

Features	DiffWave (Aishell-3)	DiffWave (ESMB)
Enhanced spectrogram	0.78	0.78
Groundtruth spectrogram	0.91	0.93

that employing a vocoder trained on an out-of-domain dataset also yields impressive results in terms of bone-conducted speech restoration, which demonstrates the generalization ability of the proposed two-stage system. Furthermore, in future research, we can leverage larger external datasets to train vocoders that possess improved generalization capabilities.

V. CONCLUSION

In this paper, we have proposed a new bone-conducted speech enhancement system to improve the quality and intelligibility of the bone-conducted speech, in particular for telecommunications in very low SNR conditions. The proposed system consists of two independent stages. In the first stage, the proposed system maps the log Mel-scale spectrogram of the bone-conducted speech to that of the of air-conducted speech. In the second stage, a speech vocoder trained on an extra multi-speaker data transforms the enhanced log Mel-scale spectrogram to the time-domain waveform. We introduced adversarial training in the first stage that can lead to a better bone-conducted/air-conducted feature mapping compared with supervised learning. The speech vocoder can guarantee the generalization to unknown speakers. Experiments on ESMB dataset validated the advantages of the proposed method over the existing models in terms of objective and subjective evaluations.

REFERENCES

[1] T. Dekens and W. Verhelst, “Body conducted speech enhancement by equalization and signal fusion,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 2481–2492, Dec. 2013.

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

[3] R. E. Bouserhal, T. H. Falk, and J. Voix, “On the potential for artificial bandwidth extension of bone and tissue conducted speech: A mutual information study,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5108–5112.

[4] M. Wang, J. Chen, X. Zhang, Z. Huang, and S. Rahardja, “Multi-modal speech enhancement with bone-conducted speech in time domain,” *Appl. Acoust.*, vol. 200, 2022, Art. no. 109058.

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

[6] H. Wang, X. Zhang, and D. Wang, “Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 3134–3143, 2022.

[7] M. Wang, J. Chen, X. -L. Zhang, and S. Rahardja, “End-to-end multi-modal speech recognition on an air and bone conducted speech corpus,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 513–524, 2023.

[8] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, “Combining standard and throat microphones for robust speech recognition,” *IEEE Signal Process. Lett.*, vol. 10, no. 3, pp. 72–74, Mar. 2003.

[9] T. Shimamura and T. Tamiya, “A reconstruction filter for bone-conducted speech,” in *Proc. IEEE 48th Midwest Symp. Circuits Syst.*, 2005, vol. 2, pp. 1847–1850.

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

[11] T. Tat Vu, M. Unoki, and M. Akagi, “An LP-based blind model for restoring bone-conducted speech,” in *Proc. IEEE 2nd Int. Conf. Commun. Electron.*, 2008, pp. 212–217.

[12] T. V. Tat, G. Seide, M. Unoki, and M. Akagi, “Method of LP-based blind restoration for improving intelligibility of bone-conducted speech,” in *Proc. Interspeech*, 2007, pp. 966–969.

[13] P. Nghia Trung, M. Unoki, and M. Akagi, “A study on restoration of bone-conducted speech in noisy environments with LP-based model and Gaussian mixture model,” *J. Signal Process.*, vol. 16, no. 5, pp. 409–417, 2012.

[14] 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, no. 4, pp. 175–178, Jul. 2020.

[15] M. A. T. Turan and E. Erzin, “Source and filter estimation for throat-microphone speech enhancement,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 2, pp. 265–275, Feb. 2016.

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

[17] M. J. Bianco et al., “Machine learning in acoustics: Theory and applications,” *J. Acoust. Soc. Amer.*, vol. 146, no. 5, pp. 3590–3628, Nov. 2019.

[18] H.-P. Liu, Y. Tsao, and C.-S. Fuh, “Bone-conducted speech enhancement using deep denoising autoencoder,” *Speech Commun.*, vol. 104, pp. 106–112, Nov. 2018.

[19] 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. Fundam.*, vol. E102.A, no. 12, pp. 2001–2007, Dec. 2019.

[20] 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.

[21] C. Zheng, L. Xu, X. Fan, J. Yang, J. Fan, and X. Huang, “Dual-path transformer-based network with equalization-generation components prediction for flexible vibrational sensor speech enhancement in the time domain,” *J. Acoust. Soc. Amer.*, vol. 151, no. 5, pp. 2814–2825, May 2022.

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

[23] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, “Image super-resolution via iterative refinement,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023 .

[24] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “DIFFWAVE: A versatile diffusion model for audio synthesis,” in *Proc. Int. Conf. Learn. Representations*, 2020.

[25] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6309–6318.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[27] K. Kumar et al., “MelGAN: Generative adversarial networks for conditional waveform synthesis,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14910–14921.

[28] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, “Linear versus MEL frequency cepstral coefficients for speaker recognition,” in *Proc. IEEE Autom. Speech Recognit. Understanding*, 2011, pp. 559–564.

[29] A. van den Oord et al., “WaveNet: A generative model for raw audio,” 2016, *arXiv:1609.03499*.

[30] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “SoundStream: An end-to-end neural audio codec,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 495–507, 2022.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comp. Comput.-Assist. Interv.*, 2015, pp 234–241.

[32] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019.

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

[34] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–9.

[35] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.

[36] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[37] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.

[39] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17022–17033.

[40] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Proc. Interspeech*, 2022, pp. 4506–4510.

[41] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proc. Interspeech*, 2019, pp. 2723–2727.

[42] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021, pp. 166–170.

[43] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[44] Z. Du, X. Zhang, and J. Han, "A joint framework of denoising autoencoder and generative vocoder for monaural speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1493–1505, 2020.

[45] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 3617–3621.

[46] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[47] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker mandarin," 2021, *arXiv:2010.11567*.

[48] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2022.

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

[50] B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi, and J. Smeyers-Verbeke, *Data Handling in Science and Technology*. New York, NY, USA: Elsevier, 1998.

[51] P. C. Loizou, "Speech quality assessment," in *Multimedia Analysis, Processing and Communications*. W. Lin, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo, and H. Wang Eds. Berlin, Heidelberg: Springer, 2011, pp. 23–34.

[52] T. H. Falk and W.-Y. Chan, "A non-intrusive quality measure of dereverberated speech," in *Proc. 14th Int. Workshop Acoust. Signal Enhancement*, 2008.

[53] D. Rethage, J. Pons, and X. Serra, "A WaveNet for speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5069–5073.

[54] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[55] 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.

[56] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.

[57] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "HiFi: A unified framework for bandwidth extension and speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023.

[58] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Fully convolutional recurrent networks for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6674–6678.

[59] J. Abel, M. Kaniewska, C. Guillaumé, W. Tirry, and T. Fingscheidt, "An instrumental quality measure for artificially bandwidth-extended speech signals," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 2, pp. 384–396, Feb. 2017.

Changtao Li received the B.S. degree in physics from the University of Chinese Academy of Sciences, Beijing, China, in 2019. He is currently working toward the Ph.D. degree in signal processing with the Institute of Acoustics, Chinese Academy of Sciences, Beijing. His research interests include speech enhancement and deep learning.

Feiran Yang (Member, IEEE) received the B.E. degree in electrical engineering from Shandong University, Jinan, China, in 2005, the M.E. degree in signal processing from Southeast University, Nanjing, China, in 2008, and the Ph.D. degree in signal processing from the Institute of Acoustics, Chinese Academy of Sciences (IACAS), Beijing, China, in 2013. From 2008 to 2010, he was with Fortemedia, Inc., as a DSP Engineer. From 2016 to 2017, he was with Ruhr-Universität Bochum, Bochum, Germany, as a Visiting Scholar. Since 2013, he has been with IACAS, where he is currently a Professor. His research interests include adaptive filtering, microphone array signal processing, and spatial audio. He was the recipient of the President's Award of the Chinese Academy of Sciences in 2013. He was also the recipient of the Excellent Doctoral Dissertation Award of the Chinese Academy of Sciences in 2016.

Jun Yang (Senior Member, IEEE) received the B. Eng. and M. Eng. degrees from Harbin Engineering University, Harbin, China and the Ph.D. degree in acoustics from Nanjing University, Nanjing, China, in 1990, 1993, and 1996, respectively. He is currently a Deputy Director with IACAS. From 1996 to 1998, he was a Postdoctoral Fellow with the Institute of Acoustics, Chinese Academy of Sciences (IACAS), Beijing, China. From 1998 to 1999, he was with Hong Kong Polytechnic University as a Visiting Scholar. From 1997 to 1999, he was with IACAS as an Associate Professor. He joined the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, as a Research Fellow, a Teaching Fellow, Assistant Professor, and Associate Professor in 1999, 2001, 2003, and 2005, respectively. Since November 2003, he has been a Professor with IACAS. From 2011 to 2020, he was the Director of the Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences. His reseach interests include communication acoustics, 3-D audio systems, acoustic signal processing, sound field control, and nonlinear acoustics. He is as an Editor-in-Chief of *Sound & Vibration*. He is a Fellow of the International Institute of Acoustics and Vibration.