Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables

Lixing He
Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong SAR China
1155170464@link.cuhk.edu.hk

Haozheng Hou
Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong SAR China
1155161507@link.cuhk.edu.hk

Shuyao Shi
Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong SAR China
ss119@ie.cuhk.edu.hk

Xian Shuai
Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong SAR China
1155118647@link.cuhk.edu.hk

Zhenyu Yan
Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong SAR China
zyyan@cuhk.edu.hk

ABSTRACT

Head-mounted wearables are rapidly growing in popularity. However, a gap exists in providing robust voice-related applications like conversation or command control in complex environments, such as competing speakers and strong noises. The compact design of HMWs introduces non-trivial challenges to existing speech enhancement systems that use microphone recording only. In this paper, we handle this problem by using bone vibration conducted through the head skull. The principle is that the accelerometer is widely installed on head-mounted wearables and can capture the clean user's voice. Hence, we develop *VibVoice*, a lightweight multimodal speech enhancement system for head-mounted wearables. We design a two-branch encoder-decoder deep neural network to fuse the high-level features of the two modalities and reconstruct clean speech. To address the issue of insufficient paired data for training, we extensively measure the bone conduction effect from a limited dataset to extract the physical impulse function for cross-modal data augmentation. We evaluate VibVoice on a dataset collected in real world and compare it with two state-of-the-art baselines. Results show that VibVoice yields up to 21% better performance in PESQ and up to 26% better performance in SNR compared with the baseline with 72 times less paired data required. We also conduct a user study with 35 participants, in which 87% participants prefer VibVoice compared with the baseline. In addition, VibVoice requires 4 to 31 times less execution time compared with baselines on mobile devices. The demo audio of VibVoice is available at https://www.youtube.com/watch?v=8_-s_C_NGRI.

MobiSys '23, June 18–22, 2023, Helsinki, Finland

ACM ISBN 979-8-4007-0110-8/23/06…$15.00
https://doi.org/10.1145/3581791.3596832

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computer systems organization** → **Embedded and cyber-physical systems**; • **Computing methodologies** → Artificial intelligence.

KEYWORDS

Speech Enhancement, Ear-Worn Wearable, Earable sensing

ACM Reference Format:
Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. 2023. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables. In *The 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys '23), June 18–22, 2023, Helsinki, Finland.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3581791.3596832

1 INTRODUCTION

Head-mounted wearables, or HMWs, are smart devices that can be put on users' heads or ears, which include True Wireless Stereo (TWS) earphones, VR/AR headsets, and smart glasses. HMWs are equipped with several sensors and run various applications like VR/AR, motion recognition, and voice assistants. Represented by TWS earphones like the Apple Airpods series, the shipment of HMWs has grown as the largest category among all wearable devices, estimated to be more than 273 million units worldwide in 2023 [39]. Manufacturers are adding various functions to HMWs. For example, some HMWs (especially earphones and headphones) support active noise cancellation (ANC) to improve the listening experience. On the speech side, voice-related applications using HMW s microphones are among the most frequently used.

For example, an increasing number of people make phone calls with TWS earphones. Through voice commands, users can also interact with voice assistants (e.g., Siri and Alexa). However, the speech quality on HMWs is unsatisfactory due to the following challenges: First, most HMWs adopt omnidirectional microphones that can receive audio from any angle, leading to extra environmental noises. Second, although most HMWs are equipped with multiple microphones to form an array and remove noises based on

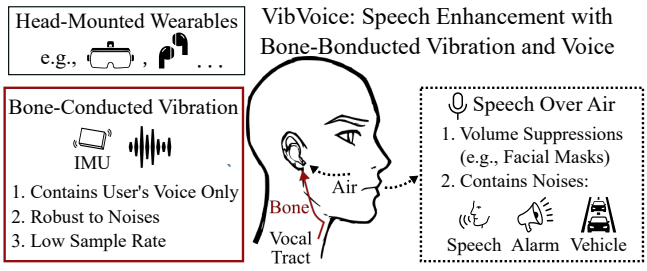


Figure 1: VibVoice enhances speech quality of head-mounted wearables by extracting the user's clear voice from the bone-conducted vibrations.

beamforming, the microphones on HMWs are too close to achieving a satisfying performance. Third, the speech audio is significantly suppressed when it reaches the HMWs as the HMWs are usually far away from the user's mouth. Lastly, the worldwide pandemic forces people to wear masks, which can further reduce the audio intensity.

Various approaches have been developed for speech enhancement. Signal processing-based methods [12] remove noises based on their statistical models. However, these approaches cannot handle complex environments. Microphone beamforming [17, 54] removes noises based on their directions. But they fail to distinguish the speaker's voice and noise when the microphones are placed too close. Several research [9, 41, 56] adopt deep neural networks (DNNs) to improve speech quality. However, their performance varies when the domain changes. Except for the audio-only solution, several works leverage other modalities like contact sensors [14], vibration sensor [25], camera [13], mm-Wave Radar [21, 22, 29], ultrasonic sensing [42, 55], and Lidar [36], which introduce additional hardware requirements or user overhead to existing HMWs.

Inertial Measurement Unit (IMU) is a common sensor on most HMWs and can capture audio with a direct connection with the speaker [7, 46]. Since HMWs are well connected to the user's head, the IMU accelerometer can sense the vibrations caused by the speaker's voice through the skull's bone conduction. The bone-conducted vibration contains speech from the user only without environmental noises or voices. However, the accelerometer on HMWs has a low sample rate (e.g., 1.6 kHz) due to the limited size and energy consumption, which is not enough to recover the human's voice (e.g., up to 3.4 kHz [48]).

We propose *VibVoice*, the first end-to-end multi-modal speech enhancement system for HMWs using the bone conduction from the vibration to the audio. The design of VibVoice fits mobile devices, considering the on-device sensor, execution latency, and communication overhead. VibVoice exploits the complementary characteristics of the two modalities: the acceleration has a low sample rate covering a partial audible spectrum but with no environmental noises. The microphone has a high sample rate covering the whole audible spectrum but is mixed with ambient voice and noises. Such a combination of modalities provides a stable reference for extracting the target speech. VibVoice uses a DNN to reconstruct clean speech with two encoders and decoders for processing the data from two modalities, respectively. VibVoice faces three major technical challenges as follows:

Lack of paired labeled acceleration and audio. Collecting a paired multi-modal dataset is labor-intensive. Although there are public datasets of either acceleration or audio, none has the paired data on head-mounted wearables. Collecting such a paired dataset on a large scale can introduce excessive overheads, such as recruiting diverse volunteers and recording hundreds of hours of audio with annotations.

Fusion of data with different modalities and sample rates. Microphone audio has much richer information than acceleration data due to the higher sample rate. As a result, the DNN is prone to overfitting and may ignore the data input of acceleration. Therefore, during the training of DNN, we require a special training strategy to avoid overfitting and two dedicated losses to balance the weights of the two modalities.

Practical concern for deployment. Deploying deep learning speech enhancement in the real world is non-trivial due to the variance among people's voice and bone conduction channels. Even though some users may tolerate a rapid data collection phase before use, the amount and quality of the collected data are not guaranteed for training a high-performance model.

To overcome the above challenges, we design VibVoice based on our extensive measurements of the bone-conduction effects on real-world data. We conclude our contributions as follows:

- We propose a novel estimation approach to model Bone Conduction Function with a limited dataset to augment paired acceleration and audio based on a large public audio dataset. The augmented paired dataset can be used for model training.
- We design a two-branch DNN that a) recovers the speech-related information from the acceleration signal with a low sample rate, b) extracts the clean speech from contaminated audio with the help of the acceleration signal, and c) has a lightweight design and can be run on the mobile device with low latency.
- We collect a two-modal dataset and evaluate VibVoice on a self-designed platform, *EarSense*, equipped with an accelerometer and a microphone to acquire paired acceleration and audio data simultaneously. The dataset and the design of data collector are open-sourced.

We evaluate VibVoice on a paired acceleration and microphone from 15 volunteers. We test VibVoice's performance on the offline synthetic noise dataset and the concurrent noise dataset in various scenarios. VibVoice achieves the best performance with 31 times less latency on mobile devices than the two strong baselines. In addition, data augmentation can reduce the requirement of paired data by more than 72 times. We recruit 35 volunteers for a user study in which VibVoice is preferred by 87% users compared to the baseline.

2 RELATED WORK

Speech enhancement is an essential function in voice communication that has received extensive attention. We categorize existing speech enhancement methods based on the input modalities, i.e.,

audio-only, *multi-modal*, as well as works sensing acoustic signals with other modalities.

2.1 Audio-only Speech Enhancement

2.1.1 Model-driven. Traditional speech enhancement either relies on assumptions of stationarity of signals, independence of speech, noise in the time-frequency domain [12], or multiple omnidirectional microphones (e.g., a microphone array) to improve audio quality. The microphone array leverages the difference of arrival times (known as beamforming) toward each microphone to determine the direction of the speaker for speech enhancement [17, 54], and speech separation [19, 53]. However, those approaches cannot adapt to dealing with dynamic noises without prior knowledge.

2.1.2 Machine Learning. Recent studies adopt DNN [9, 41, 56] for speech enhancement. The deep structure of neural networks can capture the inherited features of the target voice and potential noises based on a large training dataset. DNN-based methods work on both single-channel audio [41] and multi-channel audio with arbitrary microphones layouts [56]. ClearBuds [9] uses a DNN to process the stereo audio recorded by customized earbuds, which causes excessive communication overhead. ClearBuds filters noises based on the angles of sources, which does not work when a competing speaker stands at the same angle as the speaker. Although audio-only deep learning speech enhancement achieves good performance, they rely on the training data domain and fail to enhance the speech quality with slim-volume audio.

2.2 Multi-Modal Speech Enhancement

Speech events involve the motions of several articulatory organs, such as the tongue, teeth, lips, jaw, and facial muscles, which can be leveraged for speech enhancement.

2.2.1 Audio and Visual. Researchers in [13, 26] leverage the video from a camera in the environment to correlate the audio and visual for speech enhancement and separation. These approaches leverage deep learning and cross-modal embeddings to extract the target speech from noises. However, a visual sensor like a camera is not always available in daily usage.

2.2.2 Audio and Wireless. The authors in [42, 55] use the smartphone's speaker to transmit an inaudible acoustic signal (i.e., higher than 17 kHz) and simultaneously receive the echo reflected by moving lips, which can be used to enhance the noisy audio. [22, 29] use coupled mmWave and microphone recording for speech recognition rather than enhancing general speech contents. However, a mmWave radar is bulky and not common on HMWs, and the ultrasonic solution requires the user to fix the phone toward their mouth.

2.2.3 Audio and IMU. Some recent works have exploited the noise-less speech information from a bone conduction sensor or the accelerometer and the microphone recording for multi-modal speech enhancement. [45] proposes a multi-modal DNN for speech enhancement trained by a self-collected dataset in Mandarin. However, the bandwidth of the vibration sensor is around 5 kHz, which is much higher than the one available on commercial devices [37]. SEANet [43] uses a DNN to augment acceleration from audio with the same transformation among all users. However, the diverse bone shapes can lead to different transformations and thus affect the generalizability. In addition, both works [43, 45] focus on the design of deep learning models and fail to evaluate the systems with real-world noises and unseen users.

2.3 Sensing Acoustic Signals

IMU or piezoelectric sensor [25] can be installed on commercial devices to measure the contact sound, including unvoiced sound [18], breathing sounds [14], and eavesdropping smartphone/VR headset speaker [7, 38, 40, 46]. Although various contact sensing technologies are developed for acoustic sensing, they can classify keywords and events only rather than reconstructing the full-spectrum audio. Previous works exploit remote acoustic sensing by correlating with several non-acoustic sensors (e.g., geophones, accelerometers, and gyroscopes) [15], using the LiDAR on a vacuum robot to predict acoustic signals from the vibrations on the surrounding surfaces [36], or leveraging a mmWave Radar to achieve authentication [21]. However, they can only detect a given set of speech contents like digits and music. Measurements in [4] show that smartphone's motion sensor can only capture air-conducted acoustic vibrations remotely in limited use cases.

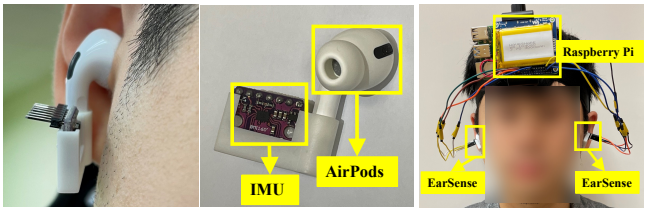
3 BACKGROUND

3.1 Audio Recording on HMWs

Although voice-related applications have been studied for decades, they usually exhibit unsatisfactory performance due to the following limitations. First, to save space and cost, most microphones installed on HMWs are omnidirectional instead of directional microphones due to their bulkiness and limitations in general use cases such as recording environmental sounds (e.g., Logitech 650e [24], around 100 USD). The omnidirectional microphones pick up audio from any angle, so they inevitably record the interference signals, especially the speech of a competing speaker close to the target user. Second, although the latest HMWs usually equip multiple microphones for beamforming, the compact layout of the microphone array can impede the acoustic beamforming since the audio from different sources arrives at the microphones with only a minor time difference. Specifically, the microphone array's interval is much smaller than $\lambda_{min}/2$, while the λ_{min} is the minimum wavelength of the interested frequency band. Third, the microphones on HMWs are farther away from the user's mouth than wired earphones or standalone microphones since they are usually inserted into the ears or mounted around the eyes. Therefore, the received intensity of the speech audio can be significantly suppressed due to the slim air transmission. Lastly, the worldwide pandemic forces people to wear facial masks, which can significantly reduce speech volume by up to 7dB [32]. Hence, seeking another available transmission channel for robust speech sensing is desirable.

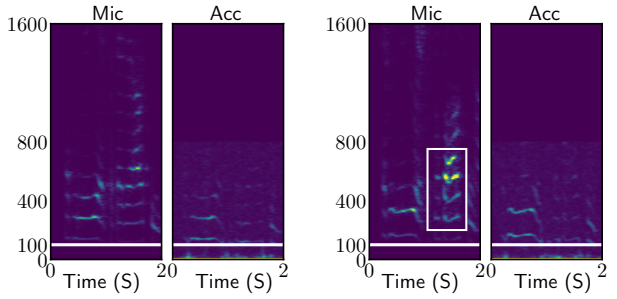
3.2 Bone-Conducted Vibrations

Existing commercial headphones leverage bone conduction [50] for inner ear hearing assistance. In this paper, we focus on the transmission from the vocal cord to the head skull, which can be detected by the contact microphone [51]. The noise-less feature of bone conduction sound has been discovered [43] with a bone



(a) *EarSense* with an Airpods Pro. (b) Test Setup.

Figure 2: *EarSense* is an open-sourced data collector attachable to commercial HMWs for vibration sensing.



(a) The user is talking with no noise. **(b) The user is talking with a competing speaker.**

Figure 3: Impact of the competing speaker.

conduction microphone. The professional bone-conducted microphone has also been applied in extremely harsh environments like battlefields or underwater applications [1, 2]. However, there are still challenges to leveraging it for speech enhancement on HMWs. Firstly, as depicted in [8, 52], the propagation through the head is complicated, resulting in an unstable response even for the exact location. Secondly, although a professional bone-conducted microphone can capture a clear voice at a high sample rate, they are too expensive (e.g., 60 US dollars [1, 2]) and not available on commercial HMWs. The current commercial-level vibration sensor may not provide sufficient sample rate and precision, leading to frequency aliasing and a bad signal-noise ratio. Some approaches tackle the problems by duplicating the signals of low-frequencies to high-frequency domain [11], utilizing the time stretching to generate high-frequency harmony [27], or also trying to unfold the acceleration signals from low sample rate using deep learning [46]. However, these approaches are far from achieving satisfying performance because the speech in high frequency contains extra information than that in low frequency. As a result, a simple recovery on the high-frequency part can amplify the noise of acceleration and degrade the quality of the generated audio.

4 METHODOLOGY

In this section, we first develop EarSense, an open-source wearable platform that has a similar setup to current commercial HMWs for data collection in Section 4.1. Then, we exploit the bone-conducted vibrations and the audio signals recorded by EarSense under various settings in Section 4.2. Finally, we overview the design of our proposed system in Section 4.3, introduce Bone Conduction Function and multi-modal network in Section 4.4 and Section 4.5, respectively.

4.1 EarSense: An Open Source Data Collection Platform for HMWs

Most commercial HMWs do not provide APIs to collect raw acceleration data. Recently, Apple provided APIs [6] to collect motion data from AirPods earphones at 100 Hz, which is too low for speech recording and doesn't fully utilize commercial IMU. Hence, it is desirable to develop a new sensing platform that can: a) collect the acceleration and acoustic data synchronously at a sample rate to recover speech information; b) have direct contact with the user's head like HMWs; c) have compatibility to test with existing commercial devices.

Fig. 2 shows the prototype of *EarSense* [23], a new open-sourced sensing platform that equips a 3D-printed enclosure with an IMU sensor (i.e., Bosch BMI-160 [37]) inserted. EarSense has a flexible design to make it attachable to any commercial HMW device (e.g., AirPods Pro). EarSense captures the same vibration as the testing HMW and has no direct contact with the user's head (shown in Fig. 2a). We connect two EarSenses to a Raspberry Pi (RPi) with a battery for data collection. Fig. 2b shows a volunteer wearing the device on the head to avoid excessive vibrations caused by the connecting wires. A Python script on the RPi to control the EarSense(s) through the I2C protocol and store the data for offline processing. We only use the three-axis accelerometer provided by the IMU chip because the gyroscope and magnetometer are less related to the vibration. The default sample rate of the microphone and the accelerometer are 16 kHz and 1.6 kHz, respectively. Since commercial earphones like AirPods Pro doesn't support two-channel audio recording, EarSense records audio in a mono channel. We apply L2-Norm to the spectrograms of three-axis acceleration to extract the vibration intensity and avoid the impact of wearing position and user's motion.

4.2 Measurements of Bone-Conducted Vibrations

We measure the difference between audio and bone-conducted vibrations with EarSense under various settings, i.e., the noises caused by the competing speaker and user motion. In each experiment, we ask a volunteer to wear Apple AirPods Pro with EarSense attached (shown in Fig. 2b) in a meeting room with the size of 10 m^2.

Competing speaker. First, we ask the user to sit in the meeting room and speak while a loudspeaker is playing a pre-recording speech one meter from the user as the competing speaker. Note that the competing speaker is one of the most challenging environmental noises for HMWs since its spectrum is similar to the speech's, making it indistinguishable from the noise suppression algorithms. We set the volume of the competing speaker to be similar to the user, in which SNR is 3 dB. It is difficult to separate the two sources according to volume since the two speeches are mixed. Fig. 3a and Fig. 3b show the spectrograms of microphone audio and acceleration when the environment is quiet or contains a competing speaker when the user is talking, respectively. Note that we only

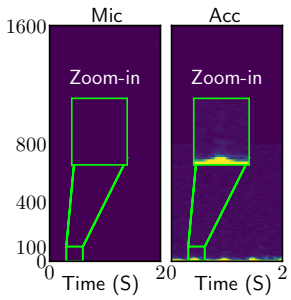


Figure 4: Spectrograms of microphone and acceleration when the user is walking.

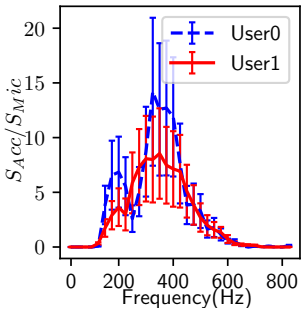


Figure 5: Frequency responses of two users' bone-conducted vibrations.

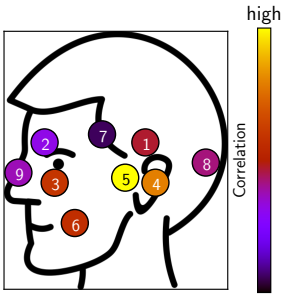


Figure 6: Intensities of received bone-conducted vibrations on ten locations of the head.

Figure 7: Potential placements on devices.

show the spectrogram up to 1.6 kHz since there is slim energy in the frequency bands higher than 1.6 kHz of the microphone, and the accelerometer can only sense the signal with a frequency band of 800 Hz. The spectrograms show that acceleration has lower signal energy than microphone audio since acoustic vibration attenuates during bone conduction, especially for high-frequency vibrations. Compared with the silent environment, the audio spectrogram of microphone audio with a competing speaker shows strong distortions, as highlighted with the white box in Fig. 3b. However, the accelerometer spectrograms do not show this phenomenon because the competing speaker cannot produce bone-conducted vibrations. In summary, the accelerometer receives the user's voice through bone conduction only, excluding environmental sounds over the air, which is ideal for speech enhancement.

User motion. We also measure the noises caused by the user's motion. We ask the volunteer to walk while talking. Fig. 4 shows that walking causes low-frequency noises in the acceleration. The green boxes show the zoom-in result of the signals below 100 Hz. We can observe clear periodic fluctuations in low frequency (i.e., < 50 Hz), which are caused by the steps of the volunteer. In summary, the user's motion only affects the acceleration at lower frequencies, which does not affect our speech enhancement. In addition, we observe slight low-frequency fluctuation in acceleration in Fig. 3, although the volunteer is asked to stand still in this experiment. Since most of the human speech is above 85 Hz [48], the removal of audio in low frequency (i.e., the white horizontal lines in Fig. 3) can effectively reduce the interference caused by human motion.

Frequency Response. We further explore the frequency response of bone-conducted vibrations among users. Specifically, we measure the frequency response of bone-conducted vibrations as follows: First, we compute the spectrums of audio and vibration data, which is denoted by S_{Mic} and S_{Acc}, respectively. Then, we compute the Bone Conduction Function denoted by the frequency response of bone-conducted vibrations by S_{Acc}/S_{Mic} at every frequency band. Fig. 5 shows two volunteers' frequency responses, in which the error bars represent the averages and variances. The acceleration shows higher sensitivity than the microphone within the frequency of 200Hz $\sim$ 500Hz, while lower sensitivity at a higher frequency ($>$ 500Hz) due to the absorption by the head skull. In addition, the frequency response keeps significant similarity among users, indicating the generalization to other users. Hence, the Bone Conduction Function should consider inter-people diversity and inter-people similarity.

Locations on the head. Bone-conducted vibration exists at various locations on the head. As shown in Fig. 6, we select ten unique locations of interest on the head and measure the intensities of received bone-conducted vibrations. In particular, the nine locations are #1 *upper ear*, #2 *eyebrow*, #3 *cheekbones*, #4 *ear*, #5 *temporomandibular joint*, #6 *cheek*, #7 *temple*, #8 *back of the head*, and #9 *nose*. In addition, we tape EarSense on the interior of the pad of an over-ear headphone, which is denoted as #10. Among those positions, some are compatible with commercial HMWs like glasses, headphones, and VR headsets. Fig. 7 illustrates the corresponding positions on HMWs. The numbers on the HMWs correspond to the locations in Fig. 6. We attach EarSense to the HMWs (when applicable) or on the face for all the following tests. We compute Pearson correlation coefficients between the audio and the acceleration on each location and mark the values using a color map in Fig. 6. We observe that bone-conducted vibrations exist at most locations of the head. In particular, locations closer to the mouth show higher correlations, indicating a larger possibility of extracting clean speech.

In summary, our measurements show that bone-conducted vibration has the following characteristics: a) Bone-conducted vibration is robust to environmental voice and only captures the user's speech; b) The user's motion only generates vibrations lower than 85 Hz; c) Bone-conducted vibration has suppressed low and high-frequency audio, which varies across users and within the same user; and d) We can receive audio from multiple locations on the head.

4.3 Overview of VibVoice

Motivated by our findings in Section 4.2, bone-conducted vibration is a promising complementary sensing modality to microphone recording for speech enhancement under environmental noises. However, the limited sample rate of HMW's accelerometer and diverse frequency responses of bone-conducted vibration introduce challenges for multi-modal speech enhancement. On the other hand, although it is possible to adapt existing audio-only DNN-based speech enhancement models [22, 26, 29, 42, 55] to take both

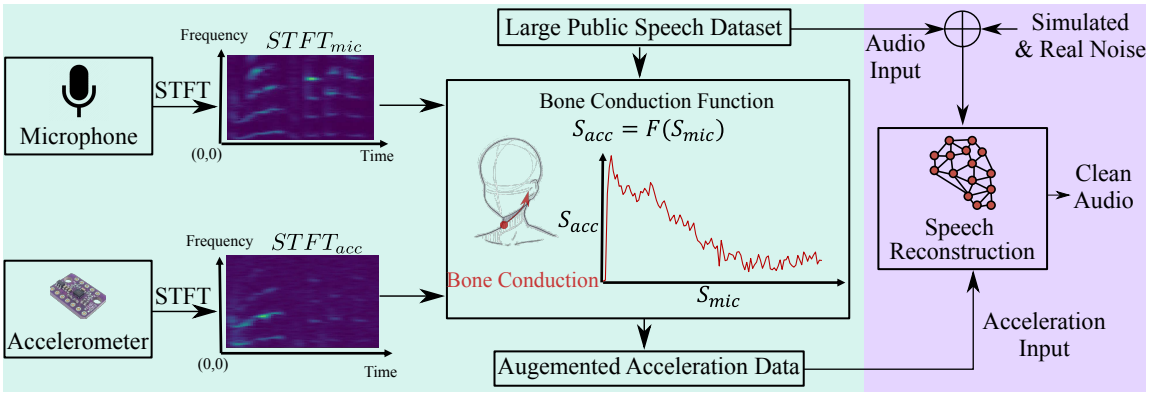


Figure 8: The overview of *VibVoice* system. We first augment the acceleration data using a large public speech dataset with estimated Bone Conduction Functions (e.g., the greed box). Then, we train a multi-modal DNN for speech reconstruction (e.g., the purple box).

inputs, there is no large dataset with paired acceleration and audio on HMWs available.

We design *VibVoice*, a speech enhancement system for HMWs that leverages both the microphone recording and the speech information from the bone-conducted vibrations. Fig. 8 overviews the design of VibVoice, which contains three components: estimation of Bone Conduction Function, data augmentation, and speech reconstruction network. First, we estimate a set of Bone Conduction Functions using a small paired acceleration-audio dataset collected from volunteers in Section 4.4.2. We note that this estimation does not resort to black-box deep learning approaches, which require a huge amount of training data. Instead, we take advantage of prior knowledge that a frequency response exists between the acceleration and audio spectrograms. Then, we augment the acceleration data using the public audio dataset *LibriSpeech* [30] with Bone Conduction Functions in Section 4.4.3. Finally, after generating a large amount of paired acceleration-audio dataset, we train a DNN model which can reconstruct the clean audio from microphone audio and the acceleration data in Section 4.5.

4.4 Bone Conduction Function

4.4.1 Function Formulation. Measurements in Section 4.2 show that the acceleration contains acoustic vibrations caused by bone conduction effects. There are two characteristics of acoustic vibrations: a limited sample rate up to 800 Hz and diverse amplitude suppression levels at frequency bands. We model the vibration sensed by the accelerometer as follows:

$$s_{acc} = f(s_{mic}) + \epsilon_{acc} = f(s_{speech} + \epsilon_{mic}) + \epsilon_{acc}, \quad (1)$$

where s_{acc} and s_{mic} are the raw data captured by the accelerometer and the microphone, respectively; s_{speech} denotes the ground-truth (clean) speech audio; ϵ_{acc} and ϵ_{mic} are environmental noises captured by the accelerometer and the microphone, respectively; and f is the Bone Conduction Function.

4.4.2 Function Estimation. To estimate the Bone Conduction Function, we recruit volunteers and record a five-minute speech for each person with EarSense and AirPods Pro. The volunteers are asked to read from a daily conversation material [44] in a silent environment. We split paired data into 5-second windows, and each window can contribute one Bone Conduction Function.

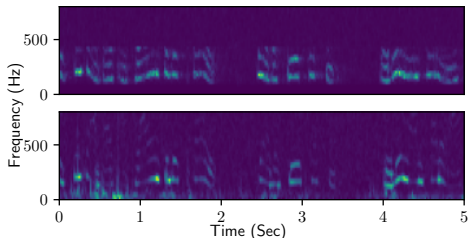


Figure 9: Augmented (upper) and real (lower) Acceleration.

First, we compute the spectrograms $STFT_{acc}$ and $STFT_{mic}$ of s_{acc} and s_{mic} for each window using short-time Fourier transform (STFT), respectively. Then, we apply Otsu's method [28] to $STFT_{acc}$ and $STFT_{mic}$ for automatic image thresholding, and then discard the bins whose value is lower than a threshold to remove outlier noises. We note that our thresholding setting can leverage temporal information, which can better diminish the noise than a constant threshold on the frequency or time domains. Furthermore, we perform the pixel-wise division between $STFT_{acc}$ and $STFT_{mic}$ to obtain a response spectrogram, selecting the lower frequency part of the audio spectrogram to align to the acceleration spectrogram. We model the Bone Conduction Function using the Gaussian distribution in the frequency domain since the frequency response (as shown in Fig. 5) has a non-trivial variance due to the complex structure of the head skeleton [8]. In particular, we compute $f = S_{acc}/S_{mic}$ for the corresponding time window in $STFT_{acc}$ and $STFT_{mic}$, respectively. The function $f \sim N(\mu, \sigma^2)$, in which μ and variance σ contribute to the contour and fluctuation of frequency response, respectively. We measure μ and σ for each time window to construct the parameter pool of Bone Conduction Functions.

4.4.3 Data Augmentation with Bone Conduction Functions. We develop a data augmentation approach with Bone Conduction Functions described in Section 4.4.2. Note that we cannot apply the

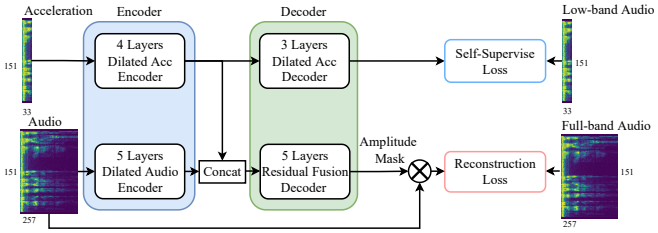


Figure 10: Overview of our proposed multi-modal network for speech enhancement.

	Encoder				Decoder		
Layer	1	2	3	4	1	2	3
Filters	16	32	64	128	64	32	16
Kernel	3	3	3	3	3	3	3
Scale	1	1	0.5	0.5	1	2	2

(a) Acceleration branch.

	Encoder					Decoder				
Layer	1	2	3	4	5	1	2	3	4	5
Filters	16	32	64	128	256	128	64	32	16	1
Kernel	3	5	5	5	5	5	5	5	5	3
Scale	0.5	0.5	0.5	0.5	0.5	2	2	2	2	2

(b) Audio branch.

Table 1: The design of DNN for multi-modal speech reconstruction.

inverse Bone Conduction Function to turn the acceleration back to audio due to the significantly limited sample rate of acceleration data. In addition, the frequency band (larger than 500 Hz) has very slim energy, which can cause unreasonable large energy after the inversion. We utilize these functions to generate acceleration signals using a large-scale audio dataset, i.e., LibriSpeech [30]. To be specific, for each audio clip, we first select a Bone Conduction Function (i.e., a list of means and variances over different frequencies) from the pool randomly. Then, we restore the frequency response from the Gaussian distribution of given parameters. Lastly, we can augment the audio to synthetic acceleration data by directly multiplying the frequency response.

Fig. 9 shows the spectrograms of augmented acceleration and real acceleration signals, respectively. The augmented spectrogram is close to the real acceleration spectrogram. We compute the similarity by calculating the mean of the absolute distance of all pixels in the whole spectrogram and divide it by the largest value of the real acceleration spectrogram. The average error for all volunteers is only 4.5%, which indicates that our proposed acceleration augmentation is reliable.

4.5 Multi-Modal Speech Reconstruction

Since the sample rate of the microphone is 16 kHz while the accelerometer is only 1.6 kHz. The reconstruction of high-quality audio from acceleration is an ill-posed task since it requires predicting the lost patterns of the high-frequency band. Moreover, as we have already generated a large-scale paired acceleration-audio dataset, we can unleash the strong feature extraction capability of DNNs for training. Compared with hand-crafted signal processing approaches, deep learning is robust to diverse inputs. Specifically, we adapt the multi-modal fusion paradigm [42, 55] and encoder-decoder architecture based on U-net [35] to build the deep learning model. Fig. 10 overviews of the DNN design, which is described in detail as follows.

4.5.1 ***Architecture.*** Table 1a and 1b show the hyperparameters of our multi-modal neural network of the acceleration and audio branches, respectively. The details are illustrated as follows.

Encoder. We use two convolutional networks (CNNs) to encode the high-level features from the two modalities, respectively. Then, we concatenate the features from two encoders along the channel dimension. We stack basic blocks to construct the encoder, where each block consists of a 2D convolutional layer, batch normalization, ReLU activation, and max-pooling. We also design a residual shortcut of the block input before the last deconvolution layer to expedite the training. The filters increase when the layer gets deeper to extract more diverse features. To increase the reception field to capture the harmonic pattern of the whole spectrogram, we use dilated convolution rather than the conventional one. Since the sample rate of audio data (i.e., 16 kHz) is ten times higher than that of the acceleration data (1.6 kHz), the size of the audio spectrogram on the frequency axis is also ten times larger. We note that the feature maps from the two modalities should have the same shape at the end of the encoders. Therefore, we discard the parts of the audio spectrogram whose frequency is higher than 6.4 kHz, making the remaining audio frequency band exactly eight times larger than the acceleration data (i.e., 0.8 kHz). In addition, the audio subnet (i.e., encoder) has three more down-sample operations than the acceleration branch to make the output sizes of the two modalities' features consistent.

Decoder. We design two decoders: a fusion decoder and an auxiliary decoder. The decoder has the same design of blocks as the encoder, but the stacked blocks have a decreasing number of filters but increasing sizes of the output tensor. The fusion decoder receives both modalities' features via a concatenation and outputs a spectrogram mask. To obtain the constructed clean spectrogram, this mask will be overlaid on the original noisy audio spectrograms by element-wise multiplication. The auxiliary decoder only receives the feature of the accelerometer and predicts the low-frequency part of clean audio. The self-supervised loss can prevent the acceleration encoder from being dominated by the audio input, given that the noisy input audio is similar to the clean audio to acceleration by nature (i.e., both are audio).

4.5.2 ***Waveform reconstruction.*** After obtaining the enhanced spectrogram from the multi-modal decoder, the last step is converting the spectrogram to the waveform using inverse short-time Fourier transform (ISTFT). The DNN of VibVoice only reconstructs the magnitude of the spectrogram. Since the phase of the speech

spectrogram is hardly predictable, we use the same phase of the noisy audio signal for the output. Our insight is that although the audio phase can be polluted in a noisy environment, our reconstructed magnitude can attenuate the unwanted time-frequency pixels in spectrograms, which restricts the interference of the noisy phase. Although several works [42, 55] try to estimate the clean phase by a subnet or a pre-defined algorithm, our tests show that the predicted phase is unstable, and their improvements are marginal and computationally expensive.

4.5.3 ***Model Training.*** We generate a pool of Bone Conduction Functions by performing cubic interpolation on existing functions from our dataset. We use Gaussian distribution to alleviate our model's overfitting and improve the generalizability among users. In addition, we augment the data with Gaussian distribution based on each frequency band to preserve the low-pass features. Thus, the limited Bone Conduction Functions can augment a large acceleration-audio dataset from an existing public dataset with diverse features. This dataset can extract the basic features of multiple modalities, but it is not enough for deployment because of the difference between the target user and our pool of Bone Conduction Functions. Therefore, we collected another acceleration-audio paired dataset from a different group of volunteers to train the model, which is also used to evaluate VibVoice's performance. Note that VibVoice does not require any data from the target user either during the Bond Conduction Function estimation or model training. All evaluations adopt leave-one-out validation. We pick one user as the testing dataset, while the others are regarded as the training dataset.

Although the training of VibVoice does not require any data from the target user, one-shot data from the target domain can further improve the overall performance. This data collection can share the recorded data during the setup of HMWs, e.g., the personalized model for calling voice assistants, which incurs no extra overhead to the user. In addition, VibVoice can record data when the user is speaking in a quiet environment. As the data collection doesn't require specific content, our system can unobtrusively record user-specific data during the long-time usage of HMWs and improve the performance of the target user continuously.

The fusion decoder uses the STFT Loss [10]. We use y and $\hat{y}$ to represent the clean and enhanced signals. The STFT loss can be denoted as follows:

$$\begin{aligned} L_{stft}(y,\hat{y}) &= L_{sc}(y,\hat{y}) + L_{mag}(y,\hat{y}), \\ L_{sc}(y,\hat{y}) &= \frac{||STFT(y)| - |STFT(\hat{y})||_F}{||STFT(y)||_F}, \\ L_{mag}(y,\hat{y}) &= \frac{1}{T}|log|STFT(y)| - log|STFT(\hat{y})||_1, \end{aligned} \tag{2}$$

where L_{sc} refers to convergence (sc) loss and L_{mag} refers to the magnitude (mag) loss. We use Mean Square Error (MSE) as the training loss for the auxiliary decoder. The fusion decoder and auxiliary decoder targets are full-band clean spectrogram and lower-band clean spectrogram, respectively. The final loss is formulated as follows:

$$L(y,\hat{y}) = L_{stft}(y,\hat{y}) + |y_{lowband} - \hat{y}_{lowband}|_2 \times \lambda, \tag{3}$$

which is the joint loss function that covers both fusion and auxiliary decoder. We set λ to 0.05 to balance the scales of two losses.

5 EVALUATION

5.1 Experiment Setup

We use EarSense introduced in Section 4.1 to collect audio and acceleration signals simultaneously[1]. The volunteers are asked to wear EarSense and speak in a meeting room with the size of 10 m^2. The reading material is selected from daily English conversations [44]. Each volunteer reads the material for 30 seconds and repeats it 20 times, generating ten minutes of data. For data augmentation, we recruit eight volunteers to generate a pool of Bone Conduction Functions and use it to augment 100 hours of data from LibriSpeech [30]. In addition, we recruit another 15 volunteers for training and testing. For each experiment, we adopt the leave-one-out validation, i.e., training the model for each user with the dataset except that user. The volume of our collected speech is between 60dB to 70dB, which is the same as regular conversations [3]. To test VibVoice's robustness to diverse noises, we use three categories of noises with balanced possibility, including environmental noises (50 classes) [31], competing speakers from another subset of LibriSpeech [30], and 20 songs with languages of English, Mandarin, and Japanese. We apply a random room impulse response from dataset [20], containing point-source noises, real isotropic noises, and simulated the noises of 600 rooms. We implement VibVoice using PyTorch and train it on a PC with an Intel i9-12900K CPU and two Nvidia RTX 3090 GPUs. The model is trained with a step learning rate scheduler with a learning rate of 0.001 and an Adam optimizer. The epoch number is 30. The length of the STFT window and the overlap for the audio are 640 and 320, while 64 and 32 are for the acceleration. We open-source the implementation and data in [23].

5.2 Baseline

We deploy two baselines, i.e., FullSubNet (FSN) [16] and SEANet (SN) [43]. FSN and SN are two state-of-the-art speech enhancement approaches using audio-only and audio-acceleration inputs, respectively. We train SN using our dataset since the one used in its paper is not available to the public. Specifically, we train SN using acceleration data with a sample rate of 1.6kHz to ensure it is the same as VibVoice. Note that SN indicates that it supports acceleration data with a wide range of sample rates.

5.3 Evaluation Metric

We use three evaluation metrics, i.e., Perceptual Evaluation of Speech Quality (PESQ), Signal Noise Ratio (SNR), and Log-Spectral Distance (LSD), which are described as follows:

Perceptual Evaluation of Speech Quality (PESQ) is a popular test standard for automatically assessing the user experience of speech quality, defined in P.862 standard by International Telecommunication Union [34]. The results generated by PESQ represent the opinion scores that range from 1 (bad) to 5 (excellent). We use the wide-band version to evaluate the full-band speech quality in the evaluations.

Signal Noise Ratio (SNR) is a metric widely used in signal processing that can be measured as follows: $SNR(x,y) = 10 * log_{10}(\frac{y}{x-y})^2$, where x and y denote the estimated and clean audio, respectively.

[1]The experiments that involve human subjects have been approved by the IRB of the authors' institution (CUHK-SBRE-21-0570).

SNR compares the desired signal and the deviation. A higher SNR value means better audio quality. We use the scale-invariant SNR in the implementation to reduce the impact of scale.

Log-Spectral Distance (LSD) measures the quality of frequencies between the reconstructed audio and the ground truth audio with: $LSD(x,y) = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{K}\sum_{k=1}^{K}(X(l,k)-\hat{X}(l,k))^2}$, where l and k denote the time and frequency index of the spectrogram, respectively; $X = log(|STFT(y)|^2)$, and $\hat{X} = log(|STFT(x)|^2)$. A higher LSD value means lower audio quality.

5.4 Overall Performance

We investigate VibVoice's performance by mixing clean speech with noise audio randomly picked from the noise dataset. We set the SNR of the noisy data ranges from 0dB to 20dB, with an average of 10dB, which covers a wide spectrum of noise levels.

Target user calibration. VibVoice can work out of the box, while data from the target user can further improve performance. Fig. 11 shows the performance of VibVoice and the two baselines with different amounts of data from the target user, in which zero means no target-user data is used during the training. The results show that longer target-user data can improve performance for all approaches, and VibVoice performs better than baselines with the same amount of calibration data. The calibration data can be collected continuously during usage without the user's inputs/operations (c.f., Section 4.5.3). VibVoice achieves the best perceptive performance according to PESQ and the best-reconstructed spectrogram according to LSD. VibVoice and FSN have similar SNR results since they output similar signals in the time domain compared to the clean one.

Noise type. We evaluate the impact of different types of noises in Fig. 12, i.e., environmental noises, competing speakers, and music. The result shows that VibVoice performs better under all noises and metrics except for the SNR with music noise. This is because the complex and dynamic spectrum of the user's speech and music's vocals introduce minor fluctuations in high frequencies, reflecting large fluctuations in SNR.

Noise level. We test VibVoice's performance under noise levels of low (10 dB), medium (5 dB), and high (0 dB with only speech noise). The results in Fig. 13 show that VibVoice has better performance improvements, especially when the noise is challenging, i.e., 21% improvement on PESQ and 26% improvement on SNR. This is because the acceleration can more robustly identify the target speech, whereas the audio-only solution is difficult to differentiate sound with a similar pattern (e.g., strong speech noise).

Noise source. We evaluate the impact of different numbers of noise sources by repeatedly mixing clean audio with random audio clips. Fig. 14 shows that VibVoice's performance degrades as the number of noise sources increases but still outperforms all the baselines.

Temporal stability. We further examine how VibVoice performs for the same user over time. Note that the offset of sensor placement and minor changes in speech can cause a slight change. We collect ten-minute data from three volunteers twice, six months apart. The results show that the performance of VibVoice has negligible changes, from 2.6 to 2.5 for PESQ, 15.7 to 15.5 for SNR, and 4.3 to 4.6 for LSD. Besides, VibVoice outperforms FSN, whose performance is 2.1 for PESQ, 15.2 for SNR, and 11 for LSD. The results affirm that VibVoice is robust to temporal changes.

Airway blockage. The blockage of air transmission caused by personal protective equipment like facial masks can significantly suppress the user's speech volume. We ask three volunteers to test VibVoice by speaking the same content with and without facial masks. The result shows that VibVoice's performance only degrades by 0.05 (< 2%) for PESQ, 0.4 (< 3%) for SNR, and 0.1 (< 3%) for LSD when the subject wears the mask. VibVoice gets a more significant margin than FSN, whose performance is 2.15 for PESQ, 13.8 for SNR, and 10 for LSD. This aligns with our expectation since VibVoice uses bone-conducted vibration that is not affected by air transmission.

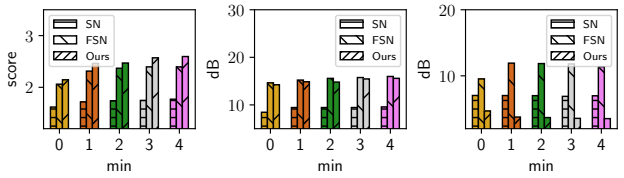
Variances among users. Speech and bone-conducted vibration can differ across users due to vocal features, head skull shapes, body fat, etc. Fig. 16 shows the performance of VibVoice across 15 users. VibVoice shows stable and significantly better performance in PESQ compared to baseline and comparable performance in SNR.

Sensing location. We test VibVoice when EarSense is placed in ten locations on the head as defined in Fig. 6, validating VibVoice's effectiveness for different HMW devices. The bars in Fig. 15 show VibVoice's performance at each location. The red line represents the performance of the baseline at #4 *ear*. The results show that VibVoice achieves satisfactory performance at all locations in PESQ. Note that locations like #1 *upper ear*, #2 *eyebrow*, #5 *temporomandibular joint*, #7 *temple*, and #10 *interior of the pad of the headphone* show similar or slightly lower SNR than the baseline, which is because the vibration intensity is slim due to their far distance to the audio source.

Data augmentation effectiveness. We evaluate how VibVoice's data augmentation reduces the amount of paired training data needed. We compare the performance of training VibVoice from a) paired data of 18 to 180 hours and b) data augmented from three hours of paired data. The dataset [45] is a large-scale Chinese acceleration and audio dataset collected by earphones, and the bandwidth of acceleration is around 5kHz. The results in Fig. 17 show that VibVoice with data augmentation can achieve better performance with ~ 24× less paired data.

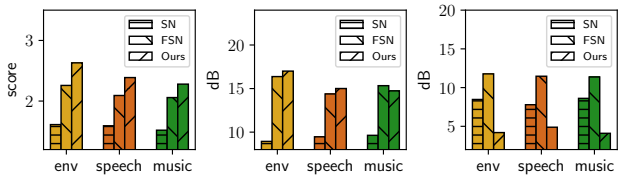
Summary. VibVoice outperforms FSN and SN by up to 21% for PESQ when the noise volume is low, where the PESQ for VibVoice and FSN are 2.7 and 2.21, respectively. VibVoice outperforms the baselines up to 26% for SNR when the noise is speech with high volume, where the SNR of VibVoice and FSN are 2.0 and 1.6, respectively. In addition, VibVoice outperforms the baselines 50~80% in LSD under most impact factors, indicating the efficiency of our multi-modal design and novel data augmentation. VibVoice has slightly lower performance in SNR for some cases as it can be biased due to the similar spectrum of the user's speech and music's vocal. The dynamic also introduces minor fluctuations. In comparison, LSD evaluates the whole band without preference, so VibVoice outperforms the two baselines by a large margin. We further evaluate the perception of real users through a user study in Section 6.

Compared with the strong baseline FSN, VibVoice achieves good performance in various cases. Compared with multi-modal baseline SN, whose network design has limited capability to recover speech information as VibVoice outperforms it by a large margin. Besides, as discussed in Section 2.2.3, the data augmentation of SN does not consider the user-specific bone conduction effects, while



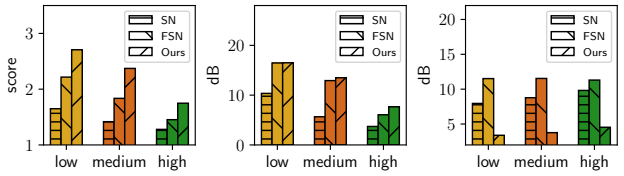
(a) PESQ. (b) SNR. (c) LSD.

Figure 11: Imapct of calibration time.



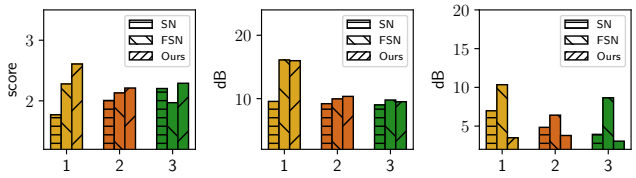
(a) PESQ. (b) SNR. (c) LSD.

Figure 12: Impact of noise types.



(a) PESQ. (b) SNR. (c) LSD.

Figure 13: Impact of noise levels.



(a) PESQ. (b) SNR. (c) LSD.

Figure 14: Number of noise sources.

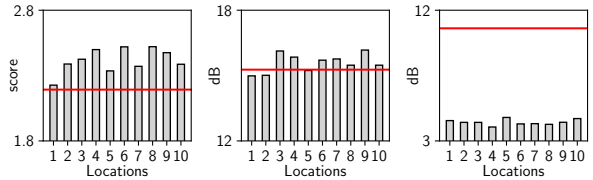
VibVoice combines the knowledge from extensive measurements and individual features from the target user.

5.5 Ablation Study

We conduct an ablation study to understand the performance of different design components in VibVoice. The performance of VibVoice without different components is listed in Table 2.

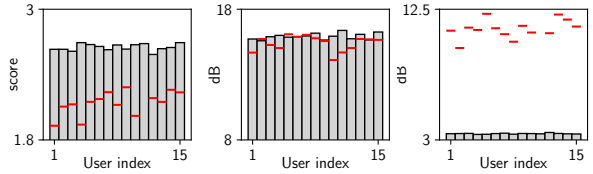
No auxiliary decoder. First, we remove the self-supervise loss, meaning the audio may dominate the model. The results indicate that the variant slightly degrades by 0.1 in PESQ, 0.3 in SNR, and 0.8 in LSD, respectively.

No data augmentation. Second, we remove the data augmentation based on Bone Conduction Function. The performance significantly degrades to 1.9 in PESQ, 14 in SNR, and 5 in LSD. According to



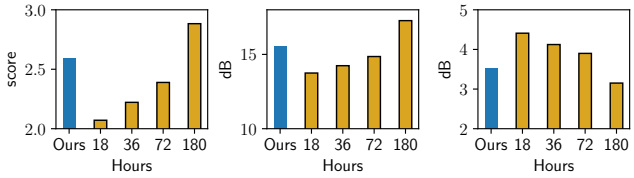
(a) PESQ. (b) SNR. (c) LSD.

Figure 15: VibVoice on different head locations. Red line: the performance of the baseline, i.e., FullSubNet.



(a) PESQ. (b) SNR. (c) LSD.

Figure 16: VibVoice on different users. Red line: the performance of the baseline, i.e., FullSubNet.



(a) PESQ. (b) SNR. (c) LSD.

Figure 17: Effectiveness of data augmentation. Blue bars: VibVoice using less than three-hour paired data with augmentation. Yellow bars: VibVoice using 18- to 180-hour paired data without augmentation.

	PESQ	SNR	LSD
VibVoice	2.6	15.6	3.5
w/o auxiliary decoder	2.5	15.1	4.4
w/o augmentation	1.9	14	5
w/o Gaussian approx	2.4	15.2	4.2
Accelerometer sample rate: 1200 Hz	2.47	14.4	4.2
Accelerometer sample rate: 800 Hz	2.45	14.3	4.5
Accelerometer sample rate: 400 Hz	2.4	14.2	4.4

Table 2: Ablation study.

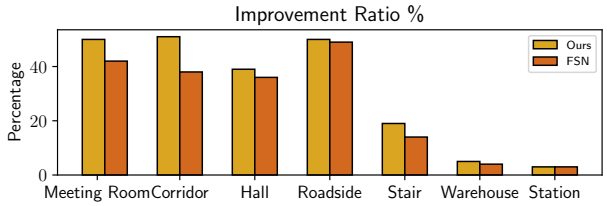


Figure 18: Different environments.

	VibVoice	FSN	SN
Desktop CPU	0.05	0.27	0.5
Desktop GPU	0.016	0.034	0.07
P30	0.16	5	1.9
Mate20	0.29	4.6	1.7
Pixel7	0.31	5.2	1.4

Table 3: Runtime analysis (second/instance).

the definition of ITU and mean opinion score (MOS), the audio quality is poor when the score drops from 2.5 to 1.9 [47]. This result aligns with our motivation that a small-scale self-collected dataset is insufficient to train a strong neural network.

No Gaussian approximation. Third, the Bone Conduction Function is modeled by only the mean, while the variance is zero. The performance drops 0.2 in PESQ, 0.4 in SNR, and 0.5 for LSD, indicating that our Gaussian approximation is close to the nature of the Bone Conduction Function.

Lower sample rate. The frequency response in Fig. 5 shows that the speech information from above 600 Hz is very limited. Hence, we evaluate the performance of VibVoice with a lower sample rate by downsampling the acceleration data to 1200 Hz, 800 Hz, and 400 Hz. The results show that VibVoice is robust to various sample rates, which consume less power in processing and communication. When the sample rate is 800 Hz, the output's PESQ only degrades 10%.

5.6 On-Site Evaluation

In the following, we conduct extensive experiments to validate the performance of VibVoice with ongoing noise in the wild. The volunteers are asked to read the same content as the experiments in Section. 5.4. Apart from the environmental noises, we use a smartphone (i.e., Huawei P30) to play recorded speech and ask a volunteer to speak one meter away as the interference. The speech content of the competing speaker is irrelevant to the target speech. We collect 10 minutes of data with about 3dB SNR at each test location. Unlike synthetic noisy speech, we can neither capture the ground truth of clean speech and evaluate the metrics like SNR and PESQ nor train the model. Instead, we use the result of Automatic Speech Recognition [33] to evaluate the quality of the speech. We use Word Error Rate (WER = $\frac{S+D+I}{N}$) as the evaluation metric, where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the reference. The higher the value of WER, the lower the audio quality. To better evaluate the performance of VibVoice, we further define the relative improvement of WER in percentage as $\frac{(WER_o - WER_v)}{WER_o} \times 100\%$, where WER_o refers to the WER of the original audio, WER_v refers to the WER of VibVoice.

5.6.1 ***Environments.*** We evaluate VibVoice in diverse indoor and outdoor environments, including the meeting room, corridor, stairs, lumber room, and railway station. We note that the experiment conducted in the real world naturally contains diverse environmental noises (e.g., train, car, construction, wind) and competing speakers. The results in Fig. 18 show that VibVoice effectively improves speech quality in most environments. Most existing works have good performance in statistical noises like machinery noises, however, they perform badly when there are competing speakers in the same room, which is a common indoor scenario of voice communication. Although speech enhancement is challenging in small spaces like a meeting room and corridor due to the strong multi-path effect, VibVoice can still improve the speech quality and achieve an improvement ratio of over 50%, and 48%, respectively. VibVoice also achieves a 37% improvement for the hall, 29% for the outdoor square, and 19% in the scene of the outdoor stair when there is intermittent construction noise. We note that in both the warehouse and station, the improvement of VibVoice is marginal. This is mainly because that AirPods already perform well in these scenarios. In comparison, VibVoice keeps outperforming FSN with much less latency.

5.6.2 ***Earphones.*** We deploy VibVoice on three popular TWS earphones, i.e., Apple AirPods Pro, Huawei FreeBuds Pro, and Samsung GalaxyBuds Pro. During the experiment, we turn on all available acoustic signal processing algorithms provided by the manufacturers, including speech enhancement and noise suppression. The results show that the WER with a competing speaker is 60% for AirPods, 66% for FreeBuds, and 160% for GalaxyBuds, respectively. Due to the different designs of the earphones, we do not implement VibVoice on other earphones. Since VibVoice can effectively reduce the WER of AirPods Pro by 50%, we envision that VibVoice can be extended to other earphones.

5.6.3 ***User movements.*** We further evaluate the performance of VibVoice when the user is moving. We ask volunteers to move in the room with a speed of around 1 m/s. The result shows the improvement for a still volunteer is 58.3 %, while the same moving volunteer is 57.1%. The standard deviation for the still and moving case is 29.9 % and 30.4%, respectively. Although the improvement ratio slightly drops, VibVoice still preserves its performance under regular movements.

5.7 Runtime Evaluation

We test the execution latency of VibVoice and the two baselines (i.e., FSN [16] and SN [43]) on a desktop PC (i.e., i7-11700k CPU and RTX 3060 GPU) and three smartphones (i.e., Huawei P30, Huawei Mate20, and Google Pixel 7). We run the inference of 5 seconds clip 100 times and record the mean latency. The results in Table 3 show that VibVoice reduces up to 31× and 12× less latency on average than FSN and SN, respectively. On the other hand, the results show that VibVoice exhibits a greater advantage in runtime

for low-end devices. In conclusion, VibVoice can support real-time voice applications with a delay of fewer than 0.6 seconds (i.e., two times the inference latency) for processing 5 seconds audio clip. The real-time factor is only 0.12, significantly less than the minimal requirement, which means less energy consumption and space to process other tasks.

6 USER STUDY

We recruit 35 volunteers to study the real perceived experience of VibVoice.

Questionnaire Design. We play a few recordings from the dataset described in Section 5.1 to volunteers. The length of each sentence is clipped or padded to 5 seconds. The first part is to evaluate the intelligibility of reconstructed speech. The participants are asked to listen to the audio clips enhanced by VibVoice and write down the sentences. We use WER to evaluate the results. In the second part of the study, participants listen to two audio clips with the same content and choose the audio with better quality in their perception. This study evaluates whether the reconstructed speech improves the user experience or not. We conduct two kinds of comparisons five times. One type is to ask participants to choose between audio enhanced by VibVoice and the original noisy audio. The second kind is to ask participants to choose between audio enhanced by VibVoice and the baseline (i.e., FSN). We use the correct ratio $\frac{P}{P+N}$ to represent the improvement of VibVoice, in which P and N are the numbers of choosing VibVoice and negative versa vice.

6.1 Study Result

According to the results of the first part, VibVoice achieves an overall WER of 21.5%, which is acceptable for understanding the audio content and confirm the effectiveness of VibVoice. According to the answers to the second question, the survey results show that 87% of the participants choose VibVoice over the baseline, and 72% of them choose VibVoice over the original audio without any enhancements. In addition, we discuss with participants why they prefer the original audio sounds over the baseline. The baseline can produce acoustic artifacts and sometimes wrongly suppress the sound of the target speaker. We note that some participants observed that the impact of artifacts and suppression is hindered after knowing the content or listening repeatedly. However, the speech generated by the baseline causes lots of misunderstanding for the first-time listener. In conclusion, the user study results show that VibVoice can enhance speech quality and improve user experience compared with the original audio and the baseline.

7 DISCUSSION

System overhead. VibVoice can transmit the raw acceleration data to the mobile device at a bit rate of 153.6 kbps (6*16 bits*1.6 kHz). The Bluetooth profiles like Hands-Free Profile (HFP) support a 16 kbps (16 kHz) data rate, which is adequate to transmit the acceleration data under Bluetooth 5.0 [49]. Besides, the data of IMU is also applicable to the compression codec, which can further reduce communication overhead.

Most head-mounted wearables run on the battery. The extra energy consumed by VibVoice is mainly ADC, which is 0.54 mW (i.e., $180\mu A * 3V$) [37]. In contrast, each AirPods Pro earbud has a 43 mAh battery, which can support 3.5 hours of talking time. Hence, VibVoice only has an extra 1.5% power consumption for earphones. Note that many earphones like AirPods Pro already have motion-related applications running (e.g., spatial audio [5]), in which VibVoice can share the data collection process.

Bone conduction function. VibVoice shows that Bone Conduction Function can be estimated using the Gaussian distribution with a fine-tuning process. We note that some finite-element models can be used for bone-conduction sound simulation, which further improves the performance of VibVoice [8].

Language. The data augmentation approach is based on the bone conduction effect of the user's skull, which contains no language-specific features, which means VibVoice can be generalized to any language.

Sensing location. Based on our experience of extensive evaluations, we summarise two guidelines about the optimal placement of the IMU sensors on the user's head for the HMW manufacturers: 1) closer to the vibrating organ, and 2) a tight contact with the head. Meanwhile, it is also important to consider human comfort and compatibility with current devices.

Customized device. We envision combining VibVoice with raw data before hardware processing. We use the same APIs to record audio from the commercial TWS microphone. The recorded audio is processed by the acoustic signal processing provided by the system, requiring the least system privileges and no extra hardware modification. However, as VibVoice is built upon the black-box output, the result can be biased sometimes due to the performance of hardware processing. For example, AirPods can wrongly suppress the sound and produce unnatural noise, which doesn't appear in our training dataset and is hardly predictable. VibVoice can perform better with access to raw data of the TWS/HMW's microphone array (i.e., used for active noise cancellation or acoustic beamforming).

8 CONCLUSION

In this paper, we leverage the bone-conducted vibration to enhance the voice recording quality on head-mounted wearables. We propose VibVoice, an end-to-end multi-modal speech enhancement approach that reconstructs clean speech audio by fusing the acceleration and audio from head-mounted wearables. Meanwhile, we extract the Bone Conduction Function to augment acceleration from a large public audio dataset. We collected a paired acceleration-audio dataset to evaluate VibVoice at various locations. Our system outperforms the state-of-art audio-based speech enhancement model up to 21% in PESQ and 26% in SNR with 30 times less latency on the mobile device. In the online user study of 35 participants, VibVoice is preferred by 87% of the participants. These results reflect the potential of bone-conducted vibration sensing on head-mounted wearables, as well as the effectiveness of VibVoice in speech enhancements.

9 ACKNOWLEDGEMENT

This work is supported by the Faculty of Engineering, The Chinese University of Hong Kong, under Direct Grant (No. 4055167) and Research Grants Council (RGC) under General Research Fund (No. 14214022).

REFERENCES

[1] 2022. Bone-Conduction Ear Microphone | Shop Motorola Solutions. https://shop.motorolasolutions.com/bone-conduction-ear-microphone-system/product/PMLN5464A. (Accessed on 08/19/2022).

[2] 2022. Ear Bone Microphone - EarHugger®. https://earhugger.com/product/ear-bone-microphone/. (Accessed on 08/19/2022).

[3] 2022. What Noises Cause Hearing Loss? | NCEH | CDC. https://www.cdc.gov/nceh/hearing_loss/what_noises_cause_hearing_loss.html. (Accessed on 03/16/2023).

[4] S Abhishek Anand and Nitesh Saxena. 2018. Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1000–1017.

[5] Apple. 2022. Listen with spatial audio for AirPods and Beats - Apple Support. https://support.apple.com/en-us/HT211775. (Accessed on 12/09/2022).

[6] Apple. 2023. CMHeadphoneMotionManager | Apple Developer Documentation. https://developer.apple.com/documentation/coremotion/cmheadphonemotionmanager. (Accessed on 05/16/2023).

[7] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kui Ren. 2020. Learning-based Practical Smartphone Eavesdropping with Built-in Accelerometer.. In *NDSS*.

[8] You Chang, Namkeun Kim, and Stefan Stenfelt. 2016. The development of a whole-head human finite-element model for simulation of the transmission of bone-conducted sound. *The Journal of the Acoustical Society of America* 140, 3 (2016), 1635–1651.

[9] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 384–396.

[10] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847* (2020).

[11] Martin Dietz, Lars Liljeryd, Kristofer Kjorling, and Oliver Kunz. 2002. Spectral Band Replication, a novel approach in audio coding. In *Audio Engineering Society Convention 112*. Audio Engineering Society.

[12] Yariv Ephraim and Harry L Van Trees. 1995. A signal subspace approach for speech enhancement. *IEEE Transactions on speech and audio processing* 3, 4 (1995), 251–266.

[13] Ruohan Gao and Kristen Grauman. 2021. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 15490–15500.

[14] Pranav Gupta, Yaesuk Jeong, Jaehoo Choi, Mark Faingold, and F Ayazi. 2018. Precision high-bandwidth out-of-plane accelerometer as contact microphone for body-worn auscultation devices. In *2018 Hilton Head Workshop*. 30–33.

[15] Jun Han, Albert Jin Chung, and Patrick Tague. 2017. Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 181–192.

[16] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li. 2021. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6633–6637.

[17] Youna Ji, Jun Byun, and Young-cheol Park. 2017. Coherence-Based Dual-Channel Noise Reduction Algorithm in a Complex Noisy Environment.. In *INTERSPEECH*. 2670–2674.

[18] Prerna Khanna, Tanmay Srivastava, Shijia Pan, Shubham Jain, and Phuc Nguyen. 2021. JawSense: recognizing unvoiced sound using a low-cost ear-worn system. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 44–49.

[19] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. 2016. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 9 (2016), 1626–1641.

[20] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5220–5224.

[21] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. 2020. VocalPrint: exploring a resilient and secure voice authentication via mmWave biometric interrogation. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 312–325.

[22] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A Noise-resistant Multi-modal Speech Recognition System Fusing mmWave and Audio Signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 97–110.

[23] He Lixing, Hou Haozheng, Shi Shuyao, Shuai Xian, and Yan Zhenyu. 2023. Code of VibVoice. https://github.com/CUHK-AIoT-Sensing/vibvoice.

[24] Logitech. 2022. Logitech H650e Business Headset with Noise Cancelling Mic. https://www.logitech.com/en-hk/products/headsets/h650e-business-noise-cancelling.html. (Accessed on 04/13/2022).

[25] Héctor A Cordourier Maruri, Paulo Lopez-Meyer, Jonathan Huang, Willem Marco Beltman, Lama Nachman, and Hong Lu. 2018. V-speech: Noise-robust speech capturing glasses using vibration sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–23.

[26] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021).

[27] Frederik Nagel and Sascha Disch. 2009. A harmonic bandwidth extension method for audio codecs. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 145–148.

[28] Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* 9, 1 (1979), 62–66.

[29] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, and KJ Liu. 2021. RadioMic: Sound Sensing via mmWave Signals. *arXiv preprint arXiv:2108.03164* (2021).

[30] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.

[31] Karol J. Piczak. 2022. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia* (Brisbane, Australia, 2015-10-13). ACM Press, 1015–1018. https://doi.org/10.1145/2733373.2806390

[32] Christoph Pörschmann, Tim Lübeck, and Johannes M Arend. 2020. Impact of face masks on voice radiation. *The Journal of the Acoustical Society of America* 148, 6 (2020), 3663–3670.

[33] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624* (2021).

[34] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, Vol. 2. IEEE, 749–752.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[36] Sriram Sami, Yimin Dai, Sean Rui Xiang Tan, Nirupam Roy, and Jun Han. 2020. Spying with your robot vacuum cleaner: eavesdropping via lidar sensors. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 354–367.

[37] Bosch Sensortec. 2021. Inertial Measurement Unit BMI160. https://www.bosch-sensortec.com/products/motion-sensors/imus/bmi160/.

[38] Cong Shi, Xiangyu Xu, Tianfang Zhang, Payton Walker, Yi Wu, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. 2021. Face-Mic: inferring live speech and speaker identity via subtle facial dynamics captured by AR/VR motion sensors. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 478–490.

[39] Statista. 2020. Wearable shipments by category 2024. https://www.statista.com/statistics/690731/wearables-worldwide-shipments-by-product-category/. (Accessed on 02/15/2022).

[40] Weigao Su, Daibo Liu, Taiyuan Zhang, and Hongbo Jiang. 2021. Towards Device Independent Eavesdropping on Telephone Conversations with Built-in Accelerometer. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–29.

[41] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. 2021. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 21–25.

[42] Ke Sun and Xinyu Zhang. 2021. UltraSE: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 160–173.

[43] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. 2020. SEANet: A multi-modal speech enhancement network. *arXiv preprint arXiv:2009.02095* (2020).

[44] Department of the State USA. 2022. Everyday Conversations: Learning American English | American English. https://americanenglish.state.gov/resources/everyday-conversations-learning-american-english. (Accessed on 02/02/2022).

[45] Heming Wang, Xueliang Zhang, and DeLiang Wang. 2022. Fusing Bone-Conduction and Air-Conduction Sensors for Complex-Domain Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 3134–3143.

[46] Tianshi Wang, Shuochao Yao, Shengzhong Liu, Jinyang Li, Dongxin Liu, Huajie Shao, Ruijie Wang, and Tarek Abdelzaher. 2021. Audio Keyword Reconstruction from On-Device Motion Sensor Signals via Neural Frequency Unfolding. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–29.

[47] Wikipedia. 2020. Perceptual Evaluation of Speech Quality - Wikipedia. https://en.wikipedia.org/wiki/Perceptual_Evaluation_of_Speech_Quality. (Accessed on 12/02/2022).
[48] Wikipedia. 2020. Voice frequency - Wikipedia. https://en.wikipedia.org/wiki/Voice_frequency. (Accessed on 07/12/2022).
[49] WikiPedia. 2022. Bluetooth - Wikipedia. https://en.wikipedia.org/wiki/Bluetooth. (Accessed on 07/15/2022).
[50] Wikipedia. 2022. Bone conduction - Wikipedia. https://en.wikipedia.org/wiki/Bone_conduction. (Accessed on 07/26/2022).
[51] Wikipedia. 2022. Contact microphone - Wikipedia. https://en.wikipedia.org/wiki/Contact_microphone. (Accessed on 07/26/2022).
[52] Sook Young Won and Jonathan Berger. 2005. Estimating transfer function from air to bone conduction using singing voice. In *ICMC*.
[53] Sean UN Wood, Jean Rouat, Stéphane Dupont, and Gueorgui Pironkov. 2017. Blind speech separation and enhancement with GCC-NMF. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 4 (2017), 745–755.
[54] Nima Yousefian, John HL Hansen, and Philipos C Loizou. 2014. A hybrid coherence model for noise reduction in reverberant environments. *IEEE Signal Processing Letters* 22, 3 (2014), 279–282.
[55] Qian Zhang, Dong Wang, Run Zhao, Yinggang Yu, and Junjie Shen. 2021. Sensing to hear: Speech enhancement for mobile devices using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–30.
[56] Siyuan Zhang and Xiaofei Li. 2021. Microphone Array Generalization for Multichannel Narrowband Deep Speech Enhancement. *arXiv preprint arXiv:2107.12601* (2021).