

## Article

# A Piezoelectric Micromachined Ultrasonic Transducer-Based Bone Conduction Microphone System for Enhancing Speech Recognition Accuracy

Chongbin Liu <sup>1</sup> , Xiangyang Wang <sup>1</sup>, Jianbiao Xiao <sup>2</sup>, Jun Zhou <sup>2</sup> and Guoqiang Wu <sup>1,3,\*</sup> 

<sup>1</sup> The Institute of Technological Sciences, Wuhan University, Wuhan 430072, China; liuchongbin@whu.edu.cn (C.L.); kent\_wxy@whu.edu.cn (X.W.)

<sup>2</sup> School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China; jianbiao\_x@std.uestc.edu.cn (J.X.); zhouj@uestc.edu.cn (J.Z.)

<sup>3</sup> Hubei Yangtze Memory Laboratories, Wuhan 430205, China

\* Correspondence: wuguoqiang@whu.edu.cn

**Abstract:** Speech recognition in noisy environments has long posed a challenge. Air conduction microphone (ACM), the devices typically used, are susceptible to environmental noise. In this work, a customized bone conduction microphone (BCM) system based on a piezoelectric micromachined ultrasonic transducer is developed to capture speech through real-time bone conduction (BC), while a commercial ACM is integrated for simultaneous capture of speech through air conduction (AC). The system enables simpler and more robust BC speech capture. The BC speech capture achieves a signal-to-noise amplitude ratio over five times greater than that of AC speech capture in an environment with a noise level of 68 dB. Instead of using only AC-captured speech, both BC- and AC-captured speech are input into a speech enhancement module. The noise-insensitive BC-captured speech serves as a speech reference to adapt the SE backbone of AC-captured speech. The two types of speech are fused, and noise suppression is applied to generate enhanced speech. Compared with the original noisy speech, the enhanced speech achieves a character error rate reduction of over 20%, approaching the speech recognition accuracy of clean speech. The results indicate that this speech enhancement method based on the fusion of BC- and AC-captured speech efficiently integrates the features of both types of speech, thereby improving speech recognition accuracy in noisy environments. This work presents an innovative system designed to efficiently capture BC speech and enhance speech recognition in noisy environments.

**Keywords:** microelectromechanical systems (MEMS); piezoelectric micromachined ultrasonic transducer (PMUT); bone conduction microphone; speech enhancement



Academic Editor: Dinko Oletić

Received: 23 April 2025

Revised: 22 May 2025

Accepted: 22 May 2025

Published: 23 May 2025

**Citation:** Liu, C.; Wang, X.; Xiao, J.; Zhou, J.; Wu, G. A Piezoelectric Micromachined Ultrasonic Transducer-Based Bone Conduction Microphone System for Enhancing Speech Recognition Accuracy.

*Micromachines* **2025**, *16*, 613. <https://doi.org/10.3390/mi16060613>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

As a crucial component of human–computer interfaces, speech recognition technology has been widely applied in various fields [1–3]. However, recognizing speech in noisy environments has long presented a challenge. Noise interference can significantly degrade the quality and intelligibility of speech. In response to this challenge, speech enhancement technology has emerged. Speech enhancement technology refers to the application of various signal processing and machine learning methods to process speech signals, thereby improving their clarity and intelligibility [4]. The primary objective is to improve speech quality in noisy environments, enabling machines to comprehend spoken content more clearly.

Speech enhancement can be categorized into digital signal processing (DSP) methods and deep learning-based methods [4,5]. In recent years, deep learning-based speech enhancement methods have brought significant advancements to this field [6–8]. By leveraging the powerful nonlinear mapping capabilities of neural networks, these methods effectively suppress transient non-stationary noise and can significantly reduce background noise from unfamiliar speakers and noise types. However, in the event of an extremely low signal-to-noise ratio (SNR), the aforementioned speech enhancement methods appear somewhat ineffective [9]. This is because noise dominates the sound signal, making it challenging to recover clear speech. These uncontrollable noises arise because traditional speech capture relies on air conduction microphones, which are susceptible to noise and contamination from the surrounding environment. In contrast, bone conduction microphones directly capture vibrations from bones, minimizing the influence of ambient noise and resulting in a higher SNR [10,11]. Consequently, researchers have explored speech enhancement technology based on bone conduction speech [12–15].

The disadvantage of bone conduction speech is the attenuation of high-frequency signals, which results in reduced audibility compared to air conduction speech [16–19]. Thus, in early research, bone conduction signals were utilized to extract auxiliary speech information under noisy conditions, such as speech activity detection, SNR estimation, and pitch extraction [20,21]. Subsequently, researchers sought to extend the bandwidth of BC signals to enhance speech quality [22,23]. These methods can be categorized into three types: equalization, analysis and synthesis, and deep neural network (DNN)-based approaches [24–29]. Recently, feature fusion algorithms for BC- and AC-captured speech have achieved remarkable performance in speech enhancement [30–34]. This method is increasingly being applied in the fields of speaker recognition and speech recognition.

Furthermore, traditional bone conduction microphones are typically based on accelerometers or modified air conduction microphones [35]. Bone conduction microphones based on accelerometers demonstrate poor performance in continuous speech pickup, while those adapted from air conduction microphones possess complex structures and packaging. At present, fiber-optic sensors also show excellent results in voice detection and recognition, even in the case of a distributed operating mode [36,37].

This work proposes a bone conduction microphone system based on a PMUT to achieve a simpler and more robust device for bone conduction, enabling effective continuous collection of speech by BC. Figure 1 provides an overview of the entire system developed in this work.

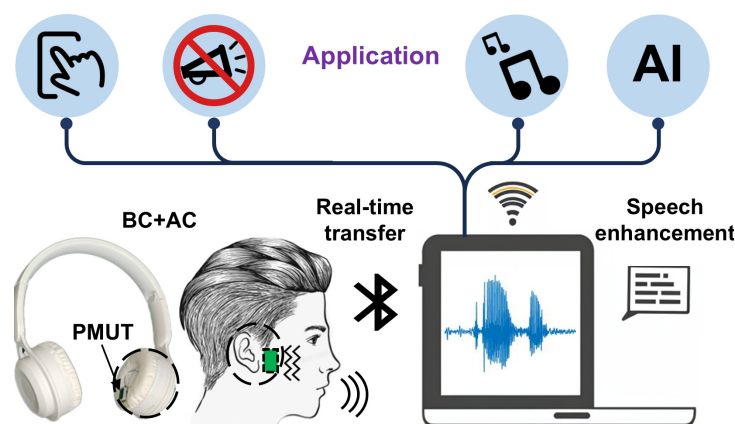


**Figure 1.** Overview of the system application developed in this work. BC: bone conduction; AC: air conduction.

PMUTs exhibit a wide bandwidth and high noise resolution at low frequencies and has been applied in the Internet of Medical Things (IoMT) [38]. Compared with MEMS

capacitive microphones, PMUTs have advantages such as simple structure, no need to control the capacitor gap, reduced process difficulty and cost, and low power consumption. The speech collected by a BCM and an ACM is transmitted in real time to the application. Following speech enhancement through attentional feature fusion (AFF) [6,39] and deep complex convolutional recurrent network (DCCRN) [40], the speech is decoded online using a trained acoustic model based on deep neural networks (DNNs). Compared to the original noisy speech, the enhanced speech results in higher speech recognition accuracy. This method exploits the noise insensitivity of BC-captured speech and the large bandwidth of AC-captured speech, thereby improving speech recognition accuracy in noisy environments. It can be used in various applications, including speech-based human interaction, far-field speech recognition, and AI technologies.

It is worth noting that the design of this study is centered on effective bone conduction speech acquisition using a PMUT and speech enhancement using the advantages of the respective acoustic features of bone conduction and air conduction speech signals. The AFF-DCCRN (attentional feature fusion + deep complex convolution recurrent network) is a suitable choice of speech enhancement neural network for validation. Although the quality of the speech can be further improved by improving the algorithms, these optimizations are not usually constrained by the hardware architecture.

## 2. Design of Bone Conduction Microphone System

In our previous work, we reported a bone conduction pickup method positioned at the throat [41]. In this work, both the front-end sensor component and the overall mold of the bone conduction microphone system have been upgraded. Additionally, a commercial microphone from STMicroelectronics is integrated into the system. The schematic diagram of the reported PMUT-based bone conduction microphone system is illustrated in Figure 2.

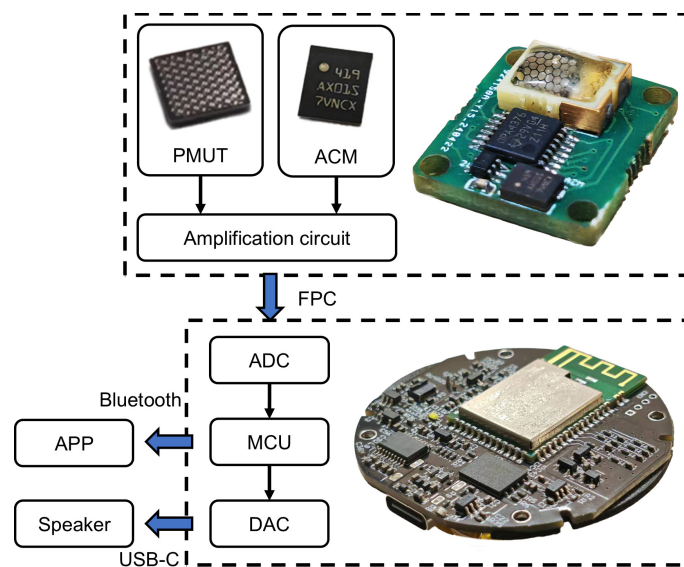


**Figure 2.** Schematic diagram of the reported PMUT-based bone conduction microphone system.

The packaged PMUT, ACM, and amplification circuit are integrated onto a single printed circuit board (PCB), designated the sensor component. The size of the PMUT device is 4 mm × 4 mm, the packaged size is 6 mm × 8 mm, and the size after integration with a microphone is 15 mm × 17 mm. The microphone measures 3.8 mm × 3.3 mm. Compared with the previous device placed on the throat [41], the system in this work demonstrated improved comfort during testing. The package material in this work is JA-2S polyurethane rubber. This material has an acoustic impedance similar to that of human skin, meeting the requirements of package materials for skin patch transducers.

The sensor component is attached to the zygomatic arch using an earphone mold to collect sound by bone conduction or air conduction; the sound is then amplified into an analog voltage [42]. The obtained voltage signals are transferred through a flexible printed circuit (FPC) to an analog-to-digital converter (ADC) for digitization and then processed by a microcontroller unit (MCU). The processed digital signals are transmitted in real time to a mobile application via Bluetooth. The processed digital signals can also be converted back to analog form by a digital-to-analog converter (DAC) and transmitted to the speaker via a USB-C interface.

As the key speech pickup sensor in the bone conduction microphone system, the PMUT is designed with a honeycomb structure [43], as illustrated in Figure 3a. Figure 3b illustrates the operating principle of the PMUT as a bone conduction microphone. The vibration of the bone induces a charge on the surface of the PMUT's piezoelectric layer, which generates a simulated signal representing speech captured through bone conduction.

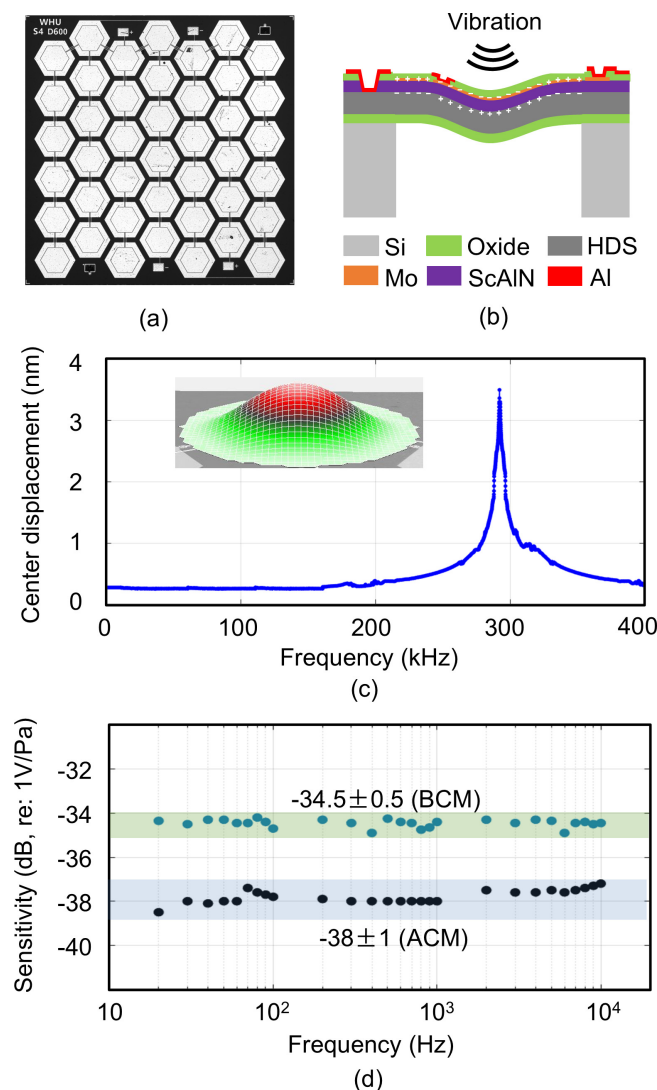


**Figure 3.** (a) Optical microscope photograph of a honeycomb PMUT. (b) Operating principle of the PMUT as a bone conduction microphone. HDS: highly doped silicon. (c) Measured amplitude–frequency response of the reported PMUT under 1 V excitation. (d) Measured sensitivity of the reported PMUT within the frequency range required for human speech pickup and frequency response of the ACM according to the data sheet.

Figure 3c illustrates the measured amplitude–frequency response of the reported PMUT under 1 V excitation. The resonant frequency of PMUT is higher than the human

voice acquisition frequency range, allowing it to maintain a stable amplitude–frequency response within the human voice acquisition frequency range, thereby ensuring that the collected sound signals are not distorted.

Figure 3d illustrates the measured sensitivity of the reported PMUT (with 60 dB amplification) alongside the typical sensitivity of the commercial ACM within the frequency range necessary for human sound detection. The results indicate that the response of the PMUT remains relatively flat between 20 Hz and 10 kHz. Under a circuit amplification of 60 dB, its sensitivity is  $-34.5$  dB (re: 1 V/Pa). Consequently, the sound signals collected by the BCM and ACM both maintain true proportions without being partially amplified. To match the greater intensity of BC vibrations compared to AC sound waves, a dynamic amplification of approximately 20 dB is added into the circuit for the ACM. It should be noted that the wearing position and method significantly impact the quality of the acquired signals, thereby affecting speech enhancement performance.

### 3. Speech Capture and Visualization

In this section, speech recordings of two different speakers were made using the reported microphone system in various noise environments. Subsequently, the temporal and frequency spectral characteristics of these speech signals, along with their spectrograms, were presented and compared.

Temporal and frequency domain analysis are two important methods for speech analysis; each has its limitations. Temporal analysis does not provide an intuitive understanding of the frequency characteristics of speech signals, while frequency domain analysis fails to capture the temporal variations of speech signals. The spectrogram combines the advantages of both analyses, clearly displaying the changes in the speech spectrum over time. In a spectrogram, the horizontal axis represents time, and the vertical axis represents frequency, with the intensity of a given frequency component at a specific moment indicated by varying shades of color. The amplitude, fundamental frequency, harmonic frequencies and envelope of the signal correspond to the three essential elements of sound: loudness, pitch, and timbre.

Firstly, the measured voltages of the spoken, BC-captured phrase “Wu Han Da Xue” (“Wuhan University” in Chinese) were collected by the reported PMUT-based system. Their temporal and frequency spectral characteristics are illustrated in Figure 4a,b. The spectrogram are presented in Figure 4c. From the frequency spectrum, the fundamental frequency of the speaker’s voice is 110 Hz, while harmonic frequencies above 1100 Hz are nearly obscured by noise.

Subsequently, the reported PMUT-based system and commercial ACM were employed to capture the speech of different speakers in various environments with different noise levels. The spectrograms of the spoken phrase “Wu Han Da Xue” collected by these two methods are illustrated in Figure 5.

The comparison of the spectrograms of the two speakers clearly indicates that the pitch of Speaker 2’s voice is higher than that of Speaker 1. It can also be observed that the amplitude, fundamental frequency, and harmonics of each character vary significantly. These distinguishing features provide a clear representation of how speech recognition differentiates between texts or speakers. Spectrograms of speech in various noise environments indicate that under quiet conditions, the SNR of the air conduction microphone is slightly superior to that of the PMUT-based bone conduction microphone. However, in noisy environments, the SNR of the air conduction microphone significantly decreases. In contrast, the PMUT-based bone conduction microphone maintains a high SNR across various noise intensities. Table 1 presents the amplitude ratios of BC and AC speech signals to noise measured in environments with different noise intensity.

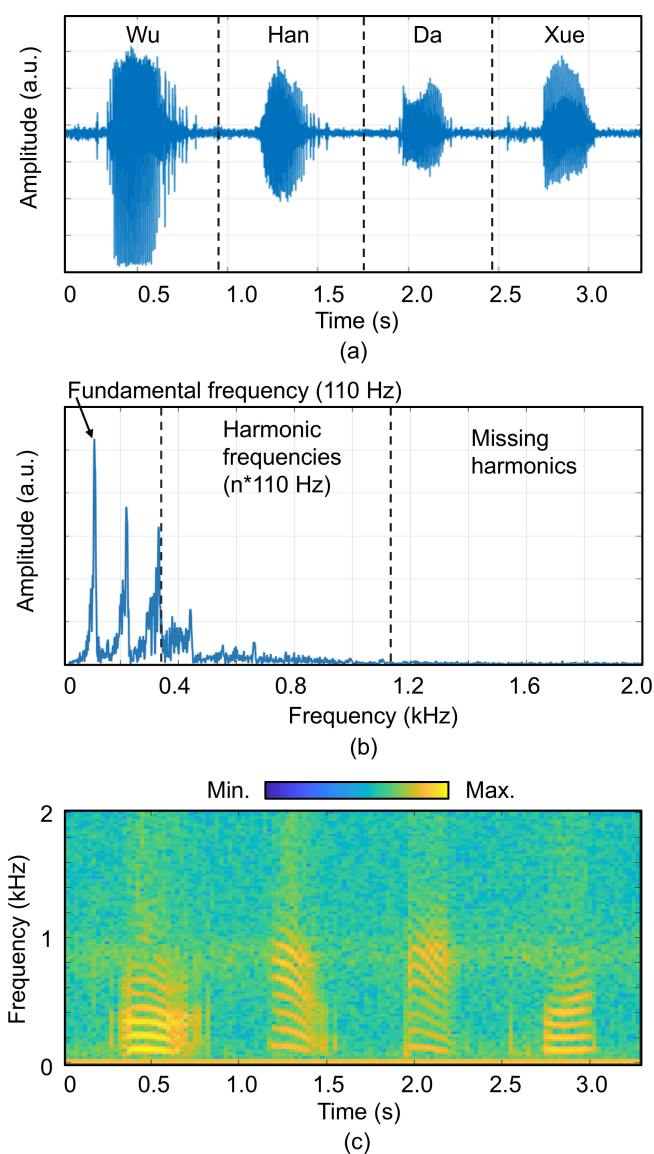


**Figure 4.** (a) Time domain, (b) frequency domain, and (c) spectrogram of the sound signal “Wu Han Da Xue” (“Wuhan University” in Chinese) collected by the reported PMUT-based system.

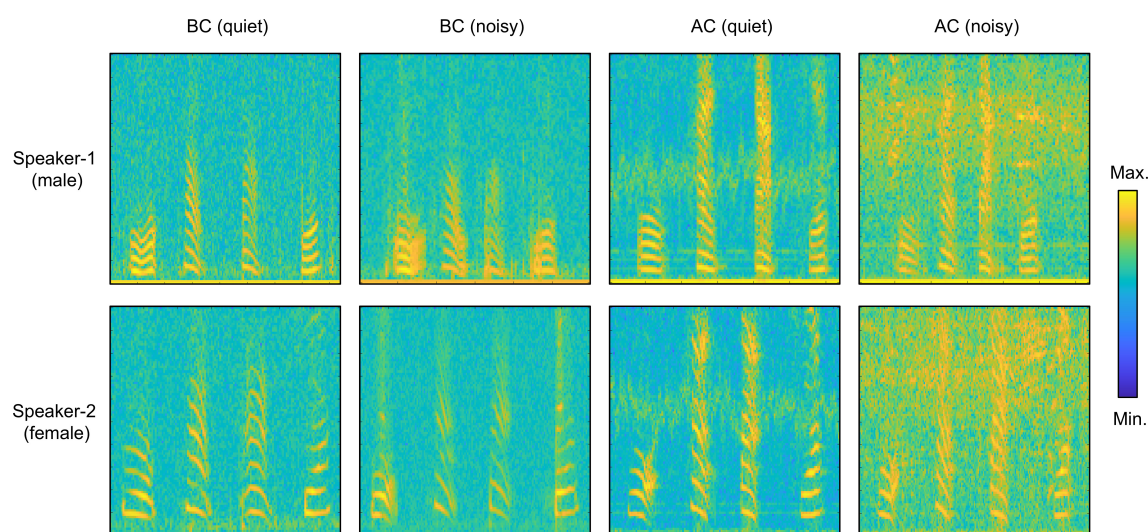


**Figure 5.** Spectrograms of the speech “Wu Han Da Xue” collected from 2 speakers using the reported BCM and ACM in different environments.

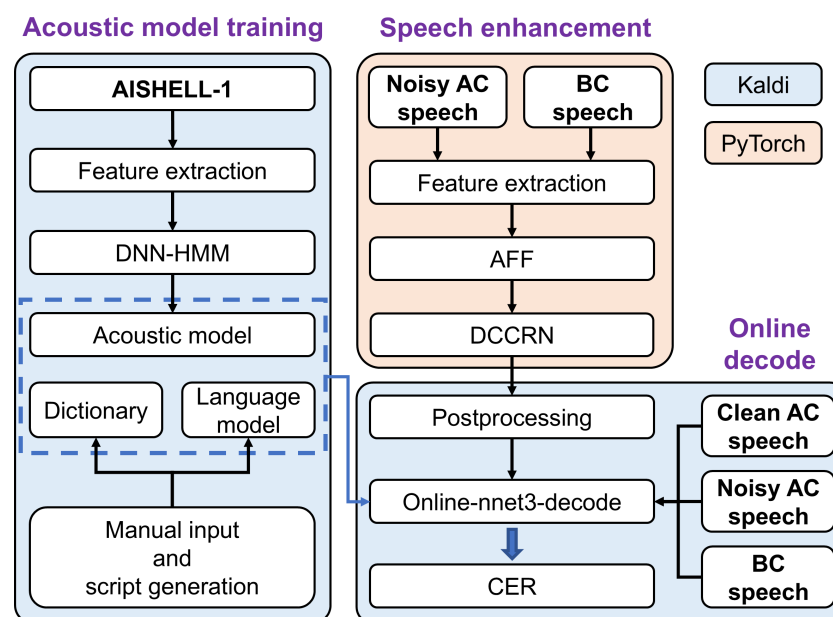
**Table 1.** Measured amplitude ratio of speech signals to noise for BC- and AC-captured speech in various environments with different noise levels.

| Noise Environment | AC  | BC |
|-------------------|-----|----|
| Quiet (~40 dB)    | 150 | 70 |
| Noisy (~60 dB)    | 20  | 60 |
| Noisy (~68 dB)    | 10  | 55 |

#### 4. Speech Recognition Accuracy Enhancement

In this section, we investigate speech enhancement, using the AC- and BC-captured speech to improve speech recognition accuracy in noisy environments. An overview of the enhanced speech recognition system is shown in Figure 6. The acoustic model training and online decoding in this work are based on the open-source project Kaldi [44]. The speech enhancement in this work is based on PyTorch 2.4.1.

The BC and AC speech signals are initially fused and denoised using the speech enhancement model. Subsequently, they are subjected to online decoding with the acoustic model trained by the speech recognition module, ultimately yielding the recognition results. The online decoding module analyzes the acoustic features of speech using a trained acoustic model, recognizing the corresponding text and outputting it in real time.

**Figure 6.** Speech recognition framework in this work. The entire work encompasses the acoustic model training, speech enhancement model, and online decoding system.

In the acoustic model training module, the open-source speech corpus AISHELL-1 [45] was subjected to MFCC feature extraction, and an acoustic model was trained using the features based on the DNN-HMM model. The lexicon and language model required for decoding were generated by manual input and Kaldi's inbuilt scripts. In the speech enhancement module, the features of datasets are extracted and then input into the AFF [39] and DCCRN [40] modules, resulting in the enhanced speech. In the online decoding module, the acoustic model, dictionary, and language model within the speech recognition module are utilized to analyze the acoustic features of the speech, recognizing the corresponding text and outputting it in real time. Character error rate (CER) is one of the

commonly used metrics for evaluating the accuracy of speech recognition [46,47], with the following calculation formula:

$$\text{CER} = (S + D + I) / N \quad (1)$$

where  $S$  denotes the number of characters replaced in error,  $D$  denotes the number of characters deleted in error,  $I$  denotes the number of characters inserted in error, and  $N$  denotes the total number of characters in the reference text. A lower value of CER indicates better performance of the speech recognition system.

#### 4.1. Speech Enhancement Model

To fully exploit the respective acoustic features of AC and BC speech signals, an attentional feature fusion approach is employed. Together with the deep complex convolutional recurrent network, these components form the speech enhancement model in this work. Figure 7a,b, respectively, illustrate the networks of AFF and DCCRN [39,40].

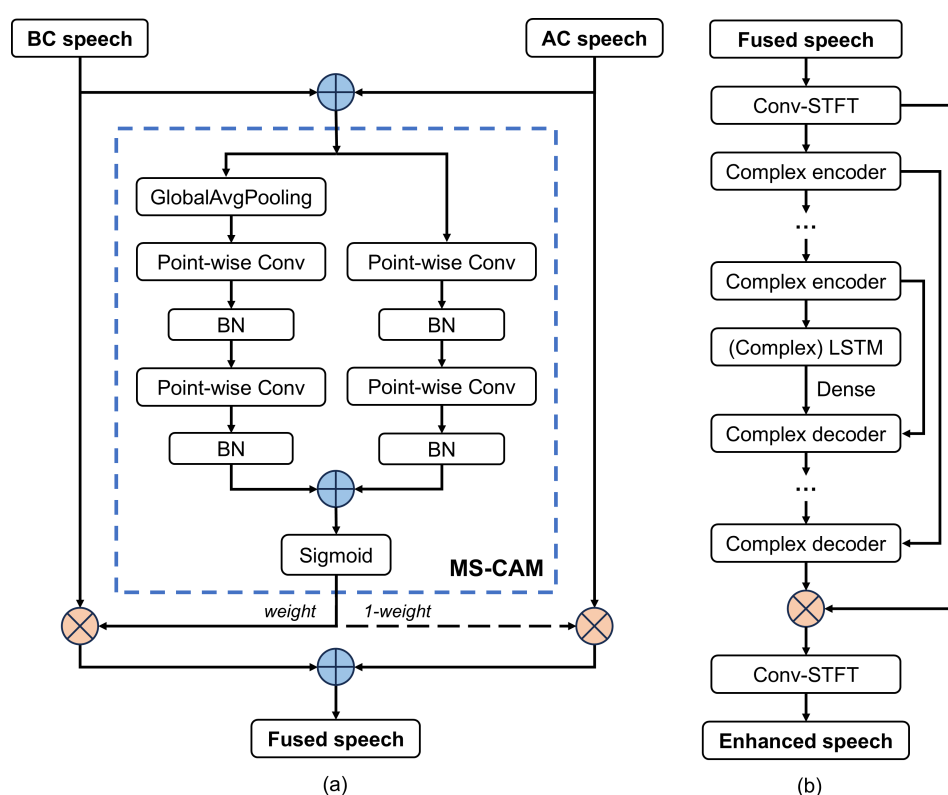


**Figure 7.** Networks of (a) AFF [39] and (b) DCCRN [40]. The first Conv-STFT is placed before the training of the AFF-DCCRN.

Firstly, the frequency spectra of AC and BC speech signals, which contain their acoustic features, are extracted using the short-time Fourier transform (STFT) and subjected to dimensionality transformation. The processed spectrograms are fed into the AFF module. The AFF module generates a fused frequency spectrum, which is then input into the DCCRN. The DCCRN trains the fused spectrum to approximate clean speech and ultimately outputs the enhanced speech.

#### 4.2. Dataset and Setup

Table 2 lists the datasets used for speech enhancement. The original clean dataset used for the speech enhancement model consists of a small dataset containing 40 speech recorded with the reported microphone system. It includes AC speech signals recorded in

a quiet environment (AW dataset) and BC speech signals recorded simultaneously with the AC signals using the reported bone conduction microphone system (BW dataset). The noise was sourced from Tsinghua University's THCH30 dataset [48] (NW dataset), which includes three types of noise, comprising cafe noise, car noise, and white noise. Of these, cafe and car noise are environmental noise.

**Table 2.** Information On the datasets used for speech enhancement.

| Dataset | Speech Source                                         |
|---------|-------------------------------------------------------|
| AW      | Captured clean AC speech                              |
| BW      | Captured BC speech                                    |
| NW      | Noise from THCH30 [48]                                |
| ANW     | Noisy AC speech obtained by mixing AW and NW datasets |

During the training and testing of the speech enhancement model, Input 1 consists of AW–NW mixtures (ANW dataset), generated by combining the AW and NW datasets at three typical noise levels: −5 dB, 0 dB, and 5 dB. Input 2 consists of the corresponding speech from the BW dataset. The ratio of the training dataset to the test dataset is 4:1.

The loss function for speech enhancement model training is SI-SNR, which is widely used as an evaluation metric. SI-SNR is defined as

$$\begin{cases} S = (\langle \tilde{s}, s \rangle / \|s\|_2^2) \cdot s \\ N = \tilde{s} - S \\ \text{SI-SNR} = 10 \cdot \log_{10} \left( \frac{\|S\|_2^2}{\|N\|_2^2} \right) \end{cases} \quad (2)$$

where  $s$  and  $\tilde{s}$  represent the clean and estimated time-domain waveforms, respectively.  $\langle \cdot, \cdot \rangle$  denotes the dot product between two vectors, and  $\|\cdot\|_2$  represents the Euclidean norm.

Each of the speech inputs has a sampling rate of 16 kHz, 16-bit depth, and a duration of 5.25 s. Therefore, the number of data points per speech input is 84,000. The frame duration is set to 25 ms, resulting in a tensor dimension of  $400 \times 210$ . The hop length is one-sixth of the frame length based on the number of convolutional layers, and the FFT (fast Fourier transform) length is 512. The learning rate is set to 0.001.

Finally, the enhanced speech, AW dataset, ANW dataset, and BW dataset are decoded using the online speech recognition system to obtain their recognition structures and CERs. The acoustic scale of decoding is set to one.

#### 4.3. Results and Discussion

Firstly, hybrid training on all noisy speech types is conducted. The CERs across different datasets are compared to evaluate the effectiveness of BC speech in improving speech recognition in noisy environments. Figure 8a illustrates the CERs of different datasets. For the validation dataset of acoustic model training, the CER is 7.4%. For the datasets in speech enhancement module, the CER for the clean AC speech is 7.9%, which is close to the validation dataset's CER, indicating that the recorded clean speech is of good quality. The CERs obtained from BW dataset decoding are both over 90%. This indicates that the absence of high-frequency features in BC speech prevents it from independently performing speech recognition tasks. After adding noise with levels of −5 dB, 0 dB, and 5 dB to the clean speech, the decoding CERs are 69.1%, 63%, and 50.5%, respectively. The CERs for the enhanced speech are 50%, 48.4%, and 42.3%, respectively, representing reductions of 19.1%, 14.6%, and 8.2% compared to the original noisy datasets. This demonstrates that the speech enhancement method based on the fusion of BC and AC speech effectively improves speech recognition accuracy in noisy environments.

To comprehensively evaluate the speech enhancement performance, perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) metrics were incorporated into the hybrid training. As shown in Figure 8b,c, compared with the noisy speech, the PESQ and STOI of the enhanced speech exhibit significant improvements at all noise levels. Specifically, under the low SNR condition of  $-5$  dB, the PESQ and STOI of the enhanced speech are improved by 1.26 and 14.4%, respectively, compared to those of the noisy speech. As the SNR increases, the gains of speech enhancement via BC and AC fusion gradually decrease. This is because under high-SNR conditions, the noise interference is minimal, weakening the advantage that BC speech is less sensitive to noise. Under SNR conditions of 0 dB and 5 dB, the PESQ of the enhanced speech is improved by 1.15 and 0.9, respectively, and the STOI is improved by 10.9% and 7.1% respectively, compared with the noisy speech.

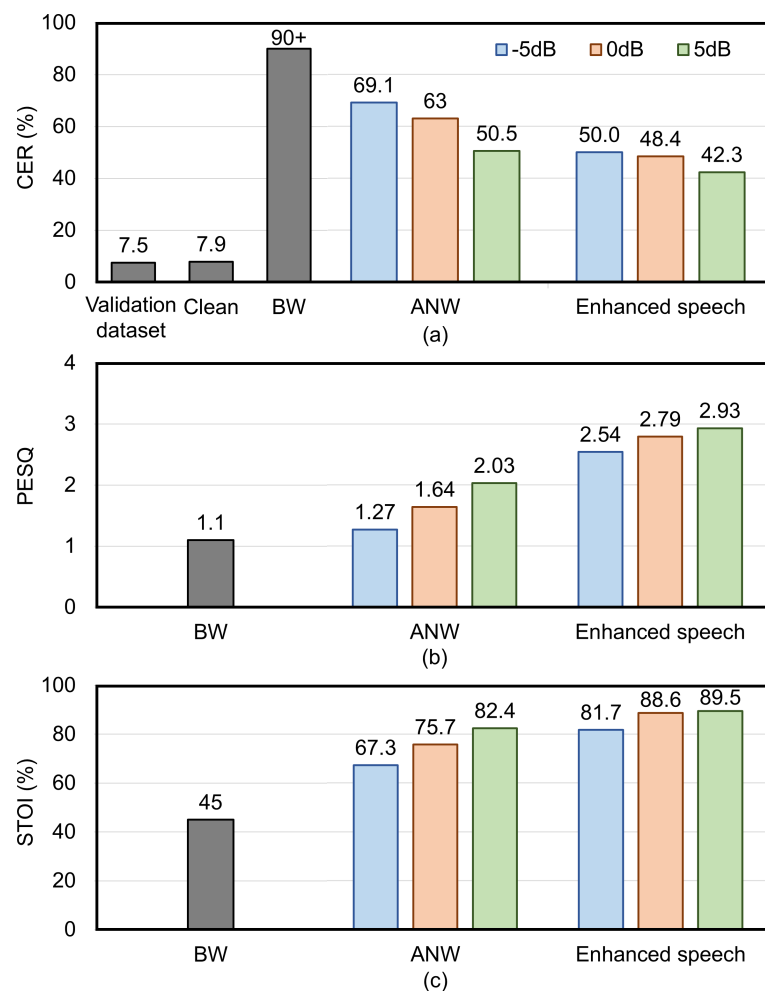


**Figure 8.** Comparison of (a) CERs, (b) PESQs, and (c) STOIs for different datasets during hybrid training, including the validation dataset of the acoustic model, the AW and BW datasets, and the ANW and enhanced speech datasets with various noise levels.

However, the CER of over 40% in speech recognition for the enhanced speech remains significantly high. Comparing the CER of enhanced speech across different types of noise, the CER for speech with added white noise (90.1%) is significantly higher than that for other noise types (35.1%). This is because white noise covers many of the acoustic features of speech, resulting in the discarding of these acoustic features during the training process. As a result, although the PESQ and STOI of the enhanced speech improve, accurate speech recognition remains unachievable.

Figure 9 illustrates a comparison of the Mel spectra of noisy and enhanced speech in environmental and white noise. It is evident that the environmental noise suppression effect is superior. In contrast, white noise causes severe contamination across the entire frequency band of speech, which leads to difficulties in recovering some acoustic features.

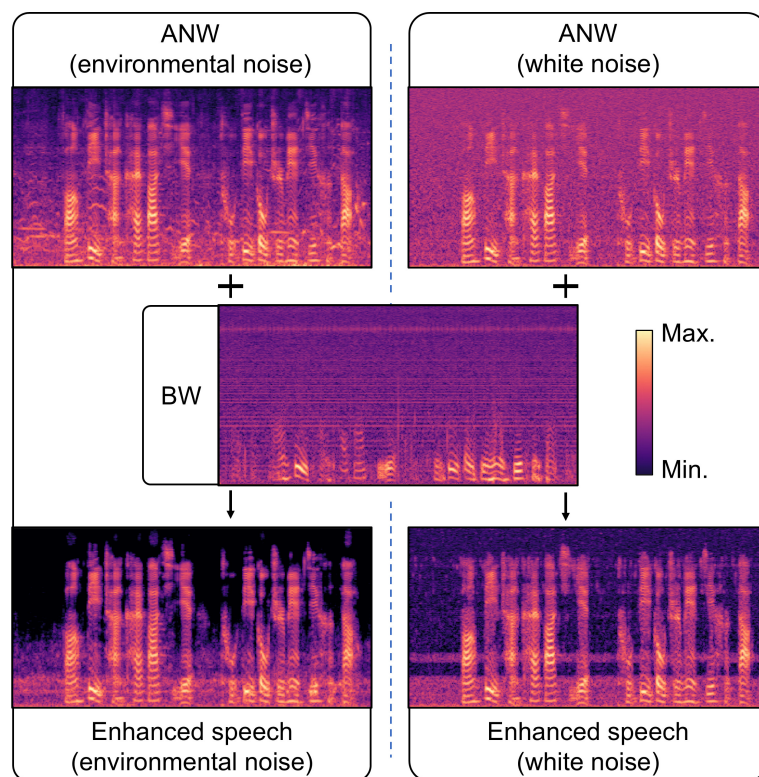


**Figure 9.** Comparison of the Mel spectra of noisy and enhanced speech in environmental and white noise.

White noise was then removed from the types of added noise, and separate training of noisy speech without white noise is conducted. The resulting CERs are shown in Figure 10a. The decoding CERs of noisy speech with  $-5$  dB,  $0$  dB, and  $5$  dB were 53.6%, 45.2%, and 25.9%, respectively. The CERs for the enhanced speech were 26.9%, 13.6%, and 8.1%, respectively, representing reductions of 26.7%, 31.6%, and 17.8% compared to the original noisy datasets. The enhanced speech at  $0$  dB and  $5$  dB noise levels has achieved speech recognition capabilities comparable to that of clean speech.

Subsequently, we conducted a separate speech enhancement for the noisy speech only with white noise. The speech enhancement reduced the CER of the noisy speech from over 99% to 84.5%. Next, by decreasing the white noise intensity, the CER of the original noisy speech is 80.2% at a noise level of  $10$  dB, while the CER of the enhanced speech is reduced to 15.4%, as illustrated in Figure 10b. It indicates that the speech enhancement method is effective in suppressing non-extreme white noise.

The problem of weak robustness to full-band white noise is preliminarily alleviated by reducing the white noise intensity to  $10$  dB, though the problem has not been completely resolved. In subsequent study, targeting the characteristics of white noise, we will adopt a method combining DSP algorithms with deep learning and optimize the speech enhancement model, so as to achieve better versatility and performance.

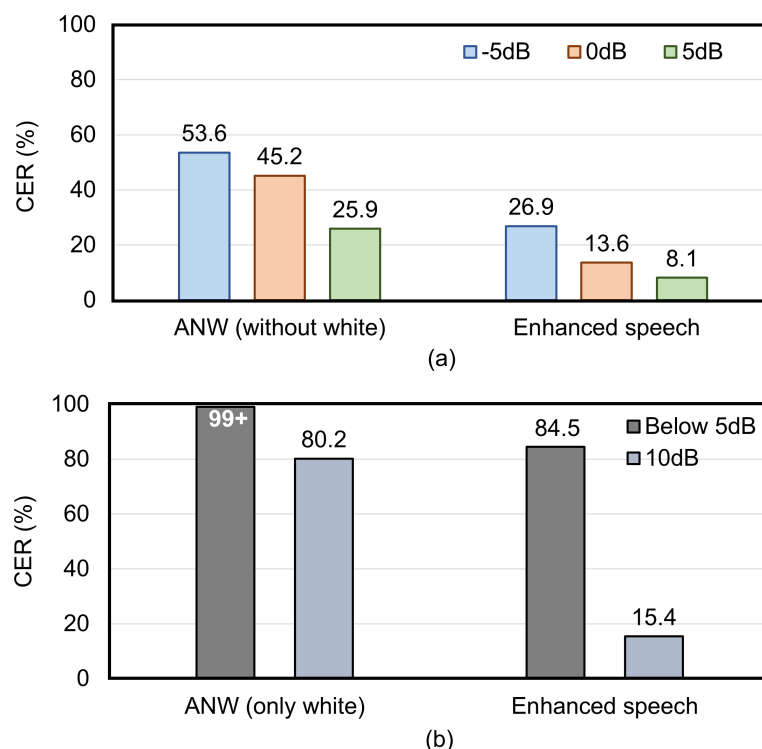


**Figure 10.** Comparison of CERs for different datasets during separate training. (a) includes the ANW and enhanced speech datasets that exclude white noise. (b) includes the ANW and enhanced speech datasets with only white noise (“Below 5 dB” represents the average value of −5 dB, 0 dB, and 5 dB).

Furthermore, the dataset size of 20 speech samples is relatively small. We used 20 samples to preliminarily verify the feasibility of the BC and AC speech fusion enhancement method and the application potential of PMUT. After expanding the samples to 40, the speech enhancement effect has slightly improved, but the overall trend remains unchanged. In subsequent experiments, we will further expand the size of the original speech dataset and include different speakers to improve the model’s generalization ability, thereby forming a more comprehensive research framework.

## 5. Conclusions

In this work, a customized bone conduction microphone system based on a PMUT is developed to capture real-time BC speech signals, and an ACM is integrated for simultaneous capture of AC speech signals. The PMUT-based BCM system achieves a simpler and more robust device for effective continuous capture of BC speech. The BC-captured speech achieves a signal-to-noise amplitude ratio over five times greater than AC-captured speech in an environment with 68 dB noise. Subsequently, an enhancement method based on the fusion of BC and AC speech signals was used to improve the accuracy of speech recognition in noisy environments, demonstrating the application potential of this BCM system.

For hybrid training with various noise types, the CERs of enhanced speech are 50%, 48.4% and 42.3% at −5 dB, 0 dB, and 5 dB noise levels, respectively. Compared with the original noisy dataset, the CERs are reduced by 19.1%, 14.6% and 8.2%, respectively. Following the removal of white noise, the CERs of the enhanced speech improve further, recording values of 26.9%, 13.6% and 8.1% at −5 dB, 0 dB, and 5 dB noise levels, respectively. These CERs represent reductions of 26.7%, 31.6% and 17.8% compared with the original noisy dataset, respectively. Additionally, in separate training at a 10 dB white noise level, the enhanced speech achieved a CER of 15.4%. Collectively, the results indicate that the speech enhancement strategy based on the feature fusion of BC and AC speech in this work

effectively suppresses environmental noise and non-extreme white noise, approaching the speech recognition accuracy of clean speech.

This work focuses on efficiently capturing BC speech using PMUT and using the acoustic feature advantages of BC- and AC-captured speech for speech enhancement, without improving the network structure of the speech enhancement used. Moreover, the speech datasets in this work consisted of unprocessed raw speech, highlighting the practicality of this system for applications. However, this also brings disadvantages in terms of SNR and recognition accuracy. In future study, we plan to increase the size of the speech dataset, preprocess the speech data, and optimize the speech enhancement model for white noise to improve the generalization ability of the model and achieve better speech recognition accuracy. Additionally, in terms of hardware, we will utilize ASIC and integrate the system with AI chips to achieve a system with higher integration and lower power consumption.

**Author Contributions:** C.L.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, writing—review and editing; X.W.: visualization, writing—review and editing; J.X.: methodology, writing—review and editing; J.Z.: supervision; G.W.: conceptualization, methodology, resources, writing—review and editing, supervision, project administration, funding acquisition. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the National Natural Science Foundation of China under Grant 62322409, in part by the Fundamental Research Funds for the Central Universities under Grant 2042025kf0028, and in part by the Hubei Provincial Science and Technology Research Project under Grant 2024BAB006.

**Institutional Review Board Statement:** The speakers in this paper are the authors themselves and their relatives, and the speech samples captured are not harmful and do not involve any sensitive personal information or commercial interests. Therefore, this study was exempted from ethical review.

**Informed Consent Statement:** Informed consent was obtained from all speakers involved in the study.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Ma, Z.; Liu, Y.; Liu, X.; Ma, J.; Li, F. Privacy-preserving outsourced speech recognition for smart IoT devices. *IEEE Internet Things J.* **2019**, *6*, 8406–8420. [\[CrossRef\]](#)
2. Gaikwad, S.K.; Gawali, B.W.; Yannawar, P. A review on speech recognition technique. *Int. J. Comput. Appl.* **2010**, *10*, 16–24. [\[CrossRef\]](#)
3. Farahani, B.; Tabibian, S.; Ebrahimi, H. Toward a personalized clustered federated learning: A speech recognition case study. *IEEE Internet Things J.* **2023**, *10*, 18553–18562. [\[CrossRef\]](#)
4. O'Shaughnessy, D. Speech enhancement—A review of modern methods. *IEEE Trans. Human Mach. Syst.* **2024**, *54*, 110–120. [\[CrossRef\]](#)
5. Putta, V.S.; Priyadharson, A.S.M. Regional language speech recognition from bone conducted speech signals through CCWT algorithm. *Circuits Syst. Signal Process* **2024**, *43*, 6553–6570. [\[CrossRef\]](#)
6. Dong, K.; Peng, H.; Che, J. Dynamic-static cross attentional feature fusion method for speech emotion recognition. In Proceedings of the MultiMedia Modeling, MMM 2023, Bergen, Norway, 9–12 January 2023; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 13834, pp. 350–361.
7. Zhang, X.; Tang, J.; Cao, H.; Wang, C.; Shen, C.; Liu, J. Cascaded speech separation denoising and dereverberation using attention and TCN-WPE networks for speech devices. *IEEE Internet Things J.* **2024**, *11*, 18047–18058. [\[CrossRef\]](#)
8. Tan, K.; Wang, D. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 380–390. [\[CrossRef\]](#)

9. Wang, M.; Chen, J.; Zhang, X.; Rahardja, S. End-to-end multi-modal speech recognition on an air and bone conducted speech corpus. *IEEE ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 513–524. [\[CrossRef\]](#)
10. Shin, H.S.; Kang, H.; Fingscheidt, T. Survey of speech enhancement supported by a bone conduction microphone. In Proceedings of the Speech Communication; 10. ITG Symposium, Braunschweig, Germany, 26–28 September 2012; pp. 1–4.
11. Hansen, C.H. Fundamentals of acoustics. *Am. J. Phys.* **1951**, *19*, 254–255.
12. Zhou, Y.; Wang, H.; Chu, Y.; Liu, H. A robust dual-microphone generalized sidelobe canceller using a bone-conduction sensor for speech enhancement. *Sensors* **2021**, *21*, 1878. [\[CrossRef\]](#)
13. Zhou, Y.; Chen, Y.; Ma, Y.; Liu, H. A real-time dual-microphone speech enhancement algorithm assisted by bone conduction sensor. *Sensors* **2020**, *20*, 5050. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Lee, C.; Rao, B.D.; Garudadri, H. Bone-conduction sensor assisted noise estimation for improved speech enhancement. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 1180–1184.
15. Hussain, T.; Tsao, Y.; Siniscalchi, S.M.; Wang, J.; Wang, H.; Liao, W. Bone-conducted speech enhancement using hierarchical extreme learning machine. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*; Lecture Notes in Electrical Engineering; Springer: Singapore, 2021; Volume 714, pp. 153–162.
16. Rahman, M.S.; Shimamura, T. Amplitude variation of bone-conducted speech compared with air-conducted speech. *Acoust. Sci. Technol.* **2019**, *40*, 293–301. [\[CrossRef\]](#)
17. Nishimura, T.; Miyamae, R.; Hosoi, H.; Saito, O.; Shimokura, R.; Yamanaka, T.; Kitahara, T. Frequency characteristics and speech recognition in cartilage conduction. *Auris Nasus Larynx* **2019**, *46*, 709–715. [\[CrossRef\]](#)
18. Toya, T.; Birkholz, P.; Unoki, M. Measurements of transmission characteristics related to bone-Conducted speech using excitation signals in the oral cavity. *J. Speech Lang. Hear. Res.* **2020**, *63*, 4252–4264. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Ishikawa, H.; Otsuka, S.; Nakagawa, S. Threshold and frequency- and temporal resolutions of distantly presented bone-conducted sound in the audible-frequency range. *Jpn. J. Appl. Phys.* **2022**, *61*, 1065. [\[CrossRef\]](#)
20. Zhu, M.; Ji, H.; Luo, F.; Chen, W. A robust speech enhancement scheme on the basis of bone-conductive microphones. In Proceedings of the 2007 3rd International Workshop on Signal Design and Its Applications in Communications, Chengdu, China, 23–27 September 2007; pp. 353–355.
21. Rahman, M.S.; Shimamura, T. Pitch characteristics of bone conducted speech. In Proceedings of the 2010 18th European Signal Processing Conference, Aalborg, Denmark, 23–27 August 2010; pp. 795–799.
22. Shimamura, T.; Tamiya, T. A reconstruction filter for bone-conducted speech. In Proceedings of the 48th Midwest Symposium on Circuits and Systems, 2005, Covington, KY, USA, 7–10 August 2005; pp. 1847–1850.
23. Bouserhal, R.E.; Falk, T.H.; Voix, J. In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension. *J. Acoust. Soc. Am.* **2017**, *141*, 1321–1331. [\[CrossRef\]](#)
24. Trung, P.N.; Unoki, M.; Akagi, M. A study on restoration of bone-conducted speech in noisy environments with LP-based model and gaussian mixture model. *J. Signal Process.* **2012**, *16*, 409–417. [\[CrossRef\]](#)
25. Huang, B.; Gong, Y.; Sun, J.; Shen, Y. A wearable bone-conducted speech enhancement system for strong background noises. In Proceedings of the 2017 18th International Conference on Electronic Packaging Technology (ICEPT), Harbin, China, 16–19 August 2017; pp. 1682–1684.
26. Singh, P.; Mukul, M.K.; Prasad, R. Bone conducted speech signal enhancement using LPC and MFCC. In Proceedings of the Intelligent Human Computer Interaction, IHCI 2018, Allahabad, India, 7–9 December 2018; pp. 148–158.
27. Zheng, C.; Yang, J.; Zhang, X.; Cao, T.; Sun, M. Bandwidth extension WaveNet for bone-conducted speech enhancement. In *Proceedings of the 7th Conference on Sound and Music Technology (CSMT)*; Lecture Notes in Electrical Engineering; Springer: Singapore, 2020; pp. 3–14.
28. Zheng, C.; Cao, T.; Yang, J.; Zhang, X.; Sun, M. Spectra restoration of bone-conducted speech via attention-based contextual information and spectrotemporal structure constraint. *Trans. Fund. Electron. Commun. Comput. Sci.* **2019**, *E102.A*, 2001–2007.
29. Nguyen, H.Q.; Unoki, M. Improvement in bone conducted speech restoration using linear prediction and long short-term memory model. *J. Signal Process.* **2020**, *24*, 175–178. [\[CrossRef\]](#)
30. Tsuge, S.; Koizumi, D.; Fukumi, M.; Kuroiwa, S. Speaker verification method using bone-conduction and air-conduction speech. In Proceedings of the 2009 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Kanazawa, Japan, 7–9 January 2009; pp. 449–452.
31. Wang, H.; Zhang, X.; Wang, D. Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement. *IEEE ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 3134–3143. [\[CrossRef\]](#)
32. Wang, H.; Zhang, X.; Wang, D. Attention-based fusion for bone-conducted and air-conducted speech enhancement in the complex domain. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7757–7761.
33. Wang, M.; Chen, J.; Zhang, X.; Huang, Z.; Rahardja, S. Multi-modal speech enhancement with bone-conducted speech in time domain. *Appl. Acoust.* **2022**, *200*, 109058. [\[CrossRef\]](#)

34. Yu, C.; Hung, K.; Wang, S.; Tsao, Y.; Hung, J. Time-domain multi-modal bone/Air conducted speech enhancement. *IEEE Signal Process. Lett.* **2020**, *27*, 1035–1039. [\[CrossRef\]](#)
35. You, B.C.; Lo, S.C.; Chan, C.K.; Li, C.S.; Ho, H.L.; Chiu, S.C.; Hsieh, G.H.; Fang, W. Design and implementation of dual pressure variation chambers for bone conduction microphone. *J. Micromech. Microeng.* **2020**, *30*, 125009. [\[CrossRef\]](#)
36. Dorize, C.; Guerrier, S.; Awwad E.; Renaudier J. Capturing acoustic speech signals with coherent MIMO phase-OTDR. In Proceedings of the 2020 European Conference on Optical Communications (ECOC), Brussels, Belgium, 6–10 December 2020; pp. 1–4.
37. Gritsenko, T.V.; Orlova, M.V.; Zhirnov, A.A.; Konstantinov, Y.A.; Turov, A.T.; Barkov, F.L.; Khan, R.I.; Koshelev, K.I.; Svelto, C.; Pnev, A.B. Detection and recognition of voice commands by a distributed acoustic sensor based on phase-sensitive OTDR in the smart home concept. *Sensors* **2024**, *24*, 2281. [\[CrossRef\]](#)
38. Jia, L.; Shi, L.; Liu, C.; Xu, J.; Gao, Y.; Sun, C.; Liu, S.; Wu, G. Piezoelectric micromachined ultrasonic transducer array-based electronic stethoscope for Internet of Medical Things. *IEEE Internet Things J.* **2022**, *9*, 9766–9774. [\[CrossRef\]](#)
39. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional feature fusion. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 3559–3568.
40. Hu, Y.; Liu, Y.; Lv, S.; Xing, M.; Zhang, S.; Fu, Y.; Wu, J.; Zhang, B.; Xie, L. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 2472–2476.
41. Liu, C.; Wang, X.; Xie, Y.; Wu, G. Bone conduction pickup based on piezoelectric micromachined ultrasonic transducers. In Proceedings of the 2023 IEEE 36th International Conference on Micro Electro Mechanical Systems (MEMS), Munich, Germany, 15–19 January 2023; pp. 949–952.
42. Li, Y.; Wang, Y.; Liu, X.; Shi, Y.; Patel, S.; Shih, S. Enabling real-time on-chip audio super resolution for bone-conduction microphones. *Sensors* **2023**, *23*, 35. [\[CrossRef\]](#)
43. Liu, C.; Jia, L.; Shi, L.; Sun, C.; Cheam, D.D.; Wang, P.; Wu, G. Theoretical modeling of piezoelectric micromachined ultrasonic transducers with honeycomb structure. *J. Microelectromech. Syst.* **2022**, *31*, 984–993. [\[CrossRef\]](#)
44. Daniel, P.; Arnab, G.; Gilles, B.; Lukas, B.; Ondrej, G.; Nagendra, G.; Mirko, H.; Petr, M.; Yanmin, Q.; Petr, S.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011.
45. Bu, H.; Du, J.; Na, X.; Wu, B.; Zheng, H. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In Proceedings of the 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul, Republic of Korea, 1–3 November 2017; pp. 1–5.
46. Fendji, J.L.K.E.; Tala, D.C.M.; Yenke, B.O.; Atemkeng, M. Automatic Speech Recognition Using Limited Vocabulary: A Survey. *Appl. Artif. Intell.* **2022**, *36*, 2095039. [\[CrossRef\]](#)
47. Wang, P.; Sun, R.; Zhao, H.; Yu, K. A New Word Language Model Evaluation Metric for Character Based Languages. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*; Sun, M., Zhang, M., Lin, D., Wang, H., Eds.; Springer: Berlin, Germany, 2013; pp. 315–324.
48. Wang, D.; Zhang, X. THCHS-30: A free Chinese speech corpus. *arXiv* **2015**, arXiv: 1512.01882.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.