

Reviewer #1: This manuscript proposes a bone conduction microphone system based on PMUT, combined with a multimodal speech enhancement algorithm, aiming to solve the problem of the decline in speech recognition accuracy in a noisy environment. This manuscript designed a PMUT with a honeycomb structure, achieving a simplified bone conduction microphone. The manuscript innovatively proposed a multimodal speech enhancement framework based on AFF and DCCRN, improving the performance of speech recognition in noisy environments. In terms of experimental verification, through comparative tests of multiple scenarios and multiple noise types, the manuscript confirmed that the SNR of BC speech in a 68 dB noise environment was more than five times that of AC speech, demonstrating strong practical potential.

Response: We thank the reviewer very much for the careful read and very encouraging comments on our manuscript. We are very happy to address reviewer's questions and revise the manuscript based on reviewer's constructive comments and suggestions.

1. In terms of writing and presentation, the manuscript has a clear structure, detailed charts, and strong logical coherence. However, attention should be paid to the uniformity of terms. For example, when abbreviations (such as AFF-DCCRN) appear for the first time, the full names should be marked.

Response: Thanks for the reviewer's kind comment. The uniformity of terms has been checked and revised.

Line 81, Page 3: "AFF-DCCRN (attentional feature fusion + deep complex convolution recurrent network)";

Caption of Figure 3(b), Page 4: "HDS: highly doped silicon";

Line 220, Page 9: "FFT (Fast Fourier Transform)".

2. The scale of the experimental dataset containing 20 self-collected voices is relatively small, limiting the generalization ability of the model. The authors may increase the numbers.

Response: Thanks very much for the reviewer's valuable comment. As the reviewer pointed out, 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. Although increasing the number of dataset samples can further optimize performance, this process is not constrained by hardware. After submitting the manuscript, we collected 40 new speech samples and performed speech enhancement, adjusting the intensity levels of mixed noise from 3 levels (-5, 0, 5 dB) to 11 levels (from -5 to 5 dB with a 1 dB interval), resulting in 1,320 speech samples to expand the dataset. The preliminary comparison of CERs for enhanced speech from hybrid training with 20 and 40 groups of raw speech

is shown in figure 1-1.

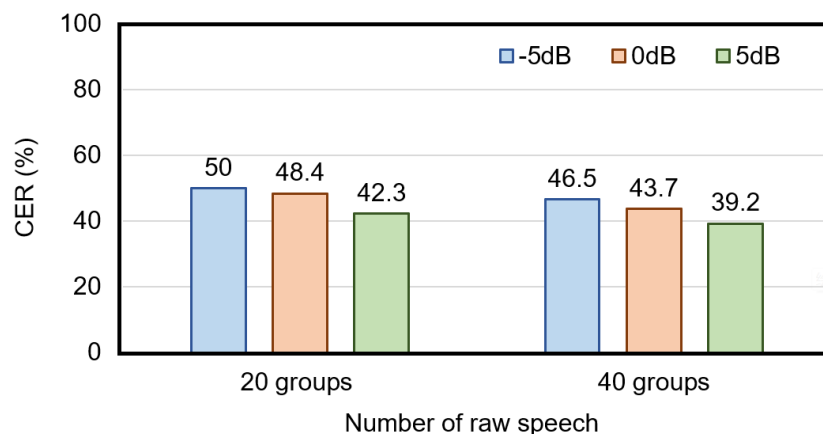


Figure 1-1. Comparison of CERs for enhanced speech from hybrid training with 20 and 40 groups of raw speech.

After expanding the dataset, the speech enhancement effect has slightly improved, but the overall trend remains unchanged. Collecting more paired speech samples and training the model require substantial time, so the data is currently incomplete, and the expanded data has not yet been updated in the original text. 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.

We have added the explanation and outlook in the discussion: “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.” (Line 281, Page 12).

3. Although the author alleviated the problem of weak robustness to full-band white noise by reducing the intensity of white noise (10 dB), targeted solutions such as frequency band selective suppression or model structure optimization were not explored. The authors may add an explanation in the results analysis and discussion part.

Response: Thanks very much for the reviewer’s valuable comment. White noise is a type of stationary noise that causes severe pollution across all frequency bands of speech. This paper uses raw audio without preprocessing to demonstrate the practicality and convenience of the application, but this also introduces disadvantages in signal-to-noise ratio (SNR) and speech recognition accuracy. In future work, we will combine traditional algorithms with deep learning and add preprocessing steps to achieve better

results.

AFF-DCCRN is a suitable choice to verify the feasibility of the BC and AC speech fusion enhancement method. Although noise suppression performance can be further improved through algorithm optimization, this approach is not constrained by hardware. In subsequent research, we will further compare the effects of different models against white noise to achieve optimal performance.

“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.” has been added to the results analysis and discussion section (Line 276, Page 11).

4. The authors may supplement the discussion on the hardware cost and power consumption of PMUT to enhance the discussion on the application value of the research.

Response: Thanks very much for the reviewer’s valuable comment. Compared with MEMS capacitive microphone, PMUT do not require bias voltage and have lower power consumption. In terms of manufacturing process, PMUT feature a simpler structure, eliminate the need for capacitor gaps, and have lower processing complexity and cost. The comparison of the structures and working principles between PMUT and MEMS capacitive microphone is shown in Figure 1-2.

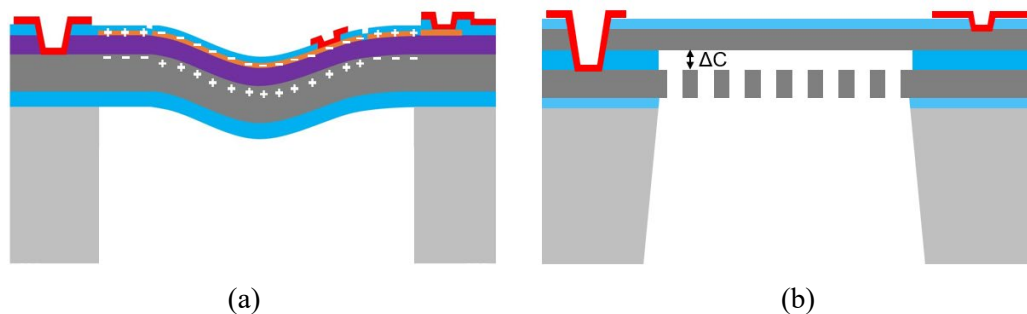


Figure 1-2. Comparison of the structures and working principles between (a) PMUT and MEMS capacitive microphone

“Compared with MEMS capacitive microphones, PMUTs have advantages such as simple structure, no need to control the capacitor gap, lower process difficulty and cost, and low power consumption.” has been added in the revised manuscript (Line 66, Page 2).

Additionally, the majority of power consumption in practical applications actually originates from hardware circuits, though this aspect is not the focus of this work. In future study, we will utilize ASIC and integrate the system with AI chips to achieve a system with higher integration and lower power consumption. This description has been added to the outlook section of the conclusion.

“In future study, we plan to optimize the speech enhancement model and preprocess the speech data. Based on this, a series of speech enhancement studies aimed at improving the generalization of small datasets, such as pre-training, fine-tuning and transfer learning, will be carried out.” has been updated to “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.” in the revised manuscript (Line 313, Page 13).

5. In line 164, the authors can provide some supporting literature for CER to explain why CER is a commonly used indicator.

Response: Thanks for the reviewer’s kind comment. Relevant references have been added to explain why CER is a commonly used indicator: “Character error rate (CER) is one of the commonly used metrics for evaluating the accuracy of speech recognition [46,47],” (Line 183, Page 8).

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.

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.; Springer: Berlin, 2013; pp. 315-324.

6. The authors should clarify which part is PMUT in Figure 1.

Response: Thanks for the reviewer’s kind comment. PMUT has been labeled in Figure 1.

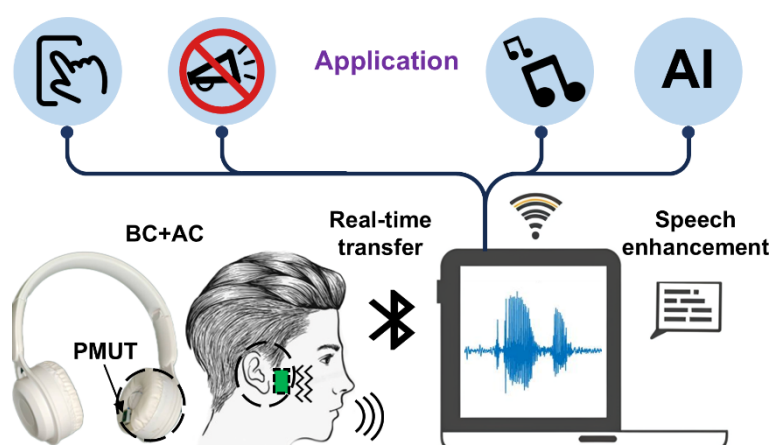


Figure 1. Overview of the system application developed in this work. BC: bone conduction, AC: air conduction.

We hope that the reviewer finds that his/her concerns have been adequately addressed in the revised article, and in this response.

Thanks again!

Reference

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.
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.; Springer: Berlin, 2013; pp. 315-324.