

Reviewer #3: Overall, I liked the work submitted for review. In this work, a customized bone conduction microphone (BCM) system based on piezoelectric micromachined ultrasonic transducer is developed to capture the real-time bone conduction (BC) speech, while a commercial ACM is integrated for simultaneous capture of air conduction (AC) speech. However, I would suggest that the author revise the manuscript taking into account my comments, which I outline below:

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

1. In a review of competing sound recording methods, it would be appropriate to mention fiber-optic sensors. These days, they show amazing results in voice detection and recognition, even in the case of a distributed operating mode:

<http://dx.doi.org/10.1109/ECOC48923.2020.9333283>

<https://doi.org/10.3390/s24072281>

Response: Thanks for the reviewer's kind comment. Descriptions of relevant sound recording methods have been added in the revised manuscript.

“These days, fiber-optic sensors also show amazing results in voice detection and recognition, even in the case of a distributed operating mode [36, 37].” has been added in the revised manuscript (Line 60, Page 2).

36. Dorize, C.; Guerrier, S.; Awwad E.; Renaudier J. Capturing acoustic speech signals with coherent MIMO phase-OTDR. in *European Conf. Opt. Commun. (ECOC)*, 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.

2. In my opinion, the work does not pay enough attention to the physical parameters of digitization. Thus, knowing the characteristic amplitude frequency response of the presented microphone, it would be possible to study various sampling frequencies and their effect on the quality of recognition.

Response: Thanks very much for the reviewer's valuable comment. In speech recognition, the sampling frequency of the test speech must be consistent with that of the dataset used in the acoustic model, AISHELL-1, which is 16 kHz.

Figure 3-1 shows the comparison of the Mel spectra of BC speech at 4 kHz and 16 kHz sampling frequencies.

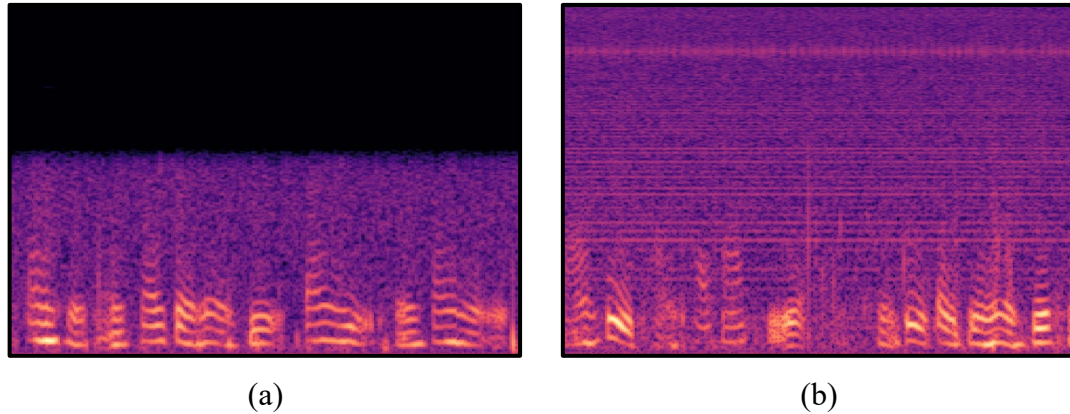


Figure 3-1. Comparison of the Mel spectra of bone conduction speech at (a) 4 kHz and (b) 16 kHz sampling frequencies.

The advantage of directly collecting BC speech at a 16 kHz sampling rate lies in its application convenience, but it is significantly affected by electromagnetic interference, which impairs its audibility. However, since electromagnetic interference is a type of steady-state noise, this kind of noise can be effectively filtered out in the speech enhancement model.

The advantage of 4 kHz speech lies in its ability to reduce the transmission burden on Bluetooth and extend recording time. However, it needs to be resampled to 16 kHz before being input into the speech enhancement model. Nevertheless, the acoustic features of BC speech above 2 kHz are attenuated to nearly zero [R1]. Therefore, resampling from 4 kHz to 16 kHz does not result in significant loss of acoustic characteristics.

In the future, we plan to adopt electromagnetic shielding measures to achieve higher-quality BC speech. Additionally, we will conduct in-depth research on how BC speech with different sampling frequencies influences speech enhancement performance and speech recognition accuracy.

3. By the way, it would be appropriate to attach the microphone's amplitude-frequency characteristic to the manuscript.

Response: Thanks for the reviewer's kind comment. In the original manuscript, Figure 3(c) shows the sensitivity-frequency characteristics of BCM and ACM. In fact, the amplitude-frequency response is positively correlated with it. To more clearly demonstrate the performance characterization of BCM, we have added the amplitude-frequency response of PMUT tested by LDV in Figure 3 and included relevant descriptions in the text.

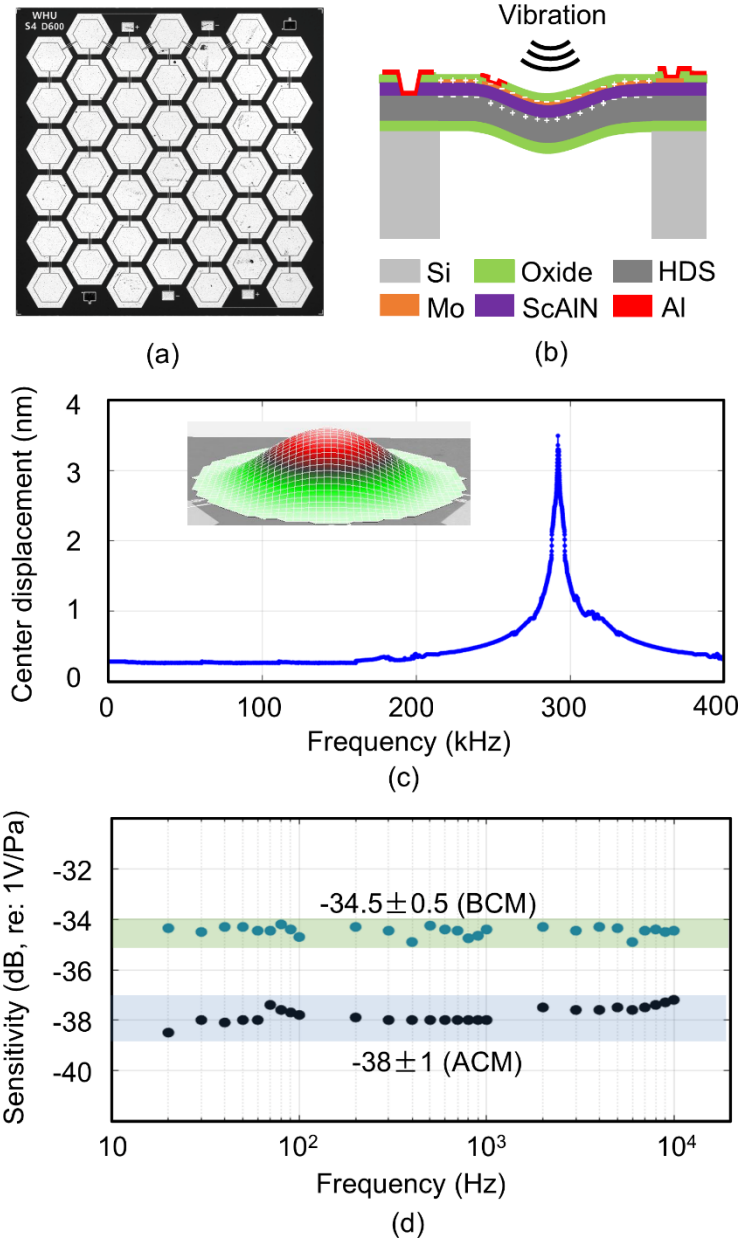


Figure 3. (a) Optical microscope photograph of a honeycomb PMUT. (b) Operating principle of the PMUT as a bone conduction microphone. **HDS: highly doped silicon.** (c) **Measured amplitude-frequency response of the reported PMUT under 1 V excitation.** (d) Measured sensitivity of the reported PMUT within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

“Figure 3(c) 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.” has been added in the revised manuscript (Line 112, Page 4).

4. It is known that preliminary data processing before recognition significantly improves the quality of this process. Why didn't the authors pay attention to such

methods as wavelet filtering, as well as empirical and variational mode decomposition?

Response: Thanks for the reviewer's kind comment. The core design of this work is to utilize PMUT for the effective acquisition of BC speech and leverage the respective acoustic feature advantages of BC and AC speech for speech enhancement. While speech quality can be further improved through algorithm refinement, these optimizations remain largely unconstrained by hardware architecture.

The speech dataset employed in this work consists of unprocessed raw speech, which highlights the system's convenience and practicality in real-world applications. However, as noted by reviewers, this approach does introduce shortcomings in terms of SNR and recognition accuracy. In future research, we plan to preprocess the speech data to provide higher-quality inputs for model training.

5. More data on word sets and formal criteria used to evaluate speech recognition quality should be presented. Harvard sentences and various metrics, such as Levenshtein distance, are usually used for this.

Response: Thanks very much for the reviewer's valuable comment. To the best of the author's knowledge, there currently does not seem to be a standard version of the Harvard Sentences in Chinese. The selected Chinese speech in this work is from the AISHELL-1 corpus, containing 20 sentences with a total of 252 characters. It covers a wide range of pronunciations and phonemes, and is expected to achieve effects similar to the Harvard Sentences. Levenshtein distance essentially serves as the numerator in the CER formula. To enable a more comprehensive evaluation of speech enhancement performance, we have incorporated perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) metrics in hybrid training.

“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 8(b) and Figure 8(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.” has been added in the revised manuscript (Line 240, Page 10).

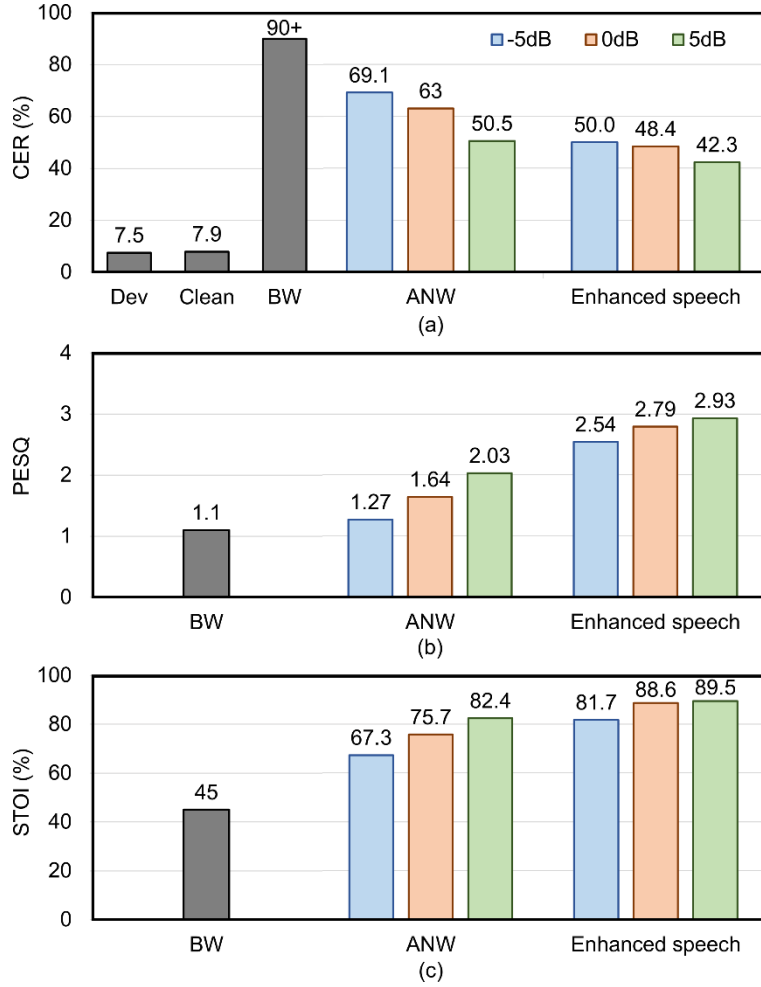


Figure 8. Comparison of (a) CERs, (b) PESQs and (c) STOIs for different datasets during hybrid training, including the validation set of the acoustic model, AW and BW datasets, as well as the ANW and enhanced speech datasets with various noise levels.

Additionally, in future study, we will expand the speech dataset to improve the generalizability of speech enhancement models.

6. Some figures are located directly at the end of sections, I would suggest moving them up, right after the first mention in the text.

Response: Thanks for the reviewer's kind comment. The location of the figures in the text has been optimized according to the suggestions.

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

36. Dorize, C.; Guerrier, S.; Awwad E.; Renaudier J. Capturing acoustic speech signals with coherent MIMO phase-OTDR. in *European Conf. Opt. Commun. (ECOC)*, 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.

R1. Shin, H. S.; Kang, H.; Fingscheidt, T. Survey of speech enhancement supported by a bone conduction microphone. in *Proc. IEEE ITG Conf. Speech Commun.*, 2012; pp. 1-4.