2024 IEEE International Conference on Consumer Electronics (ICCE) | 979-8-3503-2413-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCE59016.2024.10444140

Soft Speech, Loud World: Bone Conduction Microphones Enhance Voice Assistant Interaction

Chanel Manzanillo
BrainCo Inc.
Somerville, USA
cmanzani@gmail.com

Ralfy Chettiar
BrainCo Inc.
Somerville, USA
ralfy.chettiar@brainco.tech

Rahil Soroushmojdehi
UC Irvine
Irvine, USA
rsoroush@uci.edu

Leou Ying
UCLA
Los Angeles, USA
leouying020804@g.ucla.edu

Juewei Dong
BrainCo Inc.
Somerville, USA
juewei.dong@brainco.tech

Mostafa 'Neo' Mohsenvand
BrainCo Inc, MIT Media Lab
Cambridge, USA
mmv@mit.edu

***Abstract*—As the utilization of voice assistants becomes more widespread in daily activities, the demand for an interface capable of accurately recognizing speech at low volume levels within noisy environments is becoming increasingly important. In this study, we developed a custom device that incorporates a bone conduction microphone (BCM) by integrating a piezoelectric transducer with a noise-isolating impedance matching layer, alongside a microelectromechanical systems (MEMS) air conduction microphone (ACM). Our study aims to assess the BCM's effectiveness in facilitating soft speech communication with voice assistants in diverse noise environments and to validate its advantage over the ACM in noise reduction. We conducted experiments with participants using both the ACM and BCM to record audio samples for normal speech, soft speech, and whisper in three different noise level environments: ambient (quiet office setting), music, and loud noise. Spectrogram and signal-to-noise ratio (SNR) analysis assessed each audio file's signal quality. Additionally, we utilized an automatic speech recognition (ASR) model to estimate word error rate (WER) as a benchmark for performance comparison. In all tested scenarios, the BCM consistently demonstrated a significantly higher SNR performance compared to the ACM, as evident in the BCM's more distinguishable spectrograms for speech patterns in noisy environments. While the ACM's WER scores in the ambient environment were lower in all speech modes, the BCM outperformed the ACM in noisy conditions. Our findings confirm the custom BCM's ability to reduce background noise without requiring pre-processing or complex hardware while effectively capturing soft speech, emphasizing the potential benefits of integrating BCMs into interfaces designed for communication with voice assistants.**

***Index Terms*—bone conduction microphone, piezoelectric transducer, air conduction microphone, MEMS microphone, soft speech, speech recognition, voice assistant**

I. Introduction

In the realm of communication, microphones have served as a vital interface between humans and machines in applications such as voice assistants. Nevertheless, challenges persist in achieving accurate speech recognition, particularly in noisy environments, and the concern of privacy and inconvenience due to loud and obtrusive speech is prevalent [1]. For instance, envision a scenario where a traveler in a bustling airport terminal wishes to use their voice assistant to inquire about their flight's status. Due to the background noise, the voice assistant has difficulty recognizing the correct query and provides unrelated results. To overcome this, the user may need to raise their voice. However, not only might the user not feel comfortable sharing their flight information loudly in a crowded terminal, but the loud communication could cause unease to the nearby travelers. It is also worth noting that the requirement to speak loudly might pose challenges for certain users, such as the elderly and vocally impaired. This limitation can potentially hinder voice assistants from effectively providing assistance in these cases.

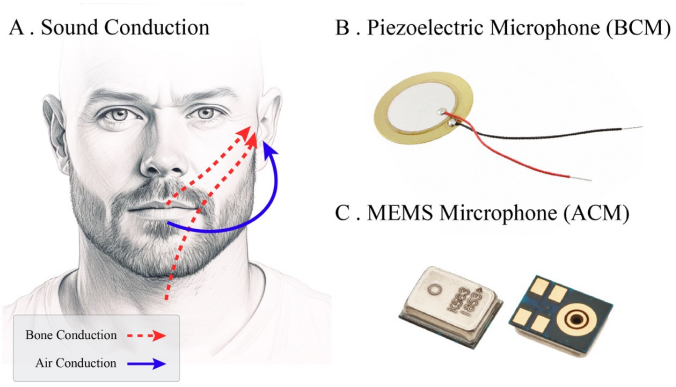

Fig. 1. A. Comparison of bone conduction and air conduction, B. A typical piezoelectric contact microphone, C. A typical MEMS microphone

A solution to these issues is to devise an interface that enables speech recognition even at lower volume levels amidst noisy surroundings. Detecting soft speech (40-50 dB [2], lower volume than conversational speech) and whisper (30-40 dB, lower volume than soft speech) in noisy environments poses significant challenges, and researchers have endeavored to address this issue through innovative approaches. Some studies have explored the use of beamforming [3] and neural speech enhancement techniques [4], [5] with conventional air

conduction microphones (ACM). While these methods show promise in enhancing speech intelligibility in noisy environments, they often require additional expensive hardware or involve computationally intensive pre-processing.

In recent years, the rise in popularity of bone conduction microphones (BCMs) can be attributed to their unique ability to detect vibrations from the user's skull, effectively minimizing interference from the surrounding environment [6], [7]. Fig. 1 illustrates the transmission of sound through both bone and air, featuring typical BCM and ACM sensors. Sectors such as the military and law enforcement [8], as well as various industrial settings [9], have adopted BCMs to ensure effective communication in loud and challenging conditions. Despite this growing adoption, a crucial aspect that remains unexplored is the potential of BCMs to recognize soft speech and whisper, particularly for interacting with voice assistants. Given the advancements in large language models and their integration with voice assistants, achieving high-quality, ubiquitous speech recognition is critical for future communication technologies.

In this study, our objective is to demonstrate the potential of BCMs as an interface for interacting with Automatic Speech Recognition (ASR) systems in challenging settings. Our methodology does not involve any pre-processing steps following data acquisition. This approach highlights the intrinsic capabilities of BCMs in comprehending spoken words in their raw form. Acknowledging the limitations of BCMs in capturing the human speech frequency spectrum [10], our aim is not to precisely replicate the ACM's wide frequency response. Instead, we focus on demonstrating the potential of the BCM's performance to detect soft speech in noisy environments without any pre-processing or noise cancellation compared to standard ACMs. By conducting various experiments, we examine users' speech in three distinct modes—normal, soft, and whisper—in the presence of three diverse noise environments, including a quiet office setting, a busy restaurant, and loud music. The results of this study could contribute valuable insights to enhance communication technologies further and extend the applications of BCMs in scenarios that demand clear and intelligible speech communication in noisy settings.

II. Materials and Methods

A. Custom Behind-The-Ear Wearable

In BCM devices, a commonly used sensor is the piezoelectric transducer. This sensor is composed of an element made from materials like quartz or specific ceramics. To enhance its performance, acoustic impedance matching is employed by adding a matching layer alongside the piezoelectric element ensuring the optimal transfer of sound waves from the skin to the sensor [11].

To collect and record speech, a custom behind-the-ear wearable prototype was created by incorporating a commercial 15 mm diameter piezo transducer [12] with an OP-AMP amplifier, a microcontroller unit, a microelectromechanical systems (MEMS) microphone [13], and a Bluetooth chip on a circuit board. The piezo transducer was paired with a custom silicon impedance-matching layer (diaphragm), 15 mm diameter, to capture vibrations due to speech. We opted not to introduce complex hardware modifications in our custom device and instead utilized the MEMS sensor without noise-canceling features typically found in commercial ACMs. This decision enabled us to conduct a direct evaluation of the BCM's performance in comparison with the ACM. The diaphragm was placed on the mastoid bone due to its high effectiveness in speech intelligibility from BCM placement studies analyzing speech transmission at different locations on the head [14], [15]. Fig. 2 displays the device and its placement behind the ear alongside a simple block diagram of the circuit board.

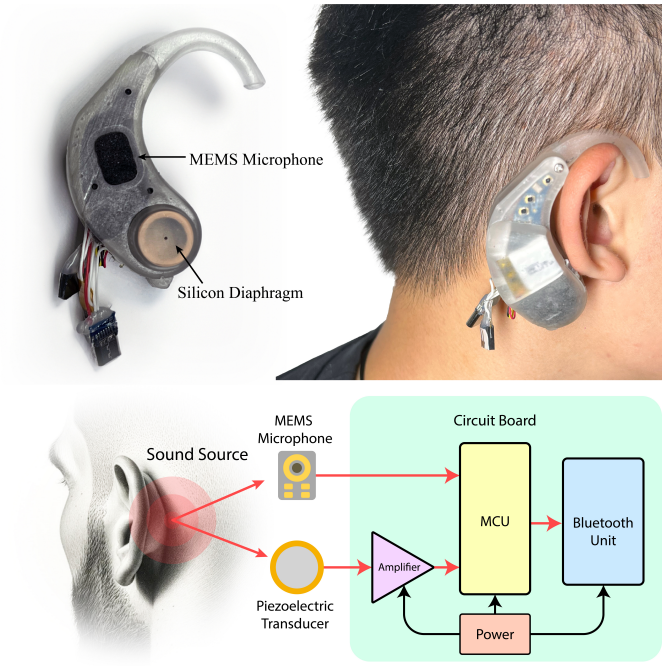

Fig. 2. Top Left: External view of the device (skin side) with diaphragm, Top Right: User wearing the device. Bottom: Block diagram of the device

B. Experimental Setup

Eleven participants (age 27.4 ± 3.2, 3 female) were recruited for the experiment. Out of the eleven speakers, eight participants were non-native American English speakers.

The test was performed in a small office room approximately 85 m^2. The participants sat in a chair in front of a desk. Two speakers [16] were placed on the desk to play background noise or music, depending on the test case. A laptop [17] played the audio through the two speakers and the video script containing the test phrases. The video script was projected on a monitor in front of the desk, visible to the participants. A decibel meter [18], placed approximately 30 cm in front of the participant's mouth monitored sound pressure levels (SPLs). The custom device was connected to a data acquisition laptop [19] installed with the proprietary recording software via Bluetooth. An image of the described setup is shown in Fig. 3, omitting the data acquisition laptop.

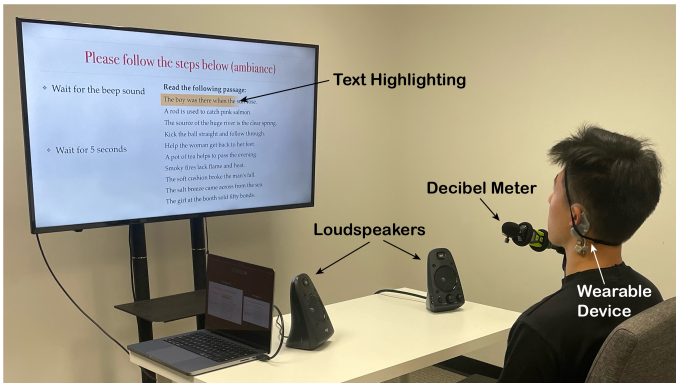

Fig. 3. The experimental setup showing the loudspeakers playing noise/music, a laptop that projects instructions to monitor, and the participant comfortably wearing the device in front of a display containing the instructions with the test phrase. SPL is monitored using a decibel meter.

The audio was collected in a static environment where participants were asked to minimize any external movement when speaking. Participants were instructed to wear the device on the left ear then place a headband over the device to ensure that the BCM maintained consistent pressure on the skin throughout the duration of the speech tasks.

To test speech intelligibility in the study, 'List 2' from the Harvard sentences [20], which features ten phonetically balanced sentences from a series of lists, was selected as the test phrase (Table I). Subjects read out the test phrase in three different modes of speech (whisper, soft, and normal) under three background noise conditions: in a quiet office environment (ambient noise), with an audio recording of piano music playing (background music), and with an audio recording of restaurant environment noise playing (background noise). SPLs of the speech modes and backgrounds are detailed in Table II. Prior to each case, participants performed a practice round for each speech mode without any background noise or music to acquaint themselves with the necessary audio range for their speech. During the test, a video was played to instruct the speaker when to begin speaking. The video highlighted the words on the screen that served as a guide for the speaking rate of each participant. A total of nine test cases were conducted resulting in a collection of 198 different audio files from eleven users with 99 BCM and 99 ACM audio recordings.

TABLE I
SPEECH PHRASES FROM HARVARD SENTENCES [20]

List 2
1. The boy was there when the sun rose.
2. A rod is used to catch pink salmon.
3. The source of the huge river is the clear spring.
4. Kick the ball straight and follow through.
5. Help the woman get back to her feet.
6. A pot of tea helps to pass the evening.
7. Smoky fires lack flame and heat.
8. The soft cushion broke the man's fall.
9. The salt breeze came across from the sea.
10. The girl at the booth sold fifty bonds.

TABLE II
AUDIO RECORDING TEST CONDITIONS

Speech Mode	SPL
Whisper	30 - 40 dB
Soft	40 - 50 dB
Normal	50 - 60 dB

Environment	SPL
Ambient	30 - 40 dB
Background Noise	65 - 75 dB
Background Music	75 - 85 dB

C. Data Evaluation

The audio recordings obtained from both the MEMS and the piezo transducer were parsed with OpenAI's Whisper.cpp (medium model) [21] without any prior pre-processing. Word Error Rate (WER) was employed based on the SpeechBrain library [22] to gauge speech intelligibility. WER quantifies the accuracy of speech recognition by comparing the number of substitution, insertion, and deletion errors between the recognized words and the reference transcript. It is a metric for evaluating the performance of speech recognition systems that assesses the precision of transcribed speech. Additionally, spectrograms were utilized to visualize the frequency content of the captured signals. Audio Signal-to-Noise Ratio (SNR), was assessed by calculating the ratio of the signal's power to the power of the background noise. For noise, a 2-second window from the initial 5 seconds of the recording was selected. Similarly, for signal, the first 2 seconds during the subject's speech were chosen. For group analysis and statistically comparing ACM and BCM measures, a Wilcoxon signed-rank test was performed as the data did not have a normal distribution.

III. RESULTS AND DISCUSSION

Upon data acquisition, the raw audio recordings were processed with the speech recognition model and WER was calculated. In Fig. 4, we present a comparison of WER across all speech modes in various noise scenarios for all subjects. The chart highlights statistically significant differences between the ACM and BCM groups. In the ambient office environment case, the ACM consistently outperformed the BCM across all speech modes. However, this trend reverses in noisy environments, where the BCM exhibits significantly higher speech intelligibility with lower WER scores. In the normal speech mode, the BCM achieves WERs under 20%. For soft speech, the average WER using the BCM was 45%, 35%, and 35% in ambient, music, and noisy environments, while the ACM had WER scores of 12%, 99%, and 100% in corresponding settings. The statistical significance of the differences between the ACM and BCM groups persists across all scenarios except for whisper mode in the music environment. In whisper mode and noisy environments, both the ACM and BCM exhibit diminished performance in speech intelligibility. This could potentially be addressed by incorporating signal pre-processing or amplification for the BCM.

In comparison to a related study using a piezo transducer as a BCM in normal speech, our experiments produced promising results. The V-speech system [23], which utilizes nasal vibration sensors in a glasses form factor, was tested

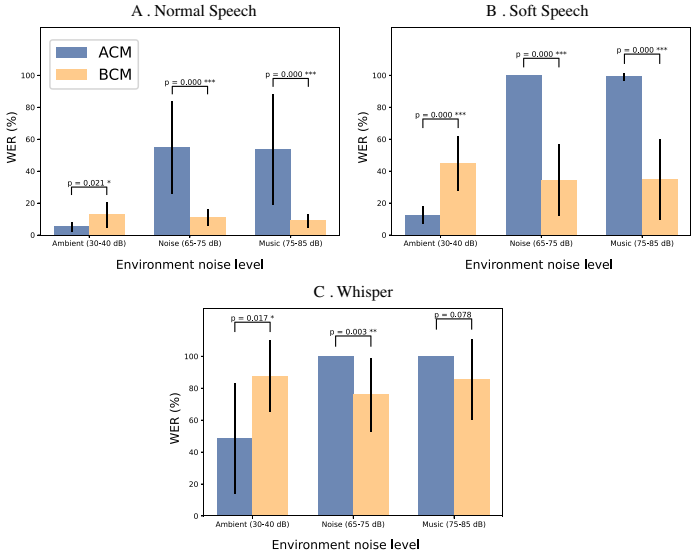

Fig. 4. WER comparison of the ACM and BCM during different modes of speech and different noise environments, A. Normal speech mode, B. Soft speech mode and C. Whisper speech mode. In all speech modes in an ambient environment, the ACM achieves lower WER scores than the BCM, while in noisy environments, the BCM outperforms the ACM in speech intelligibility. In whisper mode, both the ACM and BCM perform poorly in noisy environments.* indicates p-value <0.05, ** is for p-value <0.005, and *** is for p-value <0.0005.

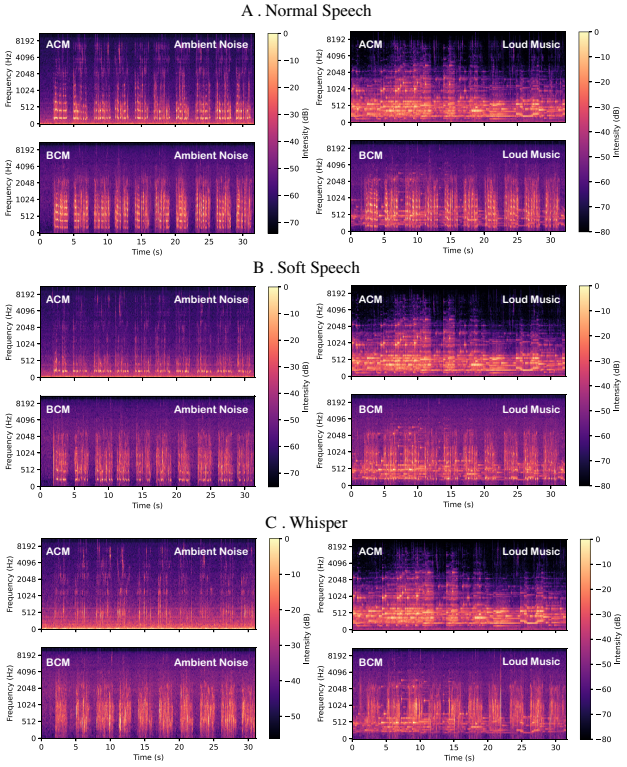

Fig. 5. ACM and BCM Spectrograms during the ambient and music environment of subject 11, A. Normal speech mode, B. Soft speech mode and C. Whisper speech mode. BCM recordings distinctly reveal speech periods even amidst background music, whereas ACM recordings fail to exhibit the same discernibility.

in both low noise (38 dB) and various high noise background conditions (79 dB, 93 dB, 81 dB, and 92 dB). When comparing the WER of their raw signal with our BCM results in low noise conditions for normal speech, our device achieved a lower WER of 12.7%, whereas the V-speech system received a WER of 21.1%. In high noise conditions, our device maintained improved performance, yielding WER scores of 9-11%, while they obtained WER scores of 22-24%. The disparity in performance may be attributed to the sensor placement as our device was positioned on the mastoid bone, while they placed their sensors on the nasal area. The V-speech study also included tests above our maximum background noise SPL (85 dB). To address nasal distortions, they employed an algorithm that improved their WER score to 14-18%, but still fell short of our results. Additionally, our implementation included an impedance-matching layer diaphragm, further enhancing signal transmission and contributing to our overall performance.

Spectrograms of the various noise and speech scenarios were analyzed to assess the audio quality captured by the BCM. Fig. 5 presents spectrograms of three speech modes in both ambient and music environments for Subject 11. Across different noise environments and speech modes, the BCM delineates visual speech periods, a contrast to the limited efficacy of the ACM in noisy settings. Based on ambient recordings (depicted in Fig. 5.A, Fig. 5.C and Fig. 5.E), the BCM covers a narrower speech frequency range than the ACM. In Fig. 5.E and Fig. 5.F during the whisper speech mode, the subject's BCM recordings effectively capture speech signals. Given this observation and the results in Fig. 4.C, introducing preprocessing measures to enhance speech intelligibility appears to be a viable and rational inference to further improve WER scores.

Fig. 6 addresses signal quality by comparing the audio SNR of all scenarios for all subjects. Although SNR decreases as the environmental noise level decreases, the SNR for the BCM had significantly higher SNR levels than the ACM in the cases with background noise and music.

IV. Conclusion

In this study, we assessed the efficacy of a commercial piezo transducer positioned on the mastoid bone, integrated into a custom behind-the-ear wearable device. We evaluated three speech modes in three diverse environmental conditions with a comparative analysis against an ACM (MEMS microphone), from audio recordings provided by eleven participants. Our findings demonstrate that BCMs have the capacity to enable intelligible communication in noisy settings, particularly for soft speech, using unprocessed audio. The BCM significantly outperformed the ACM in these scenarios for speech intelligibility with soft speech WER scores of 35% in the background music and noise environments. While the improved performance of BCMs in noisy environments is evident, there remains a larger implication for their inclusion in everyday communication technology. By incorporating the use of cost-effective microphone sensors in our custom design,

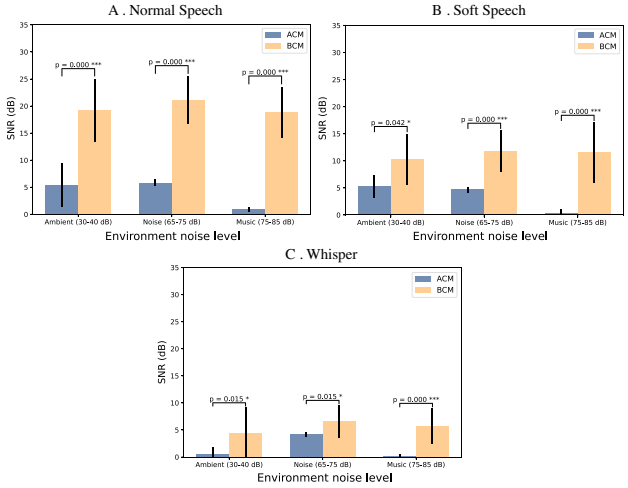

Fig. 6. SNR comparison during different modes of speech and environment noise, A. Normal speech, B. Soft speech and C. Whisper speech. In all scenarios SNR of BCM is significantly higher than ACM.

we anticipate increased accessibility to voice assistants for the broader public.

Compared with the ACM, the BCM failed to detect clear frequency signals exceeding 3000 Hz based on the generated spectrograms. Subsequent research endeavors could explore various combinations of piezo transducer geometries and impedance-matching layer materials to enhance the detection of high-frequency signals. An optimal solution may involve the integration of both an ACM and a BCM into a single device, and using the microphones in parallel to capture a broader frequency spectrum. As we explore the capabilities and limitations of both the ACM and BCM, machine learning offers another avenue for speech enhancement with possibilities ranging from discerning the most suitable microphone for varying scenarios of environmental noise to improving the audio signal in an ASR system. Algorithms trained on soft speech and whisper from BCMs can lead to the development of a robust speech detection system catering to a wide array of users. The promising results of BCM technology in improving speech intelligibility for low-volume speech are notable. Similarly, its effectiveness in background noise underscores their value in voice assistant interfaces. These developments promote more widespread integration and utilization of speech recognition in everyday activities and professional settings.

ACKNOWLEDGMENT

We thank Maoxing Liang, for his and his team's contributions in creating the circuit board used for the wearable device. We also thank Richard Ye for his assistance in executing the test protocol and collecting the recordings for data analysis, as well as the BrainCo team members who participated in our data collection procedures.

REFERENCES

[1] H. R. Marriott and V. Pitardi, *Opportunities and Challenges Facing AI Voice-Based Assistants: Consumer Perceptions and Technology Realities: An Abstract*, 2022, pp. 81–82.

[2] K. Hendrickson, J. Spinelli, and E. Walker, "Cognitive processes underlying spoken word recognition during soft speech," *Cognition*, vol. 198, 2020.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, pp. 1614–1626, 2001.

[4] D. T. Grozdić, S. T. Jovičić, and M. Subotić, "Whispered speech recognition using deep denoising autoencoder," *Engineering Applications of Artificial Intelligence*, vol. 59, pp. 15–22, 2017.

[5] N. Shah, N. Shah, and H. Patil, "Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion." ISCA, 9 2018, pp. 3157–3161.

[6] T. Letowski, T. Mermagen, N. Vause, and P. Henry, "Bone conduction communication in noise: A preliminary study," 7 2004, pp. 3037–3044.

[7] P. H. Skarzynski, A. Ratuszniak, K. Osinska, M. Koziel, B. Krol, K. B. Cywka, and H. Skerzynski, "A comparative study of a novel adhesive bone conduction device and conventional treatment options for conductive hearing loss," *Otology and Neurotology*, vol. 40, 2019.

[8] T. Letowski, P. P. Henry, and T. Mermagen, "Use of bone conduction transmission for communication with mounted and dismounted soldiers," 6 2005.

[9] S. Sugahara and M. Mori, "Feasibility of using air-conducted and bone-conducted sounds transmitted through eyeglasses frames for user authentication." IEEE, 6 2021, pp. 1–4.

[10] T. Toya, P. Birkholz, and M. Unoki, "Measurements of transmission characteristics related to bone-conducted speech using excitation signals in the oral cavity," *Journal of Speech, Language, and Hearing Research*, vol. 63, 2020.

[11] P. Henry and T. R. Letowski, *Bone conduction: Anatomy, physiology, and communication*. Aberdeen Proving Ground, MD: Army Research Laboratory, 2007.

[12] Uxcell piezo disc. [Online]. Available: https://www.uxcell.com/pcs-piezo-discs-10mm-acoustic-pickup-transducer-element-buzzer-cbg-guitar-p-1604991.html

[13] Memsensing mems acoustic sensor msm421a3729h9krmc. [Online]. Available: https://www.lcsc.com/product-detail/MEMS-Microphones_MEMS-MSM421A3729H9KRMC_C966935.html

[14] M. McBride, P. Tran, T. Letowski, and R. Patrick, "The effect of bone conduction microphone locations on speech intelligibility and sound quality," *Applied Ergonomics*, vol. 42, pp. 495–502, 3 2011.

[15] P. K. Tran, T. R. Letowski, and M. E. McBride, "The effect of bone conduction microphone placement on intensity and spectrum of transmitted speech items," *The Journal of the Acoustical Society of America*, vol. 133, pp. 3900–3908, 6 2013.

[16] Logitech z623 speaker system. [Online]. Available: https://www.logitech.com/en-us/products/speakers/z623-speaker-system-thx.980-000402.html

[17] Apple macbook pro. [Online]. Available: https://www.apple.com/shop/product/FKGR3LL/A/refurbished-14-inch-macbook-pro-apple-m1-pro-chip-with-8%E2%80%91core-cpu-and-14%E2%80%91core-gpu-silver

[18] Tadeto sound level meter. [Online]. Available: https://www.amazon.com/Tadeto-Portable-Backlight-Weighted-Factories/dp/B09T37812W?th=1

[19] Dell precision 7510. [Online]. Available: https://www.dell.com/support/kbdoc/en-us/000178440/dell-precision-7510-mobile-workstation-system-guide

[20] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[22] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[23] H. A. C. Maruri, P. Lopez-Meyer, J. Huang, W. M. Beltman, L. Nachman, and H. Lu, "V-speech," 12 2018, pp. 1–23.