## Check for updates

## WhisperMask: A Noise Suppressive Mask-Type Microphone for Whisper Speech

Hirotaka Hiraki The University of Tokyo Bunkyo, Tokyo, Japan National Institute of Advanced Industrial Science and Technology Kashiwa, Chiba, Japan hirotakahiraki@gmail.com

Manabu Yoshida National Institute of Advanced Industrial Science and Technology Kashiwa, Chiba, Japan yoshida-manabu@aist.go.jp Shusuke Kanazawa National Institute of Advanced Industrial Science and Technology Kashiwa, Chiba, Japan kanazawa-s@aist.go.jp

Masaaki Mochimaru National Institute of Advanced Industrial Science and Technology Kashiwa, Chiba, Japan m-mochimaru@aist.go.jp Takahiro Miura National Institute of Advanced Industrial Science and Technology Kashiwa, Chiba, Japan miura-t@aist.go.jp

Jun Rekimoto The University of Tokyo Bunkyo, Tokyo, Japan Sony Computer Science Laboratory Bunkyo, Tokyo, Japan rekimoto@acm.org



Figure 1: WhisperMask is a wearable mask-type microphone that captures only the user's voice even in noisy environments, such as in subway stations. The audio from WhisperMask has lower ambient noise(upper middle) than that recorded from other wearable microphones, such as Apple AirPods Pro2. In indoor setting, the WhisperMask did not pick up background noises during voice calls.

## ABSTRACT

Whispering is a common privacy-preserving technique in voicebased interactions, but its effectiveness is limited in noisy environments. In conventional hardware- and software-based noise reduction approaches, isolating whispered speech from ambient noise and other speech sounds remains a challenge. We thus propose WhisperMask, a mask-type microphone featuring a large diaphragm with low sensitivity, making the wearer's voice significantly louder



This work is licensed under a Creative Commons Attribution-NonCommercial International 4.0 License.

AHs 2024, April 04–06, 2024, Melbourne, VIC, Australia © 2024 Copyright held by the owner/author(s). ACM ISBN 979-8-4007-0980-7/24/04 https://doi.org/10.1145/3652920.3652925 than the background noise. We evaluated WhisperMask using three key metrics: signal-to-noise ratio, quality of recorded voices, and speech recognition rate. Across all metrics, WhisperMask consistently outperformed traditional noise-suppressing microphones and software-based solutions. Notably, WhisperMask showed a 30% higher recognition accuracy for whispered speech recorded in an environment with 80 dB background noise compared with the pin microphone and earbuds. Furthermore, while a denoiser decreased the whispered speech recognition rate of these two microphones by approximately 20% at 30-60 dB noise, WhisperMask maintained a high performance even without denoising, surpassing the other microphones' performances by a significant margin.

WhisperMask's design renders the wearer's voice as the dominant input and effectively suppresses background noise without relying on signal processing. This device allows for reliable voice AHs 2024, April 04-06, 2024, Melbourne, VIC, Australia

Hirotaka Hiraki, Shusuke Kanazawa, Takahiro Miura, Manabu Yoshida, Masaaki Mochimaru, and Jun Rekimoto



#### Mel spectrogram saying "Any sufficiently advanced technology is indistinguishable from magic"

Figure 2: Mel-spectrogram of the speech "Any sufficiently advanced technology is indistinguishable from magic." Displayed are two speaking styles: natural speech(left side) and whispered speech(right side) delivered in an environment with varying noise levels (30 dB and 80 dB) and recorded using different microphones. The topmost Mel spectrogram shows that the proposed microphone effectively captures whispers at 80 dB noise level.

interactions, such as phone calls and voice commands, in a wide range of noisy real-world scenarios while preserving user privacy.

## **CCS CONCEPTS**

• Human-centered computing → Mobile devices; Sound-based input / output.

## **KEYWORDS**

microphone, noise suppression, whispering, wearable devices

#### ACM Reference Format:

Hirotaka Hiraki, Shusuke Kanazawa, Takahiro Miura, Manabu Yoshida, Masaaki Mochimaru, and Jun Rekimoto. 2024. WhisperMask: A Noise Suppressive Mask-Type Microphone for Whisper Speech. In *The Augmented Humans International Conference (AHs 2024), April 04–06, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3652920.3652925

## **1 INTRODUCTION**

With the widespread use of smart devices and voice interfaces, on-the-go communication is becoming increasingly common. Microphones are now integrated into various wearable devices, such as AirPods[1], smart watches[47, 48], and smart rings[86]. However, using these interfaces in noisy environments poses technical challenges, as the wearer's voice can be interfered with by other speakers' voices and background noise[25]. The existing solutions to achieve a clear voice input in noisy environments include hardware advancements, such as the development of unidirectional microphones[4], throat microphones[33], and non-audible murmur microphones[67], as well as software-based approaches such as blind source separation (BSS)[10, 64] and speech enhancement[51, 84]. However, these solutions have limitations in terms of wearability, contact noise, and the ability to capture whispered speech (Table 1). Some products, such as HashMe[31] and Mutalk[32], aim to capture only the wearer's voice in noisy environments firmly attached around the mouth. However, they limit face-to-face conversations and they can be bulky. Microphones with built-in ventilation fans[19] have also been proposed, but they generate significant noise for those nearby. To address these issues, we propose WhisperMask, a lightweight mask-type microphone with a large, soft diaphragm made of conductive fabric. Unlike the existing mask-type microphones that are firmly attached around the mouth. WhisperMask allows for natural breathing and speaking while effectively suppressing background noise without relying on signal processing. We evaluate WhisperMask using three metrics, namely, signal-to-noise ratio (SNR), audio quality, and speech recognition rate, to demonstrate its superior performance compared with that of conventional noise-suppressive microphones and software-based solutions. Our main contributions are as follows:

- Development of WhisperMask, a wearable, noise-suppressing microphone that captures whispered speech in environments with up to 80 dB ambient noise.
- Acoustic characterization, which revealed that WhisperMask provides a 10 dB advantage to the wearer's voice over any external noise ranging from 200 Hz to 5 kHz.

Microphone	Principle of Noise Reduction	No background noise	Wearable	No contact noise	Whispering
Unidirectional Mic	Directional Vibration	Δ	$\checkmark$	$\checkmark$	$\checkmark$
Array Mic	Beamforming	Δ		$\checkmark$	$\checkmark$
Throat Mic	Piezoelectric	Δ	$\checkmark$		
Nam Mic	Contact Microphone	0	$\checkmark$		
Earbuds Mic	Deep Learning	Δ	$\checkmark$	$\checkmark$	
WhisperMask	Large Electlet Diaphragm	0	$\checkmark$	$\checkmark$	$\checkmark$

Table 1: Types of microphones based on design principle. WhisperMask is wearable, eliminates background or contact noise, and capable of capturing whispered voices even in noisy places.

- SNR evaluation, which showed that WhisperMask outperforms the existing microphones by 10 dB in environments with 70 dB noise.
- Audio quality assessment, which demonstrated Whisper-Mask's superiority over the other microphones and its comparability to noise reduction software.
- Speech recognition evaluation, which showed that Whisper-Mask achieved a 30% higher recognition rate than a noise suppression software for both normal and whispered speech in noisy conditions.

In this study, we explore a new microphone diaphragm shape and a mask-type interface that captures only the user's voice without sealing the mouth, enabling reliable voice interactions in noisy real-world scenarios while preserving user comfort and privacy. As shown in Fig 2, the proposed WhisperMask effectively captures both normal speech and whispered speech in high-noise environments, demonstrating its potential for various voice interaction applications.

## 2 RELATED WORK

Voice interaction is one of the most important interaction modalities that people can engage in, and it is used in various applications such as telephone and online calls, voice command input, and interactive operation with smart assistants. In these applications, clear speech input is important; equally important is a stable voice interaction even in environments where other people are talking or where a high noise level exists, such as in subways or construction sites. To achieve clear speech input, many approaches have long been proposed in the field of speech, signal processing, and interaction, both from hardware and software perspectives.

One particular approach is increasing the number of microphones, allowing devices to become more directional and select toward a user's voice [14]; however, such devices are not suitable to become wearable due to their large size. In this section, we present the existing approaches that facilitate voice interaction designed to be wearable and operate in real-time.

#### 2.1 Wearable noise-suppressing microphones

In terms of hardware technology, noise suppression is achieved by combining the principles of sound pickup around which various microphones are designed, including unidirectional microphones [4], throat microphones [33], NAM microphones [67], and earphonetype microphones[9]. 2.1.1 Unidirectional microphones. Unidirectional microphones, such as those found in pin microphones and headsets, bear one of the most readily available noise reduction technologies today. Noise reduction is achieved by blocking the direction of vibration of the diaphragm, from the back side and restricting it to one side, thereby narrowing the directivity to 180 degrees; note that the diaphragm is found inside the microphone[4]. This allows for strong recording of sound in a limited direction relative to the microphone, which is important when multiple speakers are speaking at the same time, such as in a panel discussion. However, since these microphones do not limit the distance of sound, background noise cannot be removed by the microphone by itself, making them unsuitable for use in noisy environments.

2.1.2 Array microphones. Array microphones are equipped with multiple microphones, and beamforming[10] is used to narrow down the direction of arrival of sound by taking advantage of the time difference between the arrival of emitted sound at each microphone. This allows the selection of the speaker of an utterance. However, since multiple microphones which are arranged either in a circular fashion or in a straight line, are required to narrow down the direction of sound, using this system in a wearable device is challenging. Furthermore, even if the direction of arrival of sound can be estimated, it is not possible to determine whether it is background noise or not, requiring post-processing to deal with background noise.

2.1.3 Throat microphones. A throat microphone uses a piezoelectric element to convert the vibrations that appear on the surface of the neck when a speech is uttered [33]. By attaching the microphone onto the neck and acquiring only the surface sound, only the wearer's voice is collected; environmental noise cannot cause the piezoelectric element to vibrate sufficiently, resulting in a high immunity to background noise. However, because the device must be worn tightly around the neck, noise is generated by movements such as head shaking or nodding. Moreover, because the voice travels through the skin, the formants necessary for speech recognition are deficient, and thus post-processing is required to ensure audible and accurate speech recognition[35, 79].

2.1.4 NAM microphones. A NAM microphone works when it is in direct contact with the skin behind the ear similar to a pharyngeal microphone; it is an audio input device wherein an omnidirectional microphone is directly attached to the skin through a silicon [4]. Similar to a pharyngeal microphone, a NAM microphone greatly reduces the effect of ambient sound[67], but nodding and other

sounds become noise. However, unlike pharyngeal microphones, NAM microphones acquire sound by being worn behind the ear, although both microphones require post-processing such as speech enhancement[28, 52, 83].

2.1.5 Earbuds with microphone. An earpiece-type microphone that is inserted into earphones has been proposed, to achieve clear speech input, with earphones attached to both ears. Earbuds are also used to collect health information by monitoring exercise and biometric data [60]. Earpiece microphones are beamformer systems involving the earphones mounted on both ears, allowing selective acquisition of the wearer's voice, moreover, the hardware is open source, facilitating its faster development [61]. Machine learning methods have also been proposed to reduce noise from speech detected in both ears[9]. However, these methods are difficult to implement for voices that are not produced clearly, such as whispers.

## 2.2 Software approaches for noise reduction: speech enhancement and blind source separation

In the field of speech signal processing, the use of BSS [10, 64] and speech enhancement[51, 84] have long been proposed. In BSS, the resolution of the space where voices propagate is increased by increasing the number of microphones [14], and the independence of multiple speakers from each other is used as a criterion for source separation.

In BSS, various methods have long been proposed such as increasing the number of microphones to increase the resolution of the space where voices propagate [14], separating sound sources based on independence [37], and decomposing matrices into lower dimensional matrices[11, 12]. Other approaches [40, 50] combine these methods with deep learning. Furthermore, methods for speaker separation and speech enhancement that operate on small models have been proposed [18, 68, 70] and they work in real-time. Real-time sound source separation involves a process known as short-time Fourier transform (STFT), which achieves a quick decomposition time. This process, however, reduces the frequency resolution necessary for high-quality synthesis. To address this issue, learning methods have been introduced. These methods propose an evaluation function designed to maximize sound quality, thereby enhancing synthesis quality[41].

However, these machine learning-based methods demonstrate a limited generalization performance because they are based on specific English speaker data sets or specific noise data sets. For example, if a speaker is placed in front of a person's mouth and the exact same voice is played, the voice from the speaker will be misinput. Also, because the system is optimized for normal speech data, there are deviations from data in real environments, such as input from whispered voices or from whispered voices produced in a noisy environment.

# 2.3 Silent speech interface for speech communication

Silent speech, which facilitates interaction with a non-vocal input, such as lip image recognition as well as with speech, has been

proposed [16, 23]. Silent speech uses not only the voice emitted from the vocal cords, but also lip image [54, 69, 71], ultrasound image[15, 30, 39], myoelectricity[34, 80, 81], capacitance[38, 46], acceleration[29, 36, 59], strain[43], magnetism[87], EEG[8, 55], and other modalities to measure the human activities leading to speech recognition. Since these methods do not emit sound, they are suitable for use in noisy environments (e.g., where many people are talking)

However, silent speech input has lexical challenges in that it is not conversational and it faces interaction challenges such as wearability and hands-free input. SilentSpeller [38] demonstrates a high performance in terms of vocabulary but is not very expressive given that the current vocabulary for speech recognition is 9 million words. Meanwhile, a large lip image data is available for lip image input, that can be used in dark environments or as scalable command recognition [54], but the limitations of the commands remain unclear. Moreover, users must face their smartphones to speak, making hands-free input difficult, which is possible with voice. Myoelectricity is envisioned for use in actual conversations, such as synthesizing speech from silent speech [17]. However, myoelectricity requires the attachment of myoelectric array electrodes onto the surface of the face, making it unsuitable for prolonged use, and it does not take into account the effects of walking and other movements.

Silent speech with breathing has also been proposed in the speech modality [24], but it requires learning the interaction of breathing in; moreover, its vocabulary is smaller than that of the touch sensors and updated images described above. Input by non-audible murmur has also been proposed[28, 52, 83], but it has the problem that noise is generated by the user's natural movements, such as nodding or turning around, as well as the noise produced by touching the NAM microphone, are the identified problems.

## 2.4 Mask-type interface

The increased use of masks during the COVID-19 pandemic has led to a surge of interest in exploring their potential as wearable interfaces. While masks conceal facial expressions thereby posing a challenge, researchers have proposed methods to capture and present expressions using photo reflectors, capacitive touch sensors, LEDs, and displays [26, 42, 45, 53]. Furthermore, masks have been adapted as wearable interfaces for various interactions involving the face and mouth, including breath detection [44, 74, 75], eye tracking [5], mouth shape recognition [63, 72], and mask attachment/removal detection using the straps [82].

Notably, masks have been reported to cause voice attenuation, making speech perception difficult [73, 76]. To address this issue, researchers have proposed embedding sensors in masks to enable silent speech recognition [29, 43]. However, these approaches primarily focus on recognizing specific commands for communication with smart assistants such as Alexa, limiting their applicability for broader speech recognition and conversation capabilities.

While products such as HushMe [31] and Mutalk [32] offer voice isolation in noisy environments by fully sealing around the mouth, they are bulky and are intended primarily for gaming applications, limiting face-to-face communication. Similarly, microphone devices with built-in ventilation fans [19], although lightweight and sleek,

AHs 2024, April 04-06, 2024, Melbourne, VIC, Australia



Figure 3: Overview of WhisperMask as a masked microphone. WhisperMask is a mask-type microphone that allows for hands-free, non-obtrusive input (right). The microphone is sandwiched between the fabric of two non-woven masks (center). The diaphragm is connected to a microcontroller and can be used on a PC or smartphone via USB or Bluetooth (right).

necessitate constant fan operation to ventilate the sealed area, leading to considerable noise pollution.

Our proposed solution, WhisperMask, addresses the need for voice isolation with the use of a single lightweight mask, thus offering a practical alternative. Importantly, the hands-free and nonobtrusive nature of WhisperMask makes it particularly well-suited for environments where mask usage remains prevalent, such as operating rooms[66], clean rooms, or other noisy settings where clear communication is crucial. By effectively suppressing background noise while preserving user comfort and privacy, WhisperMask enables reliable voice interactions in these challenging scenarios.

## 3 WHISPERMASK, AN ELECRET CONDENSER MICROPHONE(ECM)-BASED MASK-TYPE MICROPHONE

We propose a wearable mask-type microphone called WhisperMask, which allows for input even in noisy environments. WhisperMask is designed based on the principle of the electret condenser microphone (ECM) and incorporates a vibrating diaphragm designed using conductive fabric and film.

#### 3.1 Principle: Electret Condenser Microphone

WhisperMask is a microphone designed based on the ECM technology. An ECM consists of a power supply section and a vibrating diaphragm, where the diaphragm forms a capacitor with the electrode, giving rise to a microphone. When a sound causes the diaphragm to vibrate, the voltage of the capacitor changes, resulting in a weak voltage that is converted into sound through a field effect transistor(FET) and through analog-to-digital conversion. In this study, we propose a novel design of a vibrating diaphragm for ECM, enabling selectivity toward the speaker's voice.

## 3.2 Design of WhisperMask

The diaphragm of our microphone consists of a conductive fabric electrode and a dielectric plastic film, forming an electret condenser. The plastic film used is PFA (perfluoroalkoxy polymer), with a thickness of 12.5  $\mu$  m. For wind protection, an adhesive tape was applied around the film, and a 200  $\mu$  m silver-plated polyester cloth was fixed onto both sides of the tape. This configuration forms the structure of an electret condenser wherein the electret was sandwiched between two electrodes.

The diaphragm was connected to an FET, which outputs the audio signal. Compared with commercially available condenser microphones, the proposed microphone has a lower voltage value, which is amplified by a factor of 3 using FETs. Furthermore, the input signal is connected via Bluetooth to a Bm63 audio system-on-a-chip (SoC), which is housed in a 6 mm x 5 mm housing and weighs less than 10g.

After producing the diaphragm and the measuring circuit, we embedded them in a breathable mask created from a mesh-type fabric. The mask was designed to bear slits to allow for the positioning and fixation of the diaphragm.

#### 3.3 Patterns of the diaphragm

WhisperMask is a mask-type microphone, which is used by wearing a mask. However, if the sensor becomes heavy, the mask will shift and fall off, resulting in decreased recognition. Therefore, in this study, we prepared three sizes of microphones (20 mm×4 mm, 20 mm×2 mm, and 10 mm×2 mm, Fig. 5), and we embedded them in a mask for assessment (Fig. 3).

## 4 FREQUENCY CHARACTERISTIC OF WHISPERMASK

Frequency characteristic is one of the most critical aspects of a microphone, as it determines the microphone's ability to capture specific frequencies and to accentuate or attenuate certain frequency bands. For example, in voice recognition, which is crucial in voicebased interactions, human speech predominantly falls within the range of 5 kHz, and a microphone that does not capture this frequency range would make voice recognition extremely challenging. Additionally, it is important to consider how clearly a signal stands out against noise, which is measured in terms of SNR. A low SNR value means that a significant part of the obtained signal is noise,



Hirotaka Hiraki, Shusuke Kanazawa, Takahiro Miura, Manabu Yoshida, Masaaki Mochimaru, and Jun Rekimoto



## SweptSine signal for measuring Impulse Response Hardware Setup

Figure 4: Swept-sine wave for measuring impulse response. Five Swept-sine of length 65536 are generated(left). A dummy head with a voice simulator is provided and the microphones are worn for measurement. The speaker is placed 500 mm away from the dummy head(upper right). The impulse response is calculated by convolving the signal obtained by preprocessing to detect Swept-sine with the inverse filter of Swept-sine. (lower right)

making it difficult to capture the desired signal, such as the speaker's voice.

### 4.1 Impulse response with swept-sine

4.1.1 Characteristic of the swept-sine signal. Frequency characteristics can be measured using the microphone's impulse response, which in turn can be measured using various methods, such as using an impulse wave or white noise. However, the widely employed method involves swept-sine, a continuous sine wave that sweeps from low to high frequencies[21]. Swept-sine measurement provides information across a wide frequency range using a single signal. It is robust against noise and relatively easy to measure. Furthermore, the deterministic properties of a swept sine signal make it less noisy compared with measurement methods involving white noise. By taking measurements repeatedly, random noise can be reduced, enabling the acquisition of accurate and reproducible values. The waveform obtained from swept-sine represents the convolution of the system's impulse response and the swept-sine itself. To obtain the microphone's impulse response from the acquired signal, an inverse convolution filter can be applied during the generation of a swept-sine signal (Fig. 4, right).

4.1.2 Design of the swept-sine wave for measurement. A swept-sine is capable of resolving different frequencies based on the length of a sample, with longer samples providing higher a frequency resolution. In this study, we used a sample length of 65536 points to achieve a frequency resolution of less than 1 Hz when the sampling frequency was set at 44.1 kHz. The frequency range of the swept-sine was limited to up to 22.1 kHz, which corresponds to the upper limit of human auditory perception. As mentioned earlier, repeating the swept-sine measurement improves the SNR; thus, conducting

multiple measurements is crucial. Therefore, a swept-sine consisting of 65535 samples was created and the process was repeated five times to generate a single wav file (Fig. 4 left). However, to identify the starting positions of each swept-sine, a buffer of the same length as the swept-sine was inserted between consecutive swept-sines. By including these gaps, the input consisted of five repetitions of the swept-sine, which was then repeated 10 times, resulting in a total of 50 measurements of the impulse response. This averaging process was carried out to assess the system's noise robustness. Finally, we use 1/3 octaveband averaging [20] to smoothen the impulse response.

## 4.2 Experimental condition

4.2.1 Environmental setup for measuring impulse response. For the impulse response measurement, a dummy head (SAMAR4700M) equipped with a mouth simulator was used to mimic human speech output. The SAMAR4700 complies with the international standards IEC 60318-7 and ITU-T Rec.P51 for head shape and mouth simulator, enabling the simulation of human speech for measurements. The dummy head was positioned on a tripod 40 mm above the floor, and a speaker that replicates the emitted sound was fixed 50 mm in front of the dummy head.

The acoustic measurements were conducted in an electromagnetic anechoic chamber to minimize interference from electromagnetic waves and certain sound waves. The room was treated with porous materials to absorb sound, resulting in a room noise of 28.8 dB.

*4.2.2 Input signal.* During the measurements, a dummy head simulating human speech and a speaker imitating external noise were set up. To avoid distortion of the output sound, 80 dB swept-sine



Figure 5: Sizes of the WhisperMask diaphragm: 20 mm×4 mm(upper), 20 mm×2 mm(center), and 10 mm×2 mm(lower)



Figure 7: Impulse response of WhisperMask(20 mm×2 mm)

was output as a human voice from a mouth-simulating speaker. Two swept-sines (80 and 60 dB) were output as background noise from the external noise speaker (Fig. 4 right upper). These sound levels were measured with a precision noise system, and the obtained values fell within an error of 0.1 dB.

#### 4.3 Swept-sine detection

The output of the swept-sine is susceptible to variations in frequency characteristics depending on the properties of the receiving microphone. In other words, it is impossible to acquire the full range of frequency bands, and in some cases, only a partial representation is obtained. However, to accurately measure the impulse response, determining the timing at which the swept-sine signal begins is crucial. To estimate the start and end times of the swept-sine from the information obtained within certain frequency bands, the following approach was employed. Firstly, the swept-sine was decomposed into individual frequencies at 1 kHz intervals, and the envelope was acquired. As the swept-sine was output five times with an interval of one pulse, five rising edges appeared in the frequency regions where the swept-sine was well represented. By capturing these, five points were recorded. Dividing the frequencies at intervals of 1 kHz, a maximum of 22 points (22 kHz/1000) was recorded. Using



Figure 6: Impulse response of WhisperMask(10 mm×2 mm)



Figure 8: Impulse response of WhisperMask(20 mm×4 mm)

these points, a linear regression was performed, with the point at frequency 0 representing the start time and the point at frequency 22000 representing the end time.

#### 4.4 Result for impulse response

The impulse response results are shown in Fig. 6, 7, 8. The blue line represents the impulse response of the assumed human speech, played from the mouth-simulating speaker of the dummy head, with 80 dB swept-sine. The orange and green lines are both assumed to be noise and were output at 80 dB and 60 dB, respectively, from the external speaker.

In each of the three patterns, especially in the frequency band between 200 Hz and 5 kHz, the output of the dummy head was approximately 10 dB higher than the outside noise, indicating that when the same waveform at the same sound pressure is input inside (dummy head) and outside (noise simulating speaker) the microphone, the microphone captures the inside sound in more easily, that is, it reduces noise.



Figure 9: SNR result for different microphones at 60 dB input

## 5 EVALUATION: NOISE SUPPRESSING MICROPHONE

## 5.1 Evaluating the effect of noise suppression: SNR measurement

The proposed microphone can capture a speaker's voice in noisy environments. To demonstrate this, the SNR was measured. The SNR was computed by recording the output (N) of the microphone without supplying input signals from the speaker and by recording the output (S) when input signals were generated. The root mean square (RMS) values were then calculated for each microphone output, and the SNR was determined using  $20log10(S_{RMS}/N_{RMS})$ [7]. SNR evaluation is commonly used to assess the performance of microphones in noisy environments and is also utilized in array microphones [6].

#### 5.2 Environmental setup for SNR measuring

For SNR measurement, a dummy head(SAMAR4700M) equipped with a mouth simulator was used to mimic human speech output. The SAMAR4700 complies with the international standards IEC 60318-7 and ITU-T Rec.P51 for head shape and mouth simulator, enabling the simulation of human speech for measurements. The dummy head was positioned on a tripod 40cm above the floor, and a speaker that replicates the emitted sound was fixed 50cm in front of the dummy head.

Acoustic measurements were conducted in an electromagnetic anechoic chamber to minimize interference from electromagnetic waves and certain sound waves. The room was treated with porous materials to absorb sound, resulting in a room noise of 28.8 dB.

The input signal used for the measurement was a 20 Hz–20 kHz swept-sine to also perform calculations in the full range. The sound pressure of the signal output from the dummy head was 60 dB, which is close to that of human speech. Environmental noise was output from an external speaker in 10 dB increments from 30 dB to

90 dB, with an upper limit of 90 dB based on the characteristics of the speaker's output.

Three devices were used for measurements: WhisperMask (10 mm x 4 mm), pin microphone(PinMic), and throat microphone(ThroatMic). The earbuds with microphone (Airpods) were optimized for human voices and could not input noise-like waveforms similar to those of the swept-sine. Each microphone was mounted on a dummy head in an ideal position on the dummy head.

5.2.1 SNR results for noisy environment. The inputs in the noisy environment were as follows: the SNR for PinMic and ThroatMic decreased as the ambient noise increased, whereas that for Whisper-Mask hardly changed from 30 dB (SNR: 22.2) to 60 dB (SNR: 21.3) of ambient noise. SNR was calculated as the ratio of the power of the signal to the power of the noise:  $SNR = 20log10(S_{RMS}/N_{RMS})$ . And since the input signal( $S_{RMS}$ ) is almost constant at 60 dB swept-sine, from 30 dB to 60 dB. WhisperMask picked up almost no ambient noise because  $N_{RMS}$  is nearly constant.

The impact of noise on WhisperMask's performance was greater when the external noise was higher than 70 dB; at 70 dB, the SNR for WhisperMask was 17.83, which was 10 dB higher than the SNR values for PinMic(2.3) and for ThroatMic(5.83).

## 6 EVALUATION: QUALITY OF THE RECORDED VOICES

Widely used for voice input and calls, microphones require not only noise resilience but also high-quality sound during recordings. We evaluated the sound quality in noisy conditions to compare WhisperMask not just against existing microphones but also against the performance of conventional microphones following the use of a noise reduction software.

## 6.1 Record conditions

The audio recordings used in this study were of five phrases extracted from the Text Entry dataset[49], and read aloud by four proficient English speakers (one male and three females). These WhisperMask: A Noise Suppressive Mask-Type Microphone for Whisper Speech



Figure 10: The WebUI used to evaluate the quality of the recorded audio clips; the metrics were based on MUSHRA[2]. The participants rated each audio clip on a scale of 0 to 100. The reference audio clip (top row) had a high quality and was used as a criterion for selecting responses. Four of the five test items are recordings captured in a noisy environment by each device. One test item is the same as the reference, any participants who rated this test item as having lower quality than the reference audio will be judged as less faithful respondents.

speakers consented consented to participate in this study and become subjects for data collection, for which they received \$20. The noise level during the recording was 80 dB, and the four microphones used for recording were WhisperMask, unidirectional pin microphone, earbuds with microphone (Airpods Pro2), and throat microphone. The reference audio is a clear sound recorded in a quiet environment (30 dB) with a pin microphone.

To compare with software-based noise reduction, we created recordings by applying noise removal software to the captured audio clips. There are two main approaches to noise removal in noisy environments: noise removal and speech enhancement. These approaches are not always explicitly compared in speech research. Therefore, for comparison, we choose one method from each approach that utilizes a learning-based embedding model for comparison. For speech enhancement, we used a denoiser [18]. Denoiser is an extension of U-Net [62] and is trained on the noisy speech dataset [57, 77]. For speech separation, we use a waveformer [78]. A waveformer is a model that extends CNN to handle sequential data and is trained to synthesize and separate sounds from a selected speech dataset [22] and an environmental sound dataset [27].

## 6.2 MUSHRA: a metrics for evaluating audio quality

For subjective evaluation of sound quality, the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) has been proposed. Compared with Mean Opinion Score(MOS), MUSHRA is defined in ITU-R BS.1534 and is evaluated on a scale of 0 to 100, allowing for

spoc

Quality Rating

20

WhisperMask

Figure 11: Assessment results for the quality of audio recorded by WhisperMask and other devices in an environment with 80 dB background noise. "Pinmic," "pods," and "throat" in the x-axis represent unidirectional pin microphones, Airpods Pro2, and throat microphones, respectively. The reference is a clear sound recorded in a quiet environment(30 dB) with a pin microphone.

the evaluation of subtle differences. Multiple test items were provided to the evaluators among these items is an "anchor" wherein the sound quality was intentionally reduced to provide a reference for the quality evaluation. This approach ensures consistent evaluations [2].

MUSHRA evaluations have been used in a wide range of applications in speech processing, including noise reduction [3], text-tospeech synthesis [85], and voice transformation [58]. It was also been used for evaluations by the participants of an online experiment [65].

### 6.3 Experimental proceedure

We recruited 25 native English speakers (13 males and 12 females) aged 18 and above through Mechanical Turk. The audio quality was evaluated using WebUI, as shown in Fig. 12.

In the WebUI, the participants would rate each audio clip on a scale of 0 to 100; a high-quality audio clip was provided as a reference. One of the five choices was the same as the reference, enabling the identification of less faithful respondents. The participants may listen to the sound clips as many times as they wished. The average completion of the participants in the experiment was 49 minutes. They were compensated \$7 for their participation.

#### 6.4 Result of the audio quality evaluation

The results are shown in Fig. 11. "Pinmic," "pods," "throat" represent unidirectional pin microphone, Airpods Pro2, and throat microphone, respectively. The t-test results show that WhisperMask was superior to Pinmic, pods, and throat at p=5.06E-22, 2.95E-10, and

:hroat

Reference



Figure 12: Audio quality of the recording captured by WhisperMask following post-processing with denoiser [18], a speech enhancement software.

1.50E-28 (p<0.05), respectively, and with corresponding effect sizes of 1.20, 0.75, 1.43 [13].

## 6.5 Comparison of sound quality with and without noise reduction

For the experiment, we used MUSURA[2] as a metrics, and we recruited 25 native English speakers (13 males and 12 females) aged 18 and above through Mechanical Turk. The results are as shown in Fig. 11. As mentioned above, "pinmic", "pods", and "throat" represent unidirectional pin microphones, Airpods Pro2, and throat microphones, respectively. The t-test result in Fig. 12 show that WhisperMask was superior to pinmic, pods, and throat, at p=1.31E-4, 4.51-08, and 1.83E-22 (p<0.05), respectively with corresponding effect sizes of 0.455, 0.659, and 1.257 [13].

By contrast, when sepformer was used (Fig. 13), pin microphone was superior to WhisperMask at p=0.02 in t-test (p<0.05) and 0.27 in effect size. The performance of the other devices (AirPods and throat microphone) did not statistically differ from that of WhisperMask (p=0.78, 0.45 in t-test).

Furthermore, the quality of recording captured by WhisperMask with and without denoiser or sepformer did not significantly differ (p=0.575>0.05), and it was better than the sound quality obtained using sepformer (p=8.74E-07<0.05; effect size 0.583

## 7 EVALUATION: SPEECH RECOGNITION ACCURACY

Speech input is widely used not only in telephony but also for interactive tasks, such as operating smart assistants and interactive searching using speech recognition. In this study, we performed speech recognition using two noise-robust speech recognition methods. Whisper [56] employs an encoder-decoder transformer model trained through supervised learning.

## 7.1 Data Collection

Data were collected from 9 participants (mean age 26.2 years; 4 males and 5 females). The participants were asked beforehand to



Figure 13: Audio quality of the recording captured by WhisperMask following post-processing with sepformer [70], a speech separation software.

rate their English proficiency on a 5-point scale, with 3 being the mean. During data collection, each participant read 20 pre-prepared phrases collected from the Mackenzie and Soukoref dataset [49].

In the experimental environment, white noise was varied at 40, 60, and 80 dB (denoted as w40, w60, and w80 dB in Fig. 14, respectively). a30 dB indicates the noise level in the space when no white noise was being played. During the experiment, two methods of delivery were used: natural speech and whispered speech. The three microphones evaluated were WhisperMask, pin microphone, and earbuds with microphone (AirPods Pro2).

The background noise was played from a laptop and was output in a stereo; it was measured using a noise system to ensure that the desired sound pressure was reaching the user's mouth and then adjusted to a difference of 0.5 dB or less. Measurements were taken in a soundproof room where the normal noise level was 30 dB. In each microphone and noise environment, two types of speech were used: natural speech and whispered speech

#### 7.2 Analysis

In speech research, speech enhancement methods have been proposed to reduce background noise [51, 84]. In recent years, real-time noise reduction systems have become highly accurate [51, 84], and it is already possible to apply noise reduction in microphones, such as pin microphones and AirPods, to obtain the desired audio quality. This study examines how much speech recognition accuracy can be improved relative to the audio quality obtained after applying real-time denoising to speech recorded with a pin microphone or AirPods. In a noise-free environment (30 dB), the participants' average recognition rate for natural speech recorded using pin microphones and AirPods was over 90%. The voice recognition rate for throat microphone was significantly lower at 64% compared with that for the other devices. This discrepancy was attributed to improper fitting for some participants, leading to their exclusion from this consideration. Speech recognition was performed using Whisper large [56], a transformer-based speech recognizer with a strong language model. Speech recognition was evaluated based on the percentage of correct answers per character

WhisperMask: A Noise Suppressive Mask-Type Microphone for Whisper Speech

AHs 2024, April 04-06, 2024, Melbourne, VIC, Australia



Figure 14: Results for speech recognition in environments with different noise levels. a30dB indicates a recording environment without white noise, and w40dB, w60dB, and w80dB indicate a recording environment with a white noise output of 40, 60, and 80 dB, respectively. In natural speech (left), recognition accuracy improved in the following decreasing order: WhisperMask followed by pin microphone and then by earbuds. A difference of approximately 30% was observed for whispered speech in an environment with 80 dB background noise (right). Noise reduction by Denoiser [18] did not result in considerable changes in the recognition accuracy for natural speech but it significantly reduced that for whispered speech.

## 7.3 Result

The results are shown in Fig. 14; the results for normal and whispered speech are shown on the left and right sides, respectively. For natural speech recorded with 80 dB background noise, the recognition accuracy for WhisperMask was over 20% higher than that for pin microphone and earbuds. Notably, under the same noise condition, the recognition rate for WhisperMask without denoiser was higher by over 20% than that for the microphones with denoiser. The recognition accuracy for whispered speech recorded in an environment with 80 dB background noise was higher for WhisperMask by over 30% than that for pin microphone and earbuds. After denoiser application, the recognition rate for whispered speech decreased by approximately 20% in the 30, 40, and 60 dB environments. This is a significant decrease compared with that observed for natural speech, indicating that the denoiser is not well-suited for whispered speech. Moreover, the recognition accuracy for whispered speech recorded in an environment with 80 dB background noise was higher by over 30% for WhisperMask without denoiser compared with that for pin microphone and earbuds, suggesting the superiority of WhisperMask for recognizing whispered speech in noisy environments.

## 8 DISCUSSION

### 8.1 Mechanism of noise reduction

The difficulty of vibration may be one of the reasons behind the enhanced ability of WhisperMask to input speech in noisy environments. As shown in Figs. 6–8, despite the 80 dB sound coming from the dummy head speaker and the noise coming from the speaker, the maximum input on the microphone side was approximately 40 dB, indicating that a loud sound input is required. This phenomenon is likely because normal human speech is produced at

approximately 60–80 dB, but when a sound source is closer to a microphone, louder voices are picked up. In fact, when measured at a distance of about 3 cm using a sound level meter, a normal voice becomes louder by approximately 80–90 dB. One of the key factors contributing to noise reduction is that voices are captured more loudly at a close proximity.

### 8.2 Issues for daily use

*8.2.1 Reusability.* Microphones are devices intended for daily use and must be designed for durability. In this study, we asked nine users to use the proposed device; they were instructed to put on and take off their masks for each input session. We observed no performance issues related to mask usage. Moreover, the users wore the device over a woven mask, meaning the device can operate without direct contact with the mouth. Consequently, the proposed device was associated with fewer concerns in relation to contamination and hygiene compared with devices that are attached directly to the face. Furthermore, the vibrating component of the microphone can be detached from the circuitry and may be embedded in a protective material, making it washable without causing any issues.

8.2.2 Noise when walking. We did not evaluate the impact of walking and other movements on the performance of WhisperMask. Motion artifacts may possibly introduce noise and affect the device's performance. Further investigation is needed to assess the impact of user movements and develop strategies to avoid any associated noise.

8.2.3 Blowing wind. iven that voice enters the microphone through air transmission, strong winds or turbulent airflow between the mouth and the microphone can disrupt voice recording. This issue also affects the performance of ordinary pin microphones, and a windshield may be necessary. Although we have not evaluated the effects of strong winds in this study, we have confirmed that sound can be collected even when the sensor is placed inside a non-woven fabric mask, which could provide some protection against wind.

In the future, we intend to further evaluate the impact of user movements, such as walking, and environmental factors, such as wind, on the performance of WhisperMask.

## 9 CONCLUSION

We propose WhisperMask, a mask-type electret condenser microphone that can clearly capture a user's voice even in noisy environments compared with conventional microphones. We demonstrated WhisperMask's acoustic characteristics by measuring its impulse response using swept-sine signals. Furthermore, we evaluated WhisperMask based on three key metrics: SNR, quality of recorded voices, and speech recognition rate. Across these metrics, WhisperMask significantly outperformed the conventional noise reduction methods, which involve either hardware- or software-based approaches.

The recognition rate for whispered speech recorded in an environment with 80 dB background noise was notably higher by over 30% for WhisperMask than that for pin microphones and earbuds. Moreover, while a denoiser software decreased by approximately 20% the other microphones' recognition rate for whispered speech recorded with 30–60 dB background noise, WhisperMask maintained a high performance even without denoising, surpassing the performance of the other microphones by a large margin. These results highlight WhisperMask's overwhelming superiority in capturing whispered speech under noisy conditions.

In conclusion, WhisperMask represents a significant advancement in the wearable microphone technology. By effectively addressing the challenge of capturing clear voice input, especially whispered speech, in high-noise environments, WhisperMask opens new possibilities for enhanced communication and interaction across a wide range of voice-based applications while preserving user privacy. Its lightweight mask-type form factor and exceptional noise suppression capabilities make it a promising tool for various real-world scenarios requiring reliable voice input.

### ACKNOWLEDGMENTS

This work was supported by JST ACT-X Grant JPMJAX23KG, JST Moonshot R&D Grant JPMJMS2012, JST CREST Grant JPMJCR17A3, and the commissioned research by NICT Japan Grant JPJ012368C02901.

#### REFERENCES

- [1] [n. d.]. https://appleinsider.com/articles/21/03/30/apple-airpods-beats-dominatedaudio-wearable-market-in-2020.
- [2] B. Series. Recommendation ITU-R BS. 1534-3. 2014. method for the subjective assessment of intermediate quality level of audio systems. In *International Telecommunication Union Radio Communication Assembly*.
- [3] M. Aubreville, K. Ehrensperger, A. Maier, T. Rosenkranz, B. Graf, and H. Puder. 2018. Deep Denoising for Hearing Aid Applications. In 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). 361–365. https://doi.org/ 10.1109/IWAENC.2018.8521369
- [4] Benjamin B Bauer. 1962. A century of microphones. Proceedings of the IRE 50, 5 (1962), 719–729.
- [5] Christopher Beach, Nazmul Karim, and Alexander J. Casson. 2019. A Graphene-Based Sleep Mask for Comfortable Wearable Eye Tracking. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 6693–6696. https://doi.org/10.1109/EMBC.2019.8857198

- [6] Reuven Berkun and Israel Cohen. 2015. Microphone array power ratio for quality assessment of reverberated speech. EURASIP Journal on Advances in Signal Processing 2015, 1 (2015), 49. https://doi.org/10.1186/s13634-015-0233-y
- [7] Colin Breithaupt, Timo Gerkmann, and Rainer Martin. 2008. A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing. In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. 4897–4900. https://doi.org/10.1109/ICASSP.2008.4518755
- [8] Jonathan S. Brumberg, Alfonso Nieto-Castanon, Philip R. Kennedy, and Frank H. Guenther. 2010. Brain-computer interfaces for speech communication. Speech Communication 52, 4 (2010), 367–379. https://doi.org/10.1016/j.specom.2010.01. 001 Silent Speech Interfaces.
- [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 (Portland, Oregon) (MobiSys '22). Association for Computing Machinery, New York, NY, USA, 384–396. https://doi.org/10.1145/3498361.3538933
- [10] Seungjin Choi, Andrzej Cichocki, Hyung-Min Park, and Soo-Young Lee. 2005. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews* 6, 1 (2005), 1–57.
- [11] Andrzej Cichocki and Anh-Huy Phan. 2009. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals* of electronics, communications and computer sciences 92, 3 (2009), 708–721.
- [12] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. 2006. New algorithms for non-negative matrix factorization in applications to blind source separation. In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Vol. 5. IEEE, V–V.
- [13] J. Cohen. 1962. The statistical power of abnormal-social psychological research: A review. The Journal of Abnormal and Social Psychology 65, 3 (1962), 145–153. https://doi.org/10.1037/h0045186
- [14] Ryan M Corey and Andrew C Singer. 2018. Speech separation using partially asynchronous microphone arrays without resampling. In 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, 1–9.
- [15] Tamás Csapó, Tamás Grósz, Gábor Gosztolya, László Tóth, and Alexandra Markó. 2017. DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface. In *interspeech 2017*. 3672–3676. https://doi.org/10.21437/Interspeech.2017-939
- [16] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg. 2010. Silent speech interfaces. Speech Communication 52, 4 (2010), 270 – 287. https: //doi.org/10.1016/j.specom.2009.08.002 Silent Speech Interfaces.
- [17] Lorenz Diener, Mehrdad Roustay Vishkasougheh, and Tanja Schultz. 2020. CSL-EMG\_Array: An Open Access Corpus for EMG-to-Speech Conversion. In INTER-SPEECH 2020 – 21st Annual Conference of the International Speech Communication Association.
- [18] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real Time Speech Enhancement in the Waveform Domain. In Proc. Interspeech 2020. 3291–3295. https://doi.org/10.21437/Interspeech.2020-2409
- [19] LG Electronics. 2024. LG PuriCare™ Wearable Air Purifier (w/ VoiceON™) | LG Philippines. https://www.lg.com/ph/air-care/lg-ap551awfa. Accessed: 2024-02-10.
- [20] K Elenius. 1980. Long time average spectrum using a 1/3 octave filter bank
- [21] angelo farina. 2000. simultaneous measurement of impulse response and distortion with a swept-sine technique. *journal of the audio engineering society* (february 2000).
- [22] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Favory, Jordi Pons, and Xavier Serra. 2018. General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018). 69–73. https://arxiv.org/abs/1807.09902
- [23] João Freitas, António Teixeira, Miguel Dias, and Samuel Silva. 2016. An Introduction to Silent Speech Interfaces.
- [24] Masaaki Fukumoto. 2018. SilentVoice: Unnoticeable Voice Input by Ingressive Speech. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (Berlin, Germany) (UIST '18). Association for Computing Machinery, New York, NY, USA, 237–246. https://doi.org/10.1145/3242587.3242603
- [25] Yifan Gong. 1995. Speech recognition in noisy environments: A survey. Speech Communication 16, 3 (1995), 261–291. https://doi.org/10.1016/0167-6393(94) 00059-J
- [26] Zengrong Guo and Rong-Hao Liang. 2023. TexonMask: Facial Expression Recognition Using Textile Electrodes on Commodity Facemasks. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 627, 15 pages. https://doi.org/10.1145/3544548.3581295
- [27] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. 2019. TAU Urban Acoustic Scenes 2019, Development dataset. https://doi.org/10.5281/zenodo.2589280
- [28] Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. 2010. Silent-speech enhancement using body-conducted vocal-tract resonance signals. Speech Communication 52, 4 (2010), 301–313. https://doi.org/10.1016/j.specom.2009.12.001 Silent Speech

WhisperMask: A Noise Suppressive Mask-Type Microphone for Whisper Speech

AHs 2024, April 04-06, 2024, Melbourne, VIC, Australia

Interfaces

- [29] Hirotaka Hiraki and Jun Rekimoto. 2021. SilentMask: Mask-Type Silent Speech Interface with Measurement of Mouth Movement. In Augmented Humans Conference 2021 (Rovaniemi, Finland) (AHs'21). Association for Computing Machinery, New York, NY, USA, 86–90. https://doi.org/10.1145/3458709.3458985
- [30] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone. 2007. Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Vol. 1. I–1245–I–1248. https://doi.org/10.1109/ ICASSP.2007.366140
- [31] Hushme Inc. 2021. Hushme The World's First Voice Mask for Smartphones. https://gethushme.com/. Accessed: 2024-02-10.
- [32] Shiftall Inc. 2023. mutalk Leakage voice suppression microphone. https://en. shiftall.net/products/mutalk. Accessed: 2024-02-10.
- [33] Robert Ingalls. 1987. Throat microphone. The Journal of the Acoustical Society of America 81, 3 (03 1987), 809–809. https: //doi.org/10.1121/1.394659 arXiv:https://pubs.aip.org/asa/jasa/articlepdf/81/3/809/12095011/809\_1\_online.pdf
- [34] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In 23rd International Conference on Intelligent User Interfaces (Tokyo, Japan) (IUI '18). Association for Computing Machinery, New York, NY, USA, 43–53. https://doi.org/10.1145/3172944.3172977
- [35] Junki Kawaguchi and Mitsuharu Matsumoto. 2022. Noise Reduction Combining a General Microphone and a Throat Microphone. Sensors 22, 12 (2022). https: //doi.org/10.3390/s2212s4473
- [36] 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 (Virtual, United Kingdom) (HotMobile '21). Association for Computing Machinery, New York, NY, USA, 44–49. https://doi.org/10.1145/ 3446382.3448363
- [37] Taesu Kim, Torbjørn Eltoft, and Te-Won Lee. 2006. Independent vector analysis: An extension of ICA to multivariate components. In *International conference on independent component analysis and signal separation*. Springer, 165–172.
- [38] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Alex Olwal, Jun Rekimoto, and Thad Starner. 2021. Mobile, Hands-Free, Silent Speech Texting Using SilentSpeller. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3451552
- [39] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300376
- [40] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. 2015. Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 276–280.
- [41] Yuma Koizumi, Kenta Niwa, Yusuke Hioka, Kazunori Kobayashi, and Yoichi Haneda. 2018. DNN-Based Source Enhancement to Increase Objective Sound Quality Assessment Score. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26, 10 (2018), 1780–1792. https://doi.org/10.1109/TASLP.2018.2842156
- [42] Ryoga Kumazaki and Akifumi Inoue. 2020. Development and Evaluation of a Mask-Type Display Transforming the Wearer's Impression. In Proceedings of 31st Australian Conference on Human-Computer-Interaction (Fremantle, WA, Australia) (OzCHI '19). Association for Computing Machinery, New York, NY, USA, 568–571. https://doi.org/10.1145/3369457.3369533
- [43] Yusuke Kunimi, Masa Ogata, Hirotaka Hiraki, Motoshi Itagaki, Shusuke Kanazawa, and Masaaki Mochimaru. 2022. E-MASK: A Mask-Shaped Interface for Silent Speech Interaction with Flexible Strain Sensors. In Augmented Humans 2022 (Kashiwa, Chiba, Japan) (AHs 2022). Association for Computing Machinery, New York, NY, USA, 26–34. https://doi.org/10.1145/3519391.3519399
- [44] Takahiro Kusabuka and Takuya Indo. 2020. IBUKI: Gesture Input Method Based on Breathing. In Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology (Virtual Event, USA) (UIST '20 Adjunct). Association for Computing Machinery, New York, NY, USA, 102–104. https: //doi.org/10.1145/3379350.3416134
- [45] Hyein Lee, Yoonji Kim, and Andrea Bianchi. 2020. MAScreen: Augmenting Speech with Visual Cues of Lip Motions, Facial Expressions, and Text Using a Wearable Display. In SIGGRAPH Asia 2020 Emerging Technologies (Virtual Event, Republic of Korea) (SA '20). Association for Computing Machinery, New York, NY, USA, Article 2, 2 pages. https://doi.org/10.1145/3415255.3422886
- [46] Richard Li, Jason Wu, and Thad Starner. 2019. TongueBoard: An Oral Interface for Subtle Input. In Proceedings of the 10th Augmented Human International Conference 2019 (Reims, France) (AH2019). Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3311823.3311831
- [47] Daniyal Liaqat, Salaar Liaqat, Jun Lin Chen, Tina Sedaghat, Moshe Gabel, Frank Rudzicz, and Eyal de Lara. 2021. Coughwatch: Real-World Cough Detection using Smartwatches. In ICASSP 2021 - 2021 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP). 8333–8337. https://doi.org/10. 1109/ICASSP39728.2021.9414881

- [48] Daniyal Liaqat, Robert Wu, Andrea Gershon, Hisham Alshaer, Frank Rudzicz, and Eyal de Lara. 2018. Challenges with Real-World Smartwatch Based Audio Monitoring. In Proceedings of the 4th ACM Workshop on Wearable Systems and Applications (Munich, Germany) (WearSys '18). Association for Computing Machinery, New York, NY, USA, 54–59. https://doi.org/10.1145/3211960.3211977
- [49] I. Scott MacKenzie and R. William Soukoreff. 2003. Phrase Sets for Evaluating Text Entry Techniques. In CHI '03 Extended Abstracts on Human Factors in Computing Systems (Ft. Lauderdale, Florida, USA) (CHI EA '03). Association for Computing Machinery, New York, NY, USA, 754–755. https://doi.org/10.1145/765891.765971
- [50] Naoki Makishima, Shinichi Mogami, Norihiro Takamune, Daichi Kitamura, Hayato Sumino, Shinnosuke Takamichi, Hiroshi Saruwatari, and Nobutaka Ono. 2019. Independent deeply learned matrix analysis for determined audio source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 10 (2019), 1601–1615.
- [51] 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 29 (2021), 1368–1396.
- [52] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03), Vol. 5. V–708. https://doi.org/10.1109/ ICASSP.2003.1200069
- [53] Hye Yeon Nam, Iyleah Hernandez, and Brendan Harmon. 2020. Unmasked. In Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology (Virtual Event, USA) (UIST '20 Adjunct). Association for Computing Machinery, New York, NY, USA, 111–113. https://doi.org/10.1145/ 3379350.3416137
- [54] Laxmi Pandey and Ahmed Sabbir Arif. 2021. LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445565
- [55] Anne Porbadnigk, Marek Wester, Jan-P Calliess, and Tanja Schultz. 2009. EEGbased Speech Recognition - Impact of Temporal Effects. 376–381.
- [56] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [eess.AS]
- [57] Chandan Reddy, Ebrahim Beyrami, Harishchandra Dubey, Vishak Gopal, Roger Cheng, Ross Cutler, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke. 2020. The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework. In *Interspeech 2020*.
- [58] Jun Rekimoto. 2023. WESPER: Zero-Shot and Realtime Whisper to Normal Voice Conversion for Whisper-Based Speech Interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, Article 700, 12 pages. https://doi.org/10.1145/3544548.3580706
- [59] Jun Rekimoto and Yu Nishimura. 2021. Derma: Silent Speech Interaction Using Transcutaneous Motion Sensing. In Augmented Humans Conference 2021 (Rovaniemi, Finland) (AHs'21). Association for Computing Machinery, New York, NY, USA, 91–100. https://doi.org/10.1145/3458709.3458941
- [60] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena. 6, 3, Article 135 (sep 2022), 57 pages. https://doi.org/10.1145/3550314
- [61] Tobias Röddiger, Tobias King, Dylan Ray Roodt, Christopher Clarke, and Michael Beigl. 2022. Openearable: Open hardware earable sensing platform. In Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers. 246-251.
- [62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241.
- [63] Mose Sakashita, Keisuke Kawahara, Amy Koike, Kenta Suzuki, Ippei Suzuki, and Yoichi Ochiai. 2016. Yadori: Mask-Type User Interface for Manipulation of Puppets. In ACM SIGGRAPH 2016 Emerging Technologies (Anaheim, California) (SIGGRAPH '16). Association for Computing Machinery, New York, NY, USA, Article 23, 1 pages. https://doi.org/10.1145/2929464.2929478
- [64] Hiroshi Sawada, Nobutaka Ono, Hirokazu Kameoka, Daichi Kitamura, and Hiroshi Saruwatari. 2019. A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF. APSIPA Transactions on Signal and Information Processing 8 (2019), e12. https://doi.org/10.1017/ATSIP.2019.5
- [65] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. 2018. webMUSHRA – A Comprehensive Framework for Web-based Listening Tests. Journal of Open Research

Software (Feb 2018). https://doi.org/10.5334/jors.187

- [66] Antonia Schulte, Rodrigo Suarez-Ibarrola, Daniel Wegen, Philippe-Fabian Pohlmann, Elina Petersen, and Arkadiusz Miernik. 2020. Automatic speech recognition in the operating room - An essential contemporary tool or a redundant gadget? A survey evaluation among physicians in form of a qualitative study. Ann Med Surg (Lond) 59 (2020), 81–85. https://doi.org/10.1016/j.amsu.2020.09.015
- [67] Shota Shimizu, Makoto Otani, and Tatsuya Hirahara. 2009. Frequency characteristics of several non-audible murmur (NAM) microphones. Acoustical Science and Technology 30, 2 (2009), 139–142. https://doi.org/10.1250/ast.30.139
- [68] Daniel Stoller, Sebastian Ewert, and Simon Dixon. 2018. Wave-u-net: A multiscale neural network for end-to-end audio source separation. arXiv preprint arXiv:1806.03185 (2018).
- [69] Zixiong Su, Shitao Fang, and Jun Rekimoto. 2023. LipLearner: Customizable Silent Speech Interactions on Mobile Devices. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–21.
- [70] 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). 21–25. https://doi.org/10.1109/ICASSP39728.2021.9413901
- [71] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (Berlin, Germany) (UIST '18). Association for Computing Machinery, New York, NY, USA, 581–593. https://doi.org/10.1145/3242587.3242599
- [72] Yutaro Suzuki, Kodai Sekimori, Yuki Yamato, Yusuke Yamasaki, Buntarou Shizuki, and Shin Takahashi. 2020. A Mouth Gesture Interface Featuring a Mutual-Capacitance Sensor Embedded in a Surgical Mask. In Human-Computer Interaction. Multimodal and Natural Interaction, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 154–165.
- [73] Linda M Thibodeau, Rachel B Thibodeau-Nielsen, Chi Mai Quynh Tran, and Regina Tangerino de Souza Jacob. 2021. Communicating during COVID-19: The effect of transparent masks for speech recognition in noise. *Ear and Hearing* 42, 4 (2021), 772–781.
- [74] Vishal Varun Tipparaju, Di Wang, Jingjing Yu, Fang Chen, Francis Tsow, Erica Forzani, Nongjian Tao, and Xiaojun Xian. 2020. Respiration pattern recognition by wearable mask device. *Biosensors and Bioelectronics* 169 (2020), 112590. https: //doi.org/10.1016/j.bios.2020.112590
- [75] Vishal Varun Tipparaju, Xiaojun Xian, Devon Bridgeman, Di Wang, Francis Tsow, Erica Forzani, and Nongjian Tao. 2020. Reliable Breathing Tracking With Wearable Mask Device. *IEEE Sensors Journal* 20, 10 (2020), 5510–5518. https: //doi.org/10.1109/JSEN.2020.2969635
- [76] Joseph C Toscano and Cheyenne M Toscano. 2021. Effects of face masks on speech recognition in multi-talker babble noise. *PloS one* 16, 2 (2021), e0246842.
- [77] Cassia Valentini-Botinhao. 2017. Noisy speech database for training speech enhancement algorithms and TTS models. In University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR). https://doi.org/10. 7488/ds/2117.
- [78] Bandhav Veluri, Justin Chan, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. 2023. Real-Time Target Sound Extraction. In *ICASSP 2023* - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094573
- [79] Amritha Vijayan, Bipil Mary Mathai, Karthik Valsalan, Riyanka Raji Johnson, Lani Rachel Mathew, and K. Gopakumar. 2017. Throat microphone speech recognition using mfcc. In 2017 International Conference on Networks & Advances in Computational Technologies (NetACT). 392–395. https://doi.org/10.1109/NETACT. 2017.8076802
- [80] Michael Wand., Christopher Schulte., Matthias Janke., and Tanja Schultz. 2013. Array-based Electromyographic Silent Speech Interface. In Proceedings of the International Conference on Bio-inspired Systems and Signal Processing - BIOSIG-NALS, (BIOSTEC 2013). INSTICC, SciTePress, 89–96. https://doi.org/10.5220/ 0004252400890096
- [81] Michael Wand and Tanja Schultz. 2011. Session-independent EMG-based Speech Recognition., In Proceedings of Biosignals 2011. BIOSIGNALS 2011 - Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing, 295–300.
- [82] Takumi Yamamoto, Katsutoshi Masai, Anusha Withana, and Yuta Sugiura. 2023. Masktrap: Designing and Identifying Gestures to Transform Mask Strap into an Input Interface. In Proceedings of the 28th International Conference on Intelligent User Interfaces (Sydney, NSW, Australia) (IUI '23). Association for Computing Machinery, New York, NY, USA, 762–775. https://doi.org/10.1145/3581641.3584062
- [83] NAKAJIMA Yoshitaka, KASHIOKA Hideki, CAMPBELL Nick, and SHIKANO Kiyohiro. 2005. Non-Audible Murmur (NAM) Recognition. IEICE TRANSACTIONS on Information and Systems E89-D, 1 (2005).
- [84] Asri Rizki Yuliani, M Faizal Amri, Endang Suryawati, Ade Ramdan, and Hilman Ferdinandus Pardede. 2021. Speech enhancement using deep learning methods: A review. Jurnal Elektronika dan Telekomunikasi 21, 1 (2021), 19–26.
- [85] Frank Zalkow, Prachi Govalkar, Meinard Müller, Emanuël A. P. Habets, and Christian Dittmar. 2023. Evaluating Speech-Phoneme Alignment and its Impact

on Neural Text-To-Speech Synthesis. In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1–5. https: //doi.org/10.1109/ICASSP49357.2023.10097248

- [86] Cheng Zhang, Anandghan Waghmare, Pranav Kundra, Yiming Pu, Scott Gilliland, Thomas Ploetz, Thad E. Starner, Omer T. Inan, and Gregory D. Abowd. 2017. FingerSound: Recognizing Unistroke Thumb Gestures Using a Ring. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1, 3, Article 120 (sep 2017), 19 pages. https://doi.org/10.1145/3130985
- [87] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing Acoustic-Based Imaging for Issuing Contact-Free Silent Speech Commands. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 37 (March 2020), 26 pages. https://doi.org/10.1145/3381008