# Enabling Hands-Free Voice Assistant Activation on Earphones

Tao Chen[†], Yongjie Yang[†], Chonghao Qiu[†], Xiaoran Fan[+], Xiuzhen Guo[¶], Longfei Shangguan[†]
[†]University of Pittsburgh, [+]Google, [¶]Zhejiang University

## ABSTRACT

We present the design and implementation of EarVoice, a lightweight mobile service that enables hands-free voice assistant activation on commodity earphones. EarVoice comprises two design modules: one for joint speech detection and primary user identification that explores the attributes of the air channel and in-body audio pathway to differentiate between the primary user and others nearby; and another for accurate wakeup word enhancement, which employs a "copy, paste, and adapt" approach to reconstruct the missing high-frequency component in speech recordings. To minimize false positives, enhance agility, and preserve privacy, we deploy EarVoice on a dongle where the proposed signal processing algorithms are streamlined with a gating mechanism to permit only the primary user's speech to enter the pairing device (e.g., a smartphone) for wakeup word recognition, preventing unintended disclosure of ambient conversations. We implemented the dongle on a 4-layer PCB board and conducted extensive experiments with 23 participants in both controlled and uncontrolled scenarios. The experiment results show that EarVoice achieves around 90% wakeup word recognition accuracy in stationary scenarios, which is on par with the high-end, multi-sensor fusion-based Airpods Pro earbud. EarVoice's performance drops to 84% on mobile cases, slightly worse than Airpods (around 90%).

## CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile devices**; • **Hardware** → **Emerging technologies**.

## KEYWORDS

Voice Activation, Bone Conduction, Earable Computing

**ACM Reference Format:**
Tao Chen[†], Yongjie Yang[†], Chonghao Qiu[†], Xiaoran Fan[+], Xiuzhen Guo[¶], Longfei Shangguan[†]. 2024. Enabling Hands-Free Voice Assistant Activation on Earphones. In *The 22nd Annual International Conference on Mobile Systems, Applications and Services (MOBISYS '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3643832.3661890

## 1 INTRODUCTION

Voice assistant (VA) has become an indispensable part of mobile systems [7, 27]. It serves as a natural means of communication that transcends language barriers, making mobile applications more accessible and inclusive for a diverse range of users [34]. The rapid growth of generative AI [42], fueled by the sheer size of computation resources in the cloud, has been transforming the voice assistant into a more seamless and user-friendly user interface.

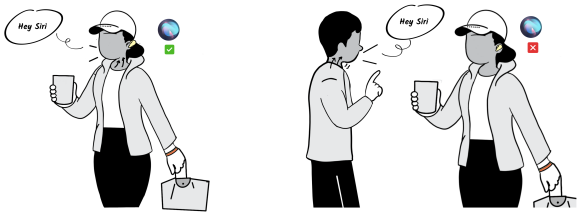


**Figure 1: A few representative examples of EarVoice.** (left): EarVoice allows mobile users to activate their voice assistant without hand intervention. (right): EarVoice can automatically detect the primary speaker, avoiding false alarms.

While the voice assistant offers flexibility to mobile users, the process of activating it remains inconvenient due to its heavy dependence on hand interventions, particularly on earphones [2]. Taking Siri [70] as an example, the user has to press and hold the talk/answer button on earphones for a few seconds until hearing the Siri beep[1]. This precaution is taken to avoid unintended activation of Siri by someone else nearby. Yet, this would divert the user's attention from their current focus, negatively impacting the user experience. This is especially notable in situations where the user's hands are occupied, as illustrated in Figure 1(a).

Notice that, in this paper we ask a simple question: *is it possible to enable hands-free VA activation on earphones*? An affirmative answer would enhance the accessibility of voice assistants by enabling individuals occupied with other tasks to interact with their devices conveniently. In addition, it can improve safety by reducing the need for hands-on device manipulation, particularly in situations where manual interaction may be risky such as driving or cycling.

Nevertheless, to harvest the aforementioned benefits, we have to take into account the following system requirements.

- **Low False Positive Rate**. A hands-free voice activation service stays in idle listening mode continuously, responding whenever a voice command is initiated. To achieve a good user experience, this service should minimize false positives, ensuring that it doesn't get triggered by ambient voices.
- **Agile and Low-Power**. The proposed service should respond to human speech agilely, with minimum or unnoticeable latency. Moreover, as an always-on service running on power-constrained mobile devices, the proposed system design should be low-power.
- **Privacy-preserving**. Voice data should be stored securely, and users should have control over their data. Besides the necessary voice commands for awakening corresponding

[1] Similarly, current wireless earbuds, including Google pixel-bud, Apple AirPods, and Bose's QC35 [6, 69, 70], all require users to activate the voice assistant either by taping a touch sensor [1] or holding an action button [73].


*MOBISYS '24, June 3–7, 2024, Minato-ku, Tokyo, Japan*

ACM ISBN 979-8-4007-0581-6/24/06…$15.00
https://doi.org/10.1145/3643832.3661890

services, other audio data should avoid being recorded on the smartphone to minimize the risk of privacy leaks.

We present EarVoice, a mobile service that explores the distinction between the acoustic air channel and the in-body bone-conduction pathway formed in human speech to enable accurate, agile, and low-power hands-free voice activation, all in a privacy-preserving way. Our system works with everyday earphones (*e.g.*, those earphones cost a few US dollars) without breaking their structures and requires neither in-ear microphones [22, 29, 33, 40, 41] nor dedicated IMU sensors that are only available on those pricey ANC earphones.

Motivated by HeadFi [20], EarVoice repurposes the earphone speaker into a microphone for wakeup words (*e.g.*, *"Hey Siri"*) detection. This allows mobile users to wake up their voice assistant using earphones even without a microphone[2]. To detect whether the recorded sounds are human speech or ambient noise, and furthermore, to distinguish if the detected speech originates from the primary user (*i.e.*, who wears the earphone), EarVoice explores an observation that the speech of the primary user reaches the earphone's speaker transducer through not only the conventional air channel but also via the human body channel, whereas the nearby speaker's speech solely propagates through the air channel to the earphone speaker transducer, with significant attenuation. This discrepancy in the audio pathways is reflected in the recorded audio spectrum, with low-frequency signals originating from the primary speaker's vocal cord vibrations being present, while the low-frequency voice components of a nearby speaker are not. EarVoice takes advantage of this unique frequency disparity to detect whether it is the primary user or someone else speaking nearby.

However, the distinct in-body bone-conduction pathway, coupled with the suboptimal frequency response of speaker transducers functioning as microphones, leads to a significant power loss in the higher-frequency speech components. The occurrence of such high-frequency deafness distorts spoken wakeup words severely, consequently diminishing the accuracy of wakeup word recognition. To address this challenge, we propose a wakeup word enhancement design to compensate for the high-frequency energy loss in the speech recording. This approach takes a MEMS microphone recording of the wakeup word (*e.g.*, "Hey Siri") as the template, extracting its high-frequency components ranging from 2 to 8 kHz, and pasting it to the voice recording. As wakeup recognition systems are primarily designed to interpret content-dependent elements of human speech such as vowels and consonants as opposed to human speaker-dependent features like tones, prosody, and intonation, the combined signal can be successfully recognized even though its frequency components come from different individuals.

Nevertheless, as different individuals speak the wakeup word at different speeds, frequencies, and loudness, blindly copying and pasting without considering the discrepancy between the speech recording and the template can lead to the misalignment of critical formants in the combined audio signal and further undermine the wakeup word recognition. To address this issue, we propose an efficient signal processing algorithm to align these two signal components along the time, frequency, and amplitude domain, ensuring two frequency components are aligned in their combined form.

[2] The line-in/boom microphones may appear on some conventional earphones and can pick up human voices. However, they usually have no direct contact with the human head and thus cannot differentiate between the primary user and nearby individuals.

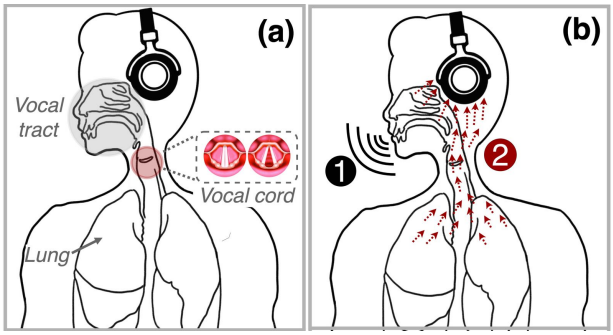


**Figure 2: (a): human speech production. (b): two human speech transmission channels.** ① air channel, ② in-body bone-conduction audio pathway.

EarVoice functions as a hybrid signal-processing pipeline with primary functions running on a low-power dongle while the wakeup word recognition runs on the smartphone. The dongle transforms the earphone speaker into a microphone, detects the human voice, distinguishes whether it originates from the primary user, and further enhances the speech quality. By exclusively forwarding only the legitimate voice commands from the dongle to the smartphone, this gating approach not only prevents inadvertent disclosure of ambient conversations but also minimizes unnecessary wakeup word recognition on the pairing device, thereby conserving power.

We have implemented a prototype of EarVoice's dongle on a 4-layer printed circuit board (PCB). It consists of a low power ESP32 MCU, an audio codec chip, and other peripherals to enable the functionality. The total cost for this dongle is around 8.3 US dollars.

We summarize our contributions below:

- We identified that the close contact between the earphone speaker transducer and the human skin offers a unique opportunity to sense the vocal cord vibrations of the user who spoke, enabling us to tell whether the voice is coming from the primary user or others in the vicinity. We then proposed a lightweight signal processing algorithm that explores this opportunity to enable hands-free voice assistant activation.
- We designed a gated signal-processing pipeline that can accurately detect, differentiate, and further enhance the incomplete voice command captured by the earphone speaker transducer, all in a low-power and privacy-preserving way. This design holds the potential to be deployed on different types of earphones.
- We implemented EarVoice on a PCB board and conducted extensive experiments in both controlled and uncontrolled environments. The results demonstrated that EarVoice achieves an overall wakeup recognition accuracy of 90% across different real-world scenarios, which is on par with the high-end, multi-sensor fusion-based Airpods Pro earbud.

## 2 SPEECH PRODUCTION PRIMER

Before we describe the potential of the earphone's speaker transducer for hands-free voice assistant activation, we first explain how human speech production works.

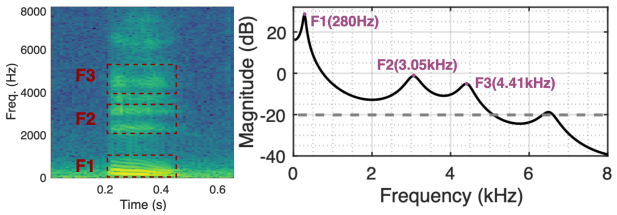


**Figure 3: Spectrogram (left) and spectral envelope (right) of the vowel sound /i/.** The first three formants are denoted as F1, F2, and F3. This audio signal is recorded by a MEMS microphone.

As illustrated in Figure 2(a), the production of human speech involves intricate coordination between multiple articulatory organs in the vocal system, including lungs, vocal cords (*a.k.a.* vocal folds), and vocal tract[3]. Specifically, the lungs provide the essential air source required for vocalization. This air subsequently passes through the vocal folds to generate a voice source and is then modulated by the vocal tract to produce output speech [38]. Vocal folds generate speech signals that are voiced by dynamically controlling the airflow originating from the lung, alternatively blocking and permitting it. On the contrary, if vocal folds do not vibrate, airflow from the lungs will be manipulated directly by the vocal tract to produce unvoiced signals, such as consonant sounds like */f/*, */r/*, *etc.*

The voiced signals consist of two components. i): vowels and some consonants that own high energy pulses in the frequency domain [58]; ii): the fundamental pitch F0 and its harmonics. The frequency components that determine the intelligence of speech words are called **formants** (spectral resonances) [35]. The first formants in a sentence are usually within 300–2800Hz frequency band, forming the pronunciation of vowels. The follow-up formants stay in a higher frequency band above 3000Hz, as shown in Figure 3.

## 3 OPPORTUNITIES AND CHALLENGES

Facilitating hands-free voice assistant activation on earphones requires the *agile detection of human voice*, *precise identification of the primary speaker*, and *robust recognition of the wakeup word* hereafter. We have two observations contribute to achieving these goals: the first a reflection on recent research, the second a consequence of unique voice channels:

(1) Recent work has demonstrated that the speaker transducer on commodity earphones can be used as a microphone for acoustic signal reception [9, 12, 20]. This leaves us an opportunity to capture spoken words on all types of earphones without requiring a microphone.
(2) The primary user's voice reaches the earphone via both an air channel and an in-body channel, while a nearby user's voice only travels through the air channel. Due to the earphone's obstruction, only a small fraction of the voice energy from the nearby user reaches the earphone's speaker. In contrast, the primary user's voice arrives at the earphone speaker with less attenuation through the in-body channel, providing us with an opportunity to distinguish the speaker (§3.1).

In the following sections, we assess the practicality of these opportunities and identify potential challenges.

[3] Vocal tract is the area from the nose and the nasal cavity down to the vocal cords, including the throat, mouth (*e.g.*, tongue, teeth, lip), nasal cavity, and facial movement [79].

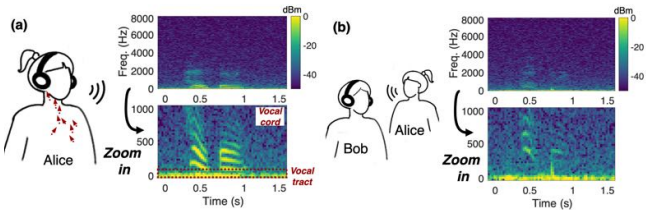


**Figure 4: Feasibility study: speech measurement from (a): a primary speaker; and (b): a nearby speaker.**

**Table 1: Wakeup words recognition accuracy on five mainstreaming voice interfaces.** Ten volunteers are invited to articulate three wakeup words 10 times each.

| ASR | Earphone speaker transducer | MEMS Microphone |
|---|---|---|
| **Google API [25]** | 9% | 82% |
| **DeepSpeech [3]** | 1% | 58% |
| **iFLYTEK [30]** | 18% | 76% |
| **SpeechBrain [57]** | 1% | 66% |
| **Whisper [56]** | 31% | 93% |

### 3.1 Identifying the Primary Speaker: An Opportunity

Voice fingerprint [23] is proposed to identify the registered primary user and might help determine whether the primary user is interacting with Siri or if someone else nearby is speaking. However, such a mechanism is prone to various security threats in real life, including impersonation, voice synthesis [47], and replay attacks [21, 51].

Instead of applying fingerprint technology, we found that the distinct speech propagation channels between the primary speaker and nearby speakers offer us another opportunity to distinguish speakers using earphones. Specifically, the speech of the primary user reaches the earphone's speaker transducer through not only the conventional air channel but also via the human body channel, as depicted in Figure 2(b). In contrast, when it comes to human speech from a nearby non-primary speaker, it solely propagates through the air channel to the earphone speaker transducer. Below we elaborate on these two channels:

(1) **Air channel for voice propagation**. For both the primary speaker and nearby speakers, the voice signal emanating from their mouth will propagate through the air channel. The earphone's speaker transducer captures this signal when the sound reaches the earphone, as denoted by ① in Figure 2(b).
(2) **Body channel for the propagation of articulatory organ vibrations**. For the primary speaker, the vibrations from her articulatory organs, such as the vocal cord and tract, would travel through the human body and ultimately reach the ear canal. Given the fact that the earphone transducer maintains close contact with the human ear, the speaker transducer is highly likely to detect these vibrations through bone conduction. Prior works have demonstrated so on in-ear microphones [5, 21] and IMUs [26].

We conducted benchmark studies in a controlled environment to assess whether human speakers are differentiable based on these two channel propagation characteristics.

**Setups**. We invited two volunteers, Alice and Bob, to conduct the experiment. As shown in Figure 4(a), Alice wears the earphones and

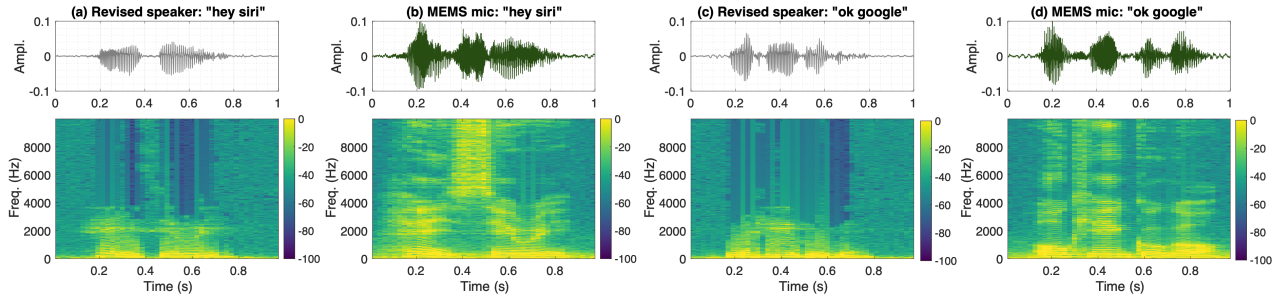


**Figure 5: We record two distinct wakeup words "Hey Siri" and "OK Google" using the pseudo-microphone and a MEMS microphone, plotting the spectrogram of the audio recordings.** Pseudo-microphone recordings of (a) "Hey Siri" and (c) "OK Google". MEMS microphone recordings of (b) "Hey Siri" and (d) "OK Google".

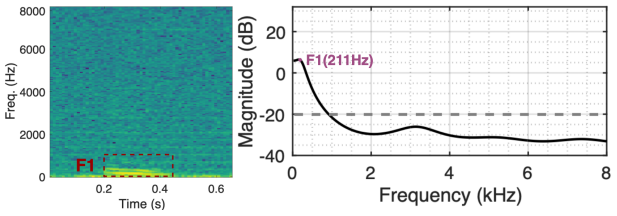


**Figure 6: The spectrogram and formants of the vowel sound /i/ captured by the earphone speaker.**

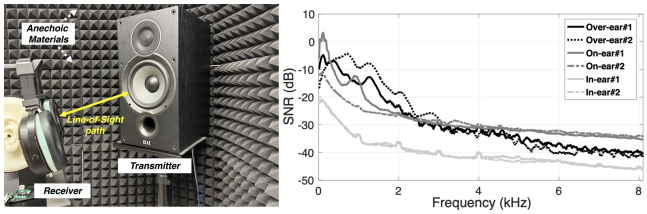


**Figure 7: Measurement setup (left) and Frequency response curve of six pairs of earphones (right).** We follow [10, 19] to play a probing signal across the frequency band to the earphone with a loudspeaker in an anechoic chamber.

acts as the primary user to activate the voice service by uttering "Hey Siri" at her preferred pace and intensity. We plot the frequency spectrogram of the signal recorded by the earphone speaker transducer in the range between 0 and 8kHz. In Figure 4(b), Bob takes on the role of the primary user, wearing the earphones and remaining stationary, while Alice acts as a nearby speaker, uttering "Hey Siri" at the same pace and intensity. To maintain a consistent over-air signal attenuation, the distance between Alice and Bob is kept identical to the distance from Alice's mouth to her ear. The earphone captures the voice from Alice via only the air channel.

**Results**. Upon comparing these two spectrograms, we observe distinct energy gaps (around 20dB), especially when we zoom in to the 0–1000Hz frequency range. This frequency range is where vibrations originating from the articulatory organs are prominent. More specifically, these articulatory organ vibrations are primarily stemming from the vocal cords and vocal tract. Vibrations related to the vocal tract, such as movements of the lips, tongue, and facial features, typically fall within the 0 to 100Hz range [24, 48]. In contrast, vocal cord vibrations span the frequency range of 100 to 1000Hz, with variations depending on genders, *i.e.*, around 90–500Hz for males while 150–1000Hz for females [68, 71].

The result indicates that the speaker transducer can capture the low-frequency signals stem from the primary speaker's vocal tract vibrations, but not from the nearby speaker. This is reasonable as both the vocal cord and tract activity travel through the body channel (in the form of bone conduction) to the earphone diaphragm, which suffers less attenuation compared with the air channel [32].

## 3.2 Wakeup Word Recognition: Challenges

The preceding section highlights the potential for distinguishing the primary speaker with dumb earphones. However, when we tested these captured wakeup words with five mainstreaming voice assistant systems, we discovered that all of them achieved very low word recognition accuracy[4], ranging from 1% to 31%. In contrast, the speech recorded by a commercial MEMS microphone achieves a recognition accuracy between 58% and 93%, as shown in Table 1.

To understand the performance gap, we examine the waveform and spectrogram of these voice recordings. As shown in Figure 5, the high-frequency components beyond 2000Hz are largely absent over our speaker recordings, whereas a MEMS microphone preserves good frequency component of the signals on the high frequency. We found the absence of high-frequency components significantly impacts the perception of formants of the wakeup words. For example, in the case of the vowel sound /i/ shown in Figure 6, due to the high-frequency deafness, only the first formant below 2kHz is observed in the earphone speaker recording while the subsequent formants above 2kHz are absent (§2). Compared with the MEMS microphone recording in Figure 3, the absence of these critical formants in the earphone recording leads to confusion in the input feature for speech recognition, ultimately causing wakeup word recognition failures.

A follow-up question arises – *what is the reason behind the absence of high-frequency components beyond 2000Hz in our speaker recordings*? Inspired by previous works [10, 58], we suspect that the speaker hardware imperfection is the root cause of this high-frequency deafness. Hence we measure the frequency response of the earphone speaker when using it as a microphone in an anechoic chamber shown in Figure 7(a).

Figure 7(b) shows the frequency response of six pairs of earphones across over-, on-, and in-ear types. We observed that the frequency response of all six pairs of earphones declines as the frequency increases. Within the 0-2000Hz frequency range, the

[4] We calculate the word level recognition accuracy $ACC$ according to the Equation $ACC = 1 - \frac{D+S+I}{D+S+C}$, where $D$, $S$, $I$ and $C$ represent the number of deletions, substitutions, insertions, and correct words.

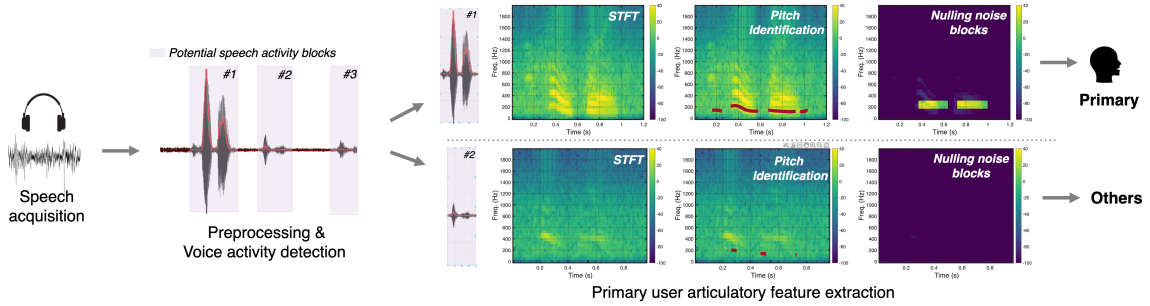


**Figure 8: An illustration of the enhancement of the joint speech detection and primary user identification.**

speaker maintains a high frequency response, which facilitates the accurate capture of vocal cord vibrations. However, as the frequency continues to rise, the speaker's frequency response decreases significantly, with an average attenuation of 30 dB. Consequently, the speech in this frequency range experience substantial attenuation, leading to reduced speech recognition accuracy.

## 4 DESIGN

We propose EarVoice to harvest the opportunities aforementioned and tackle the technical challenges identified in the preceding section. EarVoice consists of two primary functionalities, namely, speech detector and primary user identification (§4.1), and wakeup word enhancement (§4.2).

### 4.1 A Lightweight Speech Detector

This design component strives to promptly detect the presence of human speech from the audio recordings and determine whether it is its own user speaking or someone else nearby.

Existing speech detectors such as webrtc-vad [63] work in two steps. It first sends the audio recording to an energy detector to locate potential human speeches, and then feeds these high-energy pitches to a GMM model to tell whether they are human speeches or ambient noises. Although the energy detector is low-power [44], it analyzes energy levels of audio recordings across a wide frequency range spanning from 80Hz to 4000Hz, in which ambient noise frequently manifests and our pseudo-microphone (*i.e.*, using the earphone speaker as a microphone) conceals (§3.2). This can result in frequent false-triggering of the succeeding GMM-based speech detector and lead to an increase in system power consumption. Furthermore, existing speech detectors lack the capability to identify whether it is its own user talking but instead transmit all detected speech to the subsequent speech recognition module, which leads to energy wastage.

*4.1.1 Joint speech detection and primary user identification.* EarVoice instead leverages the unique in-body signal propagation channel to simultaneously identify human speech and the primary speaker through the use of only the power detector. It achieves so by detecting energy peaks specifically within the lower 1000Hz frequency band. This particular frequency range is primarily associated with the articulatory organs [71], making a strong energy peak within this band a reliable indicator of human speech presence. Furthermore, since speech from a nearby speaker propagates through an in-air channel, resulting in significant attenuation within this lower frequency band (as discussed in §3.1), we can distinguish whether the detected speech belongs to the primary speaker or someone else speaking nearby by analyzing the energy peaks within the frequency range of 0 to 1000Hz.

Our low-frequency energy detector proceeds in two steps: preprocessing and energy profiling.

**Pre-processing**. Let $x(t)$ be the audio signal recorded by the earphone's speaker transducer. We first filter $x(t)$ with a second-order Butterworth low pass filter (LPF) with a cutoff frequency of 1000Hz to eliminate the out-band noises which are largely likely to be polluted by the ambient environment noises [11]. As the user's motion noise (primarily below 50Hz [41, 74]) may still be preserved in the filtered signal, We thus adopt another Butterworth high pass filter with a cutoff frequency of 50Hz to remove human motions in that frequency band. Furthermore, due to the recorded speech energy being varied across different earphones and users, we normalize the energy of the filtered $x(t)$ by scaling it up to the range of [-4000, 4000] (dtype=int16), following the same energy normalization parameter utilized in webrtc-vad [63]. The signal normalization would not affect the relative amplitude and frequency distribution of the speech signal.

**Per-frame energy profiling**. We next locate possible voice activity on the time domain by dividing speech signals into time frames. Due to speech signals being quasi-stationary within a short time(2-50ms) [81], we divide $x(t)$ into 20ms frames and calculate the energy of each frame $i$ as follows:

$$S_i = \sum |x(n)|^2, n \in frame_i$$

where $x(n)$ are the data samples within frame $i$. EarVoice monitors the fluctuations in energy between consecutive frames and sends the audio frame(s) to the primary user identification module if their energy surpasses 1.2 times the average energy, denoted as $S_i > 1.2 \cdot S_{avg}$. The value of $S_{avg}$ is regularly updated by incorporating new frames while excluding those that have been identified as containing speech. The hyper-parameter 1.2 is obtained through our benchmark studies in various noise level settings.

*4.1.2 Enhancement.* The aforementioned procedure can detect the primary user's speeches with high accuracy because in most cases only the speech from the primary user can cause high energy peaks in the frequency below 1000Hz. However, we also noticed cases where the strong ambient noises that occupy a wide frequency band (*e.g.*, engine, wind, and road noises while driving) can fool this energy detection module, leading to false triggers of the succeeding wakeup word recognition module that is usually power hungry.

To minimize the occurrence of false activation of the wakeup word recognition module and reduce the associated power consumption, we propose to extract articulatory features from the

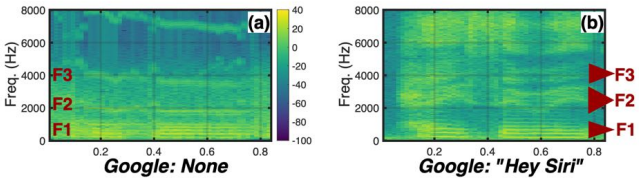


**Figure 9: (a) Reconstructed F1-F3 formants through harmonic reconstruction. Google API cannot recognize this keyword. (b) The groundtruth F1-F3 formants recorded by a MEMS microphone. Google API can successfully recognize it as "Hey Siri".**

audio recording to validate whether the detected signal represents human speech rather than mere background noise. More precisely, we segment the audio into discrete frames, where we detect the $F0$ pitch (*i.e.*, the fundamental pitch) frequency within each frame and assess the consistency of $F0$ pitch across successive frames. If the signal corresponds to human speech, the $F0$ pitch should exhibit relatively stable continuity across these frames.

We choose $F0$ pitch as our focus for several reasons. Firstly, $F0$ pitch is the essential articulation frequency determined by the rate at which the vocal cord vibrates [67] and is controlled by the tension and length of the vocal cords. As these vibrations emanate from the articulatory organs and travel through to the ear canal, the $F0$ pitch carries the most potent reference of audible energy. Secondly, the frequency of $F0$ pitch is less susceptible to certain types of interference compared with other vocal frequencies. For instance, low-frequency vocal tract resonances may be confounded by motion artifacts, and high-frequency harmonics can be masked by ambient noise.

**F0 pitch detection**. Motivated by [8, 14], we first obtain the spectrogram of the audio signal using Short Time Fourier Transform (STFT) and then detect the $F0$ pitch on the spectrogram by measuring the maximum coincidence of harmonics. The key insight is the spectrogram of a speech will exhibit prominent peaks at frequencies that are integer multiples of the $F0$ pitch, stemming from the harmonics present in the speech signal. Building on this, we establish a range of potential $F0$ pitches, ranging from 90Hz to 250Hz[5]. We then aggregate the power associated with each of these candidate pitches and its corresponding harmonics within the 1000Hz frequency range. In each time frame, we identify the pitch with the highest cumulative power as our estimated $F0$ pitch. Figure 8 illustrates this process.

Finally, we remove the noise on other frequencies to improve the SNR of the primary articulatory feature ($F0$ pitch) and feed the nullified spectrogram to a Support Vector Machine(SVM) for classification. Because the classifier focuses on detecting the continuity of the F0 pitch, a feature that doesn't vary significantly between different users, there's no necessity to amass a diverse set of training data from a large population. Moreover, the SVM's lightweight design ensures that it is computationally efficient.

It's important to note that this enhancement module is not in a constant state of activation. Instead, its activation is determined by per-frame energy profiling (§4.1), which calculates the ambient

[5] Studies show that the $F0$ frequency is around 90–180Hz for males and 165–255Hz for females [58]. We thus set the frequency band of candidate pitches to [90Hz, 250Hz] for running the $F0$ estimator.

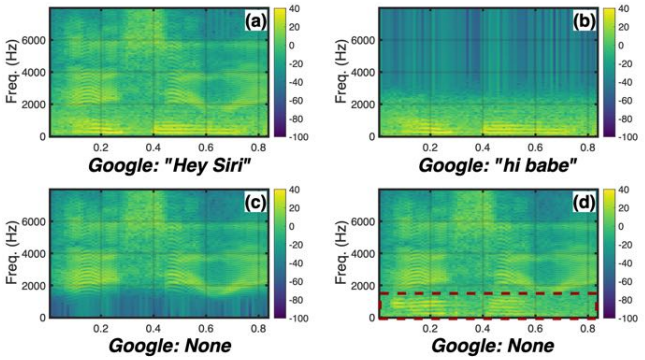


**Figure 10: Spectrogram and recognized word of each audio clip.** (a) the combined signal can be successfully recognized by Google API. (b) the speech recording with high-frequency deafness was falsely recognized as "hi babe" by Google API. (c) The high-frequency component from a template cannot be recognized by Google API. (d) The combination of a non-wakeup word and the high-frequency template cannot be recognized by Google API.

environmental energy level of each time frame. The enhancement module is activated only when the ambient energy level exceeds a predefined threshold, established based on a computation over five frames. This strategic approach allows EarVoice to activate the enhancement module in noisy environments to bolster accuracy, while also deactivating it under quieter conditions to conserve power.

## 4.2 Accurate Wakeup Word Enhancement

Once the audio speech is detected coming from the primary user, it will be sent to the wakeup word recognition module. However, as demonstrated in §3.2, directly sending the voice recording to the wakeup word recognition module associated with existing voice assistants encounters significant errors due to the absence of critical high-frequency components. We propose a lightweight wakeup word enhancement algorithm to address this issue.

*4.2.1 The failure of harmonics reconstruction.* Our initial attempt is to reconstruct the audio's high-frequency spectrogram (2–8 kHz) using their low-frequency (0–2 kHz) components that are available on the audio recordings. The opportunity here is the fundamental frequencies (*e.g.*, $F0$ pitch) in human speech manifest in higher frequency bands as harmonics (*e.g.*, $2 * F0$, $4 * F0$, ...). Following prior works [53, 58], we synthesize harmonics on 2-8kHz using the fundamental frequency components and further decay the energy across frequencies, ensuring their smoothness. However, as we sent the reconstructed audio to Google API for recognition, we found the wakeup word recognition accuracy did not get improved, maintaining at around 7%. We also fed the reconstructed audio clips released by [58] to Google API and found that these audio clips achieve similarly low accuracy.

After carefully comparing the reconstructed signal spectrogram shown in Figure 9(a) with the groundtruth shown in Figure 9(b), we found harmonics reconstruction struggles to reconstruct the formants within the higher frequency band of 2-8 kHz. This is because the formants are not solely determined by the fundamental

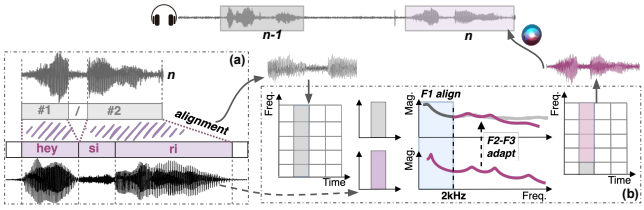


**Figure 11: (a) syllables and (b) formants alignment.**

frequency or its harmonics. It is also closely related to the physical shape and size of the user's vocal tract (§2). Accurate reconstruction of formants would require detailed information about the vocal tract's shape and size, which are typically achieved through complex acoustic modeling or data-driven approach [37] that are computationally intensive.

*4.2.2 Our solution: copy, paste, and adapt.* To mitigate the high-frequency deafness observed in the speech recording, we propose to use a MEMS microphone's pre-recording of the wakeup word (*e.g.*, "Hey Siri") as the template, extracting its high-frequency components ranging from 2 to 8kHz, and pasting it to the speech recording, as shown in Figure 10(a). This is based on an observation that when the speech recording is a wakeup word, the combined speech signal can trigger the voice assistant even though its low- and high-frequency components originate from different human speakers.

The rationale is that speech recognition systems are primarily designed to interpret content-dependent elements of human speech, such as vowels and consonants, which are characterized by these crucial formants. These systems are tuned to focus less on human speaker-dependent features like tones, prosody, and intonation, aiming to enhance the scalability of speech recognition performance [50].

Conversely, due to the lack of fundamental pitches and frequency components below 2kHz, the high-frequency component from the MEMS microphone's recording alone, as shown in Figure 10(c), cannot be successfully recognized by the wakeup word recognition module. Similarly, due to the mismatch between the low-frequency and high-frequency components, the combination of a non-pickup word speech recording and a pickup word template, also fails to trigger the voice assistant, as shown in Figure 10(d).

Yet, implementing the copy-and-paste approach poses a considerable challenge because of the diverse nature of human speech, including variations in pace, pitch, intensity, and vocal patterns. Additionally, a single user might pronounce the same wake-up word very differently at different occasions. Blindly pasting the high-frequency component of the template keyword to the speaker's speech recording can disrupt the alignment of critical formants in the combined audio signal, lead to the mismatch of the energy component in the low- and high-frequency component, and further undermine the wakeup word recognition.

To address this challenge, we propose to align the speech recording and the keyword template across three distinct dimensions: *time*, *frequency*, and *energy*. This alignment ensures that the harmonics as well as the formants in the high-frequency band are well aligned with the audio components in the low-frequency band. Next, we detail this alignment.

**Step 1. syllables alignment in time domain**. A syllable is a fundamental unit in organizing speech sounds for pronunciation in linguistic [4]. Variations in speech pace among different users can lead to discrepancies in voice duration and the number of syllables. EarVoice first aligns captured speech signals with the template by stretching/squeezing the template audio on a syllable basis. The primary challenge in this process lies in accurately detecting the boundaries of syllables in the speech recording and adjusting the template's voice speed to match that of the user, especially in the presence of background noise.

**Table 2: Comparision of word recognition accuracy.** (a): without copy-paste-adapt; (b): with copy-paste, no adapt; (c): with copy-paste-adapt; (d): with copy-paste-adapt on non-wakeup word.

| **Setup** | (a) | (b) | (c) | (d) |
|---|---|---|---|---|
| **Recog. Acc.** | 11% | 15% | 89% | 20% |

To overcome this challenge, we first calculate the energy of the ambient background noise in the speaker's audio recording and then subtract this noise to enhance the speech signal SNR, making the boundary more distinct. After that, we apply a pitch identification algorithm [8] to the speech recording to pinpoint the F0 fundamental pitch. This F0 pitch information is used to determine the number and location of syllables and the stretch ratio. The voice stretch is applied on a per-syllable basis. If EarVoice detects discrepancies in the number of syllables between the speech recording and the template (due to variations in speech pace and pronunciation habits), EarVoice merges syncopal syllables (*e.g.*, */si-ri/*) into a single syllable for alignment, as depicted in Figure 11(a).

**Step 2. formants alignment across the audible band**. After syllable alignment on the time dimension, we next align the formant components on the spectrogram. Users differ in their vocal cords and vocal tract structures, and this discrepancy can result in distinct formant location relationships in the spectrogram. For example, females typically possess a higher $F0$ pitch compared to males, causing their $F1$, $F2$, and $F3$ formants to be noticeably higher. Directly pasting the $F2$-$F3$ formants template from a female to the speech recording from a male can result in frequency misalignment, disrupt the inherent relationships among the formants, and ultimately result in errors in wakeup word recognition.

We propose to align the frequency formants on an STFT basis. As illustrated in Figure 11(b), we divide the audible band signal into a 2D time-frequency matrix. Each time frame in the matrix spans 20 ms as the audio sound is quasi-stationary over a 2-50 ms period [11]. Following the segmentation, we extract the spectral envelope of each time frame. As shown in Figure 3, the spectral envelope is an important cue for the identification of voice sounds and the characterization of formants (spectral resonances) [55]. We then align the location of the $F1$ formant (< 2kHz) in the spectral envelope by determining a shift factor. This shift factor is then adapted to the higher $F2$-$F3$ formants in the template signal. Subsequently, the adapted formant signal is copied onto the speech recording for replacement. EarVoice adopts the linear prediction spectral envelope [43] in the implementation.

**Step 3. Energy alignment**. The last step is to align the energy between the template and the speech recording. The speech loudness may change over individuals – combining the template and the speech recording in different loudness would inevitably harm the wakeup word recognition accuracy. To solve the issue, we first calculate the average energy level of the high-frequency component,

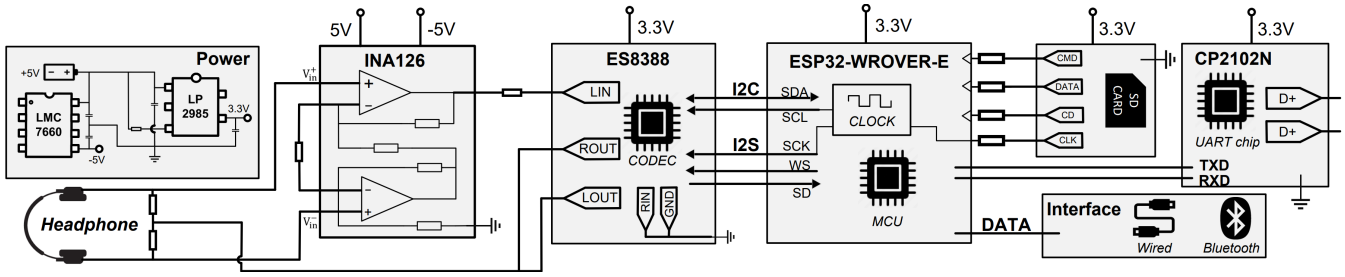


Figure 12: The schematic of EarVoice.

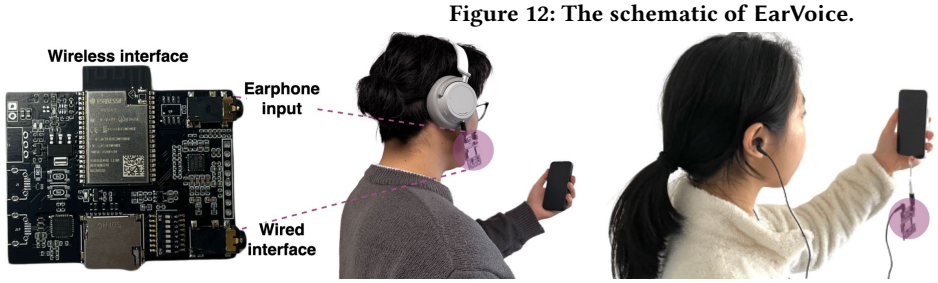


Figure 13: EarVoice supports wireless (left) and wired (right) connection.

Figure 14: Earphones.

denoted as $P_{high}$, and the low-frequency component, denoted as $P_{low}$, within the template audio. We next compute the energy level of the filtered speech recording in the low-frequency band $P'_{low}$. Finally, we adapt the high-frequency component of the combined signal using the following equation: $P'_{high} = P_{high} * (P'_{low}/P_{low})$.

**Result**. We invite a volunteer to evaluate the effectiveness of this algorithm. The volunteer is instructed to speak the wakeup word "Alexa" 100 times and random non-wakeup words 100 times at her normal communication loudness. The word recognition accuracy is shown in Table 2. We observe that our algorithm, denoted as (c), can effectively activate voice assistants with an 89% successful rate. In contrast, the success rate drops to only 11% without applying our algorithm, denoted as (a). For comparison, direct copy-and-paste has a relatively low SR recognition rate (15%) as directly applying the template on a high frequency brings in misalignment, as shown in (b). We also conducted experiments on applying the template to other non-wakeup words, denoted as (d). We found that these non-wakeup words cannot efficiently activate the SR, which demonstrates the effectiveness of our algorithm.

## 5 IMPLEMENTATION

EarVoice's signal processing includes a light-weight hardware circuit that transforms the earphone speaker into a microphone, an energy-efficient algorithm that detects human speech and distinguishes whether it is the primary user speaking, as well as a signal enhancement algorithm that improves the quality of wakeup word. All these signal modules run on a dongle. Figure 13 shows the EarVoice prototype, which supports both wireless connection (through Bluetooth) and wired connection (through a 3.5mm TRRS audio cable).

This implementation possesses two advantanges. First, because the voice detection and primary user identification features are implemented in the plug-in dongle, the earphone transducer doesn't send all captured audio streams directly to the pairing device (such as a smartphone or laptop) for further processing. Instead, the audio data is processed locally on the dongle, and only legitimate voice commands from the primary user are forwarded to the backend for further processing. Second, this gating approach not only helps prevent unintended disclosure of ambient conversations but also unnecessary acoustic signal processing on smartphones, and thus reduces power consumption.

**Hardware integration**. The EarVoice dongle comprises two 3.5 mm audio jacks, resistors in the form of a Wheatstone bridge, a power amplifier INA126, an audio codec chip ES8388, an onboard computation MCU ESP32-WROVER-E with BLE radio, a UART chip CP2102N for programming, and other peripheral electronic components. The detailed schematic is shown in Figure 12. The size of the current prototype is 6cm×4.5cm. It costs approximately 8.3 USD. Its form factor can be further reduced by adopting a stretchable PCB. We anticipate that this design can be seamlessly incorporated into mainstream True Wireless Stereo (TWS) earbuds by placing the miniaturized circuitry between the transducer and the audio chip, as suggested by previous work [46].

## 6 EVALUATION

**Data collection**. We recruited 23 volunteers (16 males, and seven females, between the ages of 18–54 years old) for the experiment under the approval of the university's Internal Review Board (IRB) protocol. The volunteers include three native speakers and 20 foreign nationals from different countries with different native languages, including Chinese, Hindi, and French, respectively. The volunteer wears EarVoice and speaks three types of wakeup words, including "Alexa", "ok Google", and "Hey Siri". The audio sampling rate is set to 16kHz. We adopt Google speech recognition API [25] as the keyword spotting model in the evaluation.

**Earphone configurations**. Voice data are collected using 13 pairs of earphones with different types (*e.g.*, over-ear, on-ear, and in-ear), prices (12-300 US dollars), and transducer sizes. Figure 14 shows the snapshot of these 13 pairs of headphones.

**Baseline**. We evaluate EarVoice against the Airpods Pro to assess its usability. The Airpods Pro takes leading position among commodity earbuds, particularly excelling in speaking sound quality. This superiority is achieved through the utilization of advanced

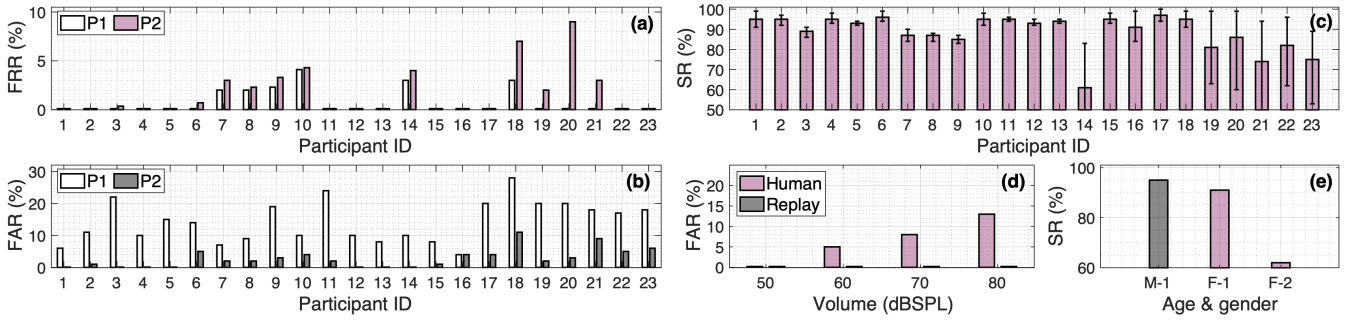


**Figure 15: (a) FRR and (b) FAR across 23 subjects (P1 refers joint speech and primary speaker detection (§4.1.1); P2 refers pitch detection-based enhancement (§4.1.2)). (c) Word recognition accuracy for 23 individuals. (d) FAR of human-based and machine-based spoofing attacks, respectively. (e) Impact of age and gender.**

sensor modalities, including the voice accelerometer and multi-microphone-based beamforming. In contrast, EarVoice only adopts the speaker transducer as the basic signal receiver. In our evaluation, we connect Airpods Pro to back-end voice assistant Siri, Google Assistant, and Amazon Alexa to evaluate the success rate for each keyword.

**Metrics**. We adopt three metrics to evaluate EarVoice:

- **False Acceptance Rate (FAR)**. This metric quantifies the frequency that EarVoice erroneously activates the voice assistant over the total number of attempts. A high FAR score can lead to an unsatisfactory user experience and inadequate privacy preservation [80].
- **False Rejection Rate (FRR)**. This metric evaluates the frequency that EarVoice does not activate the voice assistant when the primary user intent to invoke it, over the total number of attempts. A high FRR suggests EarVoice may encounter difficulties in freely accessing the voice assistant service.
- **Success Rate (SR)**. This metric quantifies the rate of successful execution over all attempts. One successful execution is counted only when the corresponding wakeup word is successfully recognized by the ASR.

## 6.1 In-lab Study

We first examine the effectiveness of EarVoice's front-end and back-end design in a controlled environment.

**Experimental procedure**. The study is divided into two sessions. In the first session, the primary subject (who wears the earphone) is instructed to utter the wake-up words at her preferred pace and intensity. Each command was uttered 20 times per user with different earphones. We then compute the false rejection rate (FRR). In the second session, we let the primary subject stay silent and invite another volunteer to speak the same wake-up word near the primary subject, playing the role of a nearby individual shown in Figure 4 (b). We then calculate the false acceptance rate (FAR). Each session takes around 30 minutes. We train the SVM model in §4.1.2 with the collected two-session dataset. Specifically, we use subject 1's voice for training the SVM and test it on the other 22 unseen participants. And we train a second SVM on another unseen user (e.g., subject 2) to evaluate the FAR and FRR of subject 1. The input is the nullified spectrogram of the voice signal and the output is the classification result (*i.e.*, 0/1: represent primary user/others). All experiments are conducted in a quiet lab environment with an ambient noise level at 45 dBSPL on average.

• **Primary speaker identification**. We examine the overall accuracy of the primary speaker identification in EarVoice. The evaluation is conducted in two phases. In the first phase (P1), we only apply the time framing identification method (§4.1.1) and examine the FRR and FAR results. As shown in Figure 15 (a) and (b), we observe a consistently low average FRR (0.8%) but a higher average FAR (14.3%) across the 23 subjects. This outcome is expected since time framing primarily detects energy presence, not specific user identification. Afterward, we incorporate the pitch detection ( §4.1.2) and observe significant improvements. The FAR drops to 2.8%, while the FRR slightly increases to 1.7%. These findings demonstrate the effectiveness of our pitch detection algorithm.

Taking a further scrutiny of these results, we find that subjects 9, 10, 18, and 21 exhibit relatively higher FRR and FAR (*e.g.*, >3%). This discrepancy can be attributed to the inadequate contact of earphones with the subjects' skin, impacting the propagation of vocal cord vibrations through bone conduction and resulting in an increased FRR. Simultaneously, this lack of close contact allows the speaker transducer to capture speech from nearby users, contributing to a higher FAR. Additionally, subjects 14, 19, and 20 exhibit a higher FRR but maintain a lower FAR in comparison to others. Further investigation into the raw audio recording of these subjects reveals that their voice volume is lower than that of other subjects, consequently leading to more frequent rejections by the EarVoice.

**Spoofing attacks**. Safeguarding against voice attacks and eliminating false positives is crucial for voice assistants. To further verify the effectiveness of EarVoice on primary speaker identification and the possibility of false triggers. We emulate two types of spoofing attacks, including a human-based and a machine-based reply attack. In the human-based attack, We invite a participant to wake up voice assistants with different volumes near the true primary user who wears the earphone. In the replay attack, we pre-record the primary user's voice and play it with a loudspeaker with different volumes near the earphone. The distance between the attacker and the earphone is kept to 50 cm.

Figure 15 (d) shows the result. Overall EarVoice demonstrates a promising defensive capability against these spoofing attacks. Specifically, the human-based attack yields an average of 6% FAR across all speaker volumes. Even at the attacker's maximum volume (80 dBSPL), the FAR only rises to approximately 13%. As a

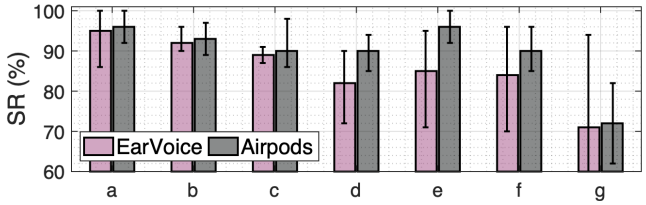


**Figure 16: Success rate of EarVoice in seven scenarios.**

comparison, a machine-based reply attack never survives to awake EarVoice. This disparity in outcomes may be attributed to the inherent differences between human and machine vocal systems. Specifically, loudspeakers typically exhibit lower efficiency in reproducing lower-frequency sounds, which makes EarVoice more effective against such a voice attack.

• **Wakup word recognition**. We next evaluate the effectiveness of wakeup word recognition using our *copy, paste, and adapt* design. Figure 15 (c) shows the recognition success rate for each individual. The error bars in the figure indicate performance variations across three different wakeup words. Overall, EarVoice achieves a success rate of 89% on average. Dig deeper, subject 14 achieves the lowest SR at 61% due to her lowest voice volume. Such reduced volume adversely affects pitch detection accuracy, subsequently impacting the precision of the alignment processes. Notably, subjects 19–23 show large variations among the three wakeup words. The result might be attributed to the lower fluency in pronouncing the words compared with the others.

**Impact of age and gender**. We focus on one wakeup word (*i.e.*, Alexa) and categorize the 23 participants into three groups based on their age and genders: M-1 (male, <31 years old), F-1 (female, <31 years old), and F-2 (female, 32–55 years old), respectively. As depicted in Figure 15(e), we observe that both M-1 and F-1 groups exhibit similar recognition accuracies, with the M-1 group achieving a slightly higher accuracy (95%) compared to the F-1 group (91%). This marginal difference may be attributed to the typically stronger vocal vibrations observed in males. Furthermore, the F-2 group, particularly participants 19 and 23, demonstrates significantly lower recognition accuracy at 62%. This reduction in performance may be attributed to factors such as less familiar English pronunciations and lower vocal volumes observed in the participants.

## 6.2 Field Study

We next assess EarVoice's end-to-end performance across various real-world scenarios. As shown in Figure 17, the evaluation encompasses four stationary and three mobility scenarios to represent typical indoor and outdoor settings. In each scenario, we collected 100 utterances for each wakeup word. We then examine the overall success rate of wakeup word recognition. Airpods are adopted for comparison. Figure 16 shows the results.

• **Stationary scenarios (a)-(d)**. EarVoice achieves a success rate of 95%, 92%, 89%, and 82% for these four static scenarios, respectively. The overall accuracy is at around 90%, which is slightly worse than that of Airpods (92%). A relatively bigger gap between EarVoice and Airpods is observed in scenario (d). This suggests that severe noise artifacts, as encountered in (d), can still be perceived by earphone speakers and impact the accuracy of template matching, consequently affecting the recognition of wakeup words.

• **Mobile scenarios (e)-(g)**. We further extend our investigation to include three types of mobility. The results of these activities are shown in Figure 16 (e)-(g). Notably, during (e) driving and (f) lifting, EarVoice achieves an average success rate of 85% and 84%, respectively. The success rate is slightly lower than in stationary environments with comparable noise levels. This decline in performance is primarily attributed to the head and upper body movements during driving and lifting, which adversely affect the signal input quality. In contrast, AirPods maintain a higher average success rate of 93%. The success rate further drops to 71% while walking at a busy intersection, influenced by noise from moving vehicles nearby and motion artifacts from the individual. The success rate of AirPods falls to 72% in these conditions.

**Results discussion**. In contrast to Airpods which leverages advanced sensors and beamforming technologies to improve the voice quality, EarVoice relies solely on the earphone's speaker transducer for voice activation and a lightweight signal processing algorithm for wakeup word enhancement. The manufacturing cost of EarVoice is approximately 8 dollars, tens of times lower than Airpods, while striving to approach a comparable performance.

## 6.3 Mirco-Benchmarks

We further conduct benchmark studies to understand the effect of various factors on EarVoice's performance.

• **Impact of music playback**. EarVoice's hardware is built upon HeadFi [20] which adopts a differential circuit (*i.e.*, Wheatstone bridge) to cancel the music interference on the user voice recording. To assess the impact of music on system performance, we invite a volunteer to conduct the speech activation experiment while listening to music at volumes ranging from 5% to 60% of the maximum, in accordance with the audiology's 60-60 rule [6] for safe listening [59]. The participant is instructed to speak three types of wakeup words 100 times each at varying music volumes.

Figure 18 shows the result. We observe that EarVoice achieves an average success rate of 98% and 89% at speaker volumes increasing modestly from 5% to 20% of the maximum, respectively. These results affirm EarVoice's capability to activate voice assistants during music playback. However, a discernible decline in success rate was observed at higher volumes: dropping to 74% at 40% volume and further to 54% at 60% volume. This performance reduction could be attributed to two factors: one is the discrepancy in impedance between the left and right earphone transducers, leading to electronic music signal leakage and interference with speech commands; the other is the music echos inside the ear canal can be captured by the speaker transducer during vocal signal recording, which negatively affects the system performance.

• **Impact of different earphones**. We invite one participant to conduct the speech activation experiment by wearing six pairs of earphones (out of 13) in the lab and speaking three types of wakeup words, with each wakeup word repeating 100 times. Airpods are adopted for comparison. The result is shown in Figure 19. Overall, we observe that EarVoice achieves an average success rate of 87% over all types of earphones. Notably, over-ear and on-ear earphones achieve the highest success rate with an average of 92% and 91% SR, respectively. These results are on par with Airpods (with an average

[6] Listen at 60 percent of the maximum volume for no more than 60 minutes a day.

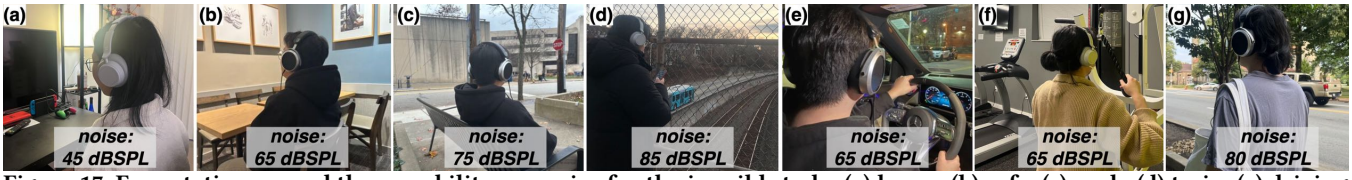


**Figure 17: Four stationary and three mobility scenarios for the in-wild study: (a) home; (b) cafe; (c) park; (d) train; (e) driving car; (f) lifting in the gym; (g) walking on a busy intersection.**

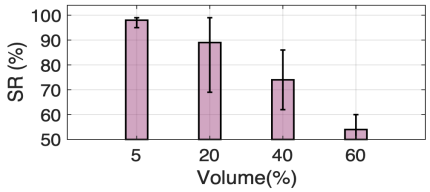


**Figure 18: Impact of music playback.**

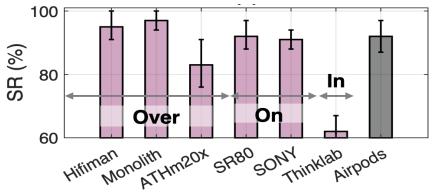


**Figure 19: Impact of earphones.**

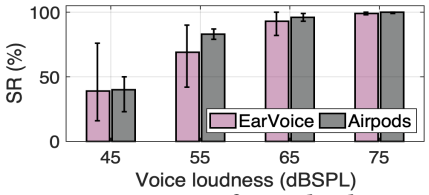


**Figure 20: Impact of voice loudness.**

**Table 3: System latency breakdown.** Joint speech and primary speaker detection (§4.1.1). Pitch detection-based enhancement (§4.1.2). Copy-paste-adapt operation (§4.2).

| **Design comp.** | §4.1.1 | §4.1.2 | §4.2 |
|---|---|---|---|
| **Latency** | 3ms | 159ms | 25ms |

success rate of 92%), demonstrating EarVoice's effectiveness across over-ear and on-ear earphones.

However, EarVoice's performance is notably lower with in-ear earphones, with a success rate of 62% on average. One reason for the better performance of over-ear and on-ear earphones can be attributed to their larger speaker transducers and inherently larger surface contact with the skull, allowing for more efficient transfer of vocal cord vibration energy. In contrast, the smaller transducers of in-ear earphones exhibit reduced sensitivity to voice commands (Figure 7). A potential solution is to adjust the speaker volume or incorporate a power amplifier into our dongle to enhance the signal strength of the speech recording.

• **Impact of different voice loudness**. We next evaluate the impact of voice loudness on EarVoice's success rate. Similarly, we invited one participant to utter the three types of wakeup words with four different loudness levels, spanning from 45 to 75 dBSPL. The range is selected based on CDC's regulation [13], Specifically, it designates approximately 40 dBSPL for a whisper, 60-70 dBSPL for a normal voice level, and 75-85 dBSPL for a loud voice conversation. As shown in Figure 20, we find that as the voice loudness increases, the success rate of EarVoice grows by 2.5× from 39% to 99%. A similar trend can be found on Airpods as well, which shows the success rate grows from 40% to 100%. Notably, the success rate of EarVoice is relatively stable (*i.e.*, 93% – 99%) when the voice loudness level surpasses 55 dBSPL. This result demonstrates EarVoice's resilience in handling normal voice conversations.

• **System overhead and latency**. We also evaluate system overhead and processing latency. Table 3 details the processing delay of the front-end design (§4.1.1 & §4.1.2), and *copy, paste, and adapt* design (§4.2), respectively. The measurement is conducted on a 2-second audio sample extracted from the audio stream. We observe that joint speech and primary speaker detection (§4.1.1) takes around 3ms for processing the 2s audio sample. The pitch detection-based enhancement §4.1.2 takes 159ms. The *copy, paste, and adapt* design (§4.2) takes around 25ms to process a 2-second audio sample. The overall signal processing delay is around 200ms, demonstrating the capability of real-time operations. We anticipate the delay will drop further through multi-thread processing.

**Table 4: Power consumption breakdown.**

| **Hard. comp.** | Sensing | Codec | MCU |
|---|---|---|---|
| **Power** | 0.2mW | 60mW | 152mW |

Table 4 summarizes the power consumption of each component. Given a supply voltage of 5V, the sensing module, audio codec, and MCU consume 0.2mW, 60mW, and 152mW, respectively. The total power consumption of EarVoice is approximately 212 mW in the active mode. An 820 mAh lithium battery can be used to provide up to 19.3 hours of continuous running of EarVoice. The battery life could be further optimized with duty-cycles.

## 7 RELATED WORK

**Voice Assistant Activation Technologies**. Existing general purpose voice activity detection (VAD) modules, *e.g.*, Google's webrtc-vad [52], GPVAD [18], and Kaldi-VAD [50], have been well-studied and integrated into many mobile applications. Nevertheless, applying these designs to earphones face challenges as voice communication on earphones can be plagued by environmental noise and more severely, the speech commands from nearby individuals.

To solve the issue, personalized VAD [16, 17, 60, 61, 72, 77] with identifying the target user's voice fingerprint has been proposed. But these personalized solutions are generally power intensive and struggle to counteract spoofing attacks. Hence they are not widely adopted by consumer devices.

Besides, various research approaches [31, 54, 75] have also been developed for simplifying voice activation by involving hand gestures. For example, Raise to Speak [80] enables Apple Watch being able to activate the voice assistant by detecting the raising hand gesture. ProxiMic [54] explores the close-to-mic voice characteristics (*e.g.*, pop noise) and enables voice activation by placing the microphone close to the user's mouth. PrivateTalk [75] activating voice input with user-defined hands-on-month gestures for earphone devices. Although these approaches guarantee low false positives, they inevitably require the involvement of hand gestures and thus bring extra burden for the users.

Different from the aforementioned works, EarVoice takes advantage of an opportunity hidden in the earphone transducer and develops a hands-free voice activation system while guaranteeing low false positives towards environmental noise and false triggering voice commands from nearby people. The proposed signal-processing algorithm could run efficiently on mobile and embedded devices without complex computation requirements.

**Bone Conduction Channels**. Recently, bone conduction sensors [5, 21, 26, 76], such as IMU [26], in-ear microphone [40], voice pickup sensor (VPU) [60], non-audible murmur (NAM) and throat microphone [45], have been explored for speech enhancement, voice activation, and speaker verification [21, 39, 62]. For example, WhisperMask [28] designs a new interface that catches the user's whispering speech with an embedded condenser microphone woven hidden in a non-woven mask to reduce the noise interference from the environment. In-Ear-Voice [60] developed a low-power personalized VAD system for hearables by exploring the bone conduction sensor. VibVoice [26] utilized the bone conduction response from IMU sensors to enhance speech quality in a noisy environment. These pioneer works demonstrate promising results, but they cannot be deployed on existing earphones due to the lack of such onboard sensors. In contrast, our study explores the bone conduction effect on the speaker transducer which pervasively exists on every earphone.

HeadFi [20] explores the reciprocal principle of earphones and demonstrates the capability of using the earphone transducer for user identification, physiological sensing, touch gesture recognition, *etc.* Our hardware dongle builds upon HeadFi but extends it to a software-hardware system that explores two different voice channels to enable hands-free voice activation. Moreover, the high-frequency deafness associated with speaker transducers introduces unique challenges to activating voice assistants and motivates us with the copy-paste-adapt keyword enhancement design to thoroughly improve the activation accuracy and enhance speech quality.

**Whisper or Silent Speech Interface**. Researchers also explore novel silent speech interface technologies [36, 66] for enriching speech recognition interfaces. For example, LipLearner [65] proposes a customizable silent speech interface on mobile phones by building up the relationship between voice commands and corresponding non-verbal lip movements through a neural network model. It allows users to activate the speech service with lip motions. HPSpeech [78] creates a silent speech interface on earphones by emitting inaudible acoustic signals to detect the movement of temporomandibular joint (TMJ) for silent voice command recognition. MuteIt [64] tracks the user's jaw motion with a dual-IMU setup to infer word articulation around the ear. EarCommand [31] emits an ultrasonic signal in the ear canal and builds the relationship between the deformation of the ear canal and the movements of the articulator to infer the corresponding silent speech commands while speaking. Unlike the aforementioned works that aim to establish new paradigms for speech interaction, EarVoice adheres to the current speech recognition (SR) service, focusing on enhancing their reliability.

## 8 DISCUSSION

EarVoice leaves room for future improvement, as discussed below:

**Scale to smart ANC earbuds**. EarVoice aims to facilitate hands-free voice assistant activation across all earphone types. Leveraging the universal presence of speaker transducers in earphones, our solution is broadly applicable to different earphone models. Although our current prototypes are only tested on traditional wired earphones, we believe the proposed signal processing designs can be applied to ANC earbuds as well as their onboard accelerometer sensors also show bone-conduction properties [26]. We leave such exploration for future work.

**Wakeup words selection**. Our current evaluation focuses on the three most widely used wakeup words (*i.e.*, Alexa, Hey Siri, and OK Google), which can minimize the user's learning curve and avoid additional user effort when interacting with our system. However, we acknowledge that limiting our evaluation to these three keywords might constrain the breadth of our findings. We recognize this as a limitation in our current study. In the future, we will investigate the the system performance based on a wider range of wake-up words.

**Improving the system performance**. Our benchmark evaluations reveal that EarVoice exhibits comparatively lower performance when used with in-ear earphones, as opposed to out-ear and on-ear earphones. This performance discrepancy stems from the smaller transducer size in in-ear earphones, which limits the area of contact with the skull. Consequently, the energy perception of the vocal cord vibration is relatively lower. One potential solution to tackle the challenge is to integrate a power amplifier within the hardware dongle and bolster the strength of the signal captured during speech recording. The tradeoff, however, is the higher power consumption, which is worth further exploration.

Similarly, the presence of motion artifacts, ambient noise, and music introduces extra interference with the perceived speech commands. Such disturbances lead to a reduced Signal-to-Interference-plus-Noise Ratio (SINR), adversely affecting the perceived clarity of speech commands and impacting the accuracy of template matching, consequently, degrading the system performance. To address the challenge, one promising solution is deep neural network-based acoustic signal enhancement [49] or denoising [15]. We leave such exploration for future work.

## 9 CONCLUSION

We have presented the design, implementation, and evaluation of EarVoice, a software-hardware solution that enables mobile users to activate their voice assistant on earphones without hand gesture intervention. EarVoice contributes a plethora of low-power signal processing algorithms that take advantage of the two speech signal propagation channels to detect the human speech, differentiate the primary speaker, and further enhance the quality of the wakeup word for accurate wakeup word recognition. The experiment in different real-world scenarios demonstrated the efficacy and effectiveness of EarVoice.

## ACKNOWLEDGMENT

We thank our Shepherd Felix Lin and other reviewers for their insightful comments. This work is supported by the National Science Foundation under Grant No.2337537, the SIGMOBILE Student Community Grant, and a Google Research Scholar Award.

# REFERENCES


[1] activate Google Assistant on headphone [n. d.]. activate the Google Assistant function on your headphones. https://www.sony.com/electronics/support/articles/00202243.

[2] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 1–28.

[3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*. PMLR, 173–182.

[4] Juliette Blevins and John Goldsmith. 1995. The syllable in phonological theory. *1995* (1995), 206–244.

[5] Bone conductive microphone [n. d.]. Bone conductive microphone. https://earhugger.com/product/ear-bone-microphone/.

[6] Bose QuietComfort Earbuds [n. d.]. Bose QuietComfort Earbuds. https://www.bose.com/en_us/support/articles/HC2648/productCodes/qc_earbuds/article.html.

[7] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[8] Arturo Camacho and John G Harris. 2008. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America* (2008).

[9] Tao Chen, Xiaoran Fan, Yongjie Yang, and Longfei Shangguan. 2022. Towards Remote Auscultation with Commodity Earphones. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*.

[10] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. 2020. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In *Network and Distributed Systems Security (NDSS) Symposium*.

[11] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. 2023. The Design and Implementation of a Steganographic Communication System over In-Band Acoustical Channels. *ACM Transactions on Sensor Networks* (2023).

[12] Tao Chen, Yongjie Yang, Xiaoran Fan, Xiuzhen Guo, Jie Xiong, and Longfei Shangguan. 2024. Exploring the Feasibility of Remote Cardiac Auscultation Using Earphones. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*.

[13] Common Sources of Noise and Decibel Levels-CDC [n. d.]. Common Sources of Noise and Decibel Levels-CDC. https://www.cdc.gov/nceh/hearing_loss/what_noises_cause_hearing_loss.html.

[14] Patricio De La Cuadra, Aaron S Master, and Craig Sapp. 2001. Efficient pitch detection techniques for interactive music. In *ICMC*.

[15] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847* (2020).

[16] Shaojin Ding, Rajeev Rikhye, Qiao Liang, Yanzhang He, Quan Wang, Arun Narayanan, Tom O'Malley, and Ian McGraw. 2022. Personal VAD 2.0: Optimizing personal voice activity detection for on-device speech recognition. *arXiv preprint arXiv:2204.03793* (2022).

[17] Shaojin Ding, Quan Wang, Shuo-yiin Chang, Li Wan, and Ignacio Lopez Moreno. 2019. Personal VAD: Speaker-conditioned voice activity detection. *arXiv preprint arXiv:1908.04284* (2019).

[18] Heinrich Dinkel, Yefei Chen, Mengyue Wu, and Kai Yu. 2020. Voice Activity Detection in the Wild via Weakly Supervised Sound Event Detection. In *Proc. Interspeech 2020*. 3665–3669. https://doi.org/10.21437/Interspeech.2020-0995

[19] Tingchao Fan, Huangwei Wu, Meng Jin, Tao Chen, Longfei Shangguan, Xinbing Wang, and Chenghu Zhou. 2023. Towards Spatial Selection Transmission for Low-end IoT devices with SpotSound. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*.

[20] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard Howard. 2021. HeadFi: bringing intelligence to all headphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*.

[21] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 343–355.

[22] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 337–349.

[23] Sonia Garcia-Salicetti, Charles Beumier, Gérard Chollet, Bernadette Dorizzi, Jean Leroux les Jardins, Jan Lunter, Yang Ni, and Dijana Petrovska-Delacrétaz. 2003. BIOMET: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. In *Audio-and Video-Based Biometric Person Authentication: 4th International Conference, AVBPA 2003 Guildford, UK, June 9–11, 2003 Proceedings 4*. Springer, 845–853.

[24] Asif A Ghazanfar and Daniel Y Takahashi. 2014. Facial expressions and the evolution of the speech rhythm. *Journal of cognitive neuroscience* (2014).

[25] Google STT [n. d.]. Google STT. https://cloud.google.com/speech-to-text.

[26] Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. 2023. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 14–27.

[27] Blanca Hernandez-Ortega and Ivani Ferreira. 2021. How smart experiences build service loyalty: The importance of consumer love for smart voice assistants. *Psychology & Marketing* 38, 7 (2021), 1122–1139.

[28] Hirotaka Hiraki, Shusuke Kanazawa, Takahiro Miura, Manabu Yoshida, Masaaki Mochimaru, and Jun Rekimoto. 2023. External noise reduction using WhisperMask, a mask-type wearable microphone. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–5.

[29] Changshuo Hu, Xiao Ma, Dong Ma, and Ting Dang. 2023. Lightweight and Non-Invasive User Authentication on Earables. In *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications*. 36–41.

[30] iFLYTEK ASR [n. d.]. iFLYTEK ASR. https://global.xfyun.cn/.

[31] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: " Hearing" Your Silent Speech Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.

[32] Eugenijus Kaniusas. 2007. *Acoustical signals of biomechanical systems*. World Scientific.

[33] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Alessandro Montanari. 2018. Earables for personal-scale behavior analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89.

[34] Linus Kendall, Bidisha Chaudhuri, and Apoorva Bhalla. 2020. Understanding technology as situated practice: everyday use of voice user interfaces among diverse groups of users in urban India. *Information Systems Frontiers* 22 (2020), 585–605.

[35] Raymond D Kent and Houri K Vorperian. 2018. Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of communication disorders* 74 (2018), 74–97.

[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*. 44–49.

[37] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 17022–17033.

[38] Valeri Aleksandrovich Kozhevnikov and Liudmila Andreevna Chistovich. 1965. Speech: Articulation and perception. (1965).

[39] Rui Liu, Cory Cornelius, Reza Rawassizadeh, Ronald Peterson, and David Kotz. 2018. Vocal resonance: Using internal body voice for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–23.

[40] Dong Ma, Ting Dang, Ming Ding, and Rajesh Balan. 2023. ClearSpeech: Improving Voice Quality of Earbuds Using Both In-Ear and Out-Ear Microphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2023).

[41] Dong Ma, Andrea Ferlini, and Cecilia Mascolo. 2021. Oesense: employing occlusion effect for in-ear human sensing. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 175–187.

[42] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2023. LLM-Powered Conversational Voice Assistants: Interaction Patterns, Opportunities, Challenges, and Design Guidelines. *arXiv preprint arXiv:2309.13879* (2023).

[43] John Makhoul. 1973. Spectral analysis of speech by linear prediction. *IEEE Transactions on Audio and Electroacoustics* 21, 3 (1973), 140–148.

[44] Gabriele Meoni, Luca Pilato, and Luca Fanucci. 2018. A low power voice activity detector for portable applications. In *2018 14th conference on Ph. D. research in microelectronics and electronics (PRIME)*. IEEE, 41–44.

[45] Yoshitaka Nakajima, Hideki Kashioka, Nick Campbell, and Kiyohiro Shikano. 2006. Non-audible murmur (NAM) recognition. *IEICE TRANSACTIONS on Information and Systems* 89, 1 (2006), 1–8.

[46] ohmic technologies [n. d.]. ohmic technologies. https://www.ohmictechnologies.com/.

[47] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[48] Hyojin Park, Christoph Kayser, Gregor Thut, and Joachim Gross. 2016. Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *Elife* (2016).

[49] Santiago Pascual, Antonio Bonafonte, and Joan Serra. 2017. SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452* (2017).

[50] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[51] Swadhin Pradhan, Wei Sun, Ghufran Baig, and Lili Qiu. 2019. Combating replay attacks against voice assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.

[52] py-webrtcvad [n. d.]. py-webrtcvad. https://github.com/wiseman/py-webrtcvad.

[53] Yingyong Qi and Robert E Hillman. 1997. Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *The Journal of the Acoustical Society of America* 102, 1 (1997), 537–543.

[54] Yue Qin, Chun Yu, Zhaoheng Li, Mingyuan Zhong, Yukang Yan, and Yuanchun Shi. 2021. ProxiMic: Convenient voice activation via close-to-mic speech detected by a single microphone. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

[55] LR Rabiner and BH Juang. 1993. Fundamentals of speech Recognition (Prentice Hall PTR. *Upper Saddle River, New Jersey* (1993).

[56] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[57] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624* (2021).

[58] Nirupam Roy and Romit Roy Choudhury. 2016. Listening through a vibration motor. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*.

[59] Safe-Listening Tips [n. d.]. Safe-Listening Tips. https://hearing.health.mil/Prevention/Dangers-of-Loud-Noise/Safe-Listening/.

[60] Philipp Schilk, Niccolò Polvani, Andrea Ronco, Milos Cernak, and Michele Magno. 2023. In-Ear-Voice: Towards Milli-Watt Audio Enhancement With Bone-Conduction Microphones for In-Ear Sensing Platforms. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*.

[61] Irtaza Shahid, Yang Bai, Nakul Garg, and Nirupam Roy. 2022. Voicefind: Noise-resilient speech recovery in commodity headphones. In *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*. 13–18.

[62] Jiacheng Shang and Jie Wu. 2019. Enabling secure voice input on augmented reality headsets using internal body voice. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[63] Mayank Sharma, Sandeep Joshi, Tamojit Chatterjee, and Raffay Hamid. 2022. A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows. *Neurocomputing* 494 (2022), 116–131.

[64] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–26.

[65] 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.

[66] Phuc Nguyen Tanmay Srivastava, Shijia Pan and Shubham Jain. 2023. Jawthenticate: Microphone-free Speech-based Authentication using Jaw Motion and Facial Vibrations. In *Proceedings of the 21th ACM Conference on Embedded Networked Sensor System*.

[67] Ingo R Titze. 1976. On the mechanics of vocal-fold vibration. *The Journal of the Acoustical Society of America* (1976).

[68] Ingo R Titze and Eric J Hunter. 2004. Normal vibration frequencies of the vocal ligament. *The Journal of the Acoustical Society of America* (2004).

[69] Use gestures to control your Google Assistant on headphones [n. d.]. Use gestures to control your Google Assistant on headphones. https://support.google.com/assistant/answer/7513985?hl=en&co=GENIE.Platform%3DAndroid.

[70] Use Siri with AirPods (1st or 2nd generation) [n. d.]. Use Siri with AirPods (1st or 2nd generation). https://support.apple.com/guide/airpods/use-siri-with-airpods-1st-or-2nd-generation-dev8779e5576/web.

[71] Vocal Fold Vibration and Pitch [n. d.]. Vocal Fold Vibration and Pitch. https://med.umn.edu/ent/patient-care/lions-voice-clinic/about-the-voice/how-it-works/physiology.

[72] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. 2018. Speaker diarization with LSTM. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5239–5243.

[73] WH-1000XM4 [n. d.]. WH-1000XM4. https://helpguide.sony.net/mdr/wh1000xm4/v1/en/contents/TP0002754730.html.

[74] Zhen Xiao, Tao Chen, Yang Liu, and Zhenjiang Li. 2020. Mobile phones know your keystrokes through the sounds from finger's tapping on the screen. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*.

[75] Yukang Yan, Chun Yu, Yingtian Shi, and Minxing Xie. 2019. PrivateTalk: Activating Voice Input with Hand-On-Mouth Gesture Detected by Bluetooth Earphones. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*.

[76] Yongjie Yang, Tao Chen, Yujing Huang, Xiuzhen Guo, and Longfei Shangguan. 2024. MAF: Exploring Mobile Acoustic Field for Hand-to-Face Gesture Interactions. In *Proceedings of CHI*.

[77] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. 2019. Fully supervised speaker diarization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6301–6305.

[78] Ruidong Zhang, Hao Chen, Devansh Agarwal, Richard Jin, Ke Li, François Guimbretière, and Cheng Zhang. 2023. HPSpeech: Silent Speech Interface for Commodity Headphones. In *Proceedings of the 2023 International Symposium on Wearable Computers*. 60–65.

[79] Zhaoyan Zhang. 2016. Mechanics of human voice production and control. *The journal of the acoustical society of america* (2016).

[80] Shiwen Zhao, Brandt Westing, Shawn Scully, Heri Nieto, Roman Holenstein, Minwoo Jeong, Krishna Sridhar, Brandon Newendorp, Mike Bastian, Sethu Raman, et al. 2019. Raise to Speak: An Accurate, Low-Power Detector for Activating Voice Assistants on Smartwatches. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2736–2744.

[81] Eberhard Zwicker and Hugo Fastl. 2013. *Psychoacoustics: Facts and models*. Springer Science & Business Media.