The EarSAVAS Dataset: Enabling Subject-Aware Vocal Activity Sensing on Earables

XIYUXING ZHANG, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, China
YUNTAO WANG*, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, China
YUXUAN HAN, Department of Computer Science and Technology, Tsinghua University, China
CHEN LIANG, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, China
ISHAN CHATTERJEE, Paul G. Allen School of Computer Science and Engineering, University of Washington, United States
JIANKAI TANG, Xinya College, Tsinghua University, China
XIN YI, Institute for Network Sciences and Cyberspace, Tsinghua University, China
SHWETAK PATEL, Paul G. Allen School of Computer Science and Engineering, University of Washington, United States
YUANCHUN SHI, Department of Computer Science and Technology, Tsinghua University, China and Intelligent Computing and Application Laboratory of Qinghai Province, Qinghai University, China

Subject-aware vocal activity sensing on wearables, which specifically recognizes and monitors the wearer's distinct vocal activities, is essential in advancing personal health monitoring and enabling context-aware applications. While recent advancements in earables present new opportunities, the absence of relevant datasets and effective methods remains a significant challenge. In this paper, we introduce EarSAVAS, the first publicly available dataset constructed specifically for subject-aware human vocal activity sensing on earables. EarSAVAS encompasses eight distinct vocal activities from both the earphone wearer and bystanders, including synchronous two-channel audio and motion data collected from 42 participants totaling 44.5 hours. Further, we propose EarVAS, a lightweight multi-modal deep learning architecture that enables efficient subject-aware vocal activity recognition on earables. To validate the reliability of EarSAVAS and the efficiency of EarVAS,

*This is the corresponding author.

Authors' addresses: Xiyuxing Zhang, zxyx22@mails.tsinghua.edu.cn, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, China; Yuntao Wang, yuntaowang@tsinghua.edu.cn, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, China; Yuxuan Han, hanyx21@mails.tsinghua.edu.cn, Department of Computer Science and Technology, Tsinghua University, China; Chen Liang, lliangchenc@163.com, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, China; Ishan Chatterjee, ichat@cs.washington.edu, Paul G. Allen School of Computer Science and Engineering, University of Washington, United States; Jiankai Tang, tjk19@mails.tsinghua.edu.cn, Xinya College, Tsinghua University, China; Xin Yi, yixin@tsinghua.edu.cn, Institute for Network Sciences and Cyberspace, Tsinghua University, China; Shwetak Patel, shwetak@cs.washington.edu, Paul G. Allen School of Computer Science and Engineering, University of Washington, United States; Yuanchun Shi, shiyc@tsinghua.edu.cn, Department of Computer Science and Technology, Tsinghua University, China and Intelligent Computing and Application Laboratory of Qinghai Province, Qinghai University, China.

2474-9567/2024/5-ART83
https://doi.org/10.1145/3659616

we implemented two advanced benchmark models. Evaluation results on EarSAVAS reveal EarVAS's effectiveness with an accuracy of 90.84% and a Macro-AUC of 89.03%. Comprehensive ablation experiments were conducted on benchmark models and demonstrated the effectiveness of feedback microphone audio and highlighted the potential value of sensor fusion in subject-aware vocal activity sensing on earables. We hope that the proposed EarSAVAS and benchmark models can inspire other researchers to further explore efficient subject-aware human vocal activity sensing on earables.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Dataset, Human Vocal Activity Recognition, Subject Awareness, Active Noise Cancelling Earables, Deep Learning

ACM Reference Format:
Xiyuxing Zhang, Yuntao Wang, Yuxuan Han, Chen Liang, Ishan Chatterjee, Jiankai Tang, Xin Yi, Shwetak Patel, and Yuanchun Shi. 2024. The EarSAVAS Dataset: Enabling Subject-Aware Vocal Activity Sensing on Earables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 83 (May 2024), 26 pages. https://doi.org/10.1145/3659616

1 INTRODUCTION

Subject-aware sensing of human vocal activity on wearable devices refers to the recognition and monitoring of the wearer's distinct vocal activities, considering the inherently personal nature of wearables, designed to serve exclusively the wearer, and filtered out potential interference from bystanders' vocal sounds. Efficient subject-aware vocal activity sensing methods pave the way for numerous applications, including but not limited to personal health monitoring [2], context-aware computing [57], and ambient assisted living for elderly and disabled people [53].

However, the prevalent absence of subject-awareness in wearable human vocal activity sensing methods poses a notable challenge. Overlooking the vocal activities of non-wearers will significantly limit the application of vocal activity sensing methods in multi-person scenarios. For example, misidentifying coughs from non-subjects could lead to incorrect medical diagnoses and misinform individuals about their health status [56]. To achieve efficient distinction between vocal activities originating from subjects and non-subjects in human vocal activity sensing (a capability termed as 'subject-awareness'), researchers have delved into exploratory studies. These exploratory studies have yielded progress in subject-aware vocal activity sensing for specific tasks, such as verifying the identity of coughers in cough event detection[35, 43], and in established areas like speaker verification [12, 26]. However, enabling efficient subject-awareness sensing in a broader spectrum of vocal activity is still elusive.

With the rise of True Wireless Stereo (TWS) and Active Noise Canceling (ANC) technologies, earables have become ubiquitous end-user accessories [7]. Owing to the incorporation of an increasing number of always-on sensors, earables are endowed with enhanced continuous sensing capabilities, which has also spurred consistent research efforts into exploring sensing applications on earables [1, 6, 11, 58, 60, 61]. Characteristics such as non-invasiveness, unobtrusiveness, good fixing, and reduced motion artifact interference make earables ideal wearable devices for continuous human vocal activity sensing. Researchers have demonstrated the feasibility of leveraging earables in detecting various vocal activities, including those related to sleep [11], food and drink consumption [4, 63], and movements of hand [48], head [39], and mouth [8].

In this paper, we envision substantial potential in achieving efficient subject-aware human vocal activity sensing on earables equipped with off-the-shelf Active Noise Cancelling (ANC) technology. A modern ANC earphone usually consists of a feed-forward microphone, a feed-back microphone, and an inertial measurement unit (IMU). The feedback microphone, which is originally designed for active noise cancelling, has the capability to capture subject's vocal signal transmitted through human body at a greater signal-noise-ratio. Together with the feed-forward microphone capturing ambient sound and the IMU capturing the subject's head motion, we expected an ANC earphone could be the minimum hardware form that provide rich and distinguishable information for subject-aware vocal activity sensing. Moreover, since ANC earables serves as a type of wearable device that is increasingly ubiquitous and possesses strong personal characteristics, exploring subject-aware vocal activity sensing on ANC earables can broaden the multi-person application scenarios for sensing technologies, offering substantial improvements over current methods in healthcare and user authentication.

In the context of the limited availability of large-scale public datasets within the earable research community [46], to the best of our knowledge, there is no publicly available dataset constructed specifically for subject-aware human vocal activity sensing on the minimum ANC earable setting. To bridge this gap, we introduce the **EarSAVAS** dataset, a novel dataset specifically crafted for **EAR**able **S**ubject-**A**ware **V**ocal **A**ctivity **S**ensing. EarSAVAS dataset consists of synchronous audio and motion data of 44.5 hours, collected from 42 participants who conducted 8 different vocal activities according to the designed protocol. Specifically, the audio data consists of two-channel audio stream from the feed-forward and feedback microphones of hybrid Active Noise Cancelling (ANC) earphones. Additionally, the motion data includes 3-axis accelerometer and 3-axis gyroscope signals collected by the Inertial Measurement Unit (IMU) within the earphones. The data collected of each user also encompass interference vocal sounds captured by the user's earphones but originating from non-subjects. This inclusion enables the dataset to support subject-aware human vocal activity sensing.

Furthermore, with the anticipation of real-time and privacy-preserving on-device algorithms, we introduced EarVAS, a lightweight multi-modal deep learning model enabling subject-aware human vocal activity sensing on earables. To validate the reliability of the dataset and the efficiency of EarVAS, we re-implemented two advanced deep neural network classifiers originally designed for similar applications. Through comprehensive experiments, we validate the efficiency of EarVAS, achieving an Accuracy of 89.30% and a Macro Averaged Area Under the Curve (AUC) of 89.03% in subject-aware vocal activity recognition across nine categories (8 specific subject vocal activities + 1 'Others' category that encompasses ambient sounds and vocal activities originating from non-subjects). Despite the performance gaps compared to more advanced models, the lightweight EarVAS still outperforms the best models on certain metrics. Ablation experiments were conducted and demonstrated the advantages of leveraging feedback microphone audio, as well as highlighted the potential value of sensor fusion in achieving effective subject-aware vocal activity sensing on monaural earables. The results of the ablation study also validated EarVAS as an innovative architecture for its proficient use of complementary information through sensor fusion. We envision that the EarSAVAS dataset, complemented by the proposed EarVAS and other two benchmark models with respective advantages, will advance subject-aware vocal activity sensing on earables.

The main contributions of this paper have three folds as below.

(1) We proposed EarSAVAS, the first publicly available multi-modal dataset constructed for subject-aware human vocal activity sensing on earphones. EarSAVAS also showcases its diversity, comprising 44.5 hours of data collected from 42 participants, spanning a total of eight various vocal activities. EarSAVAS is made publicly available on Kaggle [1].
(2) We introduced EarVAS, a multi-modal lightweight deep learning architecture enabling subject-aware human vocal activity sensing on earables. To the best of our knowledge, EarVAS is the first multi-modal approach specifically designed for effcient subject-aware vocal activity sensing.
(3) Two advanced deep learning classifiers are re-implemented to validate the reliability of EarSAVAS and the efficiency of EarVAS. Results show that EarVAS achieves an accuracy of 90.84% and a Macro Averaged AUC of 89.03% in subject-aware human vocal activity recognition. Ablation experiments are conducted and the results demonstrated the effectiveness of feedback microphone audio and also highlighted potential value of sensor fusion in subject-aware vocal activity sensing. All the code are publicly available on github [2].

[1] https://www.kaggle.com/datasets/earsavas/earsavas-dataset
[2] https://github.com/THU-CS-PI-LAB/EarSAVAS

2 RELATED WORK

This section summarizes related work, including novel sensing and interactive techniques on earables in Section 2.1, as well as existing methods for human vocal activity sensing in Section 2.2. Additionally, we provide a review of datasets for human vocal activity recognition in Section 2.3, emphasizing the crucial requirements for subject-aware human vocal activity sensing on earables.

2.1 Novel Sensing and Interactive Techniques on Earables

Earables have attracted growing attention from both industry and academia, particularly in light of the proliferation of True Wireless Stereo (TWS) and Active Noise Canceling (ANC) technologies over the past decade [13]. Modern and upcoming earables transcend their traditional role as mere audio listening devices, incorporating a growing array of sensors and micro-controllers endowed with computational capabilities. Among the emerging wearable technologies, earables offer several unique technical advantages, including non-invasiveness, unobtrusiveness, good fixing, and minimal susceptibility to motion artifacts [10]. Moreover, their proximity to critical anatomical structures, such as the brain, blood vessels, and facial muscles, affords access to a rich reservoir of physiological information [46].

Concurrent with the enhanced sensing and computing capacities of earables, the research community has been actively engaged in the exploration of methods aimed at optimizing the utilization of these devices. Recent applications encompass a wide array of fields, extending from physiological and mental health monitoring to movement and activity recognition, serving as interfaces for interactive inputs, and even facilitating user authentication and identification. In the comprehensive review by Tobias [46], earables have been shown to enable monitoring cardiac metrics such as heart rate, heart rate variability and blood pressure [6, 54, 55, 62], tracking of respiratory functions [1, 36, 47], detection of mental states including stress and emotions [22, 32], evaluation of the nervous system encompassing sleep and drowsiness [11, 18], recognition of physical activities [24, 41, 49], monitoring of food and drink intake [16, 63], functioning as an interface for user interactions [30, 58, 60, 61], and even for enabling user authentication and identification through the unique anatomical structure of the human ear canal [5, 25].

2.2 Human Vocal Activity Sensing on Earables

Human vocal activities encompass human activities that generate distinctive sound events, enabling their recognition through acoustic sensing technologies [57]. Apart from speech, this domain also covers non-speech vocalizations, such as laughter, coughs, and sighs, which are unrelated to language but serve as significant indicators in numerous applications [44].

Automatic sensing of human vocal activities has emerged as a critical confluence between human activity recognition (HAR) and sound event detection (SED), boasting a diverse array of applications, including but not limited to daily activity recognition, human health and behavioral monitoring. [2]. These applications can offer vital insights into individuals' overall well-being through the detection of health-related vocal sounds like coughs and sneezes [21, 42, 45], unveil human mental states through the identification of expressive sounds such as sighs and laughter [37, 65], and even evaluate the user's daily routine and life style by identifying the daily activities [27, 51].

Earables have become a promising tool in activity recognition due to their continuously enhancing sensor suite and their widespread daily usage. Moreover, when compared to popular wearable devices like smartphones and smartwatches, earables present unique advantages in acoustic sensing [33], thereby spurring a surge in research aiming to harness earables for human vocal activity sensing. Various studies have demonstrated the potential of earables in capturing a range of vocal activities. For instance, Min et al. [33] confirmed the efficacy of acoustic recognition across a range of activities, including physical exercise, head gestures, and conversational activities. Prakash et al. [41] ventured into sensing teeth and jaw movements in real time using audio signals, thereby proposing a system capable of detecting six distinct gestures. Wang et al. [59] introduced a comprehensive continuous system for cough event detection based on microphones in ANC smart earbuds, achieving cutting-edge performance coupled with low power consumption. However, a persistent challenge in this field lies in the microphones' ability to capture not only the subject's audio but also non-subject audio from the surrounding environment. This scenario can potentially lead to misleading outcomes in downstream applications, significantly limiting the utility in public settings due to the lack of subject-awareness [29, 43]. For instance, in public scenarios, the falsely detected coughs would then be mistakenly considered for a health analysis or disease diagnosis by clinicians, which could have serious adverse consequences due to harmful medication use and increased costs for patients [56]. Specifically, Zhang et al. [64] showed that without subject awareness, the accuracy of audio-based cough event detection methods drops to a level nearly akin to random guessing.

To address the prevalent issue of subject-awareness, considerable efforts have been channeled into developing methods that can discern between the vocal activities originating from subjects and those from non-subjects. While there has been a substantial exploration in the domain of speaker identification, especially for speech vocalization recognition [12, 26], an emerging focus is on enhancing subject-awareness for non-speech vocalizations. For instance, Nemati et al. [35] introduced CoughBuddy, a multi-modal cough event detection method, which effectively utilizes bystander coughs as negative samples, reporting a 7.3% False Positive Rate (FPR). Moreover, Rahman et al. [43] formulated an audio-based subject cough event detection method suitable for smartphones, achieving a precision of 94.2% in subject cough event recognition. However, the existing literature largely encompasses highly specific and limited activity categories, thus hindering the universal applicability of these methods in fostering subject-awareness for human vocal activity recognition.

2.3 Dataset for Human Vocal Activity Recognition

Most datasets employed in human vocal activity sensing are not inherently crafted for this specific purpose. These datasets commonly originate from extractions and compilations of widely used audio datasets, including ECS-50 [40], FSD50K [15], MIVIA [14] and AudioSet [19]. Despite the wide use of these general datasets, there's a growing requirement for datasets that are intrinsically devised for human vocal activity sensing.

In order to facilitate the advancement of human vocal activity sensing, Gong et al. introduced VocalSound [21], a comprehensive and balanced dataset comprising six categories of human vocal sounds. Furthermore, In an effort to supplement the datasets available in the domain, Rashid et al. introduced the Nonspeech7k dataset [44], which consists of seven classes of human vocal activities. Both datasets are explicitly designed for human vocal activity sensing and their effectiveness in enhancing the performance of established methods within this realm has been empirically validated. However, these datasets ignored human vocal activity from non-subjects, rendering them unsuitable for achieving subject-aware human vocal activity sensing techniques. As we will demonstrate in Section 5.1, the state-of-the-art vocal activity sensing methods introduced in Vocalsound [21], when trained without considering vocal activities from non-subjects, inaccurately classify 61.6% of non-subject human vocal activities. This highlights the requirements for human vocal activity sensing dataset encompassing vocal activity from non-subjects.

In fact, publicly available datasets specifically crafted for sensing on earables are notably scarce. As highlighted by Tobias et al. [46], of the more than 260 studies in the domain, many had sparse participant involvement (less than 10 participants), with 48 studies relying solely on a single individual. Furthermore, to our knowledge, there currently exists no dataset dedicated exclusively to human vocal activity sensing on earables. The majority of the datasets used in human vocal activity sensing are constructed through crowdsourcing, with the specifics of the recording devices remaining undisclosed. A prevalent strategy for achieving human vocal activity sensing on earables involves initial model training on these general datasets, followed by fine-tuning using a smaller dataset captured via recordings collected with earables. To fully harness the unique advantages of earables in the task of human vocal activity sensing, the availability of larger-scale datasets is indispensable.

To this end, we presented a publicly available dataset called EarSAVAS, specifically constructed for **EAR**phone-based **S**ubject-**A**ware **V**ocal **A**ctivity **S**ensing. The motivation behind constructing this dataset stems from envisioning potential of earables to achieve efficient subject-aware human vocal activity sensing. Firstly, most of the earables are equipped with hybrid active noise cancellation (hybrid ANC) technology to achieve better users' listening experience by reducing environmental noises. To achieve noise cancellation, one or more feed-forward microphones (reference microphones) are positioned on the outer side of the earphone, along with a feedback microphone (error microphone) located between the speaker and the ear. Owing to the structural and functional characteristics of hybrid Active Noise Cancelling (ANC) earphones, the audio captured by feedback microphone is filtered through both passive noise reduction through the earables' physical structure and active noise cancellation, effectively minimizes interference from non-wearer's vocal sound and ambient noise. We envision the efficacy of fusing the two-channel audio from the two microphones to achieve efficient subject-aware vocal activity sensing. Additionally, human vocal activity is often accompanied by unique physical movements specific to the subject, especially head movements. We envision that motion signals captured by IMU (Inertial Measurement Unit) sensors commonly integrated into earables can also contribute to achieving subject-awareness. Given the considerations outlined above, our dataset incorporates two-channel audio signals from both feed-forward and feedback microphones in hybrid ANC earables with motion data, encompassing three-axis accelerometer and three-axis gyroscope data stream, from IMU (Inertial Measurement Unit) sensors. The dataset consists of synchronous audio and motion data of 44.5 hours, collected from 42 participants conducted 8 various vocal activities according to the designed protocol. Within the data from each user, there are also interference data captured by the user's earables but originating from non-subjects.

In summary, compared to existed dataset proposed in human vocal activity sensing, EarSAVAS exhibits several unique advantages. Firstly, to the best of our knowledge, EarSAVAS is the first publicly available dataset in the field of human vocal activity sensing to incorporate various vocal activities from non-subjects. This inclusion enables the dataset to enhance existing vocal activity sensing methods, endowing them with subject-awareness capabilities. Furthermore, EarSAVAS is specifically crafted for vocal activity sensing on earables, with data collected from 42 users totaling 44.5 hours. As a large-scale, multi-modal publicly available dataset, EarSAVAS can assist in evaluating and improving sensing methods for earable sensing works, which commonly faces data scarcity challenges as described by Tobias [46]. Additionally, the EarVAS deep learning approach, proposed based on the EarSAVAS dataset and serving as a benchmark model, enables efficient subject-aware human vocal activity sensing. This provides a novel methodological reference for achieving effective subject-aware human vocal activity sensing on earables.

3 EARSAVAS DATASET

This section introduces our proposed novel dataset EarSAVAS. We present our designed hardware platform for synchronous data collection in Section 3.1 and the collection protocol in Section 3.2. We also detailed our EarSAVAS dataset in Section 3.2 via statistical analysis.

3.1 Hardware Platform Designed for Synchronous Data Collection

We devised a robust hardware platform, optimized for the acquisition of synchronous data, anchoring on the functionality of hybrid ANC (Active Noise Cancelling) earphones. The comprehensive layout of the hardware platform is depicted in Figure 1.

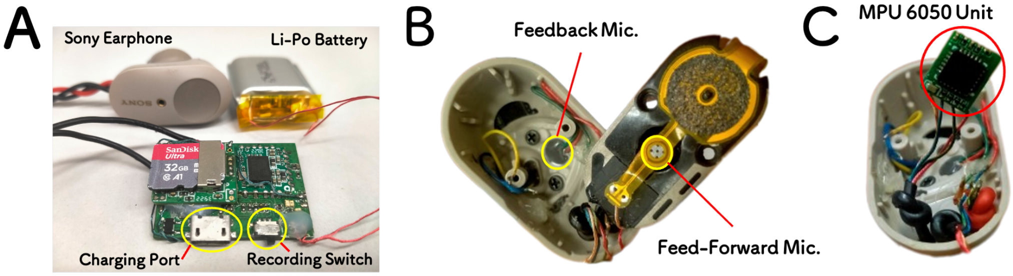


Fig. 1. The components of data collection hardware platform. A: Three components of the platform, including a microcontroller unit, a monaural Sony WF-1000XM3 earphone and a Li-Po battery. B: Feed-forward and Feedback microphones in one Sony WF-1000XM3 earphone. C: MPU6050 inertial measurement unit embedded into the earphone.

Specifically, we conducted modifications on the Sony WF-1000XM3 Wireless Noise Cancelling Earbuds [3]. Our motivation behind the modification is to leverage meticulously designed acoustic structures of commercial earables, as illustrated in Figure 1B. We removed original battery of the earbuds and embedded an MPU6050 Inertial Measurement Unit (IMU) to enable motion data collection, as shown in Figure 1C.

As shown in Figure 1A, we incorporated audio and motion signals into a popular micro-controller on off-the-shelf earables named BES2300YP [4]. It was adopted by many popular true wireless stereo (TWS) earphones, such as JBL FREE II, Samsung Galaxy Buds Live, and Huawei FreeBuds 2 Pro. In order to enable synchronous audio and motion data collection, we developed a customized program designed for the efficient synchronization of motion and sound signals. The program was installed on the BES2300YP micro-controller and periodically recorded synchronized signals onto an SD card, allowing for continuous and steady data collection.

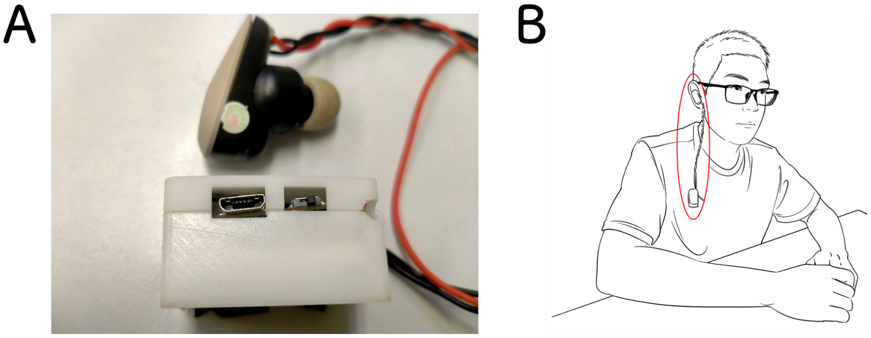


Fig. 2. Hardware platform encapsulated for data collection. A: Encase the battery and microcontroller unit to facilitate the ease of wear for the participants. B: Participants wearing the hardware platform (in red circle) during data collection process.

Finally, as shown in Figure 2, we encapsulated the data acquisition components within a box, leaving the switch and charging port accessible. This makes the entire device highly portable and easy to use, with start and stop of the collection achievable through the simple act of toggling a switch.

[3] https://www.sony.com/en-ma/electronics/truly-wireless/wf-1000xm3
[4] http://www.bestechnic.com/Home/Index/index/lan_type

3.2 Data Collection Procedure

In total, we collected audio and motion data from 46 healthy participants, consisting of 22 males and 24 females. These participants were recruited from the university campus and included both undergraduates and postgraduates. Their ages ranged from 19 to 26 with an average age of 22.23 (s.t. = 1.84). The participants involved were divided into 23 groups, each consisting of two individuals. All participants signed a consent form prior to the collection and were informed that the collected data would be made publicly available for research purposes only. Our data collection user study was approved by the Institutional Review Board (IRB).

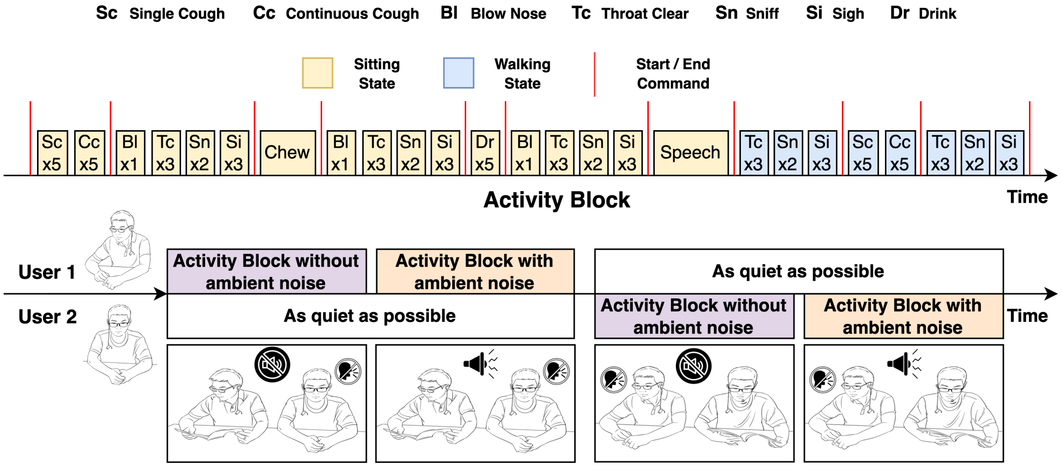


Fig. 3. Data collection protocol for each group of participants. We defined 'Activity Block', a basic unit in which the users perform several groups of activities with the same order and frequency. The groups in Activity Block are separated by the experimenter's start and end commands. Within the collection process of each pair of participants, one user (such as User 1 in figure) completed the Activity Block with and without ambient noise, while the other participant remain quiet. After User 1 completed the task, the roles are switched with User 2 executing the Activity Block while User 1 remained quiet. The positions of the two participants remained the same as before.

The data collection was conducted in a standard conference room. We opted for a controlled experiment rather than real-world scenarios to ensure more precise labeling, particularly in determining whether the data originated from the wearer or a bystander. The dilemma is further discussed in Section 6.3. Prior to data collection, each pair of participants determined the allocation of earbuds and was equipped with an earbud in one ear only. This ensures an even distribution of data from both the left and right ears within our dataset. After the allocation of earbuds, participants selected rubber ear tips that fit their ear size appropriately.

The experiment was structured in a way that had each pair of participants take turns performing specified activities according to the designed Activity Block under environments with and without ambient noise while the other remained as silent as possible, as shown in Figure 3. This design was implemented to simulate the interference issues encountered in multi-person scenarios. Upon the completion of an experiment involving a pair of participants, the earphones of each participant collected subject vocal activity data from the wearer, as well as vocal sounds originated by the other participant (non-subject vocal activity). The sound level was measured at 43-50 dB in the room without ambient noise. With the noises from the ESC-50 [40] and AudioSet datasets [19] played by Bluetooth speaker, the sound level increased to 60-65 dB in the room with ambient noise.

As depicted in Figure 3, activities within the activity block were organized into several groups, ensuring the adequate rest intervals for participants. The division between groups was indicated by the experimenter's start and end commands to facilitate the data annotation process. Participants were required to carry out activities naturally without additional guidance, in order to best reflect the individual diversity inherent in the execution of vocal activities under natural conditions. The frequency and order of activities presented in Figure 3 was determined based on feedback from participants in a pilot study. The whole collection procedure lasted approximately 90 minutes per paired user, with each participant compensated with a 25 USD gift card. A strict sanitization protocol was adhered to, involving thorough sterilization of the hardware prototype with 75% alcohol before passing it among users.

Table 1. Description and statistical results of all activities involved in data collection procedure.

Activity		Description	Tot. Dur. (hrs)	Avg Dur. (secs)	S.D. (secs)
Coughs	Single coughs	Collected during sitting and walking Continuous cough consist 2-3 sequential coughs per instance	0.105	0.454	0.146
	Continuous coughs		0.190	0.772	0.208
Chewing		Various food consumption: apple, cookie, bread, and banana	1.113	9.967	7.067
Drink		Taking small sips of water followed by swallowing	0.263	0.989	0.691
Sigh		Collected during sitting and walking	0.33	0.601	0.261
Speech		Read prepared English materials aloud or Talk with experimenter about the collection protocol	1.981	8.741	21.081
Nose blowing		Simulated nose blowing without actual nasal secretions	0.117	0.840	0.361
Throat Clearing		Collected during sitting and walking	0.344	0.551	0.222
Sniff		Collected during sitting and walking	0.190	0.516	0.204
Others	Single coughs	All data apart from vocal activity conducted by subject	0.137	0.502	0.207
	Continuous coughs		0.204	0.760	0.224
	Chewing		1.256	13.621	7.756
	Drink		0.264	1.637	1.069
	Sigh		0.386	0.703	0.351
	Speech		1.609	69.789	11.184
	Nose blowing		0.128	0.903	0.474
	Throat Clearing		0.350	0.627	0.348
	Sniff		0.239	0.654	0.287
	Ambient noise **		35.163	6.035	31.716

** Ambient noise includes silence and ambient noises recorded by our hardware.

The activities undertaken during the experiment are detailed in Table 1. These activities predominantly encompass those found in the authoritative Nonspeech7k [44] and VocalSound [21] datasets, which are datasets proposed specifically for human vocal activity sensing, though not collected with earables. We excluded activities difficult to naturally enact during data collection under a controlled environment, such as crying and sneezing. We incorporated others common in vocal activity sensing studies, like drinking, nose blowing, and eating food of various textures. To enhance data diversity from the perspective of motion data, we intentionally integrated activities known to introduce motion artifacts in Human Activity Recognition (HAR) [33, 50]. Specifically, data collection was carried out as participants engaged in a range of activities during their walking state. These activities include 'Single Coughs', 'Continuous Coughs', 'Throat Clearing', 'Sniff', and 'Sigh', as described in the 'Description' column of Table 1. The "Others" category encompasses all data recorded that is not directly associated with the participants' activities, including those activities conducted by non-subjects and ambient noise.

After data collection, we obtained two-channel audio with a sampling rate of 16000Hz, as well as synchronous 6-axis motion data with a sampling rate of 100Hz. In total, data from 42 participants was retained, excluding 4 participants whose devices suffered from unexpected interruptions during the collection process. To ensure the quality of annotations, we enlisted three professional data annotators for meticulous data labeling. The annotator manually labels the data by listening to the audio and further verifying it in accordance with our collection protocol. We labeled the data with the smallest possible window containing only the targeted event to facilitate more flexible utilization of data samples. For instance, researchers can flexibly select the minimal recognition window to enable efficient algorithms for specific activities. Following this, we undertook a comprehensive statistical analysis of the duration of all collected events. The results, showcasing a total of 44.5 hours of accumulated data, including 4.6 hours from the subjects of interest and 39.9 hours from non-subject and ambient noises, are presented in Table 1. Data samples are now publicly available [5]. To facilitate the application of the EarSAVAS dataset, we released not only segmented events labeled with annotations, but also raw data for each user, accompanied by corresponding annotation files.

4 EARVAS: PROPOSED LIGHTWEIGHT BENCHMARK MODEL FOR SUBJECT-AWARE VOCAL ACTIVITY SENSING

In this section, we proposed a lightweight multi-modal deep learning neural network for subject-aware vocal activity sensing as a reference benchmark on EarSAVAS dataset. The lightweight design is motivated by the vision of enabling real-time and privacy-preserving sensing with on-device deployment. We described data preprocessing methods in Section 4.1. Furthermore, we demonstrated the design of EarVAS-Net architecture in Section 4.2 and a series of EarVAS Variants based on EarVAS-Net in Section 4.3. We re-implemented two advanced models as described in Section 4.4 to evaluate the reliability of EarSAVAS and the efficiency of EarVAS-Net. In Section 4.5, the training protocol of EarVAS-Net and the baseline models was detailed.

4.1 Data Preprocessing

As shown in Table 1, most of the vocal activity lasts less than 1 second duration. Furthermore, Oresti et al. [3] reported that the interval 1–2 s proves to provide the best trade-off between recognition speed and accuracy. As a result, we segment the synchronized audio and motion data into 1-s windows, with no overlapping in between. Values of the audio and motion data are normalized between the ranges -1 and 1 respectively. The statistical results of the dataset after segmentation is presented in Table 2.

After segmentation and normalization, the data samples are allocated among training, validation, and evaluation datasets, with the allocation based on distinct users. These datasets include 30 users (71.5%), 4 users (9.5%), and 8 users (19.0%) respectively. This user-based division facilitates comprehensive cross-user evaluation.

We preprocess the windowed audio data into log-scaled Mel filter bank features using a 25ms Hanning window with a 10ms stride, yielding 2 x 98 x 128 filter bank features for each two-channel segmented audio instance. The audio sub-module in our proposed architecture, leveraging EfficientNet-B0 [52], induces a downsampling of both time and frequency dimensions by a factor of 32. As a result, to align with this downsampling, we implement zero-padding on the filterbank features along the time axis, thereby resulting in the final feature dimension of 2

[5] https://www.kaggle.com/datasets/earsavas/earsavas-dataset

Table 2. Statistical results of segmented activity data.

Activity		**Total samples**	**Average samples per user**
Coughs	Single coughs	832	21.3
	Continuous coughs	888	21.7
Chewing		3835	91.31
Drink		1053	25.07
Sigh		1980	47.14
Speech		7048	167.81
Nose blowing		503	11.98
Throat Clearing		2248	53.52
Sniff		1326	31.57
Others	Single coughs	982	23.38
	Continuous coughs	968	23.05
	Chewing	4358	103.76
	Drink	858	21.45
	Sigh	2003	47.69
	Speech	5753	136.98
	Nose blowing	528	12.57
	Throat Clearing	2033	48.40
	Sniff	1317	31.36
	Ambient noise**	122098	2907.10

** Ambient noise includes silence and ambient noises recorded by our hardware.

channels x 128 (time dimension) x 128 (frequency dimension). No additional preprocessing was conducted on motion signals beyond segmentation and normalization.

4.2 The Architecture of EarVAS-Net

EarVAS-Net is a lightweight multi-modal deep learning framework leveraging two-channel audio log-mel filter bank features and raw 6-axis motion signals as input.

As shown in Figure 4, the motion data sub-module of EarVAS was outlined in the yellow box. The design of this sub-module is composed of four 1D CNN layers. Each layer is defined by a kernel size of 10 and a stride of 1, with layer depths set at 128, 128, 256, and 256 respectively. ReLU activation functions are employed in each layer to facilitate non-linear data processing. Interspersed among these 1D CNN layers are batch normalization layers and max pooling layers, with the latter featuring a pool size of 2. Subsequently, the architecture integrates a dropout layer with a probability setting of 0.5 to enhance regularization. The intermediate outputs are then fed into a fully connected layer, yielding a motion embedding with a width of 256. Our selection of this architecture was inspired by its successful application in the SAMoSA project [34]. This architecture serves as a effective motion feature extraction module for SAMoSA, a multi-modal system enabling accuracy human activity recognition across 26 activities on the smartwatch. In the audio sub-module, our approach initially involves processing the filter bank features of two-channel audio through the EfficientNet-B0 [52] for feature extraction. Subsequently, we perform Mean Pooling along the frequency axis on the extracted features. Finally, following the application of a 1 x 1 convolutional layer, Mean Pooling is also executed along the time axis. This methodology integrates the robust feature extraction capabilities of EfficientNet-B0 with the simplifying aspects of pooling operations, greatly improving the computational effiency as reported in [20]. We adopted such a lightweight architecture due to its successful applications for vocal activity recognition in VocalSound [21].

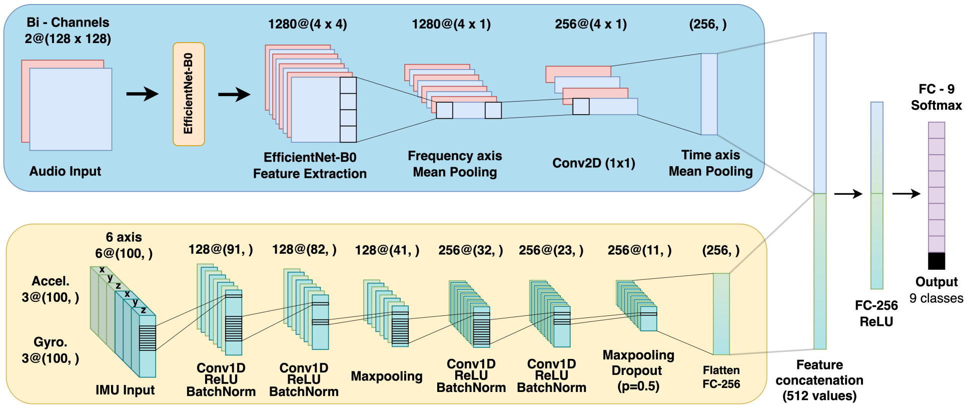


Fig. 4. Overview of EarVAS-Net multi-modal deep learning architecture. EarVAS-Net takes log-mel audio filter bank features along with a 6-DoF IMU motion instance as inputs to predict the activity. The audio sub-module is highlighted in blue and the motion one is in yellow. The output is the prediction probabilities for the 9 classes (eight dedicated to specific vocal activities of the subject (emphasized in purple) and one 'Others' category (depicted in black).

After audio and motion data streams are processed independently, the multi-modal EarVAS-Net combines the resulting two 256-dimensional embeddings. A concatenation layer merges these embeddings, directing the unified data to a fully connected layer and subsequently to a final classification layer with softmax activations.

4.3 Variants of EarVAS-Net based on Input Modality

Based on the input modality, we deconstruct EarVAS-Net and propose five variants for each architecture. The five distinct models corresponds to different data modalities: a combination of two-channel audio with motion data, two-channel audio, feed-forward microphone audio, feedback microphone audio, motion data alone. These models were proposed to systematically assess the contribution of each modality to the overall performance of the EarVAS-Net, providing insights into how these different types of data interact and influence the performance of subject-aware human vocal activity sensing.

Among the five models we proposed based on each architecture, the model structure of those multi-modal models that utilize both audio and motion data inputs is identical to EarVAS-Net illustrated in Figure 4. As for the uni-modal models, which focus solely on either audio or motion, we utilize the corresponding sub-module of EarVAS-Net to extract 256-dimensional features. Following this feature extraction, the output is directly fed into a fully connected layer with 256 nodes. The process concludes with a softmax-based classification layer, which produces the final results.

4.4 Baseline Deep Learning Classifiers

To the best of our knowledge, there are currently no methods specifically designed for vocal activity sensing. However, to further validate the reliability of the proposed dataset and evaluate the efficiency of our proposed EarVAS, we re-implemented two advanced baseline models from similar applications.

(1) **BEATs [9].** As an iterative pre-training framework for audio, BEATs aims at learning Bidirectional Encoder Representations from Audio Transformers. BEATs has been demonstrated to perform well on datasets such as the ESC-50 [40] (rank #2) and AudioSet [19] (rank #5). We re-implemented the original BEATs model,

which was designed for single-channel audio input, and modified the input channel count of the patch embeddings (Conv2D) to support two-channel input streams.

(2) **SAMoSA [34]** SAMoSA is a multi-modal deep learning model enabling effective human activity recognition across 26 activity classes on smartwatches. We re-implemented the SAMoSA model, adapting the input channel of the motion sub-module to accommodate six-axis input. Additionally, we altered the first layer (Conv2D) of audio processing to support two-channel audio input.

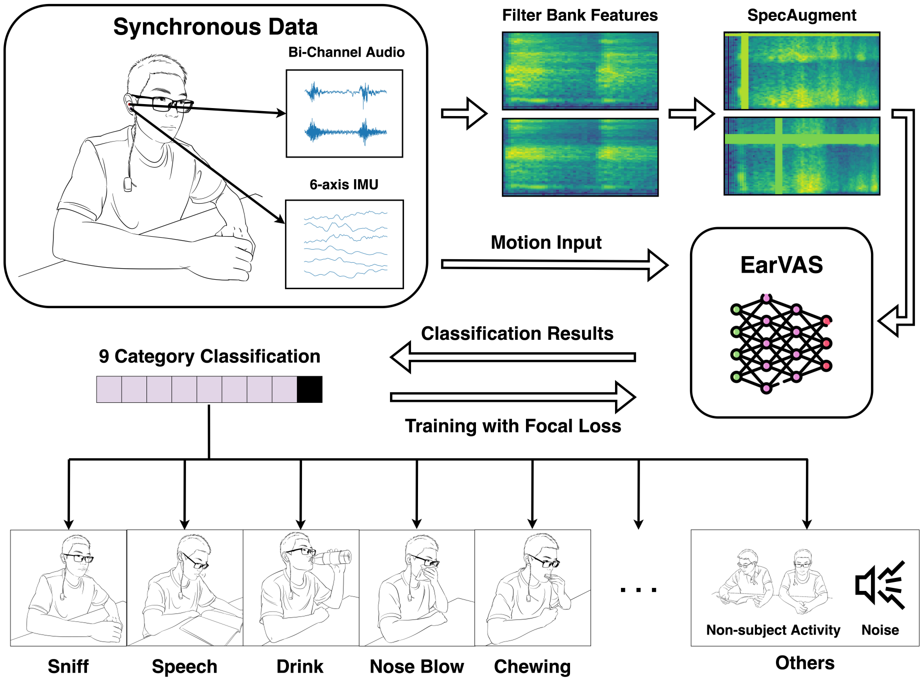


Fig. 5. Overview of training process for EarVAS model. The model takes corresponding data streams as inputs and employs supervised learning with a focal loss function to efficiently perform subject-aware human vocal activity recognition with unbalanced dataset.

4.5 Training Protocol of EarVAS and Baseline Models

In our training process of EarVAS, we employed SpecAugment [38] for data augmentation. Specifically, in our approach, we first apply consecutive masking to both the frequency and temporal dimensions of features, with a maximum extent of 37.5% proportion in each dimension. We also incorporated a temporal shift in the feature domain, which randomly shifts the feature sequence along the time axis, simulating potential variations in time alignment of the vocal activities. These augmentation methods introduces variability into the data for training, thus enhancing the model's ability to generalize from diverse and subtly altered audio representations.

Our training procedure is illustrated in Figure 5. All models were built using Pytorch 1.13.0. We used the Adam [28] optimizer with a constant learning rate of 1e-4. As illustrated in Table 2, there is a significant imbalance in our dataset. During the training process, we aim to preserve the diversity of the training data to the greatest extent and fully leverage the advantages of data volume. Focal Loss [31] has been demonstrated to effectively address the issue of imbalanced data in object detection tasks. Therefore, we adopt the Focal Loss as our loss function. All models were trained on an Nvidia GeForce RTX 3090 GPU with a batch size of 128.

During the training process, we conducted tests on the validation set at the end of each epoch, preserving the model that exhibited the best performance for the final evaluation.

For re-implemented SAMoSA, we processed the audio into a 64-bin log-scaled Mel spectrogram according to the original paper. Input streams of BEATs are the same as EarVAS-Net as reported in the original paper. All other settings during the training procedure of the two baseline models are identical to those of EarVAS-Net.

5 RESULTS AND FINDINGS

In this section, we justify the necessity of subject-awareness in Section 5.1, and establish the practical significance of the EarSAVAS dataset. We evaluated the reliability of EarSAVAS and efficiency of EarVAS from the perspectives of binary and multi-class classification, presenting the results and findings respectively in Section 5.2 and 5.3.

5.1 Justify the Necessity of Subject-Awareness

To understand the impact of subject awareness on human vocal activity sensing, we selected a benchmark model from a vocal activity dataset named VocalSound [21], which represents one of the cutting-edge methods in vocal activity sensing. As reported in the original paper, the benchmark model achieves an accuracy of 90.5 ± 0.2% in the recognition of six types of human vocal activities. We named this benchmark model EfficientNet-B0-Mod due to its derivation from the EfficientNet-B0 [52] model as described in the original paper.

EfficientNet-B0-Mod was trained on both the whole EarSAVAS dataset and a version with non-subject vocal activities removed. The input of EfficientNet-B0-Mod consisted of the feed-forward microphone audio from EarSAVAS, which is aligned with the distribution of data obtained from microphones without noise cancellation in the VocalSound dataset. Audio pre-processing and the training process replicated the VocalSound settings.

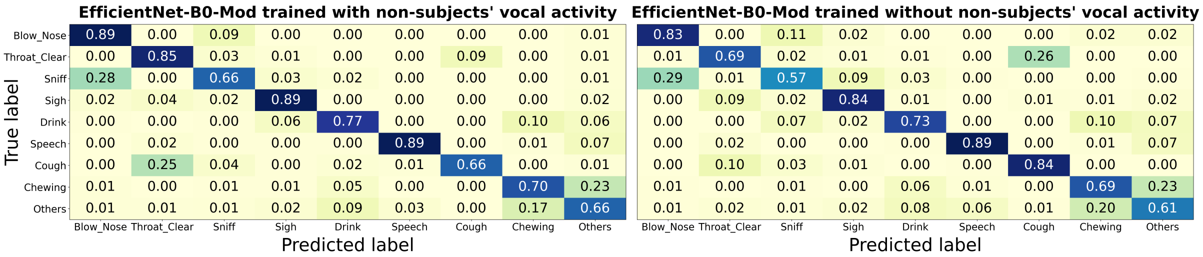


Fig. 6. Comparative subject-aware vocal activity recognition performance of EfficientNet-B0-Mod trained with and without non-subject vocal activity.

As shown in Figure 6, despite the expansion of the classification task from six to nine categories, EfficientNet-B0-Mod achieves state-of-the-art performance in recognition of various vocal activities, with an accuracy of 86.57%. However, as illustrated in Figure 7, EfficientNet-B0-Mod trained without non-subjects' vocal activity misidentified 61.6% non-subjects' vocal activities as subject vocal activity. With the inclusion of non-subjects' vocal activities as negative samples, EfficientNet-B0-Mod misidentified 32.1% non-subjects' vocal activities as subject vocal activity, a significant drop compared with the EfficientNet-B0-Mod trained without non-subjects' vocal activities. This comparison highlights the necessity of subject-awareness in human vocal activity sensing methods and underscores the necessity of including non-subjects' data in the dataset for vocal activity sensing.

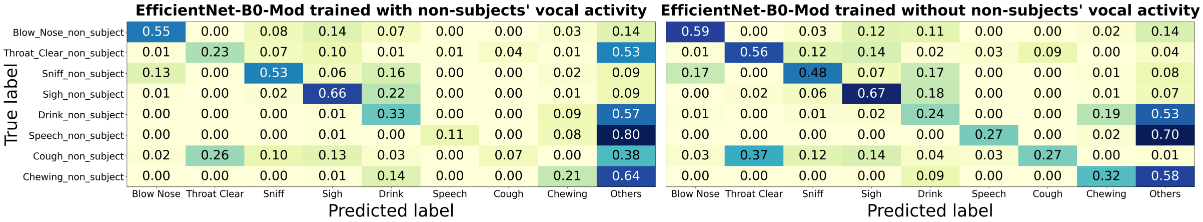


Fig. 7. Comparative subject-awareness ability of EfficientNet-B0-Mod trained with and without non-subject vocal activity.

5.2 Binary Classification Performance Evaluation: Distinguishing Subject Vocal Activity from Other Events

We initially considered all subject vocal activities as an entity, evaluating the ability of the benchmark models in distinguishing the subject's vocal activity from other events (vocal activity from non-subjects and ambient noise), which is essentially a binary classification task.

5.2.1 Evaluation Metrics and Results. Traditional binary classification metrics are involved in our evaluation, encompassing Accuracy, Sensitivity, Specificity and F1-Score. In order to enable comprehensive evaluation, we also considered Geometric Mean Score as a critical reference metric considering the unbalanced evaluation dataset. Additionally, although Specificity reflects false positive rate, it's worth noting that the negative class can be decomposed into two distinct sub-categories: activity from non-subjects and ambient noise. To offer a more nuanced view, we quantified the proportion of non-subject activities that were misclassified as subject activities, as well as the proportion of background noise identified as subject activities. The comparable performances between benchmark models are listed in Table 3.

Similar to Section 5.1, we also demonstrate the cases where vocal activity from non-subjects is misclassified as subject vocal activity, using the format of a confusion matrix illustrated in Figure 8. This serves as a more detailed reference for evaluating the model's subject-awareness capability.

5.2.2 Insights and Findings based on Comparative Binary Classification Performance. Through the analysis of the comparative results, we have concluded with the following findings:

1. The audio captured by the feedback microphone exhibits significant superiority to the audio captured by the feed-forward microphone in distinguishing subject vocal activities from other acoustic events: As indicated in Table 3, the models utilizing feedback audio as input exhibit a significant improvement in performance across various metrics when compared to models using feed-forward audio as input. Furthermore, feedback microphone audio proves to be highly effective in substantially reducing the misidentification rate of non-subject vocal activity and ambient noise, resulting in lower FPR1 and FPR2 compared to models utilizing the feed-forward audio. The advantages of feedback microphone audio are further supported by the results presented in Figure 8. Specifically, across almost all involved vocal activities, models based on feedback microphone audio demonstrate superior subject-awareness ability compared to models based on feed-forward microphone audio.

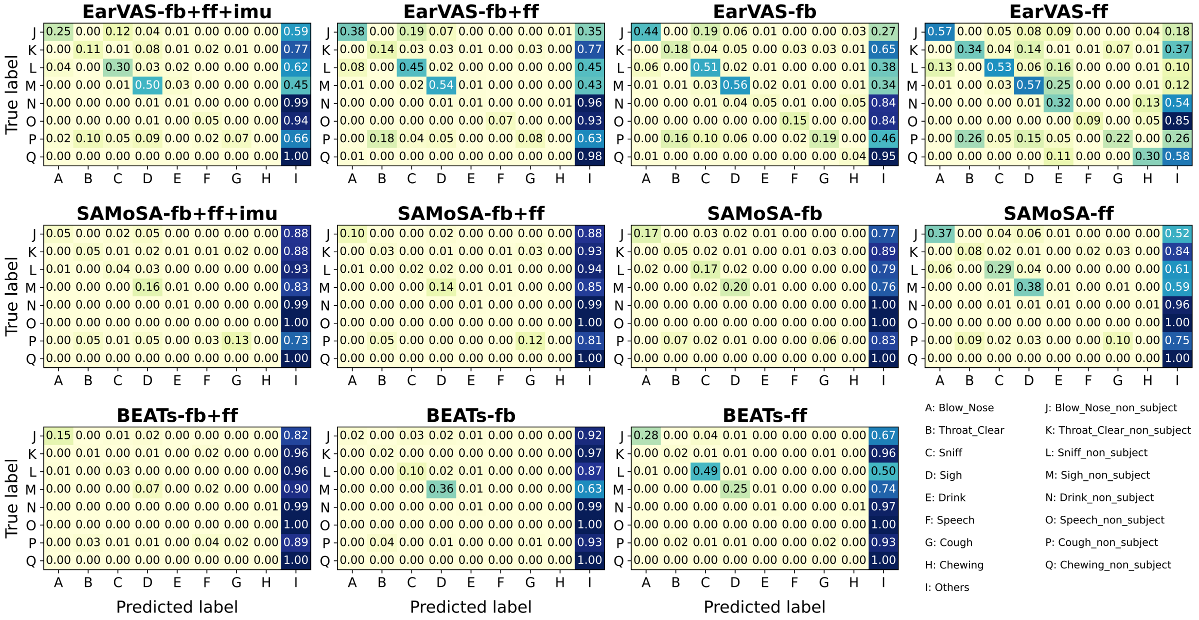
The comparative results can be largely attributed to the unique structural and functional characteristics of active noise-cancelling earables. The feedback microphone, by capturing noise-reduced sounds, effectively filters out a majority of ambient noise and non-subject vocal sounds transmitted via air conduction. Additionally, its proximity to the bone conduction pathway of the wearer's vocal activities enables it to pick up more distinct and clear vocal sounds. This results in a more distinguishable characteristic between the audio captured for non-subject and subject vocal activities by the feedback microphones.

Table 3. Comparative results of distinguishing subject vocal activity from interfering events among benchmark models

Method	**EarVAS**				**SAMoSA [34]**				**BEATs [9]**			
	Acc	Sen.	Spec.	F1	Acc	Sen.	Spec.	F1	Acc	Sen.	Spec.	F1
Input												
ff + fb + imu	93.2%	97.4%	92.5%	78.6%	95.4%	85.8%	96.9%	82.9%	-	-	-	-
ff + fb	90.8%	96.8%	89.9%	73.1%	**96.0%**	84.1%	**97.7%**	84.3%	95.3%	83.4%	97.0%	82.0%
fb	87.4%	98.1%	85.8%	66.7%	94.3%	74.6%	97.2%	77.0%	95.9%	93.9%	96.2%	**85.6%**
ff	71.8%	**98.4%**	67.9%	47.3%	90.5%	70.9%	93.4%	65.9%	92.4%	86.2%	93.3%	74.6%
imu	52.8%	96.6%	46.4%	34.6%	52.8%	69.0%	50.5%	27.4%	-	-	-	-

Method	**EarVAS**			**SAMoSA [34]**			**BEATs [9]**		
	G-Mean	FPR1*	FPR2+	G-Mean	FPR1*	FPR2+	G-Mean	FPR1*	FPR2+
Input									
ff + fb + imu	94.9%	18.5%	6.1%	91.2%	7.9%	2.4%	-	-	-
ff + fb	93.3%	21.6%	8.6%	90.6%	6.5%	**1.6%**	90.0%	**4.3%**	2.7%
fb	91.7%	30.1%	12.7%	85.2%	8.8%	1.9%	**95.1%**	6.4%	3.4%
ff	81.7%	49.8%	31.6%	81.4%	15.7%	5.2%	89.7%	10.9%	6.0%
imu	67.0%	69.4%	55.4%	59.0%	64.2%	47.2%	-	-	-

*: The proportion of non-subject vocal activity misclassified as subject vocal activity; +: The proportion of ambient noise misclassified as subject vocal activity; ff=feedforward mic. audio; fb=feedback mic. audio



EarVAS-fb+ff+imu

	A	B	C	D	E	F	G	H	I
J	0.25	0.00	0.12	0.04	0.01	0.00	0.00	0.00	0.59
K	0.00	0.11	0.01	0.08	0.01	0.02	0.01	0.00	0.77
L	0.04	0.00	0.30	0.03	0.02	0.00	0.00	0.00	0.62
M	0.00	0.00	0.01	0.50	0.03	0.00	0.00	0.00	0.45
N	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.99
O	0.00	0.00	0.00	0.01	0.00	0.05	0.00	0.00	0.94
P	0.02	0.10	0.05	0.09	0.00	0.02	0.07	0.00	0.66
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

EarVAS-fb+ff

	A	B	C	D	E	F	G	H	I
J	0.38	0.00	0.19	0.07	0.00	0.00	0.00	0.01	0.35
K	0.00	0.14	0.03	0.03	0.01	0.00	0.03	0.00	0.77
L	0.08	0.00	0.45	0.02	0.00	0.00	0.00	0.00	0.45
M	0.01	0.00	0.02	0.54	0.01	0.00	0.00	0.00	0.43
N	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.96
O	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.93
P	0.00	0.18	0.04	0.05	0.00	0.00	0.08	0.00	0.63
Q	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98

EarVAS-fb

	A	B	C	D	E	F	G	H	I
J	0.44	0.00	0.19	0.06	0.01	0.00	0.00	0.03	0.27
K	0.00	0.18	0.04	0.05	0.00	0.03	0.03	0.01	0.65
L	0.06	0.00	0.51	0.02	0.01	0.00	0.00	0.01	0.38
M	0.01	0.01	0.03	0.56	0.02	0.01	0.00	0.01	0.34
N	0.00	0.00	0.01	0.04	0.05	0.01	0.00	0.05	0.84
O	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.84
P	0.00	0.16	0.10	0.06	0.00	0.02	0.19	0.00	0.46
Q	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.95

EarVAS-ff

	A	B	C	D	E	F	G	H	I
J	0.57	0.00	0.05	0.08	0.09	0.00	0.00	0.04	0.18
K	0.00	0.34	0.04	0.14	0.01	0.01	0.07	0.01	0.37
L	0.13	0.00	0.53	0.06	0.16	0.00	0.00	0.01	0.10
M	0.01	0.00	0.03	0.57	0.25	0.00	0.00	0.01	0.12
N	0.00	0.00	0.00	0.01	0.32	0.00	0.00	0.13	0.54
O	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.05	0.85
P	0.00	0.26	0.05	0.15	0.05	0.00	0.22	0.00	0.26
Q	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.30	0.58

SAMoSA-fb+ff+imu

	A	B	C	D	E	F	G	H	I
J	0.05	0.00	0.02	0.05	0.00	0.00	0.00	0.00	0.88
K	0.00	0.05	0.01	0.02	0.01	0.01	0.02	0.00	0.88
L	0.01	0.00	0.04	0.03	0.00	0.00	0.00	0.00	0.93
M	0.00	0.00	0.00	0.16	0.01	0.00	0.00	0.00	0.83
N	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.99
O	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
P	0.00	0.05	0.01	0.05	0.00	0.03	0.13	0.00	0.73
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

SAMoSA-fb+ff

	A	B	C	D	E	F	G	H	I
J	0.10	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.88
K	0.00	0.03	0.00	0.01	0.00	0.01	0.03	0.00	0.93
L	0.01	0.00	0.02	0.02	0.00	0.01	0.00	0.00	0.94
M	0.00	0.00	0.00	0.14	0.01	0.01	0.00	0.00	0.85
N	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.99
O	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
P	0.00	0.05	0.00	0.00	0.00	0.00	0.12	0.00	0.81
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

SAMoSA-fb

	A	B	C	D	E	F	G	H	I
J	0.17	0.00	0.03	0.02	0.01	0.00	0.00	0.00	0.77
K	0.00	0.05	0.02	0.01	0.00	0.01	0.03	0.00	0.89
L	0.02	0.00	0.17	0.02	0.00	0.00	0.00	0.00	0.79
M	0.00	0.00	0.02	0.20	0.01	0.00	0.00	0.00	0.76
N	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
O	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
P	0.00	0.07	0.02	0.01	0.00	0.00	0.06	0.00	0.83
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

SAMoSA-ff

	A	B	C	D	E	F	G	H	I
J	0.37	0.00	0.04	0.06	0.01	0.00	0.00	0.00	0.52
K	0.00	0.08	0.02	0.01	0.00	0.02	0.03	0.00	0.84
L	0.06	0.00	0.29	0.04	0.00	0.00	0.00	0.00	0.61
M	0.00	0.00	0.01	0.38	0.01	0.00	0.01	0.00	0.59
N	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.96
O	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
P	0.00	0.09	0.02	0.03	0.00	0.00	0.10	0.00	0.75
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

BEATs-fb+ff

	A	B	C	D	E	F	G	H	I
J	0.15	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.82
K	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.00	0.96
L	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.96
M	0.00	0.00	0.00	0.07	0.00	0.02	0.00	0.00	0.90
N	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.99
O	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
P	0.00	0.03	0.01	0.01	0.00	0.04	0.02	0.00	0.89
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

BEATs-fb

	A	B	C	D	E	F	G	H	I
J	0.02	0.00	0.03	0.02	0.01	0.00	0.00	0.00	0.92
K	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.97
L	0.00	0.00	0.10	0.02	0.01	0.00	0.00	0.00	0.87
M	0.00	0.00	0.00	0.36	0.01	0.00	0.00	0.00	0.63
N	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.99
O	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
P	0.00	0.04	0.00	0.01	0.00	0.00	0.01	0.00	0.93
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

BEATs-ff

	A	B	C	D	E	F	G	H	I
J	0.28	0.00	0.04	0.01	0.00	0.00	0.00	0.00	0.67
K	0.00	0.02	0.01	0.01	0.00	0.00	0.01	0.00	0.96
L	0.01	0.00	0.49	0.01	0.00	0.00	0.00	0.00	0.50
M	0.00	0.00	0.00	0.25	0.01	0.00	0.00	0.00	0.74
N	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.97
O	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
P	0.00	0.02	0.01	0.01	0.00	0.00	0.02	0.00	0.93
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Fig. 8. Confusion matrices demonstrating subject-awareness ability of benchmark models across all activity types. The numerical values presented represent the proportions of specific non-subject vocal activities that are either misidentified as subject vocal activities (left eight columns) or accurately classified into the ’others’ category (rightmost column).

2. The integration of feedback microphone audio with feed-forward microphone audio has enhanced the model's capability in rejecting interference events: As illustrated in Table 3, the model incorporating two-channel audio input demonstrates improved performance across most metrics, surpassing the models with feedback or feed-forward audio input in the SAMoSA architecture. However, within the BEATs and EarVAS model architecture, the combination of two-channel audio input results in a notable decrease in sensitivity compared to the feedback audio based model. This performance decline may be attributed to the model's limitation in effectively utilizing two-channel audio, leading to interference from feed-forward microphone audio that blurs the distinction between subject vocal activity and other categories. Despite the observed decreases in sensitivity, as indicated by comparable FPR1 and FPR2 values in Table 3 together with a more detailed analysis in Figure 8, it is evident that the integration of two-channel audio input across all models enhances the capability to effectively reject false recognitions of non-subject vocal activities and ambient noise.

3. The incorporation of motion data as an auxiliary factor to potentially enhance the effectiveness in distinguishing between subject vocal activity and other events: While Table 3 indicates that the performance of models relying solely on motion inputs only marginally surpasses random guessing, we have found that the inclusion of motion inputs has the potential to improve binary classification task performance. As depicted in Table 3, the SAMoSA model without IMU data exhibits higher specificity and lower sensitivity. However, the integration of motion data assists in balancing the sensitivity and specificity of SAMoSA, leading to an enhanced G-Mean metric compared to the SAMoSA with two-channel audio as inputs. Furthermore, for the EarVAS model, the incorporation of IMU data leads to significant improvements across all the mentioned metrics, resulting in an overall model effectiveness that approaches or even surpasses that of the other two more complex models.

However, the influence of IMU data on overall performance still requires further exploration. Although there is a modest improvement in macro-level metrics for SAMoSA upon the inclusion of motion data, the enhancement remains relatively minimal. Furthermore, there is an observable decrease in the F1-Score, which may be attributed to the introduction of potentially confounding features.

4. Achieving subject-awareness on 'Blow Nose', 'Sniff', and 'Sigh' presents distinct challenges: As depicted in Figure 8, even for the models with relatively higher specificity, such as SAMoSA and BEATs, the misidentification rates for categories like 'Blow_Nose', 'Sniff', and 'Sigh' are higher compared to other non-subject vocal activities. This issue becomes more pronounced in the case of the lightweight EarVAS model. However, these results also demonstrate that the integration of two-channel audio and the inclusion of motion data have contributed to a reduction in these misclassification rates. This observation highlights the potential advantages of employing multi-modal data fusion to further enhance the ability of subject-awareness.

5. Effectiveness of EarVAS in subject-awareness ability: Compared with more advanced models, EarVAS, with the lightweight architecture, exhibits an inferior ability to reject interfering events, resulting in relatively poorer specificity and F1-Score performance. However, EarVAS exhibits superior performance in sensitivity and geometric mean score (G-Mean). Furthermore, among the benchmark models, EarVAS effectively leverages the complementary information shared by audio and motion modalities, leading to an overall enhancement in performance with the assistance of motion data.

5.3 Subject-Aware Vocal Activity Recognition Performance Evaluation

In this section, we revisit subject-aware vocal activity sensing from a multi-class perspective. We focus on the efficacy of benchmark models in recognizing each specific type of human vocal activity originating from subjects.

5.3.1 Evaluation Metrics and Results. We have chosen Accuracy, Macro-averaged Area Under the Curve (Macro-AUC), Macro-averaged F1 Score (Macro-F1) and Matthews Correlation Coefficient (MCC) as our primary evaluation metrics. These metrics have proved to be efficient in imbalanced classification performance evaluation due to their ability to provide a holistic and nuanced view of the model's performance across various aspects [23].

From the perspective of efficiency, we have also incorporated size of parameters and FLOPS (Floating Point Operations per Second) as additional parameters. We illustrated the comparative performance of benchmark models in Table 4.

Table 4. Comparative results on subject-aware vocal activity recognition among benchmark models

Method	**EarVAS**				**SAMoSA [34]**				**BEATs [9]**			
	Acc	AUC*	F1$^{\#}$	MCC	Acc	AUC*	F1$^{\#}$	MCC	Acc	AUC*	F1$^{\#}$	MCC
Input												
ff + fb + imu	89.3%	**89.0%**	64.5%	69.1%	91.7%	85.2%	68.0%	73.3%	-	-	-	-
ff + fb	88.8%	88.5%	59.3%	64.5%	90.3%	87.5%	75.2%	78.9%	93.1%	85.3%	73.3%	78.0%
fb	89.5%	88.9%	56.2%	60.0%	90.0%	83.5%	68.8%	71.4%	93.1%	87.4%	**80.1%**	**83.6%**
ff	87.6%	86.1%	49.3%	41.5%	91.2%	79.9%	58.9%	59.8%	**94.2%**	81.5%	71.5%	76.0%
imu	28.3%	53.5%	10.8%	2.9%	26.6%	55.7%	12.8%	7.2%	-	-	-	-

Method	**EarVAS**		**SAMoSA [34]**		**BEATs [9]**	
	FLOPS (M)	Params (MB)	FLOPS (M)	Params (MB)	FLOPS (M)	Params (MB)
Input						
fb + ff + imu	169.74	24.22	404.78	73.09	-	-
fb + ff	129.22	16.78	362.24	55.20	4398.98	90.4
fb or ff	128.04	16.78	358.63	55.20	4048.72	90.4
imu	40.52	7.44	43.17	20.29	-	-

*: AUC=Macro-averaged Area Under the Curve (Macro-AUC). $^{+}$: F1: Macro-averaged F1 Score (Macro-F1); ff=feedforward mic. audio; fb=feedback mic. audio

As shown in Figure 9, confusion matrices have been utilized to provide a detailed presentation of the performance of benchmark models in subject-aware vocal activity recognition.

5.3.2 Insights and Findings based on Comparative Subject-Aware Vocal Activity Recognition Performance. From the results illustrated in Table 4 and the Figure 9, we concluded the following findings:

1. Feedback microphone audio are more efficient in subject-aware vocal activity recognition compared to audio from feed-forward microphone: As shown in Table 4, models relying on feedback microphone audio consistently outperformed those leveraging feed-forward microphone audio. This finding is corroborated when comparing the results in Figure 6 and 9. Despite sharing the same architecture, the EfficientNet-B0-Mod utilizing feed-forward microphone audio is surpassed by the EarVAS model with feedback microphone audio as input. An inspection of Figure 9 provides us with a micro perspective. Through the analysis of the Figure 9, it becomes evident that the improvement in macro-level metrics can largely be attributed to the model's enhanced recognition accuracy for 'Drink', 'Chewing', and all negative samples. 'Drink' and 'Chewing' activities are often characterized by weaker vocalizations. As half of our audio data are recorded with ambient noise played in the background, these activities are more likely to be misclassified as 'Others' when using feed-forward microphone audio. However, with the noise-reduced feedback audio, the models prove to be more effective in discerning the subtle vocalizations from background noise.

2. The fusion of two-channel audio potentially enhances the model's performance in subject-aware vocal activity recognition: The results presented in Table 4 demonstrate that the SAMoSA model, when utilizing two-channel audio input, exhibits a substantial performance improvement compared to its counterparts relying on mono-channel audio. However, the EarVAS model with two-channel audio inputs shows a comparable

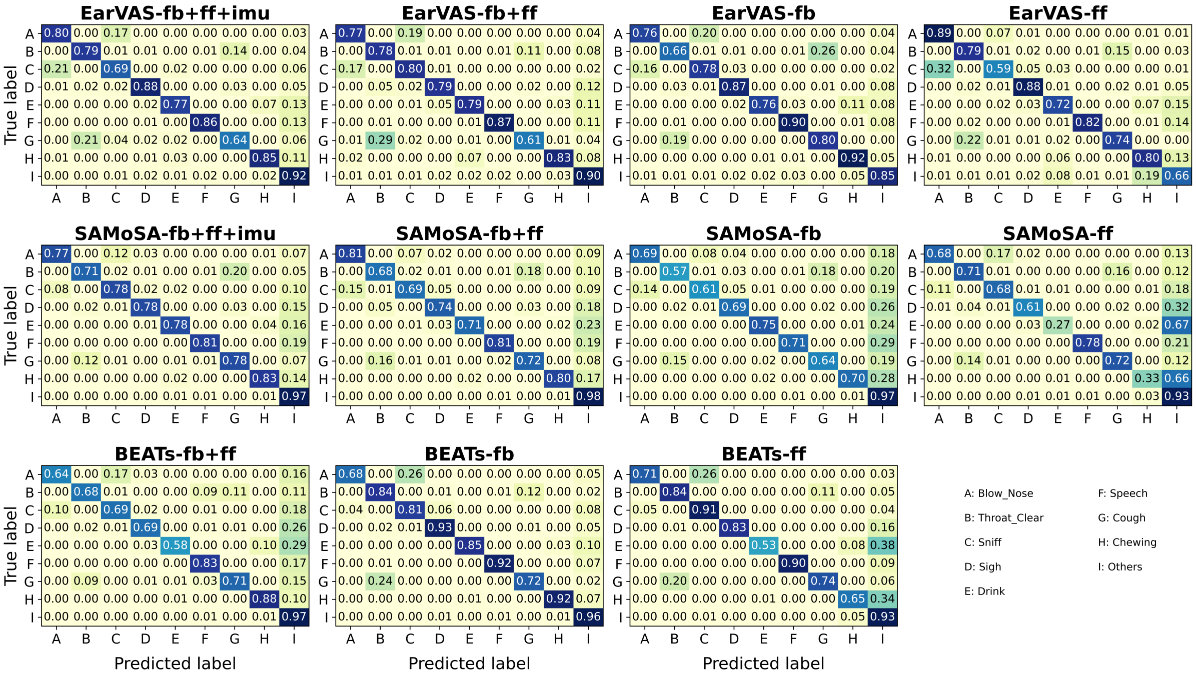


Fig. 9. Confusion matrices of benchmark models on subject-aware vocal activity recognition.

performance to the model utilizing feedback microphone audio, with the latter even exhibiting an advantage in Macro-AUC and Accuracy. Similar findings can be observed in the BEATs architecture.

According to the micro-level results presented in Figure 9, the superiority performance in macro-level metrics of SAMoSA with two-channel audio can be largely attributed to the enhanced subject-awareness capability, consistent with the insights discussed in Section 5.2.2. However, for the EarVAS and BEATs models, although they prove to be more effective in rejecting false recognitions of non-subject vocal activity and ambient noise with two-channel audio inputs, the introduction of feed-forward audio also brings interference, resulting in a reduction of model sensitivity. This observation aligns with the results presented in Table 3. Additionally, as demonstrated in Figure 9, the fusion of two-channel audio has a limited impact on distinguishing between different subject vocal activities.

In summary, building upon the second insight presented in Section 5.2.2, the fusion of two-channel audio proves to be effective in rejecting interference events originating from non-subjects and ambient noise. Despite the observed decrease in sensitivity for EarVAS and BEATs models, we envision SAMoSA to be a highly effective architecture for the successful integration of two-channel audio. Further exploration of effective fusion methods is necessary to investigate whether two-channel audio fusion can bring enhanced performance in distinguishing between various subject vocal activities.

3. The incorporation of motion data potentially offers benefit to subject-aware vocal activity recognition: For the EarVAS architecture, the model incorporating two-channel audio and IMU data as inputs demonstrates significant enhancements in Macro-F1 and MCC, along with a slight improvement in Macro-AUC and Accuracy compared to the EarVAS model based solely on two-channel audio. This improvement in macro-level metrics primarily stems from enhanced binary classification performance achieved after integrating motion data, as supported by a higher True Positive Rate (TPR) for the 'Others' category, as observed in Figure 9.

However, with the inclusion of IMU data, the SAMoSA model experiences a decrease in certain metrics, particularly in MCC, Macro-F1 and Macro-AUC. As depicted in Figure 9, the introduction of IMU data in the SAMoSA architecture has led to improved recognition accuracy for each subject vocal activity, while the accuracy for the 'others' category has decreased. This finding further supports that the introduction of motion data assists in balancing sensitivity and specificity as demonstrated in SAMoSA architecture as discussed in Section 5.2.2.

Due to the existing gaps in the field, we consider the feasibility demonstrated by EarVAS in enabling effective subject-aware vocal activity sensing as a promising starting point. There is a critical need for a more comprehensive exploration of the potential benefits associated with integrating IMU and audio data in subject-aware vocal activity sensing. We envision EarVAS as an efficient sensor-fusion subject-aware vocal activity sensing on EarSAVAS.

4.Challenges in distinguishing between similar subject vocal activities: As illustrated in Figure 9, all benchmark models exhibit relatively poor performance in discriminating between the nasal vocal activities 'Blow Nose' and 'Sniff', as well as in differentiating 'Throat Clear' from 'Cough'. For these two pairs of activities, the features are extremely similar in the audio domain. Optimization of proposed EarVAS and further exploration of efficient subject-aware human vocal activity sensing are necessary.

5. Validated efficiency of lightweight EarVAS model: Compared to BEATs and SAMoSA, our proposed EarVAS boasts a smaller size and demands fewer computational resources. Although there is still a gap compared to advanced methods, its effectiveness has been validated, even surpassing the BEATs model in terms of the Macro-AUC metric. Moreover, among the three architectures, EarVAS proves to be the most proficient model in leveraging the benefits of audio and motion data fusion. As a result, we envision EarVAS architecture as an inspiration for further exploration of efficient approaches, and even on-device algorithms enabling real-time and privacy-preserving subject-aware vocal activity sensing on consumer-grade earables.

Despite the relatively higher computational demands, BEATs and SAMoSA have showcased impressive capabilities on the EarSAVAS, validating the reliability of the dataset. With their respective advantages, BEATs and SAMoSA also serve as valuable benchmark models, providing a foundation for researchers to delve deeper into high-performance algorithms. As the most complex architecture, BEATs exhibits the most comprehensive performance, particularly noticeable in the variant employing feedback microphone audio. The medium-sized SAMoSA architecture effectively integrates two-channel audio streams, achieving recognition performance comparable to BEATs through the benefits of fusion. As a result, the architecture of SAMoSA may provide valuable insights for achieving effective subject-aware vocal activity sensing by combining feedback and feed-forward microphone audio.

6 DISCUSSION

In this section, we discuss applications of the EarSAVAS Dataset, our major findings, limitations, and future works.

6.1 Applications of EarSAVAS Dataset

The findings discussed in Section 5.1 emphasize that ignoring non-subject vocal activity as negative instances leads to a significant number of misidentifications of non-wearer's events by EfficientNet-B0-Mod, despite the promising performance of the model on VocalSound dataset reported in the original paper. In healthcare applications, the falsely detected events would then be mistakenly considered for a health analysis or disease diagnosis by clinicians, which could have serious adverse consequences due to harmful medication use and increased costs for patients [56]. However, audio-based health-related event recognition methods commonly regard only ambient noise as a negative example, requiring non-subject vocal activity samples during the training process to enhance the capability of subject-awareness. To the best of our knowledge, EarSAVAS is the first dataset supporting various subject-aware vocal activity sensing on earables, encompassing a substantial volume of non-subject vocal activities. The EarSAVAS dataset could potentially facilitate the evaluation of methods' capacity for subject-awareness in healthcare contexts, including but not limited to behavior monitoring (e.g. eating and drinking behaviors) and the identification of abnormal events (e.g. coughs, throat clearing, nose blowing, etc.). Additionally, with non-subject vocal activities within EarSAVAS as negative instances, relevant approaches can enhance the subject-awareness ability and expand the application scenarios. The inclusion of subject vocal activities and ambient noise samples from the EarSAVAS dataset as supplementary data may also bolster the model's robustness in detecting health-related events.

Beyond the healthcare domain, we also envision the potential value of EarSAVAS in user authentication applications. By augmenting the ability of subject-awareness via training on EarSAVAS, acoustic-based user authentication methods may efficiently figure out subject vocalization, enabling robust user enrollment in crowded scenarios. Lightweight methods proposed on EarSAVAS can also serve as a detector to avoid frequent awakening complex user authentication algorithms for acoustic-based feature matching, thereby reducing power consumption.

6.2 Unique Advantages of Subject-Aware Vocal Activity Sensing on Monaural Earables

In Section 5, we have demonstrated the feasibility of implementing subject-aware vocal activity sensing on a monaural earphone. We substantiated that compared to the feed-forward microphone, the feedback microphone in hybrid active noise cancellation earables capture subject vocal activity signals with a higher signal-to-noise ratio, which can provide more distinctive information and yield better performance in subject-aware vocal activity sensing. Incorporating the additional information provided by the feed-forward microphone further enhances the performance by augmenting the ability to reject negative samples. With the assistance of motion data gathered by the IMU sensor, the approach demonstrates enhanced capability in EarVAS architecture. However, further explorations are required to fully investigate the effectiveness of two-channel audio fusion and integration of motion data.

Despite binaural earphones providing richer information and even potentially enabling lightweight effective solutions such as beamforming, we envision that subject-aware vocal activity sensing on a monaural earphone has distinct advantages. Firstly, we envision deploying the sensing model on earables for offline inference in the future, enabling privacy-preserving and real-time applications. However, this becomes challenging with binaural data streams, which require aggregation for computational inference, potentially necessitating sophisticated federated learning algorithms, leading to increased power consumption and computational load. Furthermore, the time required for data exchange may impede the real-time functionality of the technology. Utilizing monaural information to achieve subject-aware vocal activity recognition also enables sensing ability even when the user wears only one earpiece, expanding application scenarios compared to the binaural approach.

6.3 Limitations and Future Works

This paper introduced EarSAVAS, the first dataset enabling subject-aware vocal activity sensing on smart earphones and explored the feasibility of subject-aware vocal activity sensing on commodity earables. Ablation experiments were conducted to provide insights into how different modalities of data interact and influence the performance of subject-aware human vocal activity recognition. However, this work also has several limitations.

6.3.1 Limitations of EarSAVAS Dataset. Although we constructed the first dataset enabling subject-aware vocal activity sensing on smart earphones based on a well-designed data collection protocol, there are three limitations regarding EarSAVAS: 1) All the data utilized in EarSAVAS is gathered within a controlled environment, potentially restricting the practical applications of the methods proposed in real-world scenarios; 2) All recruited participants are young students from campus, constraining the generalizability of the techniques to a diverse population.; 3) Data samples are collected only from hardware prototypes based on Sony WF-1000XM3 earphones, imposing limitations on the cross-device ability.

To comprehensively enhance the dataset from multiple perspectives. In the near future, we will recruit participants with various social backgrounds to engage in the collection, with their ages spanning a range from 18 to 60 years, thereby encompassing diverse age groups. We plan to undertake in-the-wild data collection with hardware platforms based on various noise-canceling earables, including but not limited to Huawei FreeBuds 2 Pro and Bose QC 20. During the collection process, participants will wear our collection hardware platforms during their regular routine for a period ranging from one week to a month. With the real-world data collection, EarSAVAS will not only be enhanced to incorporate data samples from real-world scenarios, but it will also encompass a broader variety of vocal activities, including those that can not be spontaneously exhibited by participants in controlled environments such as laughing or screaming.

Under in-the-wild environment, it is quite challenging to label whether the vocal activity is from the subject or non-subjects. As a result, to obtain precise labels, we conducted data collection under controlled environments. To overcome the annotation challenge for real-world scenarios, we intend to introduce the assistance of a portable camera with the function of video monitoring, specifically targeting the mouth area, to facilitate precise data annotation. We envision the feasibility of obtaining more precise labels by observing the wearer's facial expressions, mouth movements, and interactions between the hands and face.

Although EarSAVAS facilitates basic healthcare applications including subject-aware health behavior monitoring and abnormal event detection, the dataset still lacks practical utility in intricate healthcare tasks deeply integrated with medicine. For example, the diagnosis and assessment of diseases require authentic data collected from patients, and specialized annotations must be conducted by medical experts.

6.3.2 Subject-Aware Vocal Activity Detection on EarSAVAS Dataset. In this paper, we define subject-aware vocal activity sensing as a task of event recognition. However, a more practical approach not only identifies the type of activity but also determines their temporal boundaries, which was defined as subject-aware vocal activity detection. We have considered subject-aware vocal activity detection as a direction for future work. As described in Section 3.2, we have also released the complete collection data for each user in our dataset, along with the corresponding annotation files. These files include the start and end times of each event of interest, thereby aiding researchers in further exploring algorithms for effective subject-aware vocal activity detection.

6.3.3 Optimization and On-device Deployment of EarVAS. Despite demonstrating proven efficacy, EarVAS still holds great potential for advancements in subject-aware vocal activity recognition tasks, especially the ability to reject interfering events including non-subject vocal activities and ambient noise. Additionally, all the benchmark models share some common deficiencies. As we demonstrated in Section 5.2, considering the performance in the binary classification, all models still exhibit a relatively high misidentification rate for non-subjects vocal activities, especially for 'Sniff', 'Blow Nose', and 'Sigh'. Additionally, as shown in Section 5.3, the models struggle to distinguish between 'Throat Clear' and 'Cough', as well as between 'Sniff' and 'Blow Nose'. In the future, we intend to optimize EarVAS to enable more effective subject-aware vocal activity sensing, especially from the perspective of efficient multi-modal representations [17].

On-device subject-aware human vocal activity recognition is essential, as it offers significant benefits in facilitating continuous, real-time, and privacy-preserving sensing capabilities. However, this also imposes requirements on the complexity of the sensing methods. BES2300YP, as one of the most popular micro-controllers on earables, only has dual ARM-Cortex M4F processors with up to 300 MHz CPU, 992 kB SRAM, and 4 MB flash storage. With a vision for on-device deployment, EarVAS was created as the smallest and most computationally efficient model on EarSAVAS. However, the currently proposed EarVAS still comprises 6.347 million trainable parameters.

In order to enable on-device deployment of EarVAS, model compression techniques, such as quantization and distillation, will be explored and achieve a balance between performance and model complexity.

6.3.4 Privacy Issues. Although the EarSAVAS we made public consists of raw audio data, its publication has been confirmed by the participants, and the entire collection experiment was approved by the Institutional Review Board (IRB). As for the most sensitive audio in 'Speech' category, all conversations involved only reading specified materials or discussing issues related to the collection protocol. We confirmed that the speech data does not contain any private information.

Furthermore, in this paper, we did not investigate the privacy issues related to the subject-aware vocal activity sensing approach. However, subject vocal activities include speech and also events pertaining to individuals' health privacy. As a result, there is a substantial necessity for technologies that preserve privacy. As discussed in 6.3.3, we aim to achieve on-device deployment and enable offline inference to avoid data exchange with the external environment, thereby protecting the privacy of users. This is also one of the primary reasons for proposing the lightweight architecture of EarVAS. We consider real-time and privacy-preserving offline inference as one of our future research and hope that the dataset will inspire researchers to propose other effective algorithms for privacy preservation.

7 CONCLUSION

In this paper, we presented EarSAVAS, the first multi-modal dataset crafted for subject-aware human vocal activity sensing on earables. EarSAVAS consists of feedback and feed-forward microphone audio, together with 6-axis IMU data collected from earables. With 44.5 hours of data collected from 42 participants, EarSAVAS encompasses a total of eight various vocal activities. We also proposed EarVAS, a lightweight multi-modal deep learning architecture enabling efficient subject-aware human vocal activity sensing on earables. To evaluate the reliability of EarSAVAS and the efficiency of EarVAS, we introduced BEATs [9] and SAMoSA [34] as two advanced benchmark models. Results show that EarVAS achieves an accuracy of 90.84% and a Macro-AUC of 89.03%. Although lightweight EarVAS exhibits performance disparities when compared to more sophisticated models, it surpasses the best model in specific metrics. Additionally, EarVAS can serve as an innovative architecture, particularly in leveraging complementary data through sensor fusion. With ablation experiments based on the three benchmark models, we demonstrated the effectiveness of feedback microphone audio and the potential value of sensor fusion methods in subject-aware vocal activity sensing. We hope that the proposed EarSAVAS and EarVAS benchmark model encourage other researchers to further explore on more efficient subject-aware human vocal activity sensing on earables.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (NSFC) under Grant No. 62132010, Young Elite Scientists Sponsorship Program by CAST under Grant No.2021QNRC001, Tsinghua University Initiative Scientifc Research Program, Beijing Key Lab of Networked Multimedia, Institute for Artificial Intelligence, Tsinghua University, Undergraduate / Graduate Education Innovation Grants, Tsinghua University, and Beijing National Research Center for Information Science and Technology (BNRist).

REFERENCES

[1] Tousif Ahmed, Md Mahbubur Rahman, Ebrahim Nemati, Mohsin Yusuf Ahmed, Jilong Kuang, and Alex Jun Gao. 2023. Remote Breathing Rate Tracking in Stationary Position Using the Motion and Acoustic Sensors of Earables. 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 325, 22 pages. https://doi.org/10.1145/3544548.3581265

[2] Yang Bai, Li Lu, Jerry Cheng, Jian Liu, Yingying Chen, and Jiadi Yu. 2020. Acoustic-based sensing and applications: A survey. *Computer Networks* 181 (2020), 107447. https://doi.org/10.1016/j.comnet.2020.107447

[3] Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. 2014. Window Size Impact in Human Activity Recognition. *Sensors* 14, 4 (2014), 6474–6499. https://doi.org/10.3390/s140406474

[4] Abdelkareem Bedri, Gregory Abowd, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Beh, Mayank Goel, and Thad Starner. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1 (09 2017), 1–20. https://doi.org/10.1145/3130902

[5] Hongliang Bi, Yuanyuan Sun, Jiajia Liu, and Lihao Cao. 2022. SmartEar: Rhythm-Based Tap Authentication Using Earphone in Information-Centric Wireless Sensor Network. *IEEE Internet of Things Journal* 9, 2 (2022), 885–896. https://doi.org/10.1109/JIOT.2021.3063479

[6] Assim Boukhayma, Anthony Barison, Serj Haddad, and Antonino Caizzone. 2021. Earbud-Embedded Micro-Power mm-Sized Optical Sensor for Accurate Heart Beat Monitoring. *IEEE Sensors Journal* 21, 18 (2021), 19967–19977. https://doi.org/10.1109/JSEN.2021.3098861

[7] Canalys. 2020. Global smart device shipment forecasts 2020 to 2023. https://www.canalys.com/newsroom/canalys-worldwide-smart-device-shipments-2023 Last accessed September 2023.

[8] Yetong Cao, Huijie Chen, Fan Li, and Yu Wang. 2021. CanalScan: Tongue-Jaw Movement Recognition via Ear Canal Deformation Sensing. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*. 1–10. https://doi.org/10.1109/INFOCOM42981.2021.9488852

[9] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. BEATs: Audio Pre-Training with Acoustic Tokenizers. arXiv:2212.09058 [eess.AS]

[10] Romit Roy Choudhury. 2021. Earable Computing: A New Area to Think About *(HotMobile '21)*. Association for Computing Machinery, New York, NY, USA, 147–153. https://doi.org/10.1145/3446382.3450216

[11] Kenneth Christofferson, Xuyang Chen, Zeyu Wang, Alex Mariakakis, and Yuntao Wang. 2022. Sleep Sound Classification Using ANC-Enabled Earbuds. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. 397–402. https://doi.org/10.1109/PerComWorkshops53856.2022.9767394

[12] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622* (2018).

[13] Xiaoran Fan and Trausti Thormundsson. 2023. Design Earable Sensing Systems: Perspectives and Lessons Learned from Industry. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2023 ACM International Symposium on Wearable Computers.*

[14] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2015. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters* 65 (2015), 22–28. https://doi.org/10.1016/j.patrec.2015.06.026

[15] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. FSD50K: An Open Dataset of Human-Labeled Sound Events. arXiv:2010.00475 [cs.SD]

[16] Yang Gao, Ning Zhang, Honghao Wang, Xiang Ding, Xu Ye, Guanling Chen, and Yu Cao. 2016. iHear Food: Eating Detection Using Commodity Bluetooth Headsets. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. 163–172. https://doi.org/10.1109/CHASE.2016.14

[17] Ziqi Gao, Yuntao wang, Jianguo Chen, Junliang Xing, Shwetak Patel, Xin Liu, and Yuanchun Shi. 2023. MMTSA: Multi-Modal Temporal Segment Attention Network for Efficient Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 96 (sep 2023), 26 pages. https://doi.org/10.1145/3610872

[18] Agustina Garcés Correa, Lorena Orosco, and Eric Laciar. 2014. Automatic detection of drowsiness in EEG records based on multimodal analysis. *Medical Engineering & Physics* 36, 2 (2014), 244–249. https://doi.org/10.1016/j.medengphy.2013.07.011

[19] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*. New Orleans, LA.

[20] Yuan Gong, Yu-An Chung, and James Glass. 2021. PSLA: Improving Audio Tagging With Pretraining, Sampling, Labeling, and Aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3292–3306. https://doi.org/10.1109/taslp.2021.3120633

[21] Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A Dataset for Improving Human Vocal Sounds Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. https://doi.org/10.1109/icassp43922.2022.9746828

[22] Unsoo Ha, Yongsu Lee, Hyunki Kim, Taehwan Roh, Joonsung Bae, Changhyeon Kim, and Hoi-Jun Yoo. 2015. A Wearable EEG-HEG-HRV Multimodal System With Simultaneous Monitoring of tES for Mental Health Management. *IEEE Transactions on Biomedical Circuits and Systems* 9, 6 (2015), 758–766. https://doi.org/10.1109/TBCAS.2015.2504959

[23] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications* 73 (2017), 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

[24] Ghena M. Hammour and Danilo P. Mandic. 2021. Hearables: Making Sense from Motion Artefacts in Ear-EEG for Real-Life Human Activity Classification. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 6889–6893. https://doi.org/10.1109/EMBC46164.2021.9629886

[25] Feiyu Han, Panlong Yang, Shaojie Yan, Haohua Du, and Yuanhao Feng. 2023. BreathSign: Transparent and Continuous In-ear Authentication Using Bone-conducted Breathing Biometrics. 1–10. https://doi.org/10.1109/INFOCOM53939.2023.10229037

[26] John HL Hansen and Taufiq Hasan. 2015. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine* 32, 6 (2015), 74–99.

[27] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking* (New Delhi, India) *(MobiCom '18)*. Association for Computing Machinery, New York, NY, USA, 289–304. https://doi.org/10.1145/3241539.3241548

[28] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[29] Yang Li, Yang Guanci, Zhidong Su, Shaobo Li, and Yang Wang. 2022. Human activity recognition based on multienvironment sensor data. *Information Fusion* 91 (10 2022), 47–63. https://doi.org/10.1016/j.inffus.2022.10.015

[30] Zisu Li, Chen Liang, Yuntao Wang, Yue Qin, Chun Yu, Yukang Yan, Mingming Fan, and Yuanchun Shi. 2023. Enabling Voice-Accompanying Hand-to-Face Gesture Recognition with Cross-Device Sensing. 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 313, 17 pages. https://doi.org/10.1145/3544548.3581008

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal Loss for Dense Object Detection. arXiv:1708.02002 [cs.CV]

[32] Nick Merrill, Max T. Curran, Jong-Kai Yang, and John Chuang. 2016. Classifying mental gestures with in-ear EEG. In *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. 130–135. https://doi.org/10.1109/BSN.2016.7516246

[33] Chulhong Min, Akhil Mathur, and Fahim Kawsar. 2018. Exploring Audio and Kinetic Sensing on Earable Devices. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications* (Munich, Germany) *(WearSys '18)*. Association for Computing Machinery, New York, NY, USA, 5–10. https://doi.org/10.1145/3211960.3211970

[34] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing Activities with Motion and Subsampled Audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 132 (sep 2022), 19 pages. https://doi.org/10.1145/3550284

[35] Ebrahim Nemati, Shibo Zhang, Shibo Zhang, Tousif Ahmed, Tousif Ahmed, Md. Mahbubur Rahman, Jilong Kuang, and Alex Gao. 2021. CoughBuddy: Multi-Modal Cough Event Detection Using Earbuds Platform. *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)* (2021). https://doi.org/10.1109/bsn51625.2021.9507017

[36] Nhan Nguyen, Avijoy Chakma, and Nirmalya Roy. 2021. A Scalable and Domain Adaptive Respiratory Symptoms Detection Framework using Earables. In *2021 IEEE International Conference on Big Data (Big Data)*. 5620–5625. https://doi.org/10.1109/BigData52589.2021.9671796

[37] Maja Pantic and Léon Rothkrantz. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE* 91 (10 2003), 1370 – 1390. https://doi.org/10.1109/JPROC.2003.817122

[38] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*. ISCA. https://doi.org/10.21437/interspeech.2019-2680

[39] Felix Pfreundtner, Jing Yang, and Gábor Sörös. 2021. (W)Earable Microphone Array and Ultrasonic Echo Localization for Coarse Indoor Environment Mapping. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4475–4479. https://doi.org/10.1109/ICASSP39728.2021.9414356

[40] Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. https://doi.org/10.7910/DVN/YDEPUT

[41] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: Earphones as a Teeth Activity Sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (London, United Kingdom) *(MobiCom '20)*. Association for Computing Machinery, New York, NY, USA, Article 40, 13 pages. https://doi.org/10.1145/3372224.3419197

[42] Kun Qian, Chenshu Wu, Fu Xiao, Yue Zheng, Yi Zhang, Zheng Yang, and Yu Liu. 2018. Acousticcardiogram: Monitoring Heartbeats using Acoustic Signals on Smart Devices. 1574–1582. https://doi.org/10.1109/INFOCOM.2018.8485978

[43] Md Juber Rahman, Ebrahim Nemati, Mahbubur Rahman, Korosh Vatanparvar, Viswam Nathan, and Jilong Kuang. 2019. Efficient Online Cough Detection with a Minimal Feature Set Using Smartphones for Automated Assessment of Pulmonary Patients.

[44] Muhammad Rashid, Guiqing Li, and Chengrui Du. 2023. Nonspeech7k dataset: Classification and analysis of human non-speech sound. *IET Signal Processing* 17 (06 2023). https://doi.org/10.1049/sil2.12233

[45] Yanzhi Ren, Chen Wang, Jie Yang, and Yingying Chen. 2015. Fine-grained sleep monitoring: Hearing your breathing with smartphones. 1194–1202. https://doi.org/10.1109/INFOCOM.2015.7218494

[46] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 135 (sep 2022), 57 pages. https://doi.org/10.1145/3550314

[47] Tobias Röddiger, Daniel Wolffram, David Laubenstein, Matthias Budde, and Michael Beigl. 2020. Towards Respiration Rate Monitoring Using an In-Ear Headphone Inertial Measurement Unit. In *Proceedings of the 1st International Workshop on Earable Computing* (London, United Kingdom) *(EarComp'19)*. Association for Computing Machinery, New York, NY, USA, 48–53. https://doi.org/10.1145/3345615.3361130

[48] Camilo Rojas, Niels Poulsen, Mileva Van Tuyl, Daniel Vargas, Zipporah Cohen, Joe Paradiso, Pattie Maes, Kevin Esvelt, and Fadel Adib. 2021. A Scalable Solution for Signaling Face Touches to Reduce the Spread of Surface-Based Pathogens. *Proc. ACM Interact. Mob.*

Wearable Ubiquitous Technol. 5, 1, Article 31 (mar 2021), 22 pages. https://doi.org/10.1145/3448121

[49] Siddharth Rupavatharam and Marco Gruteser. 2020. Towards In-Ear Inertial Jaw Clenching Detection. In *Proceedings of the 1st International Workshop on Earable Computing* (London, United Kingdom) *(EarComp'19)*. Association for Computing Machinery, New York, NY, USA, 54–55. https://doi.org/10.1145/3345615.3361134

[50] Xue Sun, Jie Xiong, Chao Feng, Wenwen Deng, Xudong Wei, Dingyi Fang, and Xiaojiang Chen. 2023. Earmonitor: In-Ear Motion-Resilient Acoustic Sensing Using Commodity Earphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 182 (jan 2023), 22 pages. https://doi.org/10.1145/3569472

[51] Akira Tamamori, Tomoki Hayashi, Tomoki Toda, and Kazuya Takeda. 2017. An investigation of recurrent neural network for daily activity recognition using multi-modal signals. 1334–1340. https://doi.org/10.1109/APSIPA.2017.8282239

[52] Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs.LG]

[53] Nian Chi Tay, Tee Connie, Thian Song Ong, Andrew Beng Jin Teoh, and Pin Shen Teh. 2023. A Review of Abnormal Behavior Detection in Activities of Daily Living. *IEEE Access* 11 (2023), 5069–5088. https://doi.org/10.1109/ACCESS.2023.3234974

[54] Timo Tigges, Thomas Büchler, Alexandru-Gabriel Pielmus, Michael Klum, Aarne Feldheiser, Oliver Hunsicker, and Reinhold Orglmeister. 2018. *Assessment of In-ear Photoplethysmography as a Surrogate for Electrocardiography in Heart Rate Variability Analysis.* 293–297. https://doi.org/10.1007/978-981-10-9038-7_54

[55] Hoang Truong, Alessandro Montanari, and Fahim Kawsar. 2022. Non-Invasive Blood Pressure Monitoring with Multi-Modal In-Ear Sensing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6–10. https://doi.org/10.1109/ICASSP43922.2022.9747661

[56] Korosh Vatanparvar, Ebrahim Nemati, Viswam Nathan, Mahbubur Rahman, and Jilong Kuang. 2020. CoughMatch – Subject Verification Using Cough for Personal Passive Health Monitoring. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2020). https://doi.org/10.1109/embc44109.2020.9176835

[57] Wei Wang, Fatjon Seraj, and Paul J. M. Havinga. 2020. A Sound-Based Crowd Activity Recognition with Neural Network Based Regression Models. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Greece) *(PETRA '20)*. Association for Computing Machinery, New York, NY, USA, Article 18, 8 pages. https://doi.org/10.1145/3389189.3389196

[58] Yuntao Wang, Jiexin Ding, Ishan Chatterjee, Farshid Salemi Parizi, Yuzhou Zhuang, Yukang Yan, Shwetak Patel, and Yuanchun Shi. 2022. FaceOri: Tracking Head Position and Orientation Using Ultrasonic Ranging on Earphones. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 290, 12 pages. https://doi.org/10.1145/3491102.3517698

[59] Yuntao Wang, Xiyuxing Zhang, Jay M. Chakalasiya, Xuhai Xu, Yu Jiang, Yuang Li, Shwetak Patel, and Yuanchun Shi. 2022. HearCough: Enabling continuous cough event detection on edge computing hearables. *Methods* 205 (2022), 53–62. https://doi.org/10.1016/j.ymeth.2022.05.002

[60] Xuhai Xu, Haitian Shi, Xin Yi, WenJia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K. Dey. 2020. EarBuddy: Enabling On-Face Interaction via Wireless Earbuds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376836

[61] 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* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 1013–1020. https://doi.org/10.1145/3332165.3347950

[62] Qingxue Zhang, Dian Zhou, and Xuan Zeng. 2017. Hear the heart: Daily cardiac health monitoring using Ear-ECG and machine learning. In *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. 448–451. https://doi.org/10.1109/UEMCON.2017.8249110

[63] Shijia Zhang, Yilin Liu, and Mahanth Gowda. 2022. Let's Grab a Drink: Teacher-Student Learning for Fluid Intake Monitoring using Smart Earphones. In *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*. 55–66. https://doi.org/10.1109/IoTDI54339.2022.00014

[64] Xiyuxing Zhang, Yuntao Wang, Jingru Zhang, Yaqing Yang, Shwetak Patel, and Yuanchun Shi. 2023. EarCough: Enabling Continuous Subject Cough Event Detection on Hearables. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 94, 6 pages. https://doi.org/10.1145/3544549.3585903

[65] Xiaodan Zhuang, Xi Zhou, and Mark Hasegawa-Johnson. 2010. Real-world acoustic event detection. *Pattern Recognition Letters* 31 (09 2010), 1543–1551. https://doi.org/10.1016/j.patrec.2010.02.005