GrooveMeter: Enabling Music Engagement-aware Apps by Detecting Reactions to Daily Music Listening via Earable Sensing

Euihyeok Lee
Korea University of Technology and Education
Republic of Korea

Chulhong Min
Nokia Bell Labs
United Kingdom

Jaeseung Lee
Korea University of Technology and Education
Republic of Korea

Jin Yu
Korea University of Technology and Education
Republic of Korea

Seungwoo Kang*
Korea University of Technology and Education
Republic of Korea

ABSTRACT

We present GrooveMeter, a novel system that automatically detects vocal and motion reactions to music and supports music engagement-aware applications. We use smart earbuds as sensing devices, already widely used for music listening, and devise reaction detection techniques by leveraging an inertial measurement unit (IMU) and a microphone on earbuds. To explore reactions in daily music-listening situations, we collect the first-kind-of dataset containing 926-minute-long IMU and audio data with 30 participants. With the dataset, we discover unique challenges in detecting music-listening reactions and devise sophisticated processing pipelines to enable accurate and efficient detection. Our comprehensive evaluation shows GrooveMeter achieves the macro F_1 scores of 0.89 for vocal reaction and 0.81 for motion reaction with leave-one-subject-out (LOSO) cross-validation (CV). More importantly, it shows higher accuracy and robustness compared to alternative methods. We also present the potential use cases.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

music-listening engagement, reaction detection, earable sensing

ACM Reference Format:
Euihyeok Lee, Chulhong Min, Jaeseung Lee, Jin Yu, and Seungwoo Kang. 2023. GrooveMeter: Enabling Music Engagement-aware Apps by Detecting Reactions to Daily Music Listening via Earable Sensing. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3611968

1 INTRODUCTION

Listening to music is an integral part of our life. According to a study [1], in 2022, we listened to music for more than 2.87 hours daily, equivalent to listening to about 57 songs. While listening to songs, we often nod our heads, tap our feet, and sing along to the songs simultaneously. These are the natural responses [8, 9, 27, 37], which is considered a characteristic showing our engagement with music [3, 14]. The reactions are compelling to enable interesting *music engagement-aware applications*. For example, music player apps can leverage listeners' on-the-fly reactions to provide an engagement-aware automatic music rating and recommendation by observing which part of a song listeners often react to.

Observing responses to music listening has widely been investigated in music psychology studies. They provided insightful findings to understand the characteristics of responses to music listening. However, adopting these methods for real-life applications is almost impossible because they mostly rely on self-report or bulky experimental equipment in controlled environments. For example, the previous studies measured brain activity using positron emission tomography or functional magnetic resonance imaging [7, 25], observed music-induced movement using motion capture systems [8, 15, 28], and measured physiological responses such as an electrocardiogram and galvanic skin response [29].

We propose *GrooveMeter*, a novel mobile system that tracks reactions to music listening to support music engagement-aware applications. As an initial attempt, we focus on readily observable bodily reactions which people usually experience while listening to music [16]. Specifically, we target singing along, humming, whistling, and head motion because they are common reactions from our in-the-wild dataset presented in §4. To this end, we use earbuds as sensing devices and devise reaction detection techniques by leveraging an IMU and a microphone on earbuds.

While research efforts have been made to recognize human activities and gestures, detecting music-listening reactions with sensor-equipped wearables has yet to be studied. Our observation reveals unique challenges in detecting the reactions accurately and robustly using audio and motion sensing. First, there are often reaction-irrelevant events with similar signal characteristics, which can cause false positive errors (e.g., mumbling to talk to themselves or looking at monitors and a keyboard alternately). Second, since listening to music is often a secondary activity, audio and motion signals can be affected by background noise (e.g., sound of talking nearby) and other motion artifacts, respectively. Models trained with data from a lab environment show poor performance in daily situations. Third, running reaction detection models on mobile devices incurs considerable overhead for continuous execution.

To address the challenges, we devise sophisticated processing pipelines for vocal and motion reaction detection with three main

[*] Corresponding author (swkang@koreatech.ac.kr)

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611968

features. First, we investigate the signal characteristics of data segments that can be *certainly* labeled as non-reaction and devise a method to effectively filter out those segments at the beginning of the pipeline, not only for cost-saving but also for improving robustness. Second, we elaborate multi-step reaction detection pipelines, reflecting the unique patterns of reactions. Third, we leverage the semantic similarity between sensor data and *musical structure information* retrieved from a song. The intuition is that the reactions correlate with the song, e.g., a listener's humming would naturally follow the song's pitch pattern. We correct ambiguously labeled audio segments based on the prosodic similarity at the last stage.

To build and evaluate GrooveMeter, we collect the first-kind-of dataset, called *MusicReactionSet*, from 30 participants under four situations (resting in a lounge, working at an office, riding in a car, and relaxing at a cafe). It contains 926-minute-long IMU/audio data with manually-labeled accurate annotation. Our extensive evaluation shows that GrooveMeter achieves the macro F_1 scores of 0.89 and 0.81 with LOSO CV for vocal and motion reaction detection, respectively. More importantly, it achieves higher accuracy and robustness compared to alternatives. Especially in noisy situations, we observe a significant performance enhancement, e.g., an increase in F_1 score by up to 0.21 and 0.09 for vocal and motion reactions in the case of relaxing at a cafe, respectively. To demonstrate its usefulness, we develop a prototype on Android phones and earbuds, along with example applications. We present a case study showing the feasibility of automatic music rating and familiarity detection.

We summarize the contribution of this paper as follows. First, we collect MusicReactionSet, containing IMU/audio data with 30 participants to explore reactions in daily music-listening situations. Second, we develop GrooveMeter, a novel system that detects reactions to music listening via earables sensing. To the best of our knowledge, this is the first to present an earable sensing solution specialized for automatically detecting vocal and motion reactions to music. Third, we propose a novel technique to detect reactions efficiently and robustly by filtering out reaction-irrelevant data segments and leveraging music information retrieved from a song. We also present a comprehensive evaluation using MusicReactionSet.

2 RELATED WORK

Reaction sensing: Several works attempted to monitor consumers' reaction to multimedia content or performing arts, e.g., the responses of users watching movies [4], audience responses in live performances [31], the experience of audiences in the play [43], the frisson of audience during music performances via physiological sensing [17]. They share a high-level goal with ours, detecting content consumers' reactions at runtime. However, due to different characteristics of content-dependent reactions, the required sensor modality, devices, and techniques should be different. We aim at detecting music-listening reactions by reflecting unique signal characteristics of the reactions. We also discover an opportunity for robust detection in noisy conditions. We develop a novel technique to exploit the semantic similarity between sensor data and music information. Note that the previous works do not utilize the characteristic of contents, but rely on sensor data only.

Human sensing using earables: Recent works tried to sense diverse human contexts using IMUs and microphones in earbuds. They use IMUs to recognize physical activity [32], facial expression [26, 41], jaw movement [22], and gait posture [20]. Microphones are used for eating detection [32], motion tracking [10], gait sensing [11], and human activity recognition [30]. MusicalHeart [34] monitors heart rate and activity level. However, we focus on novel reaction sensing for daily music-listening situations.

Understanding music listening behavior and contexts: Some studies tried to understand music-listening behavior and contexts. One of the initial attempts was made in [35], which used text messages to analyze the music people heard. A smartphone-based tool to collect music-listening behavior was also developed, e.g., surrounding contexts, and user activity [45]. For contextual music recommendation, a study tried to understand the users' intent for listening to music and its relationship to common daily activities with an online survey [42]. A music recommendation model based on activities, e.g., working, studying, and sleeping, was also proposed [44]. Our work differs from them in two aspects. While most of the existing studies focused on a user's behavior and contexts, i.e., what/when/why/where people listen to music, we focus on *how people react* to music, which has been rarely studied in the field. Next, we address unique challenges in detecting reactions to music in real-life situations and devise novel reaction detection methods.

3 REACTION SENSING

Monitoring on-the-fly music listening reactions in unconstrained mobile environments opens a broad spectrum of applications. Figure 1 shows the high-level process of how GrooveMeter supports music engagement-aware applications. First, as a common basis for any applications, GrooveMeter focuses on detecting vocal and motion reactions in real time using IMU and audio signals, i.e., which type of reaction was made, and when and for how long. By combining this primitive information, GrooveMeter further provides music engagement-aware applications with high-fidelity information, e.g., which part of a song listeners most sang along to or moved. We discuss the potential scenarios with the case study in §5.3.

In this work, we target *singing along*, *humming* and *whistling* as vocal reactions, and *head motion* as motion reactions. Note that these are commonly observed reactions from our in-the-wild dataset with 30 participants presented in §4.

We develop GrooveMeter based on three design considerations.

- Unobtrusive sensing: Tracking reactions to music in real-life situations should not rely on neither infrastructure-deployed nor excessive on-body sensors.
- Accurate and robust detection: Reaction detection should be accurate and robust against reaction-irrelevant behaviors similar to reactions, background noise, and other motion artifacts in daily music-listening situations.
- Low overhead: Although GrooveMeter runs only while a user listens to music, users would not prefer to consume much battery.

3.1 Vocal Reaction Detection

3.1.1 Challenges. **Vocal reaction events:** A straightforward way of detecting sound events like vocal reactions is to use pre-trained models or develop new ones. However, simply adopting existing audio models does not fit our purpose. Recently, many sound classification models have been released, which are pre-trained with an

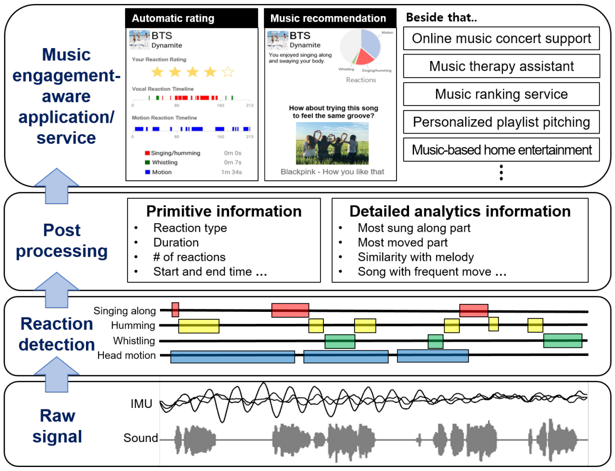


Figure 1: High-level process to support music engagement-aware applications.

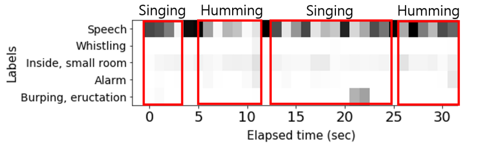


Figure 2: YAMNet result for vocal reactions

enormous amount of data and predict a large number of real-life events. For example, YAMNet [19] classifies 521 audio event classes.

The pre-trained models show satisfactory accuracy for daily events. However, none of them includes our target vocal reaction events yet, and accordingly they show poor performance. For example, YAMNet's *singing* label is mostly derived from video clips where a song is played with instruments. However, when listeners make *singing* reaction while listening to music via earbuds, the captured audio includes only their singing voice. Figure 2 shows our preliminary study with YAMNet; it shows the output of the softmax layer. YAMNet hardly selects the *singing* label, but it classifies singing reactions mostly as the *speech*. Developing a custom model with newly collected data could address these issues. However, it still requires collecting large-scale real-life sound events to reflect user variability and avoid errors due to noises in daily situations.

Mixed with background noise: Background noise (e.g., a sound of nearby people talking or background music in a cafe) makes it more complicated to detect vocal reactions correctly. We observe that YAMNet often classifies singing and humming reactions mixed with diverse noise in a cafe as the *music* label. Moreover, under a noisy condition, YAMNet's softmax scores of the reaction-relevant labels tend to be relatively low and sometimes even lower than reaction-irrelevant labels. From our data collected in noisy places such as a cafe and a car, only 5.3% of singing/humming segments result in greater than 0.95 of YAMNet softmax score; in less noisy places such as lounge, 11.7% does so.

Intermittence and alternation: When a vocal reaction is made, it does not continue ceaselessly within a session, but sporadically and sometimes alternately with other reactions. For example, when listeners sing along, we observe intermittent, short-period pauses that the listener makes to breathe. Also, they often make different types of reactions alternately. For example, while listeners are singing along, they often switch to humming or whistling momentarily if they do not know the lyrics and come back to sing along again in the part where they know the lyrics.

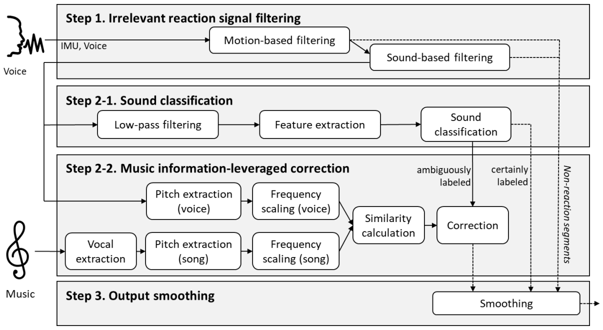
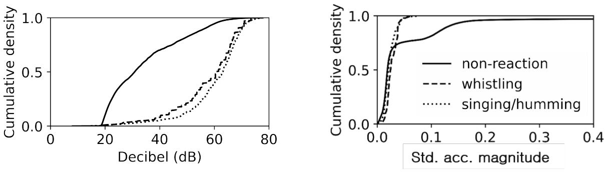


Figure 3: Vocal reaction detection pipeline



(a) Sound level **(b) Movement level**

Figure 4: CDF of sound and movement levels

Processing cost: Sound classification often involves processing-heavy operations such as MFCC computation and deep neural networks. While today's models provide optimization for on-device processing, e.g., MobileNet architecture of YAMNet, it still incurs significant overhead for continuous execution. For example, continuously performing audio classification with YAMNet on Galaxy S21 while playing songs incurs 3% drop in battery level for one hour.

3.1.2 Overview. To address the aforementioned challenges, we devise a novel pipeline that detects vocal reactions efficiently and reliably. Figure 3 shows its overview with three major operations. First, we adopt the early-stage filtering operation to save cost. Its key idea is to filter out data segments that can be certainly labeled as *non-reaction* (§3.1.3). Second, we initially classify sound events with the YAMNet model (§3.1.4) and correct ambiguous labels by leveraging music information retrieved from a song being played (§3.1.5). Last, it smooths the final outputs to cope with the momentarily introduced short, intermittent events (§3.1.6).

3.1.3 Certain Non-reaction Signal Filtering . A filtering operation aims to identify data segments that can be certainly labeled as *non-reaction*, and avoid the processing-heavy operations for those segments. It is based on two observations. First, vocal reactions would make sound events above a certain volume due to a short distance between the earbud's microphone and the wearer's mouth. Thus, sound events below a certain volume threshold can be confidently labeled as *non-reaction*. Figure 4a shows the cumulative distribution function (CDF) of one-second decibel numbers for sound events and validates our hypothesis. Second, vocal reactions also incur a certain level of kinetic movement of an earbud because the mouth movement for vocal reactions activates the *Zygomaticus* muscle located between the mouth and ear [22, 26]. If no motion is detected, the corresponding audio signal is unlikely to be the *reaction* label. Interestingly, large motion is also associated with non-reaction since listeners hardly make vocal reactions when they walk or run.

Table 1: Mapping from YAMNet to GrooveMeter labels

YAMNet	GrooveMeter
humming, singing	singing/humming
whistling, whistle	whistling
speech, music	ambiguous (candidate for singing/humming or non-reaction)
others	non-reaction

Figure 4b shows the CDF of the standard deviation of one-second accelerometer magnitude, which supports this observation.

Based on the findings, we design a two-step filtering component. It first monitors the level of movement defined as the standard deviation of accelerometer magnitude values, and filters out data segments out of the threshold range (we set the range to 0.0104 and 0.12). Second, it further filters out previously unfiltered segments if their decibel is lower than a threshold (currently, 49 dB). We place the motion-based filter before the sound-based filter because it is more lightweight. From our measurement, the former consumes 113 mW on Galaxy S21, while the latter does 134 mW. It labels filtered segments as *non-reaction* and delivers them to the post-processing stage without performing the subsequent operations.

3.1.4 Sound Event Classification . **Target events:** We target three types of vocal reaction events; *singing (along)/humming*, *whistling*, and *non-reaction*. We combine singing and humming into the same class because they are often observed alternatively, even in a single reaction session, as mentioned above.

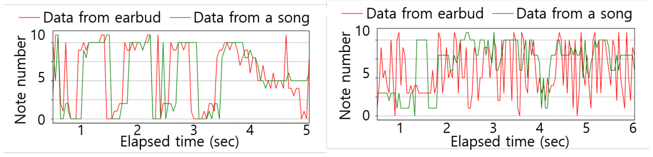
Preprocessing: Audio data are resampled at 16 kHz and divided into 1-second-long segments. We then apply a Low Pass Filter of order 1 (cut-off at 2 kHz) to reduce noise signals with frequencies higher than the major frequency range of the vocal reactions.

Audio feature extraction: The preprocessed segments are converted to spectrograms using the Short-Time Fourier Transform with a periodic Hann window; we set a window size and a window hop to 25 ms and 10 ms, respectively. We then map the spectrogram to 64 mel bins with a range between 125 and 7,500 Hz and compute log mel spectrograms. Finally, we frame the features into a matrix of 96×64 (96 frames of 10 ms each and 64 mel bands of each frame).

Classification and labeling: We use YAMNet [19] as a base component for the sound event classification. However, since the taxonomy of YAMNet labels does not fit our target classes, we construct a mapping from YAMNet to GrooveMeter labels, as shown in Table 1. As discussed, YAMNet is poorly discriminating *singing* reactions from speech and music. We thus map speech and music labels from YAMNet output to *ambiguous*. We perform further investigation for ambiguous segments with the correction operation. Other labels are directly sent to the post-processing operation.

Rank constraint relaxation: Simply relying on the YAMNet's result is insufficient for accurate detection even with applying label mapping. Due to the background noise and characteristics of vocal reactions, YANNet outputs speech or music as a top-1 classification label only for 57.8% of singing/humming segments in our dataset. The ratio increases as we include speech or music labels in a lower rank, e.g., 65.4% (top 2) and 70.1% (top 3). The whistling segments also show a similar characteristic.

To address the problem, we employ a rank constraint relaxation policy, allowing some of the segments that YAMNet does not classify as one of our target labels to go through the correction step. Here we need a balance to avoid unnecessary costs for the correction step and an increase in false positive errors due to additional non-reaction segments that can be incorrectly classified as vocal reactions. We do not apply relaxation if the quality of the YAMNet output is good enough. Otherwise, we check if lower-ranked results include some vocal reaction labels. If so, we consider it an *uncertain* label that needs further investigation.



(a) Singing along **(b) Non-reaction**

Figure 5: Notes of a chromatic scale

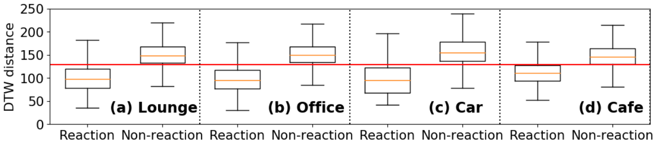


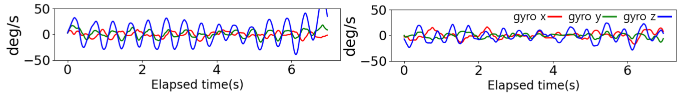
Figure 6: Similarity differences

To quantify the quality of the classification output, we adopt the strategy of uncertainty sampling [39] in the domain of active learning. More specifically, we use the least margin, which measures the uncertainty by taking the difference between the confidence values of the top two output classes. If the margin is lower than a threshold, we check top-k YAMNet output labels. If they include our target labels (the first three rows in Table 1), the segment is considered uncertainly labeled and forwarded to the correction step. We empirically set the margin threshold and k to 0.9 and 5.

3.1.5 Music information-leveraged Correction . We finalize ambiguous or uncertain segments from the previous step by leveraging music information of a played song. Based on the prosodic similarity between the audio signal and the song, we correct ambiguous segments with speech or music labels to *singing/humming* or *non-reaction*. We also deal with uncertain segments in the same way.

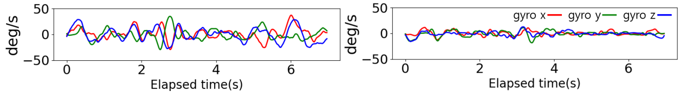
Prosodic similarity computation: We consider the melody to measure the prosodic similarity between a vocal signal and a song, which refers to a linear succession of musical tones. Our intuition is that vocal reactions would follow the sequence of notes of a played song, but reaction-irrelevant speech signals would not. Step 2-2 in Figure 3 shows the detailed procedure. To extract the sequence of a note, we first extract the pitch information every 0.1 seconds using CREPE [24], a state-of-the-art pitch tracker. We then convert it to a musical note with an octave number. We convert it again to a 12-tone chromatic scale without an octave number because we observe that vocal reactions are often made an octave higher or lower than a song. For the song's audio file, we perform vocal extraction before pitch extraction to focus on the predominant melodic line of music. This insight is based on our observation that vocal reactions mostly follow vocals rather than instruments. We use Spleeter [18] to separate a vocal source from a song.

We compute the similarity between two sequences of notes (one from a user's vocal signal and the other from a song) and make a final decision. We map 12 notes (from C, C#, to B) to twelve integers (0 to 11). Figure 5a shows a high correlation of note patterns with the song, but the non-reaction part does not (Figure 5b). We consider



(a) Nodding - case 1 **(b) Nodding - case 2**

Figure 7: Diverse patterns of motion reaction



(a) Riding in a car **(b) Chewing cookie**

Figure 8: Repetitive, reaction-irrelevant motions

dynamic time warping (DTW) [6] as a similarity function since two patterns can vary in speed.

We label a segment as *non-reaction* if the DTW distance is larger than a threshold. Otherwise, we apply label mapping and confirm its final label, i.e., speech and music to *singing/humming*, whistling and whistle to *whistling*, humming and singing to *singing/humming*. Currently, we set the threshold to 130. Figure 6 presents the distribution of DTW distance values of singing/humming reaction and non-reaction segments in our dataset. Interestingly, it shows similar distribution regardless of noise conditions.

3.1.6 Post processing . **Smoothing:** We use a Hidden Markov Model (HMM) to smooth the classification output. The key idea is to train the HMM model from the sequence of the classification outputs of the training dataset and to use the trained model for the output smoothing. We define the observation sequence as a sequence of the classification outputs and perform smoothing by estimating the optimal sequence of hidden states, which can be mapped to the smoothed sequence of reaction events. More specifically, for a given sequence of classification outputs at time t, O^t = ($o_1, \ldots\ o_t$), we extract the sequence of hidden states with the maximum probability, S^t = ($s_1, \ldots\ s_{t-1}$), from time 1 to $t-1$. Then, the smoothed value at time t, $\hat{s}_t$, is obtained as follows:

$$\hat{s}_t = \underset{s_t}{\operatorname{argmax}}\ p(s_t|O^{t+1}, \lambda) \tag{1}$$

We apply the Viterbi algorithm [13] for efficient computation of maximum probability and use the 6 second-long window as an input sequence, i.e., a sequence of recent 6 classification outputs.

3.2 Motion Reaction Detection

3.2.1 Challenges. **Periodic and repetitive, but diverse motion trajectories:** From our observation, motion reaction often exhibits repetitive, periodic patterns. For example, head nodding (Figures 7a and 7b) continues for some duration with a regular pattern, which yields a signal waveform with a certain level of periodicity. One may argue that typical pipelines for physical activity recognition can easily capture such patterns. However, we found that classifiers using widely-used features representing the signal's periodicity or statistical features were ineffective for in-the-wild reaction data, as shown in §5.2. It is mainly because real-life motion reactions to music tend to vary, unlike well-defined activities or gestures following typical motion trajectories. For example, people often move their heads to the music, sometimes up and down or side to side, and the magnitude and speed of the reaction also vary, even for the same person or listening session. Figures 7a and 7b show two different patterns of head nodding from the same participant.

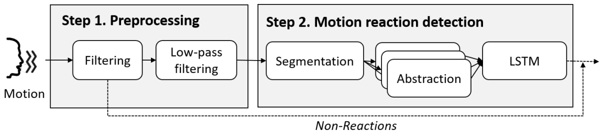


Figure 9: Motion reaction detection pipeline.

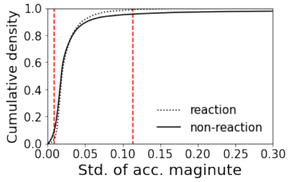


Figure 10: CDF of movement level of motion reactions.

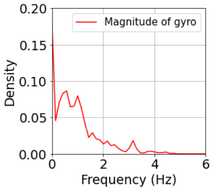


Figure 11: Frequency distribution of motion reactions.

Confusing motion trajectories from reaction-irrelevant movements: In our preliminary study, there are several reaction-irrelevant movements that show repetitive patterns and accordingly can cause classification errors in reaction detection. For example, the movement and vibration of a car during a ride can cause the repetitive IMU signal on earbuds, even though the person does not make any specific motion (Figure 8a). Similarly, chewing a cookie also generates a certain level of repetitive patterns (Figure 8b).

Affected by other motion artifacts: People often listen to music as a secondary activity, which means that their primary task (e.g., relaxing at a cafe) and corresponding movement (e.g., chewing a cookie) can affect the IMU signal. For example, when a user is nodding to the rhythm while performing another activity, the periodicity from the nodding movement is less clearly shown.

3.2.2 Overview. We design a novel pipeline for motion reaction detection that addresses the challenges. Figure 9 shows its overview with two main operations, preprocessing and reaction detection. First, we adopt a simple filter to avoid unnecessary processing for the classification operation (§3.2.3). Then, we remove the noise caused by other motion artifacts by adopting a low-pass filter (§3.2.4). Second, we extract a sequence of *motion units* from raw IMU signals to represent an abstraction of a user's motion pattern. With the sequence, we detect motion reaction using LSTM, performing binary classification (head motion vs. non-reaction) (§3.2.5).

3.2.3 Reaction-irrelevant movement filtering . We design a threshold-based filter based on our observation. It sorts out reaction-irrelevant data by looking into the movement level. We define the movement level as the standard deviation of a 1-second segment of accelerometer signal. Similarly in the motion filter of vocal reaction detection, it is obvious that no movement implies *non-reaction.* Large movement is also associated with a *non-reaction*, because it is very unlikely that listeners nod their head while doing workout, running, etc. Note that we carefully select a threshold range that can filter out non-reaction cases without missing reaction cases. We examine the CDF of the movement level (see Figure 10) and empirically set the low and high threshold values to 0.0092 g and 0.114 g, respectively.

3.2.4 Noise removal . The next step is to remove motion noise caused by other motion artifacts. The intuition behind this idea is that a listener's motion reaction tends to follow beat patterns of a song. Accordingly, we could expect that motion reactions tend to exhibit low frequency movement, considering that typical tempo of

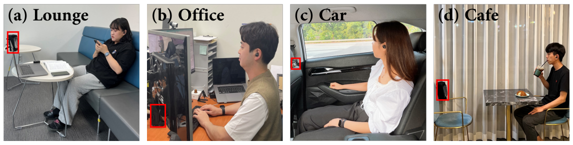


Figure 12: In-the-wild data collection in various places

Table 2: Characteristics of music listening situations

Situation	Reaction-irrelevant motions	Background Noise
Resting in a lounge	random movement while sitting on a chair	noise from air-purifier, murmuring sound outside, ...
Working at an office	motions during web search and word processing	keyboard typing, mouse clicking sound, ...
Riding in a car	bouncing along the road	various noises of driving car
Relaxing at a cafe	drinking coffee, chewing a cookie	background music, nearby conversation, chewing sound, ...

common music genres ranges between 60 and 180 beats per minute. Figure 11 shows a distribution of dominant frequency extracted from our motion reaction data. Dominant frequencies are less than 4 Hz. Thus, we process the raw IMU data with a low pass filter (LPF) of order 1 (cut-off at 5 Hz), allowing some margins.

3.2.5 Motion reaction classification . The classification operation includes two major steps. The first step is deriving temporal motion patterns from IMU signals. As mentioned, motion reactions do not show well-defined, typical movement trajectory and duration. Thus, we devise a method that abstracts IMU signals reflecting human motions and detects motion reactions accordingly. We define a *motion unit* representing a set of features derived from a short time interval of IMU data. It can be viewed as an abstraction of the user's motion pattern, which represents reactions to music, random body motion irrelevant to the reactions, or a stationary state. To extract motion units, we segment the preprocessed IMU data into 100ms and compute statistical features for each gyroscope axis, i.e., max, min, mean, range, standard deviation, and RMS.

Next, we classify a sequence of motion units into one of two classes, head motion and non-reaction. Considering that the input is a temporal sequence, we adopt an LSTM model widely used for predicting sequential data. We build a classification model consisting of an LSTM layer with 32 hidden units, a dropout layer with a drop rate of 0.5, a ReLU layer, and a softmax layer. We empirically set a window size to 7 seconds, where the F_1 score starts to saturate.

4 MUSIC REACTION DATA

To build and evaluate GrooveMeter, we create MusicReactionSet, a novel dataset consisting of audio and IMU data from a variety of music listening reactions. To the best of our knowledge, this is the first dataset targeting reactions in music listening situations. The data collection was conducted under IRB approval.

Participants: We recruited 30 participants (M: 18, F: 12) from a university campus; their ages were between 20-26 (mean 22.7). We obtained informed consent from the participants. All of them reported that they frequently listen to music in daily life. They were compensated with a gift card worth USD 18.

In-the-wild data collection: Each participant was invited to four places and listened to a set of songs while doing other activities, i.e., resting in a lounge, working at an office, riding in a car, and relaxing at a cafe (see Figure 12). We consider these places to (a) reflect diverse real-life situations where people often enjoy listening to music and (b) investigate the impact of diverse audio and motion noise. Table 2 shows the characteristics of four situations.

In each situation, the participants freely chose three songs with two different genres, i.e., exciting/up-tempo and slow/soft, from the top-50 chart in a music streaming service. They listened to the songs using earbuds and a smartphone provided. To collect data from their natural reactions, we did not give any instruction and also let them be alone in the places. Note that we did not include the data from the first song because they felt a little distracted right after they moved to a new place. Finally, 926 minute-long IMU/audio data were collected from 240 music listening sessions.

Setup: We used *earbuds* for music streaming and IMU/audio sensing, and an Android *phone* for connecting earbuds and controlling music playing. For earbuds, we used Apple AirPods Pro, but also additionally used eSense [21] to collect IMU data, i.e., one eSense unit on the left ear and one AirPod unit on the right ear; note that AirPods did not allow developers to access IMU data when we collected the data. The sampling rate of a microphone on AirPods Pro and IMU on eSense was set to 44.1 kHz and 70 Hz, respectively. For ground truth tagging, we recorded data collection session with a covered camera, marked with a red rectangle in Figure 12.

5 EVALUATION

For evaluation, we implemented the prototype of GrooveMeter as an Android service on two phones, Galaxy S21 and Galaxy S8+. We use TensorFlow Lite [2] to run our pipelines with YAMNet and LSTM. We also measure the processing latency and energy cost. Due to the page limit, we only brief the energy cost, 3.7 mJ/s (w/ filtering) and 7.3 mJ/s (w/o filtering) on Galaxy S21. The early filtering approach saves 50% of energy.

5.1 Vocal Reaction Detection

5.1.1 Overall Performance. We present the overall performance of the vocal reaction detection. We use LOSO CV with the MusicReactionSet dataset. Note that we used the original YAMNet model [19] for the classification task, thus the same model is used for testing all subjects. LOSO CV is considered to obtain the threshold values in the vocal reaction pipelines to avoid the over-fitting problem.

Figure 13 shows the averaged precision and recall of vocal reaction labels over 30 validations. The results show that GrooveMeter achieves the reasonable performance of the vocal reaction detection even for an unseen user and under a variety of real-life background noise types; the macro-averaged F_1 score is 0.90. More specifically, it detects *singing/humming* reactions with 0.85 and 0.87 of precision and recall, respectively, and *whistling* reactions with 0.93 and 0.78. The recall of whistling is relatively low compared to others because some whistling segments with weak sound or mixed with background noise are incorrectly inferred by YAMNet. The results also show that our method correctly identifies non-reaction events. The precision and recall for the *non-reaction* label are 0.99 and 0.97.

5.1.2 Effect of filtering: We examine the effect of early-stage filtering. Figure 14 shows the F_1 score and filtering ratio with different filtering strategies. For the study, we developed three different versions of vocal reaction detection pipeline; *none*, *motion*, and *sound*.

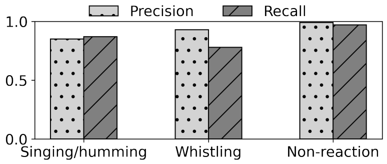


Figure 13: Overall performance

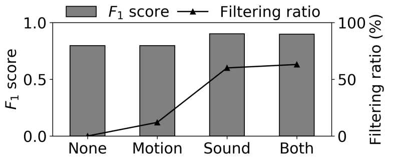


Figure 14: Effect of filtering

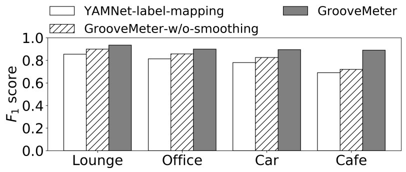


Figure 15: Robustness against noise

We define the filtering ratio as the number of filtered segments divided by the total number of segments; a high filtering ratio means less processing cost. None means the reaction detection without any filtering operation. Motion and sound refer to the pipeline when only motion- and sound-based operation is added, respectively.

Interestingly, the filtering operation is not only effective for reducing computation cost, but also helpful in improving performance by effectively filtering out reaction-irrelevant segments. While *none* achieves 0.8 of F_1 score without any filtering, *both* (applying both filtering operations) increases the F_1 score by 0.1. Specifically, the filtering ratios of motion and sound-based operation are 12% and 60%, respectively. The motion-based filtering reduces a fair amount of non-reaction data to process without compromising performance. The sound-based filtering reduces even more amount of data. Also, it effectively removes false positives of the original YAMNet, which were made due to background noise, thereby increasing the F_1 score. When both operations are used together, the filtering ratio increases up to 63% together with the increase in F_1 score by 0.1.

5.1.3 *Robustness against acoustic noise:*

We further investigate the robustness of our technique against noise in real-life situations. As presented in §4, we consider four places exhibiting different noise characteristics, and compare the F_1 scores. We break down the performance by comparing GrooveMeter with two variants: YAMNet-label-mapping and GrooveMeter-w/o-smoothing. We include the filtering operation to examine overall performance.

Figure 15 shows the macro-averaged F_1 scores for vocal reactions and non-reactions. The results show the GrooveMeter's robustness regardless of different noise characteristics. Compared to YAMNet-label-mapping (YAMNet-based classification and label mapping), it increases the F_1 score by 0.08, 0.09, 0.12, and 0.21 in lounge, office, car, and cafe, respectively, by adopting music information-leveraged correction (GrooveMeter-w/o-smoothing) and smoothing (GrooveMeter). The correction operation does contribute much in the cafe due to relatively large false positives from background noise (0.04 increase of F_1 score). Interestingly, the smoothing shows meaningful improvement by leveraging the temporal association of reaction labels (further 0.17 increase in F_1), thereby achieving comparable performance to other places.

5.2 Motion Reaction Detection

We present the performance of our motion reaction detection. For the validation, we used LOSO CV with the MusicReactionSet dataset. We implement three baselines by referring to prior works as follows.

- **RandomForest** represents feature-based sensing pipelines to recognize repetitive and periodical physical activities, e.g., [5, 23, 33]. We use time and frequency-domain statistical features [12] and auto-correlation-derived features [5, 23] from IMU data, and choose Random Forest as a classifier.
- **CNN** represents a deep learning method for activity recognition, e.g., [38], which uses convolutional neural network with 6-axes IMU data. We build a classifier consisting of 3 convolutional layers with a ReLU activation function, a max pooling layer, 2 dropout layers with a drop rate of 0.5, and a softmax layer.
- **ConvLSTM** represents a method that combines convolutional and LSTM recurrent layers for activity recognition with wearables, e,g., [36]. It uses convLSTM [40] with 6-axes IMU data. We build a model consisting of a ConvLSTM layer, a dropout layer with a drop rate of 0.5, a ReLU layer, and a softmax layer.

5.2.1 *Overall Performance.*

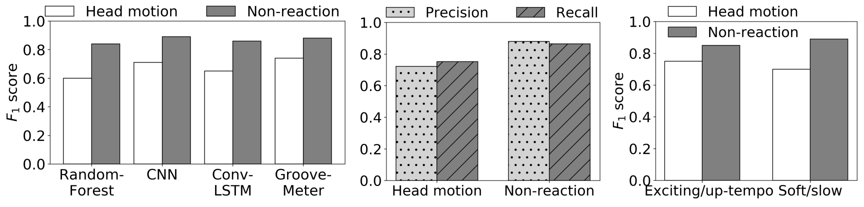
Figure 16a shows the performance comparison between GrooveMeter and the baselines. The results show that GrooveMeter detects the head motion more accurately than the baselines. While there is a marginal difference in the F_1 score of the non-reaction, GrooveMeter increases the F_1 score of the head motion by 0.09 on average, compared to the baselines. For the head motion, the F_1 score of GrooveMeter is 0.74, whereas that of RandomForest, CNN, and ConvLSTM is 0.60, 0.71, and 0.65, respectively. One may argue that the performance improvement of GrooveMeter from CNN (0.03 increase) is marginal. However, we found that GrooveMeter is more robust to motion noisy environments. We present the in-depth analysis in §5.2.2.

We take a deeper look at the performance of GrooveMeter. Figure 16b shows, even for an unseen user, our method provides reasonable performance. Specifically, it detects *head motion* with 0.72 and 0.75 of precision and recall, respectively. For *non-reaction*, it achieves 0.88 of precision and 0.87 of recall.

We look into the results depending on the genre (Figure 16c). The results of exciting/up-tempo songs show higher precision and recall for the reaction than those of soft/slow songs, resulting in a relatively large F_1 score. Specifically, the F_1 scores of *head motion* and *non-reaction* are 0.75 and 0.85, respectively, for up-tempo songs. Those for slow songs are 0.70 and 0.89, respectively. While listening to up-tempo songs, people tend to nod vigorously. Thus, motion reactions show more prominent signals and clear periodicity, which yields better performance. In contrast, with soft/slow songs, motion reactions tend to be weak, and their trajectory is small. Thus, more reaction data can be confused with non-reaction motions.

5.2.2 *In-depth Comparison .*

To better understand the difference from the baselines, we additionally collected the *controlled* dataset in a lab setting condition. We recruited 10 (M: 6, F: 4) additional participants (ages: 20-26, mean: 23.1), compensated with a gift card worth USD 9. We asked them to follow an instructed scenario: two sessions to collect music-listening reactions and one for others. In the first 2 sessions, we provided the top 100 music and let them freely select a song to listen to in every session. Then, they were asked to make a given reaction naturally, but continuously for 60 to 90 seconds. The last session's task was freely moving around the



(a) Comparison with baselines **(b) Precision/Recall** **(c) Per-genre result**

Figure 16: Motion reaction detection performance

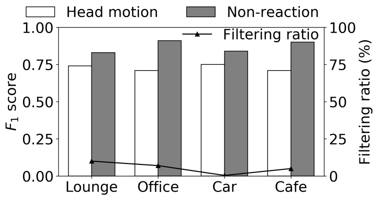


Figure 17: Effect of activities

Table 3: Comparison with the baselines

	Controlled	MusicReactionSet				
		Lounge	Office	Car	Cafe	Avg.
Head motion						
RandomForest	0.96	0.61	0.54	0.67	0.59	0.60
CNN	0.93	0.74	0.68	0.75	0.66	0.71
ConvLSTM	0.97	0.66	0.57	0.73	0.64	0.65
GrooveMeter	0.94	0.75	0.72	0.75	0.72	0.74
Non-reaction						
RandomForest	0.92	0.81	0.90	0.80	0.86	0.84
CNN	0.85	0.84	0.92	0.86	0.89	0.89
ConvLSTM	0.92	0.81	0.89	0.86	0.89	0.86
GrooveMeter	0.86	0.84	0.91	0.85	0.90	0.88

lab and making music-irrelevant motions while listening to music, which represents *non-reaction*.

Table 3 shows the F_1 scores of *head motion* and *non-reaction*, respectively, in the controlled and MusicReactionSet datasets. We first look into the performance of the head motion. Interestingly, while all the methods show similar performance in the controlled data, the performance gap is noticeable in MusicReactionSet. The F_1 scores are over 0.93 in the controlled data. In MusicReactionSet, the F_1 score generally decreases due to motion noises and diverse motion reaction patterns, but GrooveMeter shows a smaller gap than the baselines.

For the non-reaction class, CNN and GrooveMeter show lower F_1 than the others in the controlled data, but they show higher F_1 in the MusicReactionSet. The F_1 scores of CNN and GrooveMeter are 0.89 and 0.88, whereas those of RandomForest and ConvLSTM are 0.84 and 0.86, respectively. We conjecture that this is because the head motion and non-reaction segments have clearly distinguishable patterns in the controlled data. However, in MotionReactionSet, there are more confusing cases due to various patterns of natural head motions and daily motion noises, increasing false positive errors of RandomForest and ConvLSTM.

5.2.3 Effect of Activities. Figure 17 shows the macro-averaged F_1 scores and filtering ratios with four different activities. While it shows similar performance regardless of activities, F_1 scores of head motion in the office and cafe cases are slightly smaller than the others. The filtering ratio of lounge is 10%, larger than the others. The car case is only 0.4% since most of non-reaction data are within the filter range due to the movement by the car.

5.3 Application Case Study

We present application case studies to show the potential of GrooveMeter. We further recruited ten participants and asked them to listen to eight songs. We randomly selected four songs that they had never listened to before from the top 50 charts, namely *unknown* songs. We chose the other four songs from their playlist, i.e., *known* songs.

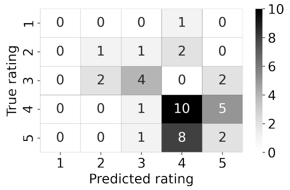


Figure 18: Confusion matrix of rating prediction.

Automatic music rating: One straightforward use case is automatic and fine-grained music rating. Today's music rating mostly relies on a user's manual input and simple statistics, e.g., the number of plays. We envision that rating can be automatically predicted in a fine-grained way by observing listeners' reaction patterns.

To study its feasibility, we asked the participants to provide ratings for eight songs they listened to, on a scale of 5; 5 means *"I like this song very much."*. To build a rating prediction model, we extract statistical features from the reaction detection output, i.e., normalized duration and the number of each reaction label, and use a decision tree as a classifier. We examine the rating prediction performance with LOSO CV. We used the data only from unknown songs since rating prediction is more useful for the songs. Figure 18 shows the confusion matrix of rating prediction. Higher values around the diagonal indicate that the predicted ratings are meaningfully close to the actual ratings (MAE: 0.22).

Familiarity detection: We investigate if the familiarity of a song can be detected using music listening reactions, i.e., to detect if a user has already listened to a song before or not. This functionality would help music streaming services accelerate to build a new subscriber's music preference without explicitly asking which songs have been enjoyed before. We build a decision-tree model trained by using statistical features of vocal and motion reaction events. Similarly to automatic music rating, we validate its performance in a LOSO manner, but using the full dataset. The F_1 score for the detection of known and unknown songs is 0.78 (precision: 0.85, recall: 0.72) and 0.81 (precision: 0.76, recall: 0.88), respectively.

6 CONCLUSION

We present GrooveMeter, a novel system to detect vocal and motion reactions to music via earable sensing. It features novel processing pipelines to make reaction detection accurate, robust, and efficient. We present extensive experiments to show its effectiveness with a dataset from 30 participants in daily music-listening situations.

ACKNOWLEDGMENTS

We thank anonymous reviewers for their valuable comments and Dongwoo Kim for his aid at the early stage of this work. This research was supported by the IITP (IITP-2020-0-01778) and the NRF (2020R1I1A3060675) grant funded by the Korea government.

REFERENCES

[1] [n.d.]. Music Listening 2022. https://www.ifpi.org/wp-content/uploads/2022/11/Engaging-with-Music-2022_full-report-1.pdf. Accessed: Apr. 22, 2023.

[2] [n.d.]. TensorFlow Lite. https://www.tensorflow.org/lite. Accessed: Apr. 22, 2023.

[3] Yudhik Agrawal, Samyak Jain, Emily Carlson, Petri Toiviainen, and Vinoo Alluri. 2020. Towards Multimodal MIR: Predicting individual differences from music-induced movement. *arXiv preprint arXiv:2007.10695* (2020).

[4] Xuan Bao, Songchun Fan, Alexander Varshavsky, Kevin Li, and Romit Roy Choudhury. 2013. Your reactions suggest you liked the movie: Automatic content rating via reaction sensing. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 197–206.

[5] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: using wearable sensors to detect eating episodes in unconstrained environments. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–20.

[6] Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*. 359–370.

[7] Anne J Blood and Robert J Zatorre. 2001. Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the national academy of sciences* 98, 20 (2001), 11818–11823.

[8] Birgitta Burger, Marc R Thompson, Geoff Luck, Suvi Saarikallio, and Petri Toiviainen. 2013. Influences of rhythm-and timbre-related musical features on characteristics of music-induced movement. *Frontiers in psychology* 4 (2013), 183.

[9] Birgitta Burger, Marc R Thompson, Geoff Luck, Suvi H Saarikallio, and Petri Toiviainen. 2014. Hunting for the beat in the body: on period and phase locking in music-induced movement. *Frontiers in human neuroscience* 8 (2014), 903.

[10] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li. 2020. EarphoneTrack: involving earphones into the ecosystem of acoustic motion tracking. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 95–108.

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

[12] Davide Figo, Pedro C Diniz, Diogo R Ferreira, and Joao MP Cardoso. 2010. Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing* 14, 7 (2010), 645–662.

[13] G David Forney. 1973. The viterbi algorithm. *Proc. IEEE* 61, 3 (1973), 268–278.

[14] Rolf Inge Godøy and Marc Leman. 2010. *Musical gestures: Sound, movement, and meaning*. Routledge.

[15] Victor E Gonzalez-Sanchez, Agata Zelechowska, and Alexander Refsum Jensenius. 2018. Correspondences between music and involuntary human micromotion during standstill. *Frontiers in psychology* 9 (2018), 1382.

[16] Susan Hallam, Ian Cross, and Michael Thaut. 2016. *The Oxford Handbook of Music Psychology (2nd ed.)*. Oxford University Press.

[17] Yan He, George Chernyshov, Jiawen Han, Dingding Zheng, Ragnar Thomsen, Danny Hynds, Muyu Liu, Yuehui Yang, Yulan Ju, Yun Suen Pai, et al. 2022. Frisson Waves: Exploring Automatic Detection, Triggering and Sharing of Aesthetic Chills in Music Performances. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–23.

[18] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software* 5, 50 (2020), 2154. https://doi.org/10.21105/joss.02154 Deezer Research.

[19] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://arxiv.org/abs/1609.09430

[20] Nan Jiang, Terence Sim, and Jun Han. 2022. EarWalk: towards walking posture identification using earables. In *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications*. 35–40.

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

[22] Prerna Khanna, Tanmay Srivastava, Shijia Pan, Shubham Jain, and Phuc Nguyen. 2021. JawSense: recognizing unvoiced sound using a low-cost ear-worn system. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 44–49.

[23] Rushil Khurana, Karan Ahuja, Zac Yu, Jennifer Mankoff, Chris Harrison, and Mayank Goel. 2018. GymCam: Detecting, recognizing and tracking simultaneous exercises in unconstrained scenes. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–17.

[24] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. CREPE: A Convolutional Representation for Pitch Estimation. arXiv:1802.06182 [eess.AS]

[25] Stefan Koelsch, Thomas Fritz, D Yves v. Cramon, Karsten Müller, and Angela D Friederici. 2006. Investigating emotion with music: an fMRI study. *Human brain mapping* 27, 3 (2006), 239–250.

[26] Seungchul Lee, Chulhong Min, Alessandro Montanari, Akhil Mathur, Youngjae Chang, Junehwa Song, and Fahim Kawsar. 2019. Automatic Smile and Frown Recognition with Kinetic Earables. In *Proceedings of the 10th Augmented Human International Conference 2019*. 1–4.

[27] Daniel J Levitin, Jessica A Grahn, and Justin London. 2018. The psychology of music: Rhythm and movement. *Annual review of psychology* 69 (2018), 51–75.

[28] Geoff Luck, Suvi Saarikallio, Birgitta Burger, Marc R Thompson, and Petri Toiviainen. 2010. Effects of the Big Five and musical genre on music-induced movement. *Journal of Research in Personality* 44, 6 (2010), 714–720.

[29] Emily Lynar, Erin Cvejic, Emery Schubert, and Ute Vollmer-Conna. 2017. The joy of heartfelt music: An examination of emotional and physiological responses. *International Journal of Psychophysiology* 120 (2017), 118–125.

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

[31] Claudio Martella, Ekin Gedik, Laura Cabrera-Quiros, Gwenn Englebienne, and Hayley Hung. 2015. How was it? Exploiting smartphone sensing to measure implicit audience responses to live performances. In *Proceedings of the 23rd ACM international conference on Multimedia*. 201–210.

[32] 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

[33] Dan Morris, T Scott Saponas, Andrew Guillory, and Ilya Kelner. 2014. RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3225–3234.

[34] Shahriar Nirjon, Robert F Dickerson, Qiang Li, Philip Asare, John A Stankovic, Dezhi Hong, Ben Zhang, Xiaofan Jiang, Guobin Shen, and Feng Zhao. 2012. Musicalheart: A hearty way of listening to music. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. 43–56.

[35] Adrian C North, David J Hargreaves, and Jon J Hargreaves. 2004. Uses of music in everyday life. *Music perception* 22, 1 (2004), 41–77.

[36] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.

[37] Alisun Pawley and Daniel Müllensiefen. 2012. The science of singing along: A quantitative field study on sing-along behavior in the north of England. *Music Perception* 30, 2 (2012), 129–146.

[38] Charissa Ann Ronao and Sung-Bae Cho. 2016. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications* 59 (2016), 235–244.

[39] Burr Settles. 2009. Active learning literature survey. (2009).

[40] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* 28 (2015).

[41] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. 2021. ExpressEar: Sensing Fine-Grained Facial Expressions with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–28.

[42] Sergey Volokhin and Eugene Agichtein. 2018. Understanding music listening intents during daily activities with implications for contextual music recommendation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 313–316.

[43] Chen Wang and Pablo Cesar. 2017. The Play Is a Hit: But How Can You Tell?. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 336–347.

[44] Xinxi Wang, David Rosenblum, and Ye Wang. 2012. Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM international conference on Multimedia*. 99–108.

[45] Yi-Hsuan Yang and Yuan-Ching Teng. 2015. Quantitative study of music listening behavior in a smartphone context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 3 (2015), 1–30.