Poster Abstract: Automatic Recognition of Vocal Reactions in Music Listening using Smart Earbuds

Euihyeok Lee
KOREATECH, Republic of Korea
euihyeok.lee@misl.koreatech.ac.kr

Dongwoo Kim
KOREATECH, Republic of Korea
dongwoo.kim@misl.koreatech.ac.kr

Chulhong Min
Nokia Bell Labs, UK
chulhong.min@nokia-bell-labs.com

Seungwoo Kang*
KOREATECH, Republic of Korea
swkang@koreatech.ac.kr

ABSTRACT

We propose an in-ear sensing method that automatically detects vocal reactions that people often exhibit when listening to music. We observe what kind of vocal reactions are often brought during music listening and investigate the challenges of applying an existing representative acoustic classification model to vocal reaction recognition. We present our vocal reaction recognition method and the preliminary evaluation to assess its performance.

CCS CONCEPTS

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

KEYWORDS

Musing Listening, Vocal Reaction, Reaction Classification

ACM Reference Format:
Euihyeok Lee, Dongwoo Kim, Chulhong Min, and Seungwoo Kang. 2020. Poster Abstract: Automatic Recognition of Vocal Reactions in Music Listening using Smart Earbuds. In *The 18th ACM Conference on Embedded Networked Sensor Systems (SenSys '20), November 16–19, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3384419.3430466

1 INTRODUCTION

Music has always been together with us in human history and is undoubtedly one of the most important entertaining activities. The study [1] reports that, in 2019, people spent an average of 2.6 hours a day listening to music. Beyond entertainment, music has also been utilized as a medium to offer several benefits in various areas, ranging from therapy to psychology, well-being, and education.

Despite such great potential, understanding music listening experiences, i.e., how a person listens to and react to music, is still very limited in scalability, granularity, and instantaneousness. The current music players understand listeners' preferences based on simple statistics, e.g., number of plays and likes. Even in situations where deeper understanding of music listening is required, e.g., music therapy, they mostly rely on experts' observation.

We envision an in-ear sensing system that understands music listener experiences automatically and unobtrusively. We chose smart earbuds as a sensing device because most of reactions in listening to music are made around a head. In this work, as an initial attempt, we target three types of vocal reactions, *singing*, *humming*, and *whistling*, which listeners often make during music listening. We investigate a set of challenges in detecting spontaneous vocal reactions and propose a novel, two-step inference pipeline. First, we collect vocal audio data in 80 music listening sessions (from 10 participants and 8 sessions per each) and build a neural network that can accurately detect vocal reactions. Second, we smooth the output of the neural network by using a HMM model. The evaluation shows that our system achieves 0.88 of F_1 score on average for classifying the vocal reactions.

2 VOCAL REACTION SENSING

2.1 Target Vocal Reactions

To build a representative set of reaction vocabularies for music listening, we design a study to observe how people listen to music. We recruited 4 participants from a university campus and video-recorded how they listen to 10 songs.We then tag reactions they made and finally select three types of reactions, *singing (along)*, *humming*, and *whistling* that are observed most frequently.

2.2 Challenges in Vocal Reaction Recognition

A simple and straightforward way of detecting vocal reactions would be to use existing, pre-trained audio models. For example, Google recently released YAMNet [2], an acoustic classification model that is trained using more than 2 million YouTube videos and classifies 521 audio event classes, including our target reactions: singing, humming, and whistling. However, we found out that YAMNet is not suitable for vocal reaction detection for two reasons. First, while YAMNet has the *singing* label, the corresponding audio data is mostly taken from video clips where the song is played with instruments. However, when a listener sings along, the earbuds can only capture the listener's voice without the music being played. According to our experiments, YAMNet misclassifies such singing-along events as *speech.* Second, YAMNet has poor annotation granularity to detect vocal reactions in music listening.

*Corresponding author

SenSys '20, November 16–19, 2020, Virtual Event, Japan

ACM ISBN 978-1-4503-7590-0/20/11…$15.00
https://doi.org/10.1145/3384419.3430466

From our observation, vocal reactions often occur for a short time, e.g., several seconds, but YAMNet is trained with 10-second long audio clips.

2.3 Our Approach

To this end, we present a novel, two-step sensing technique for automatic recognition of vocal reactions using smart earbuds.

Step 1: classification. We train a new customized classification model for our purpose. To build the model, we target three classes, Singing/Humming, Whistling, and Others. Note that we combine singing and humming into the one class because the boundary between singing and humming is generally quite ambiguous from our observation of collected data. Also, to maintain robustness against external noise, we collect various ambient noises commonly occurring in the surroundings and use them for training. That is, the Others class represents non-reaction data including quiet cases and ambient noise cases. Audio data are resampled at 16 kHz and divided into 1 second segments. Then, the segmented raw audio data are converted into a log mel spectogram and the converted spectogram is used as an input to a neural network for training. We use the same network architecture as YAMNet employing MobileNets[3].

Step 2: smoothing. From our observation, we find that vocal reactions last for several seconds to several minutes. When it is relatively long, there are cases that the model misclassifies some segments of vocal reactions. These cases usually occur when the volume of reaction is low or when listeners are breathing in the middle of the reaction and there is no vocal reaction sound. To solve this problem, we apply the Hidden Markov Model (HMM). The output of the 1-second segment from the classifier is fed as an input to the trained HMM model, and it is adjusted probabilistically according to the previous state sequence. We use HMM to compensate for the misclassified part of the vocal-related reaction.

3 EVALUATION

3.1 Experimental Setup

Data collection: We collect data from 10 participants recruited from a university campus. Each participant is invited to a lab for data collection, which is a quiet room without much ambient noise. The participant is asked to listen to a set of songs while wearing earbuds. The songs that the participants listen to consist of 8 songs, 4 of which are songs that they already have listened, and 4 are unknown. For the former, we asked them to make a list of songs themselves and for the latter, they were asked to select the songs they had not listened before from a list of songs we made. Each of the 4 songs consists of two fast songs, e.g., dance or hip-hop, and two slow songs, e.g., ballad or R&B. For vocal reaction data collection, Apple Airpods is used. We record vocal data from Airpods using a mobile application we developed. Also, the participants were video-recorded using smartphones. These collected audio and video data are used for labeling of vocal reactions.

Performance metric: We use F_1 score as a performance metric. We apply leave-one-subject-out cross validation for evaluation with the data from 10 participants.

Comparison: To examine the effectiveness of our proposed method, we prepare a set of comparison: a pre-trained YAMNet model (YAMNet), our custom-trained YAMNet model (C-YAMNet), a pre-trained YAMNet model with HMM applied (YAMNet + HMM), and our custom-trained YAMNet model with HMM applied (C-YAMNet + HMM) that is our proposed method. Note that the pre-trained YAMNet model classifies most of the singing and humming segments from our vocal reaction data as 'Speech', so we consider 'Speech' output from YAMNet as 'Singing/Humming'.

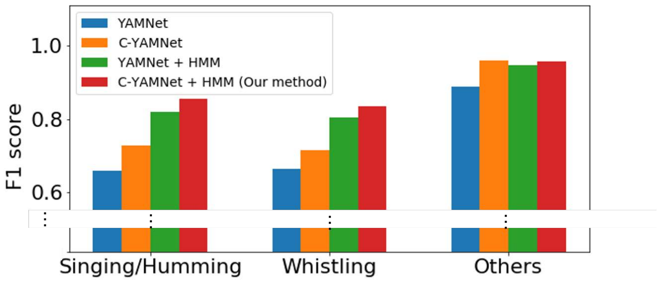

Figure 1: F_1 scores for each method for the three vocal reactions

3.2 Performance of Vocal Reaction Sensing

Overall, the average F_1 score of our method is 0.88. This is the highest score compared to other methods. As shown in Figure 1, the performance increases as we apply the proposed approach and as a result, our method shows the higher F_1 score than other methods for all classes. Our method shows 0.86 and 0.84 of F_1 score for the 'Singing/Humming' class and for the 'Whistling' class, respectively. In the case of 'Others' class, the F_1 score is the highest at 0.96 in C-YAMNet and our method. Comparing YAMNet and C-YAMNet, C-YAMNet shows the improved performance. Because YAMNet classifies inputs into 512 labels, there are more cases of misclassification compared to C-YAMNet. Also, we can see that HMM has a good effect on the performance.

4 CONCLUSION

We present a method to automatically detect the natural vocal reaction that occurs when listening to music. We collect vocal audio data in 80 music listening sessions and apply several techniques to classify the reactions and improve accuracy. We conduct a preliminary evaluation to measure the performance of our proposed method. We believe our results are a meaningful first step in understanding humans who listen to music through smart earbuds.

ACKNOWLEDGMENTS

This research was supported by National Research Foundation of Korea(NRF) and Institute of Information & Communications Technology Planning & Evaluation(IITP) funded by the Ministry of Science and ICT (2017M3C4A7066473, IITP-2020-0-01778)

REFERENCES

[1] [n.d.]. Music Listening 2019. https://www.ifpi.org/wp-content/uploads/2020/07/Music-Listening-2019-1.pdf. Accessed: September 30, 2020.

[2] [n.d.]. YAMNet. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet. Accessed: September 23, 2020.

[3] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).