CJ-002

Expression Recognition Based on Ear Canal Shape Detection Using Earbud and Ultrasound

高 天 † 董 学甫 ‡ 田谷 昭仁 § 西山 勇毅 ¶ 瀬崎 薫 ‖
Tian Gao† Xuefu Dong‡ Akihito Taya§ Yuuki Nishiyama¶ Kaoru Sezaki‖

Abstract

Earbuds are well-known for their portability, wearability, and frequent usage. This research introduces facial expression recognition using earbuds by training models with collected in-ear ultrasound data. The motivation is to enhance communication and, mainly, to help track emotional health. The strength of using earbuds for tracking is that earbuds' wearability leads to convenience and consistency, and their close position to facial muscles gives them the potential to be the best expression detectors. The possible application could have great value in the future.

1 Introduction

Every day, while the usage of phones, laptops, and smartwatches is exploited almost to its full potential, earbuds have not yet received enough attention, despite their portable size, high-frequency usage, and wearable features.

Earbuds, although currently simply being used as a tool to output sounds privately and directly to the user's ear canals, can do many more tasks across fields such as "interaction", "authentication", "physiological parameters", "movement and activity", according to [1] by Tobias Roddiger et al.

In this research, earbuds are used to recognize the user's expression considering their proximity to facial muscles and their better protection for privacy compared with cameras. The motivation of using earbuds for facial expression recognition is as follows: 1. Helping with mental health by tracking facial expressions. 2. Enhancing communication.

The basic mechanism is that, when the user wears the earbuds, a piece of ultrasound is played from the earbud, and the ultrasound bouncing back and forth in the ear canal can contain information about the ear canal shape. Using such ultrasonic data, facial expression recognition can be achieved. In this study, Mediapipe blendshape labels are used, which are extracted from videos of participants' faces, as shown in Fig.1 [2].

Similar studies, such as Eario also uses earbuds to do expression recognition. Despite researchers doing Eario emphasizing that their earbuds have limited requirements that can be easily found on some of the currently on-sell products, they still modified the earbuds in a relatively larger extent compared to this study [3]. In this study, the earbuds used are only slightly modified. The modification is so small that it signifies its feasibility to be adopted in daily usage, considering its compatibility with off-the-shelf products.

The main contributions of the study are as follows: 1. To the best knowledge, this study is the first one using off-the-shelf earbuds with ultrasonic

† 東京大学 The University of Tokyo
‡ 東京大学 The University of Tokyo
§ 東京大学 The University of Tokyo
¶ 東京大学 The University of Tokyo
‖ 東京大学 The University of Tokyo

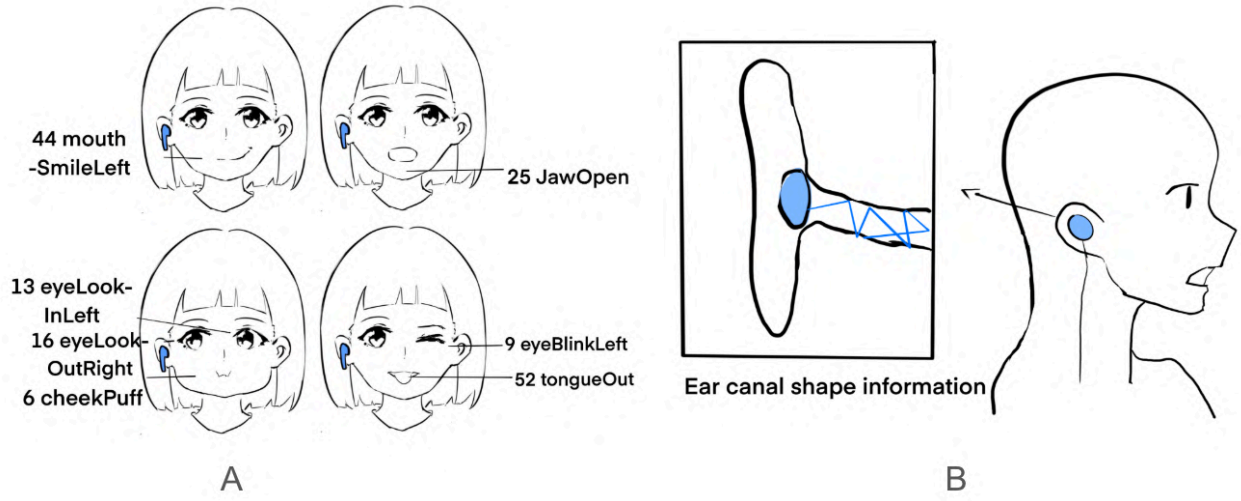

Figure1: (A)Participants make various expressions while wearing the earbuds(B)The visualization explaining how the ear canal shape information is collected through the earbud

audio to achieve facial expression recognition. 2. The result is promising and shows that this study is successful and worthy of further and deeper exploration.

In this paper, the background and related works will be introduced first. Then, the motivation of the research will be discussed. Afterward, feature and label extraction, modeling, and study design of the research will be explained. Then, experiments and results will be presented, followed by an evaluation, and a discussion. Finally, the article will end with a conclusion.

2 Related Work

Just as introduced above, earbuds can do tasks across multiple fields. Here, let's first take a look at the field of "movements" or "activities".

Earbuds can be used to track fitness, track expenditure of energy, detect posture, log consumptions, etc [1]. For example, in the article [4], a system, OESense, is used, utilizing the "occlusion effect", which helps improve the resistance to noise and enhances low-frequency signals. And the sensing and

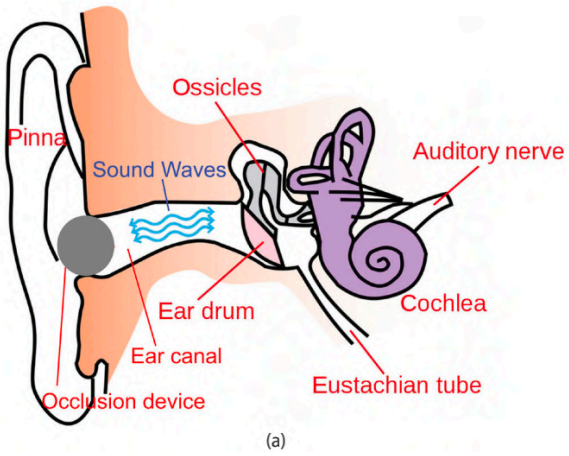

Figure2: The idea of occlusion effect. Cited from [4]

the sound delivery will not interfere one another due to their difference in frequency [4]. The basic concepts regarding occlusion effect are shown below in Fig.2 [4]. Also, another article [5] helps recognize snacking pattern.

In addition to movements or activities, people are also using earbuds to monitor health conditions. For example, in the paper [6], low-cost earbuds utilizing detections of OAE, a kind of sound generated by cochlea, has been proposed to lower the cost of

hearing screening.

Meanwhile, the research also falls under the category of facial expression. Related works mainly focus on using the image data for recognition. For example, in the article [7], Humayra A.'s research team has utilized 2 datasets: JAFFE and CK+, and both of them are image data.

Facial expression recognition using image data does have its own strengths. However, facial expression recognition using earbuds can also explore the expression in a different way: facial expression images can't identify the muscles' shape change at such a close distance and detailed way. Earbuds' detection, in adverse, can provide profound information regarding muscle movement. In addition to it, facial expressions can lead to possible risks and concerns regarding privacy and security, because images containing human faces are captured for data collection and also for actual usage. Meanwhile, earbuds generate much less concern regarding privacy. Although this study utilized video image data of the participants for labeling training, in actual usage, there's no need for facial image/video capturing (as they are just labels). During actual usage, only earbuds are needed for users. Thus, this study has its own strength over image-feature-based facial expression recognition.

3 Motivation

The research, as stated above, uses earbuds to recognize users' expressions. The motivation for this facial expression recognition research is to:

1. track detected expressions to promote the mental health conditions of users.

2. enhance hands-free remote communication by adding expression marks or labels to a conversation according to the user's expression (users can always choose when to utilize this feature and when not), which should be a simple downstream work.

The main motivation is the first: helping improve users' mental health conditions by tracking facial expressions. While mental health has been emphasized in recent years, methods are provided to improve the public's mental health. Mental therapies are provided, students in school take courses on mental health, and parents start to learn the right ways to support their children. However, a lack of an effective method to stably track mental health conditions still exists. Individuals are encouraged to log their mental state on paper or applications. However, it is still time-consuming. Missing logging or giving up logging can constantly happen.

Thus, using earbuds to help with tracking mental states of individuals has been considered as a topic worth further research. Although not strictly related, facial expressions can usually indicate an individual's emotions to some degrees. For example, if the user expresses expressions related to negative emotions most of the time, it might indicate a higher possibility of bad mental conditions such as depression. If the user expresses "anger" frequently, it can also probably reflect other mental conditions.

Using earbuds to achieve such can have several merits: Unlike diaries, facial expression recognition automatically happens, and no additional operations are needed by users. It is the most efficient, convenient way for users. In addition, earbuds are considered as the most suitable device to detect expression, considering their wearability and positions near facial muscles. Moreover, although this work used video image data as the labels during model training, during actual usage, no video/image data is needed. Unlike facial expression detection using camera, the detection using earbuds can better protect a person's privacy to some extent, by only

Figure3: The device used in the study.

collecting the users' ear canal shape data. Finally, considering earbuds' existing heavy usage, earbuds detecting expressions can be devoted to actual use in a short time at a low cost. Individuals don't need to change their lifestyle or adapt to a new device.

4 Feature and Label Extraction, Modeling, and Study Design

4.1 Overview of Approach

In brief, the research follows this flow: First, the assumption is made that ear canal shape information collected by earbuds can be used to predict facial expressions. Then, the study design has been made. Afterward, the data are from participants. Then, the data are preprocessed. Features are attracted and stored. Models are selected and trained using feature data. Models are evaluated at the end. Some codes might be refered to [8] [9] [10] [11].

4.2 Audio Data Processing and Feature Extraction

With the audio data collected during the experiment (which will be explained later), the first step is to visualize it. The waveform is plotted for better visualization. Audio data is also trimmed and cut into snippets according to their corresponding expression.

During data feature extraction, first, 45 windows are divided for each sample. Then, features including ar, short-time energy, envelope, and zero-crossing rate are extracted and combined.

The AR feature has the merit of both expression starting precise time irrelevant and sensitive to ear canal dynamic motion. To be specific, it uses the previous several samples to predict the next sample, which can enhance the robustness, because even if the expression starts at a slightly different time, the way of learning from previous samples can help against potential influences as it considers the whole window instead of single small sections; in addition, the AR feature can also lessen the impact of "re-wears" of earbuds; what's more, AR features can be extracted in a comparatively fast way. It is an innovative attempt to use the AR feature for facial expression recognition [12]. The AR feature used here has a lag of 200.

The zero-crossing rate feature has the merit of capturing any tiny and rapid movements. The short-time energy has the benefit of directly giving the "strength" of movement, as the more the ear canal changes, the more the short-time energy changes. The amplitude envelope feature has the merit of being stable and smoothing the transitions of sound, which can encounter the impact generated by small, fast shakings.

One example of processing one single snippet is shown in Fig.4, and Fig. 6.

The output of the extracted feature data of snippets is stored in a CSV file. Each row represents one data point. The First column contains the string representing the expression, while the rest of the columns represent the features mentioned above for each data point.

4.3 Label Extraction as Mediapipe Blendshapes

When coming to the label, the Mediapipe Blendshapes are chosen for quantifying the facial expres-

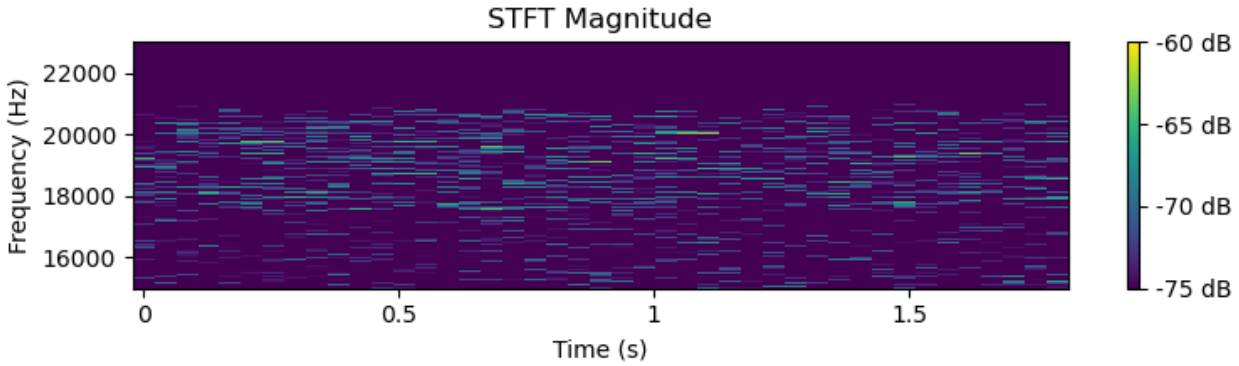

Figure4: The stft of a single expression sample

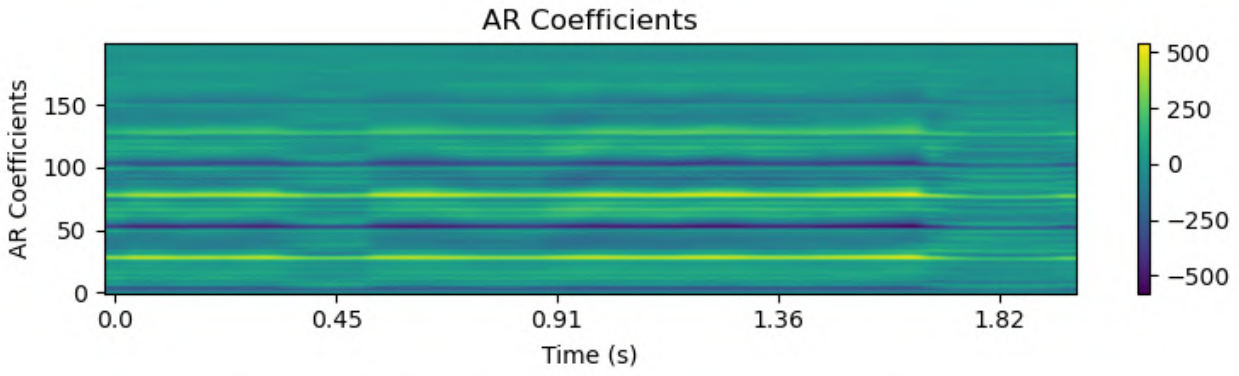

Figure5: The AR of a single expression sample

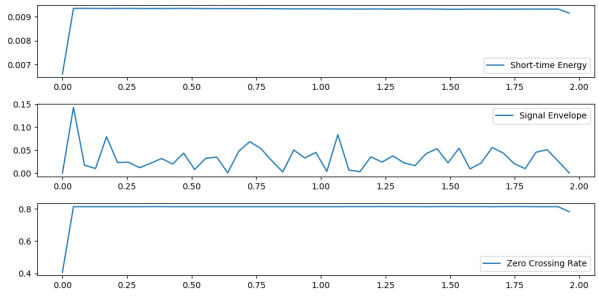

Figure6: Short-time energy, envelope, zero-crossing rate of a single expression sample

sions [2]. Mediapipe uses blendshapes as a type of reference, which are a set of data indicating critical facial movements. Shown in Fig.1A. Using Mediapipe blendshapes as labels has several merits. First, it contains much more detailed and richer measurements of facial expressions comparing to traditional labels; in addition, it is also compact comparing to other similar labels such as landmarks, as Mediapipe blendshape only contain 52 parameters per image(frame) [3].

4.4 Modeling Process

CNN-RNN model is applied with the CNN embedding extractor and 2 layers of GRU. Resnet 18 is applied to process each window separately and independently, while GRU here is used to process all windows across the temporal dimension. The network is built based on an implementation of [8] with personal additional modifications.

4.5 Device and Participants/Subjects

The device used is shown in Fig.3. Participants are young-age individuals, including one female and one male. 3 turns of experiments are conducted for

each participant, while each turn contains 8 representative types of expressions: ”happy”, ”sad”, ”angry”, ”calm”, ”boring”, ”neutral”, ”satisfied”, ”surprised”. These 8 types are repeated, forming 160 expressions in total.

4.6 Data Collection and Procedures

An ultrasound audio is prepared beforehand, and it would be played through the earbuds during data collection. Considering that some people can hear ultrasounds, the volume is adjusted to a level at which most of the participants can barely hear it. The sound volume was kept the same for all participants.

The right ear canal for the whole experiment is used, so participants only need to wear the right side earbud. During the experiment, the participants' ear canal shape information will be recorded. Basic ideas are shown in Fig.1.

During data collection, it is decided to set one turn with 160 expressions and collect 3 turns of data from each participant; each expression should have a duration of 2 seconds(asking the participant to hold each expression for 2 seconds). Adding some buffer time, each turn will last = 160*(2+2) seconds = 640 seconds = around 10 minutes. 3 turns will take around 30 minutes. Meanwhile, videos of participants' faces are also recorded, which are used as the labels during model training.

5 Evaluation

As noted in [13], in Mediapipe, there are 52 blendshape parameters in total, indicating different places of the face, where the first 22 parameters indicate the upper half of the face more, while the rest 30 parameters indicate the lower half of the face more. The result of the facial expression recognition yields the following results shown on the graph Fig.7. The upper MAE mainly reflects the MAE

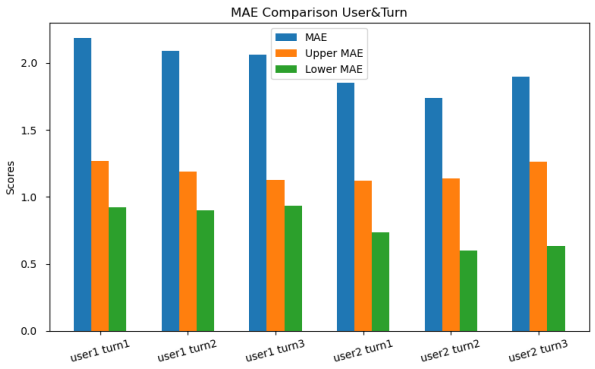

Figure7: Result of the facial expression recognition model

of the upper face blendshape parameters(the first 22 parameters), while the lower MAE reflects the MAE of blendshape parameters that indicate the lower half of the face (the remaining 30 parameters). A cross validation is conducted by excluding one turn out as the test dataset for each user. For example, the training dataset for user1 turn1 is all data of user1 except turn1. Comparing each user's own overall MAE, upper MAE, and lower MAE across their own 3 turns, it is shown that for the same user, the overall MAE does not change much. So it means that the prediction model is stable across turns for the same user.

Comparing 2 users' overall MAE, upper MAE, and lower MAE, it is easy to notice some slight changes between different users. But still, the difference is not too large, so it can still signify the stability of this prediction model across different individuals.

Comparing the upper MAE and the lower MAE, it seems that the upper MAE is larger than the lower MAE, meaning the lower half of face recognition performs better than the upper part.

6 Discussion

6.1 Artificial Expression

The expressions are not natural emotional expressions: they are made by participants following the instruction of “making certain expression”. This might affect its accuracy in real usage, considering that future applications will be used to detect expressions made in a natural context.

6.2 Valence-arousal

Quantification for each expression was tried by deciding the Valence and Arousal: Valence indicates emotions’ positivity/negativity. Arousal indicates the ’strength’ of emotions. For example, valence and arousal are both set with ranges [-1, 1]. Larger valence is defined as more positive emotions, and larger arousal as indicating more strong emotions. For specific examples, “happy” is set to a valence of 0.8 while ”sad” is set to a valence of -0.8. “Calm” is set to an arousal of -0.7 while ”surprised” is set to an arousal of 0.9. The basic concept is shown in Fig.8. However, this quantification is not used in the final, considering Mediapipe Blendshapes offer a more detailed reference for facial expression measurement. But still, such a reference can also be used in future work when a video reference is unavailable.

6.3 Additional Sensors and Data

Adding additional sensors to detect additional biological parameters, such as blood pressure, can probably help produce more accurate results: biological parameters can also help with the identification of emotions. And it is achievable using earbud devices. In addition, the current result is generated from a small dataset. In future work, additional data can be collected.

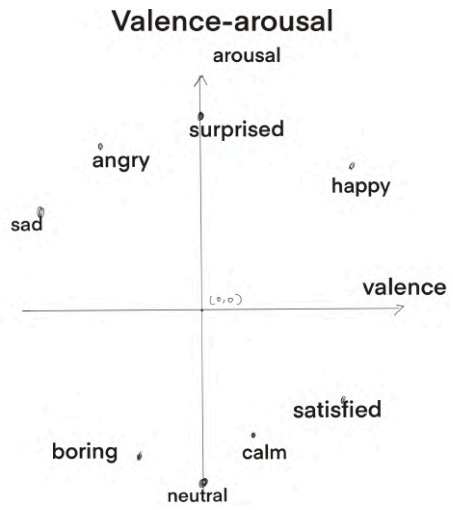

Figure8: The basic idea of valence-arousal.

6.4 Refine Data Processing and Models

Data processing can still be improved. More models can be attempted and compared to determine the best model for this expression recognition goal.

7 Conclusion

In conclusion, in this research, using earbuds with ultrasonic audio for facial expression recognition has been proposed and achieved. Focusing on facial expression recognition serves two purposes: helping to recognize mental health conditions and enhancing communication. As for the dataset, audio features are extracted from the ultrasonic audio data collected, while the labels are extracted as Mediapipe blendshapes from recorded videos of participants’ faces. The data are collected from participants wearing earbuds and performing expressions for 3 turns, where each turn contains 160 expressions. The model of modified Resnet with GRU are selected, with outputs containing rich information, guaranteeing a comprehensive prediction of users’ facial expression. Evaluations are made, which show great performance. The work of earbuds facial expression recognition can be easily

applied to existing earbuds products, rendering this research a profound potential for future application and commercial value.

8 Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP25K15350.

References

[1] T. Röddiger, C. Clarke, P. Breitling, T. Schneegans, H. Zhao, H. Gellersen, and M. Beigl, "Sensing with earables: A systematic literature review and taxonomy of phenomena," Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, vol.6, no.3, pp.1–57, 2022.

[2] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.L. Chang, M. Yong, J. Lee, *et al.*, "Mediapipe: A framework for perceiving and processing reality," Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR), 2019.

[3] K. Li, R. Zhang, B. Liang, F. Guimbretière, and C. Zhang, "Eario: A low-power acoustic sensing earable for continuously tracking detailed facial movements," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol.6, no.2, July 2022.

[4] D. Ma, A. Ferlini, and C. Mascolo, "Innovative human motion sensing with earbuds," GetMobile: Mobile Comp. and Comm., vol.25, no.4, p.24 – 29, mar 2022.

[5] M. Bin Morshed, H.K. Haresamudram, D. Bandaru, G.D. Abowd, and T. Ploetz, "A personalized approach for developing a snacking detection system using earbuds in a semi-naturalistic setting," Proceedings of the 2022 ACM International Symposium on Wearable Computers, ISWC '22, New York, NY, USA, p.11 – 16, Association for Computing Machinery, 2022.

[6] J. Chan, A. Glenn, M. Itani, L.R. Mancl, E. Gallagher, R. Bly, S. Patel, and S. Gollakota, "Wireless earbuds for low-cost hearing screening," Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services, MobiSys '23, New York, NY, USA, p.84 – 95, Association for Computing Machinery, 2023.

[7] H.B. Ali and D.M.W. Powers, "Multi-feature fusion based non negative matrix factorization: Facial expression recognition from imaging sensors," Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, MLSDA'14, New York, NY, USA, p.25 – 32, Association for Computing Machinery, 2014.

[8] elbuco1, "Elbuco1/cbam: Cbam: Convolutional block attention module for cifar10 on resnet backbone with pytorch."

[9] Gordoni, "Gordoni/aiplanner: Aiplanner is an machine learning based asset allocation and consumption planning calculator. included are sources to two other similar calculators as well as a spia pricing calculator.."

[10] aks2203, "Aks2203/easy-to-hard: Official repository for the paper "can you learn an algorithm? generalizing from easy to hard problems with recurrent networks"."

[11] gsin4455, "Gsin4455/pytorch_mod: Pytorch implementation of modulation classification."

[12] X. Dong, Y. Chen, Y. Nishiyama, K. Sezaki, Y. Wang, K. Christofferson, and A. Mariakakis, "Rehearsse: Recognizing hidden-in-the-ear silently spelled expressions," Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, New York, NY, USA, Association for Computing Machinery, 2024.

[13] I. Grishchenko, G. Yan, A. Zanr, and E.G. Bazavan, "Mediapipe blendshape v2," 2022.