FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Commodity Headphones

Xingzhe Song
University of Pittsburgh
USA
x.song@pitt.edu

Kai Huang
University of Pittsburgh
USA
k.huang@pitt.edu

Wei Gao
University of Pittsburgh
USA
weigao@pitt.edu

ABSTRACT

Facial expressions are important indicators of user needs that can be used in many interactive computing applications to adapt the system behaviors and settings. Current computing approaches to recognizing human facial expressions, however, either rely on continuous camera recordings that are energy consuming, or require custom sensing hardware that are expensive and difficult to use on commodity systems. In this paper, we present *FaceListener*, a new sensing system that recognizes human facial expressions by only using commodity headphones. The basic idea of FaceListener is to transform the commodity headphone into an acoustic sensing device, which captures the face skin deformations caused by facial muscle movements with different facial expressions. To ensure the recognition accuracy, FaceListener leverages the knowledge distillation technique to learn the subtle correlation between face skin deformation and the acoustic signal changes. Experiment results over multiple human beings demonstrate that FaceListener can accurately recognize more than 80% of different facial expressions. FaceListener is highly energy efficient, and can well adapt to different headphone models, host systems and user activities.

KEYWORDS

Human Facial Expressions, Acoustic Sensing, Headphone, Face Skin Deformation, Knowledge Distillation.

1 INTRODUCTION

In human communication and interaction, facial expressions serve as natural and effortless signals to convey ourselves besides verbal languages [23]. In many interactive computing applications such as virtual reality [14], cognitive robotics [26] and customer analytics [16], being able to timely and precisely recognize such facial expressions will be very useful for the computing system to capture the current user need and adapt the system behaviors and settings accordingly.

Most of current computing approaches to recognizing human facial expressions use images of human faces as the input [17], and can accept such images being taken from either front [25] or side [5] of the user. However, their accuracy and reliability significantly drop when the ambient light condition degrades [5, 6]. Continuous camera use will also incur very high power consumption that is not affordable on most battery-powered mobile devices.

Instead, some recent research seeks to exploit the correlation between facial expressions and facial muscle movements, which can be detected via either surface electromyography (EMG) [12, 28], on-head strain sensor [21] and in-ear air pressure sensor [2]. Facial expressions have also been considered as the outcome of human emotions, which can be monitored from different physiological signals [11, 18, 43] and human activities [27, 42]. However, all these techniques require custom sensing hardware, which are expensive and difficult to be integrated into commodity systems in practical use. The use of such custom hardware, on the other hand, also make these systems highly sensitive to the random variations of the human body, user mobility, and the surrounding environment. For example, sensing electrodes may be misplaced due to human body movements and hence result in substantial recognition errors [13, 33]. Besides, other existing approaches utilize the Inertial Measurement Unit (IMU) sensors on modern wireless earphones to recognize various human expressions from facial muscle movements. These techniques achieve fine-grained recognition on multiple facial expressions [8, 20, 36], but are only implemented on custom prototypes [19] rather than commodity devices. Given that IMU sensors are not always available on commodity off-the-shelf earphones, these existing approaches may not be used as a generic solution in practical application settings.

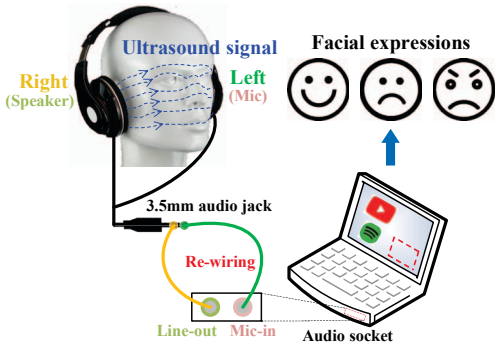


Figure 1: FaceListener recognizes facial expressions from the acoustic signal transmitted above the face skin surface.

In this paper, we aim to address the aforementioned limitations and instead propose *FaceListener*, a new sensing system that achieves precise and reliable recognition of human facial expressions by only using commodity headphones. As shown in Figure 1, FaceListener transforms the commodity headphone into an acoustic sensing device through simple audio jack re-wiring, which changes one of the headphone speakers into a microphone. With the inaudible ultrasound signal being transmitted from the headphone's speaker to microphone, our system design builds on the fact that different facial expressions correspond to different skin deformations on the human face surface, which significantly affect the propagation of ultrasound signal above skin surface. This correlation, then, allows us to precisely recognize human facial expressions from

978-1-6654-9624-7/22/$31.00 ©2022 IEEE
DOI 10.1109/IPSN54338.2022.00019

the received ultrasound signal. Furthermore, since headphones are usually attached to the user's ear, they are less likely to move and hence provide better reliability of continuous recognition over long time.

The major challenge of such recognition, however, is that different shapes of the face surface can only produce subtle changes on the propagated ultrasound signal, but the acoustic signal received by the microphone will contain the sound being produced by any other sources nearby (e.g., human voice, body vibrations, etc). To make sure that FaceListener can precisely capture the signal change caused by different facial expressions, our solution is two-fold. First, we consider the ultrasound signal's propagation above the face skin surface as an acoustic channel, and use the channel estimator as the sensing output to ensure that even the smallest change of the face surface can be captured.

Second, to ensure precise recognition of human facial expressions, one intuitive approach is to use a neural network to learn the correlation between acoustic sensory data and facial expressions. However in practice, the accuracy of such recognition may be impaired by the possible variation and uncertainty in each type of facial expression, whose impact could be largely amplified over the received ultrasound signal. Instead, our approach is to leverage the existing knowledge distillation technique, more specifically, the teacher-student learning framework [22, 39]. More specifically, during the offline training, extra facial images are taken at the same time with acoustic sensing for each facial expression, and are used to train the teacher network with higher confidence. The teacher network will then be used to supervise the training of student network that takes acoustic sensory data as input, by transferring the latent knowledge learned from facial images and using such knowledge to eliminate the ambiguity and uncertainty in the acoustic sensory data. Afterwards, when being used online, FaceListener only uses the trained student network to recognize facial expressions, from the acoustic sensory data.

FaceListener intends to provide an highly accessible solution to recognizing facial expressions in casual settings so that any low-cost commodity wired headphones can be turned into a sensor that enables facial expression related applications (e.g., in-home mental well-being tracking). Since the rewired headphones can still play sounds using one speaker, the recognition of facial expression can be operated at the same time with sound playback, with minor degradation of user experience, especially in casual application scenarios where the users are not sensitive to the audio quality.

To our best knowledge, FaceListener is the first system that recognizes human facial expressions using only commodity headphones. We implemented FaceListener over headphones that attach to human ears in different ways (e.g., on-ear, in-ear and over-ear), and evaluated its recognition accuracy over 5 student volunteers in various conditions, including lab-controlled settings and real-world scenarios. From our experiment results, we have the following conclusions:

- FaceListener is *accessible*. It can be implemented on widely available low-cost headphones without dedicated sensors. Users could capture their facial expressions and enable new applications by simply wearing a off-the-shelf headphone.
- FaceListener is *accurate*. It can achieve >80% accuracy when recognizing 7 main categories of facial expressions (neutral, joy, sad, surprise, anger, fear and disgust) over commodity headphones.
- FaceListener is *adaptive*. It is able to retain the recognition accuracy when being used over different headphone models, host systems (desktop PCs, laptops and smartphones) and ultrasound frequency bands (15kHz-23kHz). It can also well adapt to different human activities, body skin conditions or even head accessories that the users may wear.
- FaceListener is *lightweight*. On a commodity smartphone, it only takes 135ms to recognize one facial expression, and consumes <20% of smartphone battery after 3 hours of continuous use.

2 BACKGROUND & MOTIVATION

In this section, to motivate our design of FaceListener, we first demonstrate how facial expressions relate to the skin deformation caused by facial muscle movements. Afterwards, we discuss possible pathways of ultrasound signal propagation between the headphone's speaker and microphone, and demonstrate that the airborne propagation above the skin surface is strong enough for FaceListener's acoustic sensing.

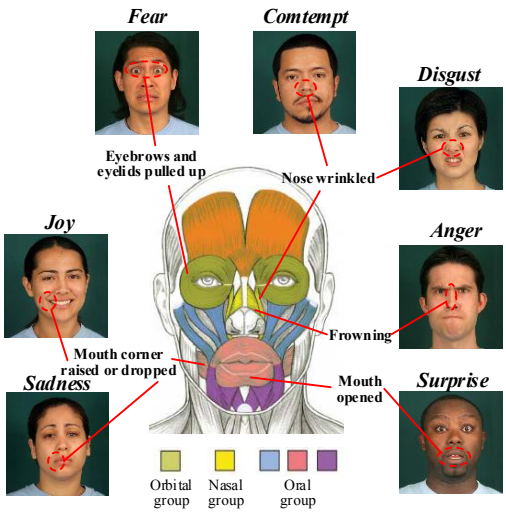


Figure 2: Facial expressions are created by different movements of facial muscles.

2.1 Facial Expressions and Facial Muscle Movements

Facial expressions are induced by the contraction of different skeletal muscles laying underneath the facial skin. Unlike other muscles on human body, facial muscles directly insert into the skin and hence their contractions produce much more evident skin deformations. As shown in Figure 2 which lists the 7 main types of human facial expressions, facial muscles are commonly divided into 3 categories: orbital, nasal and oral, which collectively produce different facial expressions [41]. For example, the orbital muscles contribute

to the expressions of sadness and fear by changing the shapes of mouth corner, eyebrows and eyelids. Nasal muscles control the movements of nose and the skin around it. They are the main contributors of disgust and contempt expressions, but also contribute to producing anger and sadness expressions. Oral muscles decide the shape of mouth and lips, from where expressions of joy, sadness and surprise are being produced.

In practice, since every facial muscle is involved in producing multiple types of facial expressions, we cannot recognize any specific facial expression by detecting the movements of individual facial muscles. This ambiguity, instead, motivates or design of FaceListener, in which the facial skin deformation as a cumulative effect of all the facial muscle movements are reflected and sensed as the variation of ultrasound signal being propagated above the skin surface.

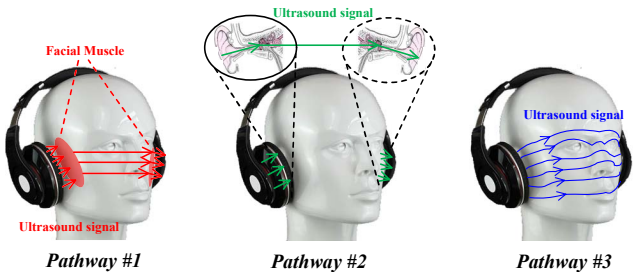


Figure 3: Three possible pathways of ultrasound signal propagation

2.2 Acoustic Signal Propagation

In FaceListener, a commodity headphone transmits ultrasound signal from one of its speakers to another that is transformed into a microphone by audio jack re-wiring (details in Section 3.1). In this case, the ultrasound signal could be possibly propagated along three different pathways as shown in Figure 3 and listed below:

(1) The signal propagates through the facial skin tissue.
(2) The signal propagates first into the ear canal of the speaker side, and then reaches the other ear canal of the microphone side through the human head.
(3) The signal propagates through the airborne pathway above the skin surface.

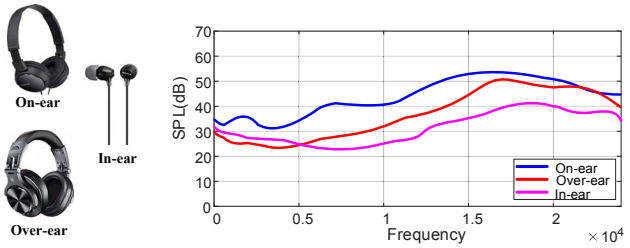
In the first pathway, the ultrasound signal will be greatly attenuated whenever it passes through the boundary between air and human tissue, in which the acoustic impedance decides such attenuation. As listed in Table 1, since the impedance in air is much smaller than that in human tissue, a significant amount of signal energy (>99%) will be absorbed by the human tissue itself and the remainder will hence have negligible impact on the microphone's received signal. Similarly, in the second pathway through the ear canals, the majority of signal energy will be absorbed by the ear drum that captures the air vibration.

To this end, we consider that the majority of signal being received at the transformed microphone is transmitted through the airborne pathway above the facial skin surface. In practice, when the headphone is being worn by the user, such propagated signal is

Media	Density (kg/m^3)	Velocity (m/s)	Impedance ($10^6 kg/m^2/s$)
Air	1.16	344	0.0004
Water	998	1482	1.48
Muscle	1041	1580	1.65
Tissue	928	1430	1.33

Table 1: Acoustic impedance of different media

the outcome of the headphone sound leakage, which widely exists on almost every commodity headphone [29]. To verify such leakage, we conducted preliminary experiments on different types of headphone models[1] shown in Figure 4(a), by transmitting a sweeping sine wave from 20 Hz to 24 kHz with the maximum volume. Results in Figure 4b show that these commodity headphones are able to greatly suppress signal leakage within the audible band, but produce much higher signal leakage in the ultrasound band. In particular, on all the headphone models, Figure 4(b) shows that the sound pressure level (SPL) above 15 kHz is always above 30 dB and at least 15 dB higher than that in the audible band. These results verify that the ultrasound signal transmitted by commodity headphones, when being received by the transformed microphone on the headphone, is strong enough for precise recognition of human facial expressions.



(a) Headphones **(b) Strength of leaked signal**

Figure 4: Measurements of acoustic signal leakage on commodity headphones

3 SYSTEM DESIGN

As shown in Figure 5, FaceListener recognizes facial expressions when a wired headphone is worn regularly by the user, in a way that one of the headphone speakers is transformed into a microphone. From the received ultrasound signal, we estimate the acoustic channel between the headphone's speaker and transformed microphone, and use such channel estimation to recognize facial expressions.

To ensure accurate recognition of facial expressions, FaceListener uses knowledge distillation to train the neural network for recognition through a teacher-student learning framework. During the training phase, a camera is used in front of the user face to capture the facial image corresponding to each facial expression being made. These images will then be used to train the teacher

[1]On-ear: Sony MDR-ZX110, Over-ear: Audio-Technica ATH-M30x, In-ear: Samsung EO-EG920BW.

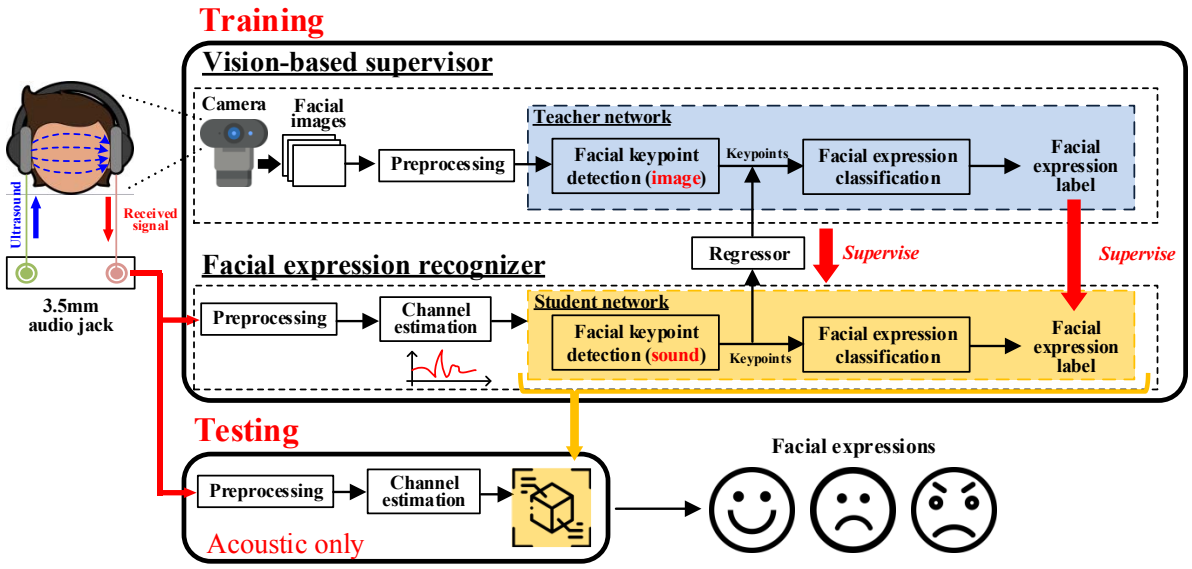


Figure 5: Overview of FaceListener system design

network. At the same time, the acoustic sensory data obtained for each facial expression will be used to train the student network, which is being supervised by the teacher network through both the facial expression labels and individual facial keypoints being detected. Afterwards, during the testing phase, only the student network will be used for online recognition of facial expressions.

3.1 Re-Wiring over Commodity Headphones

Speakers and microphones are devices that generate or capture propagated air vibrations as pressure waves. As shown in Figure 6(a), their physical structures and mechanisms share much in common. On the speaker, a diaphragm is moved by electromagnetic induction to generate such vibrations. Reversely, a microphone uses a diaphragm to capture air vibrations and then converts such vibrations into electrical signal via the magnet and coil.

Based on such similarity, the hardware of a commodity speaker can be used as a microphone, by simply using the speaker's output interface as input. In practice, existing work showed that the 3.5mm socket of audio output can be converted to input via software, because any audio port of the sound processing hardware (e.g., the PC sound card) is designed to be configurable as either output or input [31]. However, this software approach can only configure both the left and right audio channels of a headphone as output or input at the same time, and hence cannot be used to transform a commodity headphone into an acoustic sensing device that consists of both speaker (output) and microphone (input).

Instead, in FaceListener, we transform one speaker of the headphone into a microphone but retain another speaker unchanged, by re-wiring the audio jack to split the left and right audio channels as shown in Figure 6(b). On a tip/ring/sleeve (TRS) or tip/ring/ring/sleeve (TRRS) audio jack, we separate its tip and ring into two separate audio jacks, which are plugged into Line-out and Mic-in ports of the host system, separately. In this way, the sound will be played through the right speaker and then recorded by the left speaker[2].

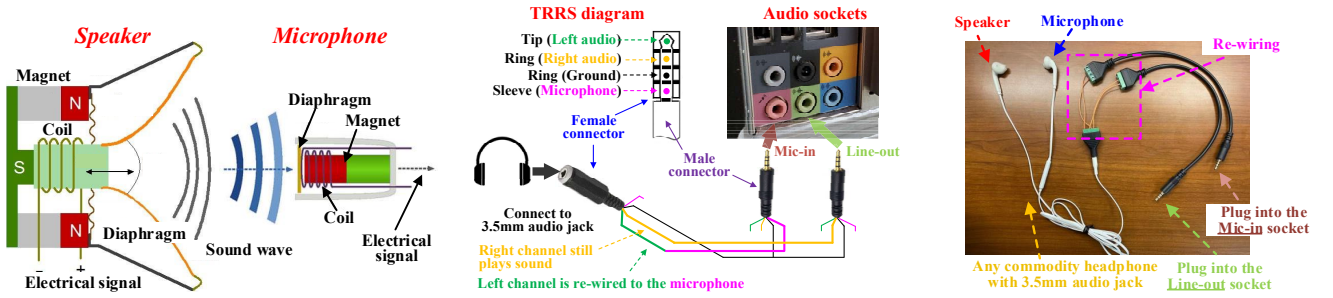
In practice, unlike existing work that requires a custom IC board [7], our approach can be easily implemented with 3.5mm male and female TRRS balun connectors, which are shown in Figure 6(c) and widely available on market with <$5 cost. On other systems with an 2-in-1 audio socket (e.g., most laptops and smartphones), an audio jack splitter can be used instead to implement such re-wiring.

3.2 Acoustic Channel Estimation

Since a speaker usually has a larger diaphragm than that on a microphone, when being transformed into a microphone, it will produce less vibrations from the incoming pressure waves, and may hence record sounds with a lower volume. The speaker's lack of noise filters and amplifiers for the incoming signal could further reduce the quality of the recorded sound. We experimentally investigated the quality of sound being recorded by such a transformed microphone, and Figure 7(a) shows that, when recording the same sound, the transformed microphone produces smaller audio volume and the received acoustic signal hence has a lower SNR.

To overcome this difficulty, in FaceListener we transmit a single-tone ultrasound wave that carries a coded sequence of 60 bits being modulated by BPSK. When the headphone is being worn by the user and the sound is being recorded by the transformed microphone, 7(b) shows that the bit sequence can always be correctly detected via autocorrelation. This detectability, then, allows us to use the received bit sequence to calculate the estimation of the acoustic channel from the headphone's speaker to transformed microphone. More specifically, received signals first go through a high-pass filter with cut-off frequency at 15kHz. This is to remove unnecessary signal components (e.g., audible sound) that may affect the estimation results. Next, we compute the correlation

[2]Similarly, the audio jacks can be re-wired in a reverse way to propagate sound from the left speaker to the right speaker.

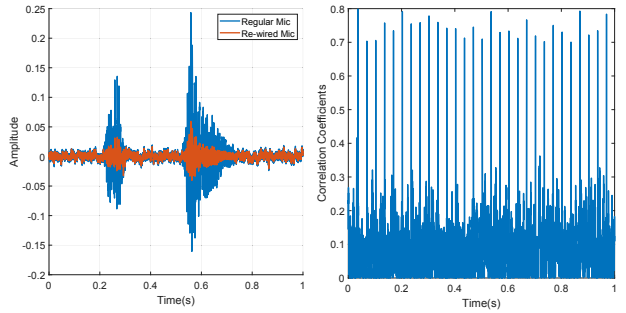


(a) Physics of speaker and microphone

(b) Re-wiring over the 3.5mm audio jack

(c) Implementation using TRRS balun connectors

Figure 6: Transforming a commodity headphone speaker into a microphone via audio jack re-wiring



(a) Raw signal recorded

(b) High correlation coefficients on the code sequence

Figure 7: Quality of sound recorded by the transformed microphone, when being worn by a user

coefficients between the received signal in a sliding window with the pre-known transmitted signal. A peak of such coefficients indicates that the microphone receives the coded sequence successfully. After demodulating the received sequence Y with the transmitted sequence X, the channel can be estimated as $H = YX^T(XX^T)^{-1}$. In practical systems, to achieve better estimation accuracy, we simultaneously transmit different bit sequences at multiple frequency bands. The corresponding channel estimations at these frequency bands, then, are collectively used as the raw features by the student neural network for recognition of facial expressions.

3.3 Vision-based Teacher Network Design

With these acoustic channel estimations, an intuitive approach to recognizing facial expressions is to apply these channel estimations and the corresponding facial expression labels to train a neural network. For example, a long short-term memory (LSTM) classification model can be trained to learn from the time series data of channel estimations. However, Figure 8 shows that using such as simple classification model lead to very low recognition accuracy. The main reason of such low accuracy is that the correlation between the acoustic channel and facial expressions is subtle, and the acoustic channel is hence highly sensitive to the small changes of face skin deformation, which could be heterogeneous among different samples of the same type of facial expression. The capacity of a simple classification model, then, is insufficient to represent such subtle correlation.

On the other hand, existing work has demonstrated that using neural networks can precisely recognize humans' facial expressions from their facial images. When using images of front faces, the accuracy of recognition using a single deep neural network (DNN) can achieve 95% [15, 38]. Based on such high accuracy, in FaceListener we intend to transfer the latent knowledge about facial expressions being learned from such vision-based teacher network to the acoustic-based student network, to supervise the training of student network and hence increase its capacity. The key insight of this supervision structure is that, unlike images that carry intuitive information about the facial muscle variations, acoustic samples are less comprehensible and interpretable for a deep learning model to extract representative features. To ensure recognition accuracy, a deep learning model based on acoustic data may need more abstraction capacity. Instead of increasing the complexity of the neural network itself as the most straightforward but unreliable approach, a vision-based teacher network can reduce the required capacity for feature extraction at the student model, which then only needs to learn those representative features processed by the teacher model.

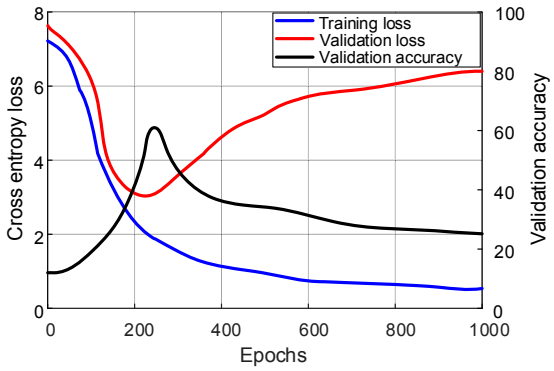


Figure 8: Accuracy of recognizing facial expressions using a simple LSTM classification model

Traditionally, such supervision is conducted by taking the teacher network's output as the target label for the student network during the training procedure. This method, however, is ineffective in FaceListener because many small changes on facial deformation may not be captured by the teacher network from facial images but could result in large variation in the acoustic channel. To address this challenge and ensure appropriate supervision, as shown in Figure 5, FaceListener divides both the teacher network and the student network into two sub-networks: the *first* sub-network detects the positions of a pre-defined set of facial keypoints (e.g., eyebrow ends, eye corners, nose and mouth corners, etc), and the *second* sub-network further recognizes the facial expression from the positions of these keypoints with respect to the human face contour.

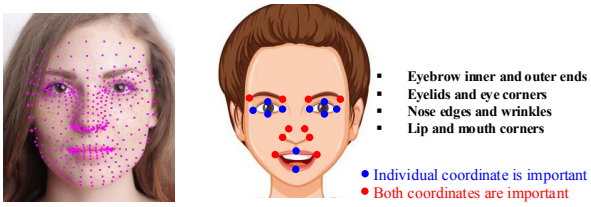
As a result, the teacher network can supervise the training of both sub-networks in the student network. In particular, the teacher network's learned knowledge on individual facial keypoints can be effectively transferred to the corresponding sub-network in the student network, hence ensuring that any small changes on facial deformation can be correspondingly mapped to the right keypoint.

3.3.1 Facial Keypoint Detection. In facial keypoint detection of the teacher network, we first preprocess each facial image by resizing and cropping it[3], to make sure that the coordinates of facial keypoints in all facial images are consistent. This consistency, then, allows us to universally compute the coordinates of facial keypoints in each image as their relative positions from the nasal tip in the image.

Afterwards, we need to decide the most appropriate set of keypoints being used, which should be the most important ones that represent facial expressions. For example, as shown in Section 2.1, the positions of nose and mouth corners are important indicators of many facial expressions including joy, sadness, anger, disgust, etc. We start from a well-established face mesh model that is developed by the Google MediaPipe project [9] and defines 468 facial keypoints as shown in Figure 9(a), and then select 30 keypoints with the highest importance in facial expression recognition to be used in FaceListener for timely online inference[4].

In FaceListener, such selection procedure is done by training a regular fully-connected neural network over the collected facial images. More specifically, we use Google MediaPipe to detect the 468 keypoints in each facial image and use their coordinates as the neural network input. The facial expression labels of the images are used as the neural network output. After that, we adopt the explainable AI technique [35] on this trained network to compute the integrated gradient (IG) of each keypoint, which quantitatively measures how this keypoint contributes to correct recognition of facial expressions. The top 30 keypoints with the highest contributions are then selected.

Our selected 30 keypoints, as shown in Figure 9(b), mainly cover the edges and corners of eyebrow, eye, nose and mouth, hence depicting the contours of these important elements on human face. Note that, the contributions of the x and y coordinates of these keypoints are separately measured, and hence some keypoints may only have one coordinate that is important to facial expression recognition. Based on such selection, Table 2 shows that, when being applied to the same set of facial images being used in training, these 30 keypoints result in similar accuracy of facial expression recognition, compared to that using all the 468 keypoints.



(a) 468 keypoints in MediaPipe **(b) 30 keypoints with the top contributions to facial expressions**

Figure 9: Keypoints on human faces

Expressions	N	J	Sa	Su	A	F	D
468 keypoints(%)	96.5	95.3	95.9	95.2	94.9	95.3	95.1
30 keypoints(%)	97.1	95.5	96.6	96.1	94.7	95.9	96.5

Table 2: Classification accuracy with different selections of facial keypoints (N=neutral, J=joy, Sa=sadness, Su=surprise, A=anger, F=fear, D=Disgust)

3.3.2 Facial Expression Classification. The second sub-network, as the classifier of facial expressions, is being trained with the coordinates of the 30 selected facial keypoints on each facial image as input and the corresponding facial expression label as the output. In particular, to ensure its reliability and accuracy, we first pre-train this network using a public dataset of facial images with known facial expression labels[5]. Afterwards, when being used on a specific user, this classifier will be further improved with this user's own facial images.

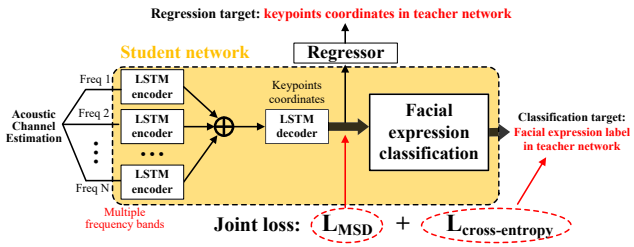


Figure 10: Student network design

[3]Such preprocessing can be done by simply using existing image processing tools such as the YOLOface model [32].

[4]In practice, different numbers of keypoints can be selected to balance bewteen the granularity and computing latency of facial expression recognition.

[5]We use the muxspace dataset: https://github.com/muxspace/facial_expressions, which contains 13,718 facial images with a resolution of 96x96. Images in this dataset will be first applied to the facial keypoint detection sub-network, and the detected facial keypoints are then used to train the facial expression classification network.

3.4 Acoustic-based Student Network Design

In order to facilitate the supervision from the teacher network, the student network in FaceListener is also designed to be consisting of two sub-networks: the first sub-network decides the positions of the 30 selected facial keypoints from the input data of acoustic channel estimation as described in Section 3.2, and then the second sub-network similarly recognizes facial expressions from the detected facial keypoints. Since acoustic channel estimations are produced as time series data, we use the LSTM model as the basic processing kernel for facial keypoint detection. When multiple acoustic channel estimations from different frequency bands are simultaneously used, they are being applied to separate LSTM encoders and then merged together.

As shown in Figure 10, the supervision on the student network is defined as a joint loss function that covers both facial keypoint detection and facial expression classification. The supervision on facial keypoint detection is designed as a regression task, which minimizes the mean square difference (MSD) between the teacher network's and student network's facial keypoint coordinates being detected. The supervision on facial expression classification, on the other hand, is done through the categorical cross entropy loss, which ensures that the student network produces similar results of facial expression recognition as the teacher network does.

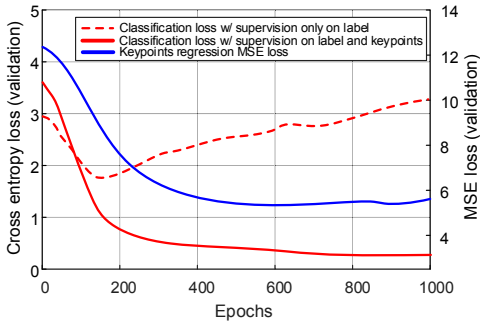


Figure 11: Validation comparison between different supervision methods

Figure 11 shows the performance of training the student network when being supervised by the teacher network. With the same neural network complexity and parameter, the training of student network cannot efficiently converge if it is only supervised by the teacher network through the facial expression label. On the other hand, the extra supervision through the regression of facial keypoints' coordinates ensures fast convergence of the student network training.

4 PERFORMANCE EVALUATION

We implement FaceListener using different types of commodity headphones, balun audio connectors and audio jack splitters, as shown in Figure 6(c), and transmit ultrasound signals in multiple frequency bands between 18 kHz and 21 kHz with an interval of 0.5 kHz. In our implementation, we allow FaceListener to monitor the user's facial expressions during normal uses of the headphones (e.g., playing music or watching videos). To allow simultaneous transmission of ultrasound signal along with audible sound playback in these cases, we use build-in audio mixers on desktop OSes such as Windows and Linux. For smartphone platforms, we realize such sound mixing using Android Audio Focus [10] or iOS duckOthers [4].

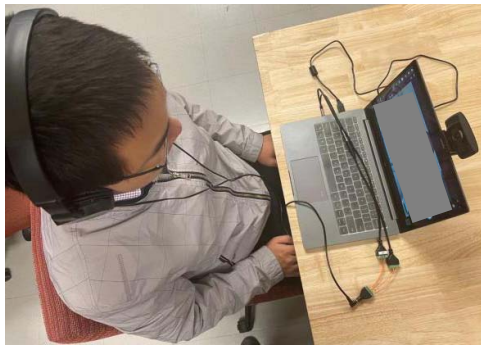
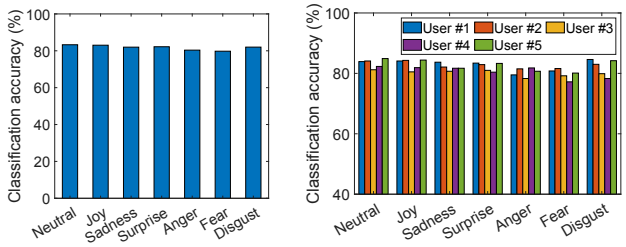


Figure 12: Evaluation setup

Based on this implementation, we evaluate the FaceListener's accuracy of recognizing the facial expressions over 5 student volunteers, in a 10m×10m lab office with regular furnitures. Due to the large variation among different human faces, the neural network models are separately trained and tested for each student volunteer, and the results are averaged over the 5 student volunteers. As shown in Figure 12, in our experiments, a volunteer sits in front of a camera and wears the re-wired headphone that connects to the host system. The volunteer then follows the host system screen's instructions to make facial expressions in 7 different categories (neutral, joy, sadness, surprise, anger, fear and disgust) and retain each facial expression for 5 seconds before switching to the next. Each experiment lasts for 10 minutes, during which the camera takes facial images at 30 FPS and acoustic sensing is conducted simultaneously. Since the acoustic signal has much higher sampling rate compared to the video frame rate, each video frame corresponds to one segment of sound recording. In total, each experiment results in 18k training samples of video frames and a similar number of acoustic recordings for both the teacher network and the student network.



(a) Average accuracy over all users **(b) Accuracies of individual users**

Figure 13: Accuracy of recognizing 7 categories of facial expressions

We train different models for each subject due to the heterogeneity of human facial muscle. The average accuracy of recognizing

the 7 categories of facial expressions over 5 student volunteers is shown in Figure 13(a), and such accuracy is above 80% for all the 7 categories. Furthermore, Figure 13(b) shows the recognition accuracy over each individual student volunteer, and demonstrates that such accuracy varies little and retains at the same level over each student volunteer.

Next, Figure 14 shows results of the leave-one-user-out model evaluation, in which we train the model using data collected from 4 users and report the recognition accuracy tested on the one left. The model still achieves 70% of accuracy even when it is not tuned to specific user. This proves that, even without user-specific training, the teacher and student model can still work collaboratively to extract effective acoustic features correlated with facial expressions.

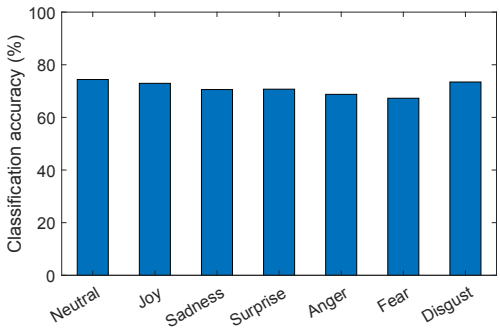


Figure 14: Leave-one-user-out accuracy of different facial expressions

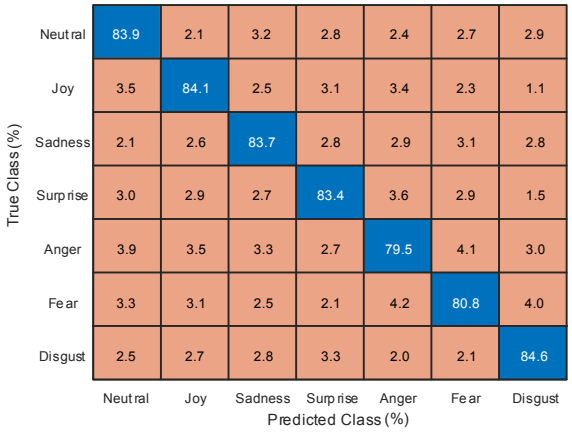


Figure 15: Confusion matrix of facial expression recognition

Furthermore, the confusion matrix of facial expression recognition is shown in Figure 15. Note that recognitions on the anger and fear expressions are less accurate when compared to those on other expressions. To investigate the reason of such inaccuracy, we inspected the acoustic channel estimations corresponding to each category of facial expressions, and found that these two categories of facial expressions result in larger variations of the acoustic channel, as shown in Table 3. The reason of such large variance is that maintaining the fear and anger expressions is more difficult for the users due to the involvement of more groups of facial muscles. As a result, different samples of these expressions may be largely heterogeneous and hence affect correct recognition of these expressions.

Expressions	N	J	Sa	Su	A	F	D
Var.	1.39	1.78	1.67	2.31	5.17	5.89	2.75

Table 3: Channel estimation variance of different categories of facial expressions. (N=neutral, J=joy, Sa=sadness, Su=surprise, A=anger, F=fear, D=Disgust)

4.1 Recognition Accuracy with Different Headphone Models and Host Systems

The recognition accuracy of FaceListener could vary over different types of headphone models (e.g., on-ear, over-ear, in-ear), which have different ways of attachments to human ears. We evaluated the recognition accuracy of FaceListener over the three headphone models shown in Figure 4(a), and the evaluation results in Figure 16 show that FaceListener has the highest recognition accuracy with on-ear headphones, but the accuracy loss on other types of headphones is within 5% for all categories of facial expressions.

More specifically, for over-ear headphones, since the soft cushion on the headphone can absorb the sound leakage and reduce the transmitted signal strength, the performance of FaceListener drops by 2%. In-ear headphones, on the other hand, lead to the lowest accuracy because of two reasons: first, in-ear headphones are inserted into the ear canal and hence result in less sound leakage to the skin surface; second, in-ear headphones have smaller sound diaphragm and hence have more difficulty in capturing the transmitted ultrasound after being re-wired into a microphone.

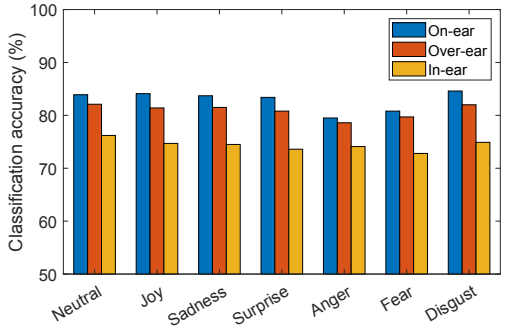
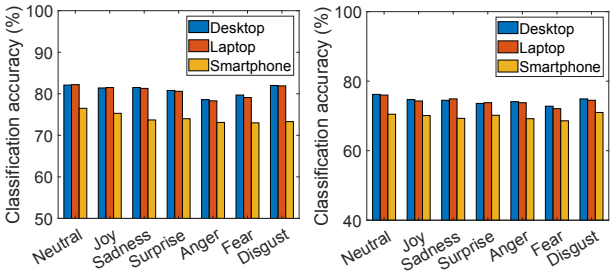


Figure 16: Impact of different headphone models

FaceListener can be used on different host systems including desktop PCs, laptops and smartphones. Figure 17 shows the recognition accuracy tested on these different host systems: a Dell Precision 5810 workstation, a Lenovo Legion 7 laptop and a Samsung S20 smartphone. Desktops and laptops have comparable performance because sound cards with similar specifications are usually integrated into their motherboards. On the other hand, FaceListener's recognition accuracy slightly drops on smartphones, due to the use

of an additional adapter that splits the 2-in-1 audio part but may introduce extra noise.



(a) Over-ear headphone (b) In-ear headphone

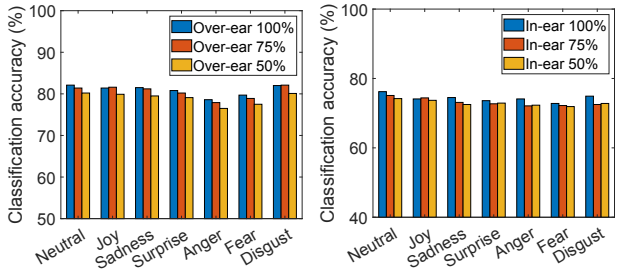
Figure 17: Accuracy on different host systems

4.2 Impact of Acoustic Signal Characteristics

The recognition accuracy of FaceListener depends on the strength of ultrasound signal being received by the transformed microphone. We evaluated the impact of such signal strength with different volumes of the transmitted ultrasound, but the results in Figure 18 show that even when the sound volume at the transmitting speaker is reduced to 50% of the maximum, the recognition accuracy only slightly drops by 1.5%. In reality, as we transmit ultrasound that is inaudible to most people, maximum volume can always be used in most cases.

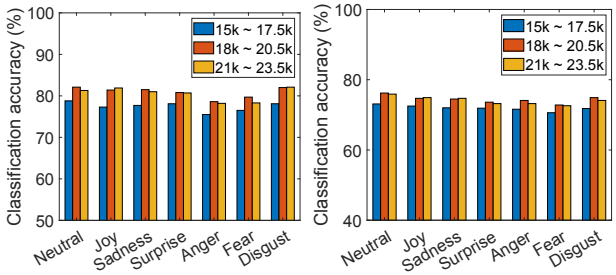
As discussed in Section 2.2, the sound leakage from commodity headphones is the basic motivation of our system design. Since such leakage varies over different frequency bands, we evaluate the system performance by transmitting acoustic signal at different frequency bands as shown in Figure 19. The results demonstrate that our system has the similar recognition accuracy over different ultrasound bands, but has slightly lower accuracy when the transmitted signal falls closer to the audible sound (e.g., around 15 kHz). Such accuracy drop is mainly because that commodity headphones are designed suppress the sound leakage within the audible bands.

Our system transmits the ultrasound signal along an one-way direction because the re-wiring only transforms one speaker into a microphone. Given that some facial expressions may be asymmetric



(a) Over-ear headphone (b) In-ear headphone

Figure 18: The impact of signal strength (measured as the percentage of sound volume)



(a) Over-ear headphone (b) In-ear headphone

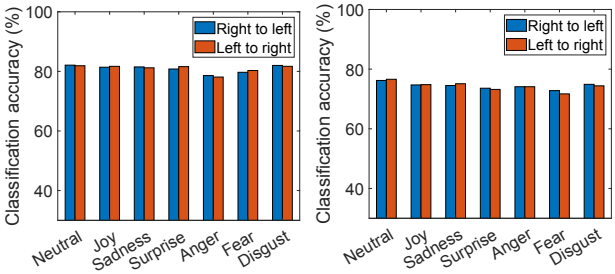
Figure 19: The impact of signal frequency band

(e.g., contempt and fear as shown in Figure 2), the direction of ultrasound propagation may affect the accuracy of recognizing facial expressions. We evaluated the impact of such direction of sound propagation. As shown in Figure 20, the impact of different propagation directions produce negligible impacts on the recognition accuracy. The main reason is that in FaceListener, facial expressions are recognized based on estimations of the acoustic channel that covers the entire human face, and such a channel hence remains the same in both directions of signal propagation.

4.3 Impact of Human Factors

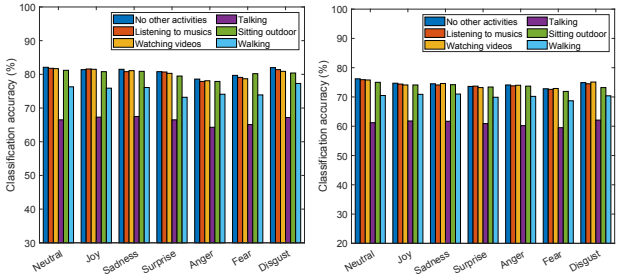
As stated above, FaceListener is designed to recognize the user's facial expressions with other concurrent user activities. We evaluated the impact of different user activities on the recognition accuracy, and evaluation results in Figure 21 show that FaceListener retains high accuracy of recognizing facial expressions for most types of user activities, even when the user is using the headphone to listen to music or watch videos that mainly involve sounds in audible bands. To verify its practical applications, we also test outdoor scenarios including sitting on a bench and walking. Because most outdoor noise can be filtered out by preprocessing, FaceListener can still achieve similar accuracy. Besides, we notice a slight performance drop when the user is walking, such difference is caused by human walking motion may add movements on the receiver side of the headphone, thus affecting the channel estimation results. However, FaceListener will experience 15% accuracy drop if the user is talking while using FaceListener, because of the extra movements of relevant facial muscles. Comparatively, we verified that vision-based recognition methods [38] will also experience 10-15% accuracy drop in similar situations, hence demonstrating the competitiveness of FaceListener to these existing methods.

Since the transmitted ultrasound signal propagates closely above the facial surface, its propagation may also be impacted by different skin conditions. To evaluate whether such factor can make an impact, we train the model with data only collected on normal faces, and test the system performance using data from other skin conditions (without any re-training or fine-tuning on these new data). We emulate these skin conditions by applying cream and water onto the user face, and the accuracy of facial expression recognition, as shown in Figure 22, remains nearly the same under these different skin conditions.



(a) Over-ear headphone (b) In-ear headphone

Figure 20: The impact of signal propagation direction



(a) Over-ear headphone (b) In-ear headphone

Figure 21: Impact of different user activities

Lastly, we evaluate some external wearing objects that can impact our system's practical usefulness. Figure 23 shows the recognition accuracy when a user wears glasses or a hoodie hat. Objects being worn may potentially change the acoustic channel characteristics by incorporating additional signal propagations. However, our results demonstrate that wearing accessories or clothes produces little impacts on the recognition accuracy. Essentially, as long as the accessory does not cover the facial skin (e.g., masks) to hide the face contour, FaceListener can retain the recognition accuracy with little performance loss.

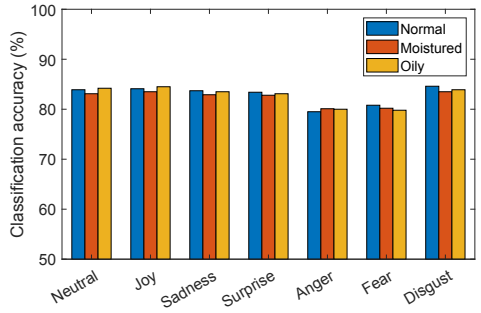


Figure 22: Impact of different skin conditions

4.4 Overhead

We compared the computational costs of FaceListener with those of a vision-based recognition system [38], both on a desktop PC with an Intel i7-8700k CPU. As shown in Figure 24, both systems take input data with a fixed interval of 33.33ms, and the computing times of FaceListener in acoustic channel estimation and neural network model inference are both limited with 100ms and comparable to those in the vision-based system.

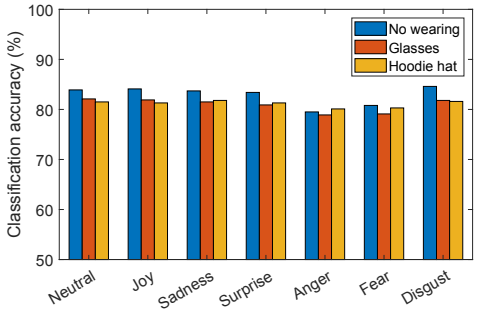


Figure 23: Impact of head accessories and clothes

We also evaluate the power consumption of FaceListener on a Samsung S20 smartphone, with the screen off and any unnecessary background apps being closed. Results in Figure 25 shows that by avoiding the expensive camera use, FaceListener is much more energy-efficient than existing vision-based recognition systems, and consumes <20% of smartphone battery after 3 hours of continuous use (being similar to normal smartphone usage such as music playing).

Process	Time (ms)
Acoustic sensing	33.33
Channel estimation	3.78
Model inference	98.23
Total	135.34

(a) Our acoustic system

Process	Time (ms)
Image capturing	33.33
Data processing	5.15
Model inference	87.26
Total	125.74

(b) A vision-based system [38]

Figure 24: Computational costs of recognizing one facial expression

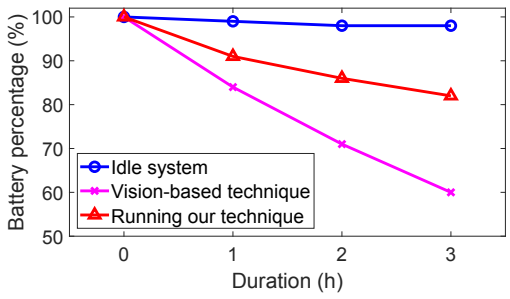


Figure 25: System power consumption

5 REAL-WORLD EXPERIMENTATION

To better examine the performance of FaceListener in real-world application scenarios, we further let the 5 student volunteers to watch two live video clips on different topics: 1) the special Apple

Event on Sept 14, 2021[6] and 2) a talkshow by Seth Meyers[7]. During the playtime of these videos, we collect the facial images and acoustic sensory data of the user using FaceListener.

First, we apply the collected facial images to two vision-based recognizer of facial expressions, namely CK+ [15] and FERPlus [38], to recognize the user's facial expressions and use these facial expressions as shown in Figure 26 to be our ground truth for evaluation. Afterwards, we compare the recognition results of FaceListener with such ground truth, and the evaluation results with respect to CK+ and FERPlus are shown in Figure 27 and Figure 28, respectively[8]. These results exhibit similar accuracy of facial expression recognition when compared with that in lab-controlled settings shown in Section 4, further verifying the usability of FaceListener in practical application scenarios.

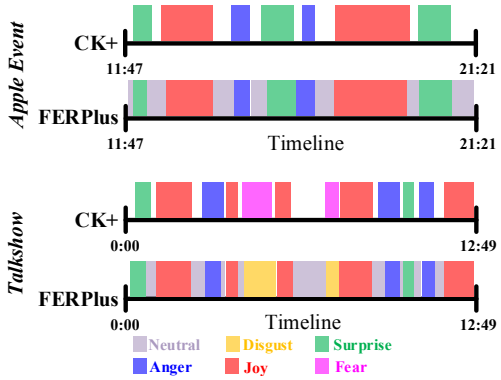
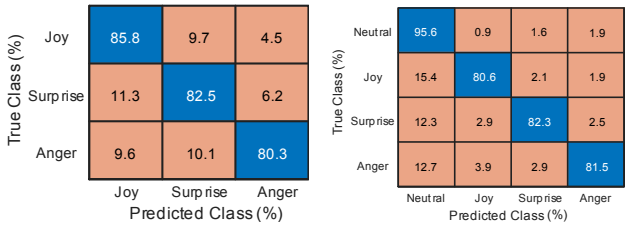
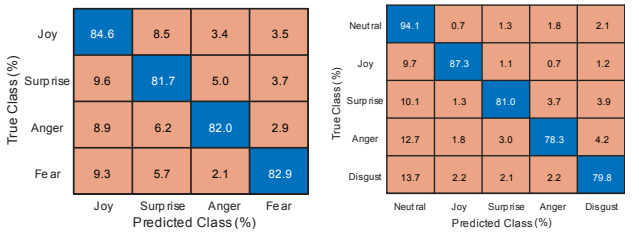


Figure 26: Ground truth of facial expressions



(a) With ground truth from CK+ **(b) With ground truth from FERPLUS**

Figure 27: Performance of recognizing facial expressions when watching the Apple Event



(a) With ground truth from CK+ **(b) With ground truth from FERPLUS**

Figure 28: Performance of recognizing facial expressions when watching the Talkshow

6 RELATED WORK

Emotion and facial expression recognition. Conventional ways to recognize human emotions require electrodes to be worn to record biosignals such as ECG, EEG, EMG and GSR [11, 12], which are inconvenient to be used on commodity systems. Besides, computer vision based DNN models [5, 15, 38], although showing solid performance after being well trained, are not always practically useful due to the need of camera in front of the user. Other recent research efforts exploit new sensing modalities such as wireless signals tracking human heartbeat [43], motion sensors recording facial muscle movements [20, 36] or acoustic signal modeling ear canal [1], but require bulky, expensive or custom hardware that are not affordable or available in many practical scenarios. In contrast, FaceListener uses only commodity headphones with the minimum audio jack re-wiring, but achieves comparable recognition accuracy with the existing techniques.

There have been existing systems that can reach higher accuracies on facial expressions recognition than FaceListener. For example, dedicated EMG and ear pressure sensors can achieve a classification accuracy of 85.2% and 87.5%, respectively. Using the IMU sensors in eSense earables [19], a recent work [36] can classify 32 facial action units with 89.9% of accuracy. By using camera captures as the input, deep learning models can also achieve higher accuracy using different online datasets [3, 15, 38]. Compared with these existing systems, although the recognition accuracy of FaceListener is lower for 5% to 10%, it significantly lower the requirement on extra hardware implementation and deployment. Instead, FaceListener allows convenient development and usage of facial recognition related applications in a more accessible and low-cost manner.

Smart health via acoustic sensing. Acoustic sensing has been widely used in smart and wearable devices for health monitoring, and can be categorized into passive sensing and active sensing. Passive sensing overhears and analyzes the sound produced during the target human activities. For example, the toothbrushing activities can be evaluated from the brushing sound to help maintain teeth hygiene [30], and body sounds that people produce from normal talking and walking can be analyzed to diagnose Parkinson diseases [40].

Our design of FaceListener is related to prior work on active acoustic sensing, in which the system transmits acoustic signals and analyzes the received signal afterwards. Existing work has demonstrated the possibility of using acoustic signals to monitor humans' heart rhythms [37], lung functions [34] and falling events [24]. FaceListener, instead, targets on a completely different but more challenging application scenario, in which recognizing the subtle changes of facial skin deformation requires new acoustic sensing methods.

[6] https://www.youtube.com/watch?v=EvGOlAkLSLw&ab_channel=Apple
[7] https://www.youtube.com/watch?v=yYnXBRo9TVA&ab_channel=LateNightwithSethMeyers
[8] Note that, since only some of our selected 7 categories of facial expressions are present in CK+ and FERPlus, only these facial expressions are used in evaluation.

7 DISCUSSIONS

Impact of head movements. Head movements have been proved to largely affect the performance of vision-based system of facial expression recognition, because the camera with a fixed position can only capture the front view and the facial area and could hence lose important information if the user head turns around. Comparatively, FaceListener suffers much less impact from head movements, because the headphones being worn move accordingly with any head movement. Such movements, on the other hand, may cause slight changes on facial muscle movements, but evaluation results in Section 4.4 have demonstrated that FaceListener can retain high recognition accuracy in these cases.

User-dependent recognition model. The acoustic channel being built and used in FaceListener may highly vary among different human beings, due to their different face shapes and facial muscle characteristics. In addition, even for the same type of facial expression, different people tend to have their own patterns that are unique from others. As a result, the neural network models in FaceListener need to be individually trained for each user. In our current system setup, each individual only needs to record a 10-minute video along with acoustic signal measurements. Since such training is one-time effort for each user, we consider this personalized approach as practical. Possible solutions to this constraint include transfer learning or self-supervised learning, which will be explored further in our future work.

Implementation on other types of headphones. The key of recognizing expressions reliably from acoustic signals is that the acoustic signal must propagate above the facial surface. However, this is not always a possible setup for some commodity headphones. For wired headphone with build-in microphones, for instance, their microphones do not locate on the ear side but always near the mouth. Such configuration cannot guarantee that the acoustic signal can propagate over the entire facial area, possibly affecting the accuracy of recognizing facial expressions. For wireless earbuds such as Airpods, they always have multiple microphones on both sides. Acoustic channel can only be ensured to cover the entire face area if we could specifically select the earphone to use the microphone on one side. However, such implementation is infeasible due to the lack of programming APIs and documentation provided by the device manufacturers.

8 CONCLUSION

In this paper, we present FaceListener, a new technique that recognizes human facial expressions using commodity headphones. Our work uses active acoustic sensing to estimate the propagation channel influenced by facial muscle deformation and classifies these changes into corresponding expressions using a machine learning network with teacher-student structural design. Our experiments show that FaceListener can achieve more than 80% accuracy on 7 common categories of facial expressions.

ACKNOWLEDGMENTS

We thank the anonymous shepherd and reviewers for their comments and feedback. This work was supported in part by the National Science Foundation (NSF) under grant number CNS-1812407, CNS-2029520 and IIS-1956002.

REFERENCES

[1] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial expression recognition using ear canal transfer function. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 1–9.
[2] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. Canalsense: Face-related movement recognition system based on sensing air pressure in ear canals. In *Proceedings of ACM UIST*.
[3] Mouath Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma, and Renaud Seguier. 2021. Learning Vision Transformer with Squeeze and Excitation for Facial Expression Recognition. *arXiv preprint arXiv:2107.03107* (2021).
[4] Apple. 2021. Apple Developers Documentation. https://developer.apple.com/documentation/avfaudio/avaudiosession/categoryoptions/1616618-duckothers
[5] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-Mounted Miniature Cameras. In *Proceedings of ACM UIST*.
[6] Shubhada Deshmukh, Manasi Patwardhan, and Anjali Mahajan. 2016. Survey on real-time facial expression recognition techniques. *Iet Biometrics* 5, 3 (2016), 155–163.
[7] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard E Howard. 2021. HeadFi: bringing intelligence to all headphones.. In *MobiCom*. 147–159.
[8] Shkurta Gashi, Aaqib Saeed, Alessandra Vicini, Elena Di Lascio, and Silvia Santini. 2021. Hierarchical Classification and Transfer Learning to Recognize Head Gestures and Facial Expressions Using Earbuds. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 168–176.
[9] Google. 2020. MediaPipe homepage. https://google.github.io/mediapipe/solutions/face_mesh.html
[10] Google. 2021. Android Developers Documentation. https://developer.android.com/guide/topics/media-apps/audio-focus
[11] Atefeh Goshvarpour, Ataollah Abbasi, and Ateke Goshvarpour. 2017. An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomedical journal* 40, 6 (2017), 355–368.
[12] Anna Gruebler and Kenji Suzuki. 2010. Measurement of distal EMG signals using a wearable device for reading facial expressions. In *IEEE Engineering in Medicine and Biology*.
[13] Richard A Harrigan, Theodore C Chan, and William J Brady. 2012. Electrocardiographic electrode misplacement, misconnection, and artifact. *The Journal of emergency medicine* 43, 6 (2012), 1038–1044.
[14] Bita Houshmand and Naimul Mefraz Khan. 2020. Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*. IEEE, 70–75.
[15] Qionghao Huang, Changqin Huang, Xizhe Wang, and Fan Jiang. 2021. Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences* 580 (2021), 35–54.
[16] Earnest Paul Ijjina, Goutham Kanahasabai, and Aniruddha Srinivas Joshi. 2020. Deep Learning based approach to detect Customer Age, Gender and Expression in Surveillance Video. In *Proc. ICCCNT*.
[17] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proc. ACM conference on multimodal interaction*.
[18] Stamos Katsigiannis and Naeem Ramzan. 2017. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics* 22, 1 (2017), 98–107.
[19] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Allesandro Montanari. 2018. Earables for personal-scale behavior analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89.
[20] 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.
[21] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–9.

[22] Jinyu Li, Michael L Seltzer, Xi Wang, Rui Zhao, and Yifan Gong. 2017. Large-scale domain adaptation via teacher-student learning. *arXiv preprint arXiv:1708.05466* (2017).

[23] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing* (2020).

[24] Jie Lian, Xu Yuan, Ming Li, and Nian-Feng Tzeng. 2021. Fall Detection via Inaudible Acoustic Sensing. *Proc. ACM IMWUT* 5, 3 (2021), 1–21.

[25] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proc. CVPR*.

[26] Zhihan Lv, Liang Qiao, and Qingjun Wang. 2021. Cognitive robotics on 5G networks. *ACM Transactions on Internet of Things* 21, 4 (2021), 1–18.

[27] Aicha Maalej and Ilhem Kallel. 2020. Does keystroke dynamics tell us about emotions? A systematic literature review and dataset construction. In *Int'l Conference on Intelligent Environments*.

[28] Denys JC Matthies, Bernhard A Strecker, and Bodo Urban. 2017. Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions. In *Proc. CHI*.

[29] Sean Olive, Omid Khonsaripour, and Todd Welti. 2018. a survey and analysis of consumer and professional headphones based on their objective and subjective performances. *Journal of the audio engineering society* (october 2018).

[30] Zhenchao Ouyang, Jingfeng Hu, Jianwei Niu, and Zhiping Qi. 2017. An asymmetrical acoustic field detection system for daily tooth brushing monitoring. In *IEEE Globecom*.

[31] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: earphones as a teeth activity sensor. In *Proceedings of ACM MobiCom*.

[32] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[33] Alain Rudiger, Jens P Hellermann, Raphael Mukherjee, Ferenc Follath, and Juraj Turina. 2007. Electrocardiographic artifacts due to electrode misplacement and their frequency in different clinical settings. *The American journal of emergency medicine* 25, 2 (2007), 174–178.

[34] Xingzhe Song, Boyuan Yang, Ge Yang, Ruirong Chen, Erick Forno, Wei Chen, and Wei Gao. 2020. SpiroSonic: monitoring human lung function via acoustic sensing on commodity smartphones. In *Proc. ACM MobiCom*.

[35] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, 3319–3328.

[36] 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.

[37] Anran Wang, Dan Nguyen, Arun R Sridhar, and Shyamnath Gollakota. 2021. Using smart speakers to contactlessly monitor heart rhythms. *Communications biology* 4, 1 (2021), 1–12.

[38] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. 2020. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing* 29 (2020), 4057–4069.

[39] Jeremy HM Wong and Mark Gales. 2016. Sequence student-teacher training of deep neural networks. (2016).

[40] Hanbin Zhang, Chen Song, Aosen Wang, Chenhan Xu, Dongmei Li, and Wenyao Xu. 2019. Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds. In *Proc. ACM MobiCom*.

[41] Ligang Zhang and Dian Tjondronegoro. 2011. Facial expression recognition using facial movement features. *IEEE transactions on affective computing* 2, 4 (2011), 219–229.

[42] Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. 2018. Moodexplorer: Towards compound emotion detection via smartphone sensing. *Proceedings of ACM IMWUT* 1, 4 (2018), 1–30.

[43] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2016. Emotion recognition using wireless signals. In *Proceedings of ACM MobiCom*.