SonicFace: Tracking Facial Expressions Using a Commodity Microphone Array

YANG GAO*†, Northwestern University, USA
YINCHENG JIN†, The State University of New York at Buffalo, USA
SEOKMIN CHOI, The State University of New York at Buffalo, USA
JIYANG LI, The State University of New York at Buffalo, USA
JUNJIE PAN, South China University of Technology, China
LIN SHU, South China University of Technology, China
CHI ZHOU, The State University of New York at Buffalo, USA
ZHANPENG JIN‡, The State University of New York at Buffalo, USA

Accurate recognition of facial expressions and emotional gestures is promising to understand the audience's feedback and engagement on the entertainment content. Existing methods are primarily based on various cameras or wearable sensors, which either raise privacy concerns or demand extra devices. To this aim, we propose a novel ubiquitous sensing system based on the commodity microphone array —- SonicFace, which provides an accessible, unobtrusive, contact-free, and privacy-preserving solution to monitor the user's emotional expressions continuously without playing hearable sound. SonicFace utilizes a pair of speaker and microphone array to recognize various fine-grained facial expressions and emotional hand gestures by emitted ultrasound and received echoes. Based on a set of experimental evaluations, the accuracy of recognizing 6 common facial expressions and 4 emotional gestures can reach around 80%. Besides, the extensive system evaluations with distinct configurations and an extended real-life case study have demonstrated the robustness and generalizability of the proposed SonicFace system.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; *Ubiquitous and mobile computing systems and tools.*

Additional Key Words and Phrases: Acoustic sensing, smart speaker, facial expression, emotion

*This work was done when the author was at The State University of New York at Buffalo.
†The first two authors contributed equally.
‡This is the corresponding author.

Authors' addresses: Yang Gao, Northwestern University, Department of Computer Science, USA; Yincheng Jin, The State University of New York at Buffalo, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA; Seokmin Choi, The State University of New York at Buffalo, Department of Computer Science and Engineering, USA; Jiyang Li, The State University of New York at Buffalo, Department of Computer Science and Engineering, USA; Junjie Pan, South China University of Technology, School of Electronic and Information Engineering, China; Lin Shu, South China University of Technology, School of Future Technology, School of Electronic and Information Engineering, China; Chi Zhou, The State University of New York at Buffalo, Department of Industrial and Systems Engineering, USA; Zhanpeng Jin, The State University of New York at Buffalo, Department of Computer Science and Engineering, USA, zjin@buffalo.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
2474-9567/2021/12-ART156
https://doi.org/10.1145/3494988

ACM Reference Format:
Yang Gao, Yincheng Jin, Seokmin Choi, Jiyang Li, Junjie Pan, Lin Shu, Chi Zhou, and Zhanpeng Jin. 2021. SonicFace: Tracking Facial Expressions Using a Commodity Microphone Array. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 156 (December 2021), 33 pages. https://doi.org/10.1145/3494988

1 INTRODUCTION

Measuring audience feedback and capturing user experience have proven to be significantly beneficial and effective in education, marketing, and advertisement. Traditional reaction or feedback measurement for consumer products primarily relies on self-reporting (e.g., Likert scale). Similarly, when it comes to the movie and video content, crowd-sourcing ratings and reviews (e.g., Nielsen TV rating [2], Rotten Tomatoes [3], IMDB [4]) are the common way to measure the audience's feedback and interest. Although most of the time, those reviews and ratings collectively can represent the overall attitude of the public, they are less effective in revealing fine-grained, instantaneous or short-time emotional reactions during the course of the movie. Collecting and gauging those fine-grained emotional responses requires active audience involvement in a real-time manner, because those first-hand experiences can not be easily retrieved or repeated. Therefore, film producers and content creators have been constantly seeking better ways to gauge the audience's reactions (e.g., "Did the audience enjoy the climax part?", "Did the audience laugh at a specific time in the movie?", or "Did the audience feel boring or tedious for the advertisements within a video stream?"). Compared with the traditional sentimental analysis via plain text reviews and ratings, capturing and recognizing the audience's emotions and facial expressions can provide more accurate and reliable feedback to assess the audience's interest level, engagement, and reactions [19, 71].

In the past decade, there has been a growing adoption of wearable body sensors to capture the viewer's physiological signals to content (e.g., galvanic skin response [46], heart rate variability, electrodermal activity [72], electromyographic (EMG) [15, 29]), which can offer rich information of continuous affective states of the audience. Even though such sensors have been gradually integrated into user-friendly, consumer electronic gadgets with compact form factors (e.g., smartwatches and wristbands), those wearable devices are not always available and accessible to every individual in a large audience population. In addition, the audience's emotional responses sometimes could be momentary and subtle to be timely captured by the aforementioned devices (e.g., getting a laugh for a joke or showing a scared face for a terrifying film shot). Furthermore, physiological signal-based methods normally need a larger time window (e.g., 30 seconds [24]) to analyze the affective profile of the audience.

The audience reaction tracking techniques based on facial expressions have been widely explored in both academia and industry. For example, Disney Research [59] developed a recognition system for group audience reactions in the theatre by utilizing the infra-red illuminator and camera to capture the audience's facial and body motions with optical flow features. Later on, Deng *et al.* [19] extracted patterns in the dynamics of facial landmarks throughout a movie using a factorized variational autoencoder (FVAE). Saha *et al.* [68] proposed an unsupervised learning approach for analyzing the facial behaviors of the audience based on a deep generative model. Other studies have also investigated multimodal analysis (i.e., physiological changes and computer vision) to estimate audience engagement and induce audience emotions [57, 58]. Recently, some companies such as Affectiva [7] and iMotions [8], have explored bringing audience reaction tracking techniques into media and advertisement research. They have provided some successful camera-based media analytic solutions to understand complex and nuanced emotions and cognitive states by analyzing the audience's facial expressions. However, most of the mentioned facial tracking techniques require audiences to install cameras. The strong reliance on recorded videos limits their usability and may raise severe privacy concerns when deployed to a large population.

Due to the recent lockdown and stay-at-home order during the COVID-19 pandemic, it is witnessed that there has been an extensive growth of the online video streaming services (e.g., Netflix, Amazon Prime Video, Twitch). Even though the growth rate gradually slowed down after the pandemic boost, there would be a media turn

Fig. 1. SonicFace: an acoustic-based emotion recognition system that senses the user's facial expressions and gestures during the entertainment. SonicFace emits inaudible acoustic waves mixed with background music from the speaker, and "illuminates" the face. The reflected echoes carry the information of facial expressions (e.g., eyebrow, cheek, mouth) and are received by the microphone array

.

that focuses on at-home streaming video services [49]. Gathering users' feedback and engagement would help video content producers (e.g., movie studios, filmmakers, and game developers) or presenters (e.g., in remote meetings and distance learning) understand and optimize with real impacts across their intended audiences. Meanwhile, it will mutually benefit users with more accurate content recommendation and customization, as well as improve the efficiency of interpersonal interactions between the presenter and audience. However, existing camera-based audience engagement solutions would inevitably brings up privacy concerns in the home settings. To provide audience feedback and capture audience engagement in this new trending scenario, an unobtrusive and privacy-preserving facial emotion recognition method is strongly demanded when the user is watching the videos. Chen *et al.* [16] designed a WiFi-based facial expressions recognition system by using the nearby WiFi router and three antennas (placed 90 cm away from the user). However, this method required additional hardware and settings (i.e., three antennas positioned at the user's head), and there was no sufficient analysis and strategy to handle the body motions, which is one of the dominant factors affecting the multi-path propagation of wireless sensing [12, 51].

In this study, we propose a novel ubiquitous sensing system — SonicFace, which is capable of capturing an individual's facial emotions, without interfering with the speaker's regular usage (e.g., playing the music or audio when watching a movie). As shown in Fig. 1, the speaker emits near ultrasound signals (16 kHz to 20 kHz) towards the user's direction (specifically towards the user's face and upper body part above the horizontal table plane), and the microphone array within SonicFace receives reflected echoes. By analyzing the fine-grained echo patterns, SonicFace aims to detect and differentiate various facial emotional expressions. The contributions of this work mainly are three-folds:

- We develop the SonicFace, an accessible, ubiquitous, and unobtrusive facial expressions recognition system based on acoustic sensing that can help recognize different emotions. Different from conventional solutions, the proposed approach doesn't rely on wearable sensors and specialized gadgets, various types of cameras, or predefined environmental setups.
- We address the challenges of filtering out emotion-irrelevant body motions from emotional gestures and propose a multi-granularity acoustic sensing pipeline to capture facial expressions, as well as emotional hand-to-mouth gestures.

- We conduct various performance evaluations of the proposed system and verify its ability to track the user's engagement.

2 RELATED WORK

This section describes research work related to our investigation, in the areas of ubiquitous acoustic sensing and facial expressions detection.

2.1 Wireless Sensing for Human Gestures and Activities

Commercial off-the-shelf wireless devices (e.g., wireless routers, smartphones, millimeter-wave radars, earphones) have been explored and used for various human behavior recognition and characterization, such as facial expressions, silent speech, and physical activities. Hof *et al.* [34] proposed an mm-Wave radar system to capture the user's face information for the verification purpose. Similarly, Zhou *et al.* [92] designed a novel user authentication system, EchoPrint, which utilized the built-in camera and the inaudible acoustic signals emitted by the smartphone's speaker. Besides user authentication, wireless sensing techniques have also been adopted for interpreting the user's lip-reading. Gao *et al.* [27] proposed a method for silent speech recognition based on acoustic sensing, which utilized STFT to recognize different speech gestures. Zhang *et al.* [88] extended the silent speech recognition, including both individual words and sentences, utilizing amplitude and phase shift to recognize different speech gestures. In addition, another popular application domain using wireless sensing is the recognition of human physical activities. For example, WiFi systems have been proven to be a cost-effective, user-friendly approach for the recognition of various human activities in indoor settings [39, 40, 52, 82, 87]. Similarly, mm-wave sensing [50, 69], acoustic sensing [47, 84], and wearable smartphones or smartwatches [28, 83] have also been explored in literature for human activity recognition.

As the closest work to SonicFace which leverages ultrasound sensing by the speaker and the microphone array, Mao *et al.* [53] proposed a hand motion tracking system by utilizing a 4-element microphones array and dual speakers to measure the propagation distance and angle-of-arrival (AoA) of reflected signals. To precisely capture the object trajectory, this solution requires a customized microphone array configuration with non-uniform microphone separations. BeamBand [35] is a wrist-worn device using ultrasound beamforming to recognize a variety of hand gestures. This system contains eight transducers to emit and receive the 40 kHz ultrasound signals. Furthermore, FM-track [44] validates their multi-target tracking system on both the customized microphone array and commercial off-the-shelf devices (i.e., smartphone, UMA-8-SP USB mic array). However, most existing ultrasound-reflectometry solutions focus on object tracking and hand gesture-based interactions. To explore the potential of recognizing fine-grained facial expressions and gestures leveraging commercial off-the-shelf devices, in this work we propose SonicFace, a customized prototype that is capable of capturing the user's facial gestures while playing the background sound, in an unobtrusive manner.

2.2 Facial Emotion Recognition

In recent centuries, researchers have developed the basic emotions theory and dimensional models of emotions. Ekman identified six basic emotions in the 1970s, including anger, disgust, fear, happiness, sadness, and surprise [23]. Contempt was later added to basic emotions to form seven universal facial emotions [22]. Other different theories and approaches to basic emotions were also introduced [37, 38]. Izard [38] argued that basic emotions have innate neural substrates and universal behavioral phenotypes, which are involved in our daily-life tasks.

A variety of physical and physiological information were utilized to recognize different emotions, such as EEG [48], heartbeats [90, 91], speech [62], body gestures [70], and facial expressions [15, 61]. Among these recognition methods, the facial expression is one of the most direct, natural, and universal cues to explicitly express the emotions by human beings [45]. In 1978, Ekman and Friesen developed the Facial Action Coding

System (FACS) [17] to model human facial expressions. FACS can code facial expressions by action units (AUs) and action descriptors (ADs), which are the fundamental action muscles and the unitary movements involving the actions of several muscle groups. Previous work has shown that people share similar facial expressions to stimuli that elicit their basic emotions [21, 23, 25].

Given the rapid development of computer vision and artificial intelligence technologies, facial expression recognition has become an increasingly popular topic in recent years. A large number of prior studies [45] have been conducted on facial expression recognition, including human-computer interaction systems, mental health monitoring, and driver fatigue surveillance. Take an example of the human-computer interaction systems, Bretan *et al.* designed a system that was used to respond to humans in an appropriate manner by detecting human facial expressions to reveal emotional intelligence [13]. In addition, detecting early symptoms of psychological or somatic problems (e.g., depression, anxiety, and bipolar disorder) through facial expressions is significant to prevent psychological problems from getting worse. Simcock *et al.* proved that youth with increased psychological or somatic problems exhibited a processing bias for anger and fear expressions, but not sadness [73]. Through the surveillance of facial expressions of seven basic emotions, people with psychological problems can get a medical service as early as possible to avoid further condition aggravation. Another promising application domain is safe driving. Fatima *et al.* proposed a low-cost solution for driver fatigue detection by placing a camera on the extreme left side of the driver and utilizing novel algorithms to facilitate accurate face and eye detection for fatigue driving. However, the limitation of these works is the privacy issues of severe concern to the public and most of the users, restricting them from practical applications. From this perspective, our proposed SonicFace system possesses superior advantages in terms of ubiquitousness, availability and accessibility, ease of use and deployment, and privacy protection, and thus can be widely adopted in daily life.

Besides watching videos, playing VR/AR games can also evoke users' emotions, in an even more immersive manner. This could be extended to emotional communications with the computer or other individuals, making it possible to design emotion-aware user interfaces or emotion regulation applications. Zhang *et al.* [89] and Xue *et al.* [85] provided an alternative way of collecting precise ground truth labels of emotions by watching videos in VR. Jicol *et al.* [41] investigated emotions induced by a virtual environment and provided implications for future VR design. In addition, various sensors can be mounted on the VR/AR devices for emotion recognition. Gupta *et al.* [32] presented a personalized real-time emotion recognition system on VR by mounting the EEG and GSR sensors. All these applications have shown that obtaining customers' emotions can bring substantial benefits to the entire human-computer interaction community.

2.3 Audience Reaction Recognition

Visual and audio content are known to draw people's attention and elicit emotions effectively [76]. Multiple studies of the audience's reactions towards videos have been conducted in recent years for the usages of writers, directors, marketers, and advertisers. Traditional methods of collecting and predicting audience reactions were largely based on users' ratings, feedback, discussions, etc. Nevertheless, recent technologies supported reaction recognition without the audience's active involvement, through collecting their direct and objective reactions like facial expressions. Joho [42] tracked the viewer's affective reactions to video content and showed that facial expression is a good feature to detect personal highlights of multimedia contents. Navarathna *et al.* [58] employed an infrared camera to obtain the audience's visual information while they were watching feature-length movies, and detected the change that conveyed audience sentiment, then created a movie rating classifier. McDuff *et al.* [54] predicted Ad liking and purchase intention based on 12,000 facial responses from 1,223 persons to 170 ads, where liking was shown by eliciting facial expressions and positive emotion drives people's purchase intention.

3 DESIGN CONCEPT AND CONSIDERATION

The SonicFace aims to recognize homebound (at-home) user engagement (i.e., different emotions and hand gestures), in order to provide better user feedback to movie producers and content creators. In this section, we first define the concept of user engagement (i.e., emotional expression) in the home-entertainment scenario. Then we discuss two major types of signals used in existing acoustic sensing applications and their limitations in the proposed engagement recognition task. Finally, we propose the customized SonicFace signal that is capable of capturing both the coarse-grained body motions and fine-grained facial expressions.

3.1 Design Concept of Emotional Engagement

As described in the book "In the Blink of an Eye" [56] written by Walter Murch, both fine behaviors in the face (e.g., smiles and yawning) and body (e.g., head-pose change and hand gestures), as well as coarse body behaviors (e.g., fidgeting and doodling), may be the indicators of engagement. For example, when the audience is watching a horror movie, the nervous system will set the body's fear response into motions, which involve wide eyes with open mouth, hands covering either the mouth or eyes, turning back, and even jumping. Given the target application scenarios, this study is limited to home entertainment only, when the user is sitting in front of the computer screen. We consider the fine-grained facial expressions and emotional hand gestures (e.g., **fear**: one hand covering the face; **amusement**: one hand covering the mouth; **sad**: hands touching the head) as the indicators of the user engagement.

However, monitoring the audience engagement level during a movie or TV show often lasts for a long period of time, during which it is inevitable to introduce extra body motions (e.g., adjusting sitting postures, turning the head, leaning back, and so on) irrelevant to emotions or engagements. It is known that acoustic sensing is sensitive to multi-path interference. Those irrelevant body motions may cause significant multi-path changes, which result in disturbance of the received echo signals and affect the extracted features. Therefore, it is imperative to distinguish facial expressions and emotional gestures from irrelevant body motions.

3.2 Design Consideration of Acoustic Signals

In this section, we introduce the Frequency Modulated Continuous Wave (FMCW) based object tracking, pure tone based motion tracking with Doppler-shift features, and their limitations. Then, to recognize and distinguish both coarse-grained body motions and fine-grained facial expressions, we introduce the SonicFace signals with the combination of FMCW and pure tone within the frequency range from 16 kHz to 20 kHz.

3.2.1 FMCW-based Absolute Tracking. FMCW is a sequence consisting of numerous repeated chirps, which has been widely used in acoustic sensing applications [15, 79]. Fig. 2(a) shows FMCW in the time-frequency domain, where the red line denotes the transmitted signals and the green line denotes the reflected signals. We utilize $f(t)$ to denote the frequency of the FMCW signal at time t; thus we get the formula: $f(t) = f_0 + \frac{Bt}{T}$, where f_0 denotes the low frequency of the chirp, B denotes the sweep bandwidth, and T denotes the chirp period. Without loss of generality, we normalize the transmitted signals as $V_{tx}(t) = cos(t)$ in the time domain. After the propagation delay of τ, the receiver would receive the reflected signal $V_{rx}(t)$ at time t, which can be represented as $V_{rx}(t) = \alpha cos(t - \tau)$, where α is the attenuation factor that refers to the amplitude decreasing of acoustic signals after the propagation. If the transmitter and receiver can be synchronized in the same clock system, τ can be calculated by the formula: $\tau = 2(R + vt)/C$, where R denotes the distance between the sound transmitter and target, v denotes the moving speed of the target, and C denotes the sound propagation speed which the approximate value is normal environment is 346 m/s.

Based on the introduction above, the distance resolution of object tracking can be defined as [44]: $\Delta R = \frac{C\tau}{2} = \frac{CT}{2B}\Delta f_b$, where C is the sound propagation speed 346 m/s and Δf_b denotes the frequency shift of the mixed signals

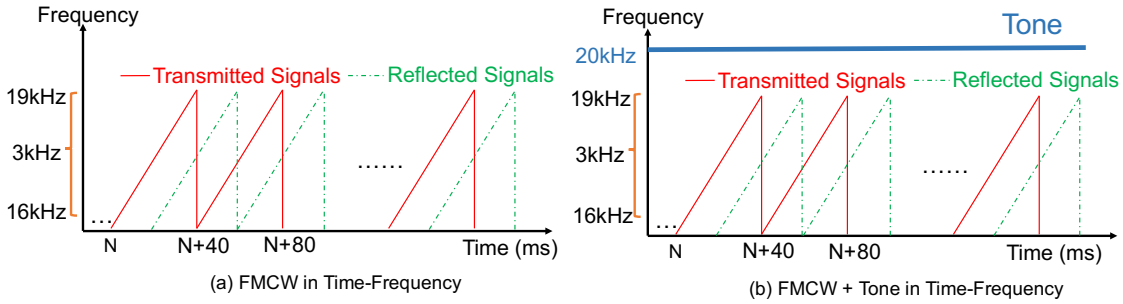


Fig. 2. FMCW Sequences (a) chirp sequences in the time-frequency domain; (b) chirp sequences combined with pure tone in the time-frequency domain.

that can last for one chirp at most, thus $\Delta f_b \geq \frac{1}{T}$. It is shown that the distance resolution depends on the sweeping bandwidth B. Given the commodity microphone and speaker settings, the bandwidth B usually ranges from 3 kHz to 6 kHz. Thus, we can calculate the minimum distance change which is 2.83 cm to 5.76 cm. Therefore, the granularity of absolute tracking is insufficient to capture the fine-grained movements of facial expression components (e.g., eye, eyebrow, and cheek).

3.2.2 Pure Tone-based Relative Tracking. Besides FMCW-based tracking methods, many prior studies [27, 88] have also utilized the pure tone (≥ 16 kHz) to capture the object's movements. The sensing techniques can be roughly divided into two methods: the first one is the Doppler Shift and the other one is the amplitude and phase extraction. Doppler Shift senses surrounding moving reflectors by calculating STFT (Short-Time-Fourier-Transform) of the reflected echos. However, the resolution of STFT time-frequency analysis limits its sensing resolution in terms of object movement speed. Another sensing technique is to extract the fine-grained amplitude and phase change of the received signal. For example, a phase change of 60° of the received signal with the wavelength of 1.73 cm (20 kHz) will map to a distance change of 1.4 mm ($\frac{1.73\times60^\circ}{2\times360^\circ} = 0.14$ cm).

Even though the phase and amplitude extraction can achieve millimeter-level tracking resolution, it can not provide the absolute position of the target. Therefore, the phase and amplitude-based methods are more suitable for single target tracking. However, as discussed in Section 3.1, user engagement is a very complicated and multifaceted process, which contains facial expressions, hand gestures, and body postures. It is challenging to recognize the user's responses solely by utilizing a pure tone based relative tracking method.

3.2.3 Dual-grained Signal Fusion. In Fig. 3, we ask a participant to do sequential behaviors: back-lean, laugh, turn-head, smile. The second row indicates the distance estimation between the participant and the microphone, based on the FMCW-based MUSIC algorithm. Given the compact microphone array size, the spatial sensing resolution is also very limited, thus, it is hard to detect subtle facial movement such as laugh and smile. The third row shows the amplitude features of reflected echo. It is observed that both body motions and facial gestures cause amplitude fluctuation, thus, it is hard to distinguish body motions from the facial expressions, especially for the large mouth movements.

As discussed above, in order to detect the multi-granularity of the body-hand gestures and facial expressions, we propose a composite signal which combines FMCW and pure tone as shown in Fig. 2(b). Specifically, we simultaneously transmit FMCW (16 kHz to 19 kHz) component and pure tone (20 kHz) component. The 1 kHz gap is designed to avoid the interference between two components. Meanwhile, in order to avoid interrupting user's normal usage of smart speaker, the background sound in video or audio content will also be mixed with

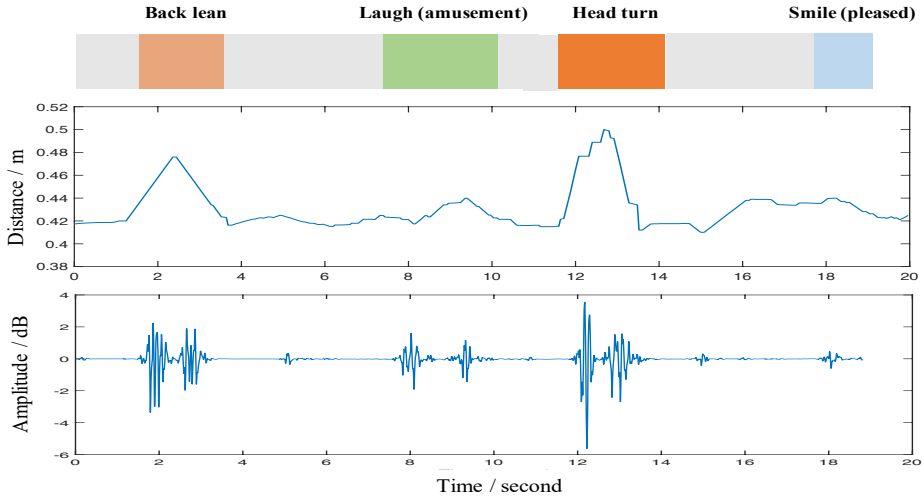


Fig. 3. An example of facial expression with body motions

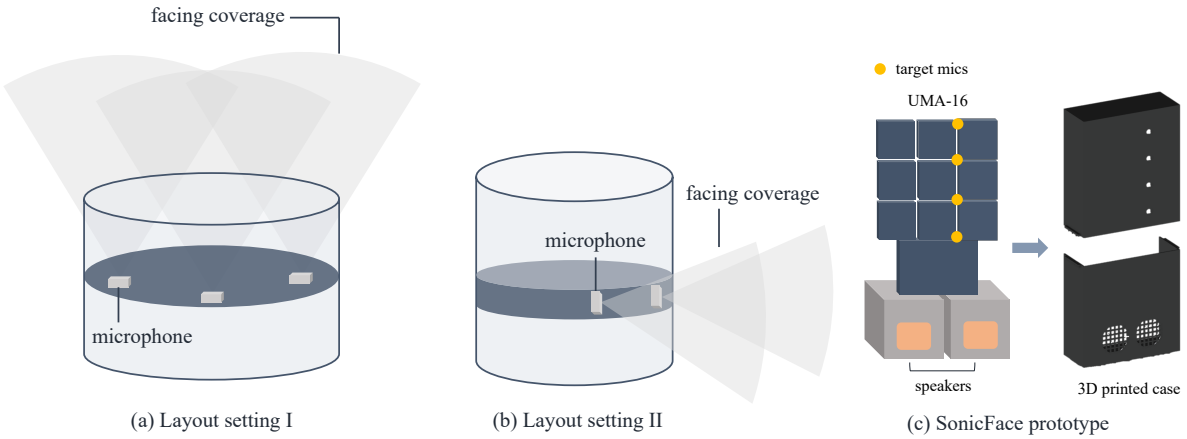


Fig. 4. (a) (b) Microphone array layouts of commercial smart speakers, and (c) layout design of SonicFace prototype.

the near-ultrasound signal with no frequency-band interference. The details of audio quality degradation caused by the multi-frequency mixture will be discussed in Section 6.5.2.

3.3 Prototype Layout and Design

In this section, we will discuss our prototype, specifically the layout of the microphone array and speaker design.

Typically, to provide good spatial awareness of sound in an open space using unidirectional microphones, existing commercial smart speaker designs usually come with a microphone array. As shown in Fig. 4, there are two typical microphone array settings in commercial smart speakers: 1) a number of vertical, upward-facing microphones are placed on the perimeter of a circle in Fig. 4(a) (e.g., Amazon Echo Studio [5], Google Nest Mini [9]); and 2) some horizontal, outward-facing microphones are placed on the perimeter of a circle in Fig. 4(b) (e.g., Apple HomePod [1]).

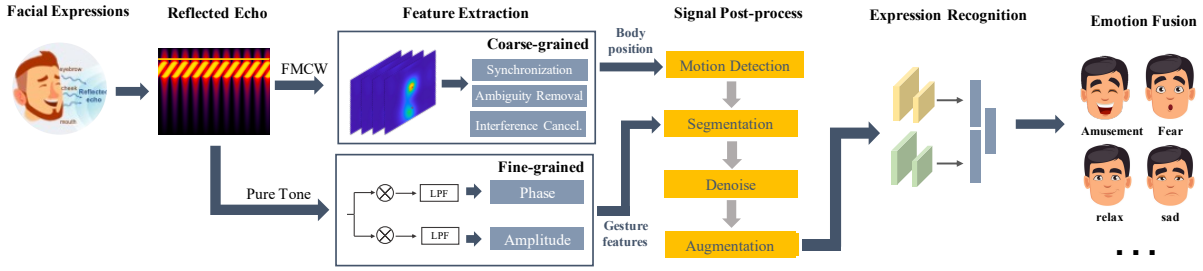


Fig. 5. System overview of SonicFace

In this work, to better capture the echoes resulting from the direct propagation path reflected by the user's facial and hand gestures, we adopt the second layout setting as shown in Fig. 4(b), with the microphone array facing outward in the horizontal plane. To further improve the spatial sensing resolution, we proposed a design prototype for SonicFace with the microphone array and speaker configured in the layout as shown in Fig. 4 (c). The emitted sound waves from the speakers will pass through the case (hollow grid) and be reflected by the user's head and body. This will significantly reduce the interference noises caused by multi-path reflections compared with the configuration in Fig. 4(a).

To build a system that can be easily deployed or adjusted in commodity smart speakers without adding any additional specialized sensor, we use two commercial microphone arrays that can provide access to the raw audio signals: MiniDSP UMA-16 and UMA-8, which best mimic today's off-the-shelf smart speakers with similar microphone configurations (i.e., mic separation and frequency response) and are widely used in many smart speaker prototype designs [10, 74, 75]. In particular, even though UMA-16 has 16 individual microphones, considering that the typical number of the microphones at each side (half-round) of commercial smart speakers design is usually 3 to 4, we thus only choose 4 microphones in UMA-16 as a linear microphone array, as shown in Fig. 4 (c). In our study, we utilize UMA-16 as the main microphone array for the SonicFace prototype and also evaluate the alternative design using UMA-8, which will be discussed in Section 6.5.1. As a future work, the size of the prototype can be further reduced with a customized microphone array.

4 SONICFACE DESIGN

4.1 System Overview

SonicFace achieves the recognition of facial expressions in three steps, following the processing flow as shown in Fig. 5. First, the speaker in SonicFace emits inaudible sound waves, and the microphone in SonicFace receives reflected echoes. Secondly, the SonicFace analyzes echoes and generates the frequency and phase shifts, which are then passed to the neural network model. The key is to separate the fine-grained facial expressions and gestures from coarse-grained body motions. Finally, the model predicts the outputs of different emotions.

4.2 Feature Extraction

In this section, we introduce the coarse-grained features based on FMCW signals to get the information of body motions, and the fine-grained features based on pure tone to extract facial expressions.

4.2.1 Coarse-grained Features. To localize the user's body, here we use the 2D MUSIC algorithm [44, 81] to jointly estimate the propagation distance (between the user and the speaker) and angle-of-arrival (AoA) of the reflected echoes. The key idea is to transform the received echoes into 2D sinusoid. As discussed in Section 3.2.3, we let the speaker emit the FMCW signals with the sweeping frequency from f_0 to $f_0 + B$ in period of T and can

be defined as :

$$C_T(t) = cos(2\pi(f_0 t + \frac{B}{2T}t^2)) \tag{1}$$

The chirps are reflected by the user's upper body with the face, and received by the microphone array. For each microphone, we multiply both the transmitted and received chirp signals, and then apply low-pass filter. The received signals are represented as:

$$C_R(n,t) = cos(-\frac{2\pi\Delta cos(\theta)}{\lambda} \cdot n + \frac{4\pi Bd}{Tv_s} \cdot t + \phi) \tag{2}$$

where n is the microphone index, t is the propagation time, θ is the AoA, and d is the distance. Δ, v_S, and ϕ are the microphone distance, sound propagation speed, and the constant delay phase. To capture the coarse-grained body motion information, we adopt the 2D MUSIC to estimate frequencies in the sum of the sinusoids ($\sum_i e^{j(\Omega_i n + \omega_i t)}$) by searching for all arrival vectors that are orthogonal to the noise subspace. To implement the search, 2D MUSIC constructs an arrival-angle-dependent power expression, pseudospectrum based on signal structure, and each peak in the spectrum indicating a frequency pair (Ω_i, ω_i).

$$P(\theta,\omega) = \frac{1}{a^H(\theta,\omega)U_n U_n^H a(\theta,\omega)} \tag{3}$$

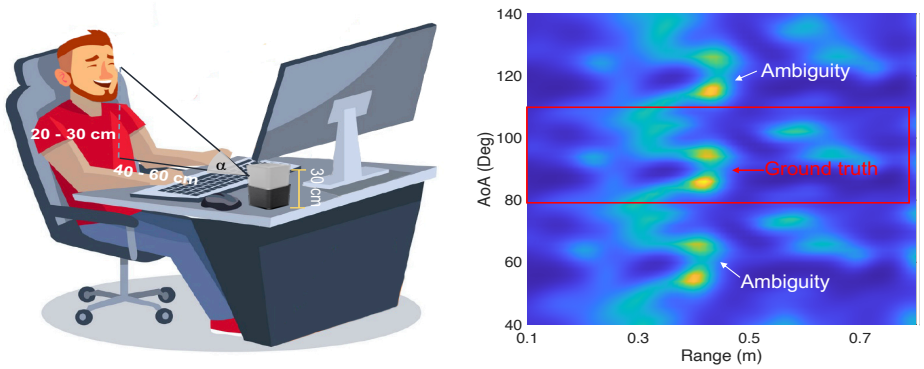
Since the frequency pairs are also determined by the AoA and distance, as in [53], we replace ($\Omega, omega$) with (d, θ), and the pseudospectrum can be represented as:

$$P(\theta,d) = \frac{1}{(v(d) \otimes u(\theta))^H \mathcal{M}(v(d) \otimes u(\theta))} \tag{4}$$

where $u(\theta) = [1, e^{-j2\pi\Delta cos(\theta)/\lambda}, ..., e^{-j2(N-1)\pi\Delta cos(\theta)/\lambda}]$, $v(d) = [1, e^{j4\pi BdT_s/(Tv_s)}, ..., e^{j(X-1)4\pi BdT_s/(Tv_s)}]$, and X is the sample length, $\otimes$ is the Kronecker product. The AoA and distance of the targets (i.e., body and head) can be estimated by searching peaks in the pseudospectrum.

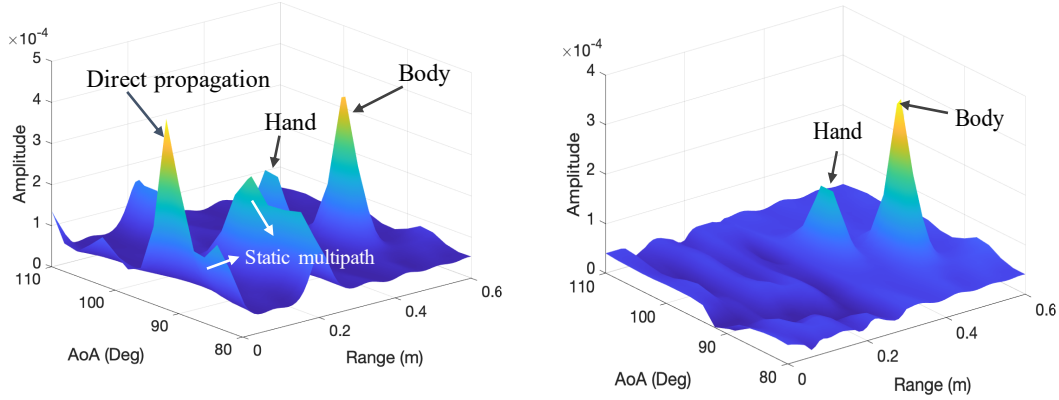
1) Synchronization. To precisely capture the location information of the body and face, a small synchronization error of 0.1 ms will result in a 3.43 cm distance prediction error. Hence, the synchronization between the speaker and microphone is critical to ensure the resolution. We estimate the propagation delay between the emitted signal from the speaker and the received signal from the microphone by examining the cross-correlation between the received signal and the original chirp signal. The maximum correlation indicates the optimal alignment. We select the time with the largest correlation as the start time of the received signal.

2) Ambiguity Removal. Most of the microphone arrays in smart speakers have the uniform separation Δ, which is initially designed for the source localization of human sound (100 Hz - 400 Hz). However, to localize the reflected ultrasound in a higher frequency band, if Δ is larger than the half of the sound wavelength λ, there will be an ambiguity in estimating the AoA of the target. Specifically, in Equation 4, when $\Delta > 0.5\lambda$, there are two angles θ_1 and θ_2, such that $-2\pi\Delta cos(\theta_1)/\lambda = -2\pi\Delta cos(\theta_2)/\lambda + 2\pi$, which makes $P(\theta_1, d) = P(\theta_2, d)$. Every peak with (θ_1, d) will have an identical pseudo-peak in (θ_2, d). In our default device setting, $\Delta = 4cm, \lambda = 2.14cm$, as shown in Fig. 6(a), the microphone array's height is 30 cm, and the distance between the user and the SonicFace device is 40 cm to 60 cm. For a user with a height of 170-180 cm (5 feet 5-11 inches), to recognize the facial expressions and some related hand gestures on the face, SonicFace needs to cover the area of face and upper chest ranging between 20 cm to 30 cm. Therefore, we can calculate $\alpha = 30$ degree with the corresponding AoA (vertical) ranging from 80 degrees to 110 degrees. In Fig. 6(b), we observe two ambiguity peaks in 115 degrees and 55 degrees with its ground truth peak at 95 degrees. To remove undesired ambiguity peaks, as indicated by the red box in Fig. 6(b), we select the zone between 80 degrees to 110 degrees to restrict the 2D profile sensing range.



(a) Device placement setting (b) Ambiguity in 2D AoA-distance profile.

Fig. 6. Ambiguity in AoA-distance profile.

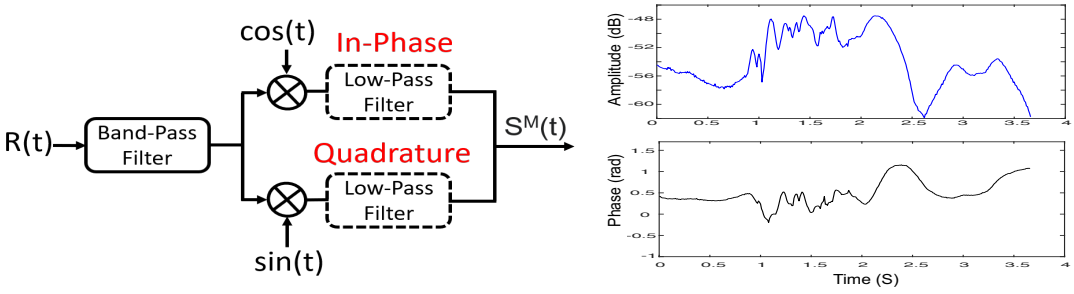


(a) Before Inference Cancellation (b) After Inference Cancellation

Fig. 7. An example of joint-estimation for AoA-distance based on 2DMUSIC.

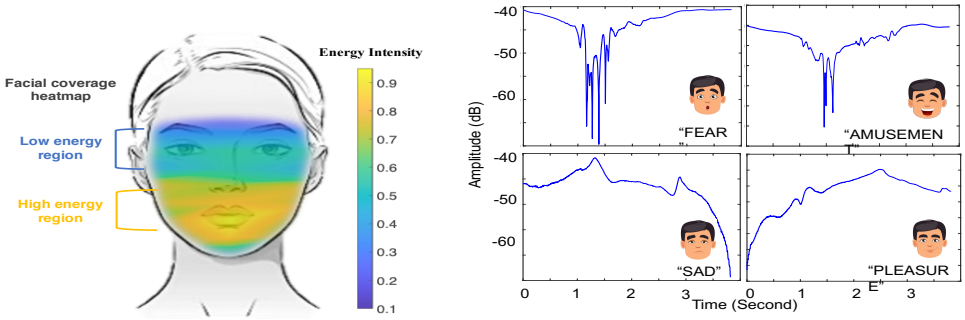
3) Interference Cancellation. When the speaker transmits acoustic waves, the signals propagate through the space omnidirectionally. Thus, sounds received by the microphone array are the reflected echoes from the body and head, as well as the direct sound propagation from the speaker. Therefore, this step seeks to remove the direct inferences of the sound transmission between the speaker and the microphone array. We first record the direct transmission when the user and other major reflections (e.g., chair) are away from the speaker (> 2 meters). Then we substrate the pre-recorded pseudospectrum from the newly received signal. An example of the AoA-distance profile without cancellation is shown in Fig. 7(a), there are multiple peaks in the 3D pseudospectrum including direct propagation source, body and hand reflectors, and static multipaths. After subtracting the pre-recorded pseudospectrum, we can obtain a clear AoA-range profile with less interference as shown in Fig. 7(b).

4.2.2 Fine-grained Features. Given the limitation of the Doppler Shift to detect fine-grained facial expressions, we extract the amplitude and phase features to classify different facial expressions. The speaker emits continuous wave (CW): $Acos(2\pi ft)$, where A represents the amplitude, f denotes the frequency. The reflected signals which



(a) (b)

Fig. 8. In-Phase and Quadrature features are extracted from the received echo signal $R(t)$. (a) Signal coherent structure; (b) Phase and amplitude for the expression "amusement."



(a) (b)

Fig. 9. (a) Heatmap of reflected echo signal intensity. (b) Quadrature features of the echo signals for different facial expressions.

received by the microphone can be denoted as:

$$R(t) = Acos(2\pi f(t - d(t)/c) - \theta) \tag{5}$$

where c is the speed of the sound, θ is the phase shift and A denotes to the amplitude of the received signal. Fig. 8(a) shows the progress of calculation of the amplitude and phase information. The received signal is multiplied by $cos(2\pi ft)$ and get the result:

$$R(t)cos(2\pi ft) = Acos(2\pi ft\phi)cos(2\pi ft) = \frac{A}{2}cos(4\pi ft\phi) + \frac{Ap}{2}cos(\phi). \tag{6}$$

We utilize a low-pass filter to remove the high frequency of $2f$. Then, we can get In-Phase (I) component: $I = \frac{A}{2}cos(\phi)$ and the Quadrature (Q) component: $Q = \frac{A}{2}sin(\phi)$. Thus, we can obtain the amplitude values of the real part and the phase values of the imaginary part accordingly by combining these two components as the real and imaginary parts of a complex signal respectively, exemplified in Fig. 8(b).

We plot the acoustic facial coverage region in Fig. 9(a), which represents the extracted IQ feature intensity of the reflected echo signal in different facial regions. We ask the participant to perform individual muscle movement in each facial region (i.e., opening the mouth, wrinkling the nose, blinking the eye, raising the eyebrow, raising the cheek) for 10 times repeatedly and record the reflected echoes. We then calculate the average signal energy of the extracted IQ features of echoes and map the values to the corresponding facial regions. Considering that the participant can not precisely control the single muscle movement on the face, this heatmap is just a

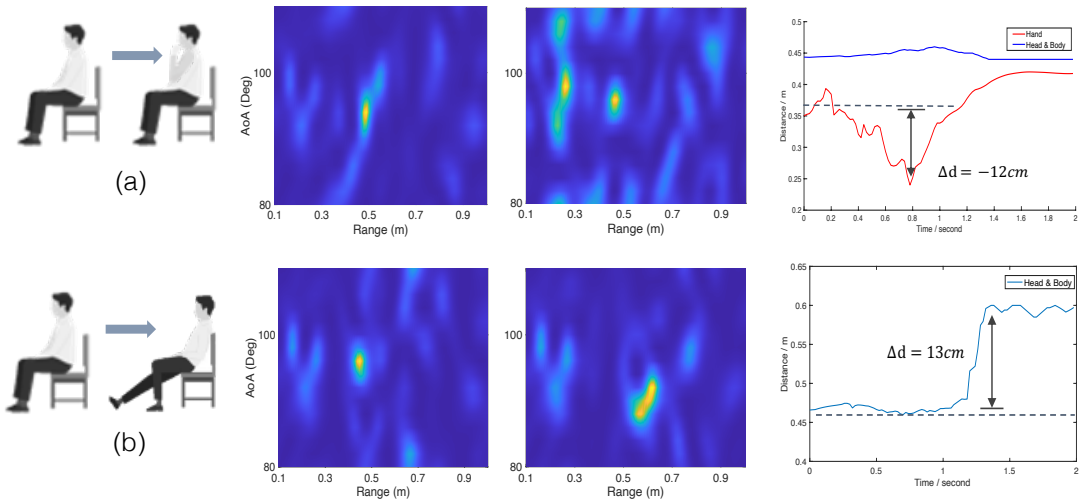


Fig. 10. (a) Emotional hand-to-face gesture: one hand covering the face; (b) irrelevant body motion: leaning backward

rough indicator to category nearby facial muscle groups. It is observed from Fig. 9(a) that the mouth and cheek regions exhibit relatively higher energy intensity compared to the eye and eyebrow regions, where the weak muscle movements are observed. In Fig. 9(b), we plot the quadrature features of reflected echo signal for different expressions. It is observed that emotion "fear" and "amusement" have relatively larger energy intensity range (high fluctuation), compared with "sad" and "pleasure". It is because the expressions of "fear" and "amusement" involve more mouth muscles which is believed to cause larger multi-path changes when reflecting the echo signal.

4.3 Feature Post-processing

4.3.1 Irrelevant Motions Filter. As discussed in Section 3.1, the phase and amplitude of the received signals are affected by the displacements of reflections. Such displacements are sometimes composite, which may contain facial muscle motions, emotional hand-to-face gestures (e.g., *fear*: one hand covering the face; *amusement*: one hand covering the mouth; *sad*: hands touching the head; *sad*: hand wiping tears), as well as other irrelevant body motions (i.e., random or non-voluntary hand and body movements that are not directly related to emotions). Given the fact that the displacements of hand and body movements are orders of magnitude larger than facial expressions, the phase and amplitude are dominated by body motions. However, by using the 2D MUSIC algorithm, we are able to roughly estimate body motions based on the AoA-distance profile, discussed in Section 4.2.1. As shown in Fig. 10, we list two common motions when the user is watching the video: (a) emotional hand-to-face gesture (one hand covering the face), and (b) emotion-irrelevant body motions (leaning backward). From the 2D AoA-range profile, it can be clearly seen that there is a stable peak in the pseudospectrum when the user sits still. When the user raises his/her hand (in Fig. 10(a)), a new peak appears at the distance of 0.3 m, because the hand now becomes a closer reflector with respect to the microphone. We plot the distance changes for two identified reflectors (i.e., the blue line indicates the hand and body; the red line shows the hand distance) within the 2-second gesture time. A face-to-hand gesture can be detected if there is a new reflection source moving towards the microphone and then moving backwards to the body in a short time (< 2 seconds). Due to the limit of displacement resolution between two objects, no new target reflector appears in the AoA-range profile for the lean-back motion. However, we can still observe that the distance between the main reflector (i.e., body and

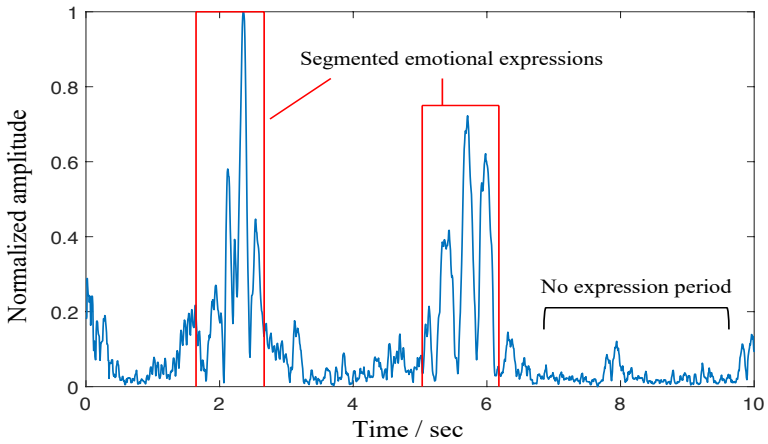


Fig. 11. An example of emotional expressions detection and segmentation by the LSTD algorithm.

head) and the microphone array is increased by the body motion, as well as slight AoA changes, which is shown in Fig. 10(b).

Considering the application scenario when the user is watching the video or live streaming, hand movements or body posture changes are occasional and unpredictable. To mitigate the impact of emotional irrelevant body motions, we filter out undesired body motions and only keep emotional gestures with hand-to-mouth movements, which can be detected by the amplitude and duration of the distance change within the restricted AoA range.

4.3.2 Event Detection & Segmentation. Another critical challenge in facial expression recognition based on acoustic signals is to filter out the undesired segments with no facial expression and then identify and quantify the occurrence of facial expression events. This essentially requires SonicFace to properly detect the amplitude changes of the received echoes incurred by different facial movements as shown in Fig. 11. The process of detecting facial expression events consists of three steps: 1) generate the amplitude changes introduced in Section 4.2.2; 2) calculate the standard deviations based on every 12 points to highlight the amplitude changes; and 3) utilize the long-term spectral divergence (LTSD) of the voice activity detector (VAD) algorithm [64] to segment the sequence correctly in noisy environments and dynamic situations. The algorithm measures the LSTD between the signal and the noises, then formulates the ratio of the signal and noise decision rule by comparing the long-term spectral envelope to the average noise spectrum.

4.3.3 Denoising & Augmentation. Since the multipath of reflected signals and surrounding noises, it is imperative to mitigate or remove noises. In this study, we utilize a moving window average with 32 points to smooth the data and make the patterns more noticeable.

Besides, it is well acknowledged that a robust neural network model shall be trained with a large amount of heterogeneous data. However, in reality, it is hard and sometimes impossible to collect sufficient data for various facial expressions from a large population. Hence, data augmentation is a common strategy to mitigate the issue of data insufficiency.

Fundamentally, facial expression detection is a time-series alignment application [43]. We utilize the time warping method, which is widely used in speech recognition [55] and human action recognition [88], to enrich the training dataset. Time-warping (TimeW) is the way to perturb the temporal location by distorting the time intervals between each sample. Here we utilize a random amplitude to generate the random sinusoidal curves.

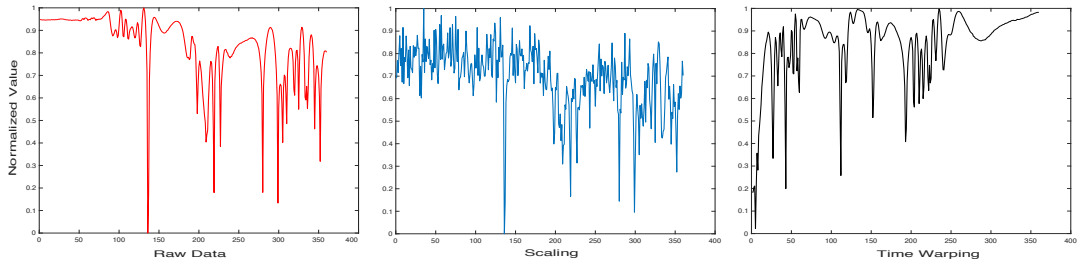


Fig. 12. Augmentation strategies applied to IQ features.

After time warping based data augmentation, we then apply the scaling data augmentation method, which changes the magnitude of data in a window by multiplying a random scalar. This approach has been traditionally used in image processing [33], wearable sensor data augmentation [78], and acoustic gesture recognition [80]. In this work, a random scalar is sampled from a Gaussian Distribution with the mean and standard deviation value of 1 and 0.1, respectively. Leveraging these two data augmentation methods, SonicFace can now cover different potential speeds of facial expressions and will be less prone to the overfitting problem. Fig. 12 shows an example of data augmentation with Scaling and Time Warping.

4.4 Recognition of Facial Expressions and Emotional Gestures

4.4.1 Beamforming. Beamforming is a signal processing technique used in the microphone array for directional signal transmission or reception; thus it has been widely used in wireless sensing applications [27, 35]. Given the 16 received acoustic signals by the UMA-16 microphone array, SoundLip [88] indicated that if combining the 16 phase profiles and 16 amplitude profiles to form the 32-channel inputs feeding into the CNN, both modalities share the same signal-noise patterns and representation capability. Therefore, we adopt the standard beamforming method to form a signal sequence and then calculate the amplitude and phase to feed into a convolution neural network for the classification task.

4.4.2 Multi-view CNN. Two popular solutions have been proposed to handle the multivariate data [63] in the past: ensemble classifiers (EC) and feature concatenation (FC). In the ensemble classifiers, separate architectures are designed for each sensing data stream to learn intra-sensor features before forming different categories, which have been adopted in many prior works [80, 88]. The feature concatenation classifier concatenates the data streams of multiple sensors at the input of the neural network to learn inter-sensor features among different streams, which have also been widely used in many applications [60, 63]. Since the amplitude (e.g., propagation attenuation) and phase (e.g., reflected angles) features have different meanings for acoustic sensing, we propose to adopt the Ensemble Classifier (EC) to learn the intra-sensor features followed by the hidden layers designed to learn different recognition information.

In addition, the 1D convolution shows an excellent performance for applications with limited labeled data and high signal variations, as well as significantly lower computational cost compared to the 2D convolutions. Thus, in our study, we first adopt two separate 1D convolutions for the amplitude and phase information to extract intra-sensor features respectively. Then, a hidden layer is connected to extract inter-sensor features of the amplitude and phase to distinguish different emotional expressions. Furthermore, a softmax activation function is used in the output layer, where the output is a probability vector indicating the likelihood of the emotional expressions to classify different emotional expressions.

4.4.3 Emotion Fusion. The outputs of the expression recognition stage are 10-class emotional expressions including 6 facial expressions and 4 emotion gestures. However, to analyze the user engagement in response of

Fig. 13. An illustrative example of typical facial expressions and emotion gestures in our dataset. For example, $\{Amusement^1, Amusement^2\}$ are separate classes in the expression recognition stage, but belongs to the same class "amusement" in the emotion fusion stage.

video-audio content, it is more practical to understand the user's emotions states instead of exact facial or hand gestures. Therefore, in the last stage of SonicFace, we combine the classes from different expressions but share the same emotion together (i.e., $\{Amusement^1, Amusement^2\}$, $\{Sad^1, Sad^2, Sad^3\}$, $\{Fear^1, Fear^2\}$).

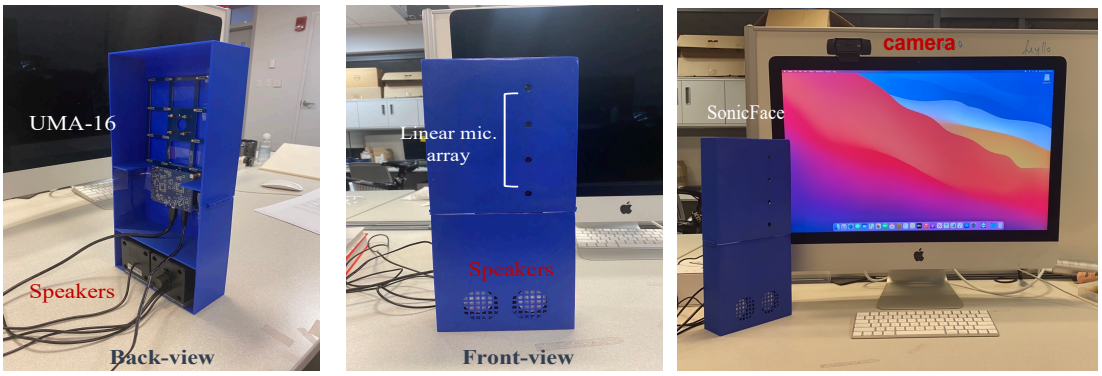
5 SYSTEM IMPLEMENTATION

5.1 Data Collection

5.1.1 Selection of Audience Engagement Events. There are six basic emotions defined by Paul Ekman [22]. In the context of at-home audience engagement analysis for movies and streaming videos, we combine both facial expressions and emotional hand-to-face gestures as the engagement-related events discussed in Section 3.1. To compose the event dataset, we select six common facial expressions adopted in [77] and four emotional hand gestures in [18] to form the fine-grained engagement events in our study, An illustration of the target facial expressions and gestures is shown in Fig. 13, which provides a visual illustrative example to the participants before the experiment, with respect to the facial emotional expressions of interest to this study. During the experiments, the participants will exhibit various facial expressions and emotional responses naturally following their own styles, while watching a wide variety of stimulating video clips (see Section 5.1.3).

5.1.2 Participants. We conducted a pilot evaluation by recruiting two groups of participants (12 in total, 2 females and 10 males, with the mean age of 26 years old) to make different facial expressions while watching the video clips. One group of participants consists of 3 students majoring in Performing Arts ("*acting participants*" in the later context), and another group is composed of 9 students from Engineering programs ("*non-acting participants*" in the later context). The rationale of recruiting and dividing two comparison groups is that, professional actors are good at managing their emotions and controlling their facial expressions, they can generate relatively consistent and repeatable expressions for different kinds of emotions. In this study, we would like to investigate the effect of facial expression consistency on system recognition performance.

5.1.3 Experimental Protocol and Data Collection. To better elicit participants' emotional states and simulate the scenario of audience engagement for films, we pre-defined a list of film clips ($2 \sim 3$ minutes) that were selected as the elicitation materials, as shown in Table A1 in the Appendix. Participants can either choose preferred clips



(a) Front and back views of the SonicFace prototype (b) Front-view of the experiment setup

Fig. 14. Experimental setup: (a) Front and back views of the SonicFace prototype using the UMA-16 microphone array. We utilize the middle 4 mics in UMA-16 to form the linear microphone array. (b) We place the SonicFace prototype next to the PC screen. The default distance between the device and participant is 40 cm. A webcam mounted on the top of monitor records the video of the participants as the groundtruth.

from the list or bring video clips prepared by themselves to help stimulate their emotional states. Before the experiment, we first asked the participant to sit in front of the desktop in a comfortable position and then watch the prepared clips for the emotion elicitation. To avoid the artifacts induced by physical muscle fatigue and ensure comfort and repeatability, we asked participants to perform the facial expressions and hand gestures 5 times with their preferable styles and styles in each session and repeat for 6 sessions. Between each session, participants were allowed to take 30 seconds break and instructed to stand up, leave the seat, and re-sit on the chair. In total, we collected 30 samples per emotional facial expression and hand gesture for each participant. The experiments were approved by the Internal Review Board (IRB) of the University at Buffalo for human subjects.

5.2 Experimental Setup

5.2.1 Hardware and Software. As shown in Fig. 14, SonicFace consists of a linear microphone array (a part of MiniDSP UMA-16 USB mic array) to receive the reflected echo signals, a dual-speaker to emit near-ultrasound, and a camera (Logitech C920) to video record participants' facial and body behaviors as the ground truth. The SonicFace prototype was placed next to the monitor that played the visual-audio content at the default distance of 40 cm. To better capture the participant's facial motions, height of the microphone array is set to 25 cm. Considering the typical form factor and hardware configurations of a commodity smart speaker, to ensure the practicability of our system, we only used a linear microphone array with 4 mics illustrated in Fig. 14(a) (Note: a detailed study investigating different numbers of mics will be discussed in Section 7.2). In the experiment, both the speaker and microphone array were connected to the desktop computer via USB connectors. SonicFace emitted custom-designed sound waves through the speaker. We developed a program in MATLAB (R2019b) to process the reflected echo signals received from the microphone array and predict the participant's emotional expressions.

5.2.2 Ground-Truth Labeling. In order to obtain the ground truth measurement of the participants' emotional states, we used the camera to video record the participant's face and body behaviors during the experiment, and utilized FacePose [6], an open-source head pose estimation and facial emotion detection tool with state-of-the-art performance. FacePose first extracted the face frames and the Practical Facial Landmark Detector (PFLD) module

[31] was then used to identify the key points of the face. The extracted key points were used to estimate the facial emotions. However, even with the SOTA performance, there were still inevitable cases of mis-annotation. Thus, we also manually checked the video clips for the outputs with low confidence score cases.

5.3 Evaluation Metrics

5.3.1 Emotion Valence and Arousal Granularity. We present six emotions in our study in 2-D coordinates based on Russell's circumplex model [66]. So that we can visualize and cluster the output based on four classes/quadrants that are separated by arousal and valence axes. The circumplex model also extends our system in different application-oriented scenarios (e.g., film makers prefer specific emotions, however, online learning providers care more about learner's overall concentration and arousal level) (the details of 2D arousal-valence model are shown in Fig. A1 in the Appendix). Given the emotional valence and arousal granularity, six basic emotions, as prototypical emotional episodes, can be located in the outer circle of the circumflex model and categorized to one of the four quadrants [67]: amusement (positive valence, positive arousal); angry (negative valence, positive arousal); sad (negative valence, negative arousal); disgust (negative valence, positive arousal); pleasure (positive valance, negative arousal); fear (negative valence, positive arousal).

To quantify the valence and arousal scales, for each prediction output, we calculate its corresponding valence and arousal scores similar to [90]:

$$S_{valence} = \max(S_{\text{amusement}}, S_{\text{pleasure}}) - \max(S_{\text{sad}}, S_{\text{disgust}}, S_{\text{angry}}, S_{\text{fear}}) \tag{7}$$

$$S_{arousal} = \max(S_{\text{amusement}}, S_{\text{fear}}, S_{\text{angry}}, S_{\text{disgust}}) - \max(S_{\text{sad}}, S_{\text{pleasure}}) \tag{8}$$

where $S_{\text{amusement}}, S_{\text{pleasure}}, S_{\text{sad}}, S_{\text{disgust}}, S_{\text{angry}}, S_{\text{fear}}$ are the classification scores of seven emotions. For example, if the prediction output has classification scores $S_{\text{amusement}} = 0.8, S_{\text{fear}} = 0.2$, and the scores for the rest emotions are zeros, we can obtain $S_{valence} = 0.6, S_{arousal} = 0.8$ from the functions above, and locate the output in the positive-valence and positive-arousal quadrant.

Therefore, to further understand the system recognition performance with respect to the emotional valence and arousal granularity, we divide all 6 emotions into 4 categories in each quadrant. Given the valence-arousal 2D plane, { "fear", "angry", "disgust" } are all in the same quadrant, and "amusement", "sad" and "pleasure" are in the rest three quadrants.

6 PERFORMANCE EVALUATION

In this section, we give a detailed description of the SonicFace's system performance. First, we present the accuracy for recognizing different emotional expressions both the user-dependent model and user-independent model with partial calibration. In details, for user-dependent model, we evaluate the performance on both standalone 10 expressions and 6 fused emotion classes, as well as valence-arousal based 2D scaling. Second, we investigate the effectiveness of coarse-grained motion filter on removing irrelevant body motions. Then, we discuss the system robustness on different device positions, environments. Lastly, we implement a long-term case study and the detailed usability analysis on different microphone array layout and audio quality measurement.

6.1 Emotional Expression Recognition Accuracy

Due to the significance of individual differences in emotional expression recognition, SonicFace first evaluates the user-dependent recognition accuracy of eight subjects. However, in real-life scenarios, it is tedious and less user-friendly to collect a large amount of data to train a robust network for each individual independently. Therefore, we present the performance of user-independent model with the calibration to mitigate the cost of data collection and accuracy improvement.

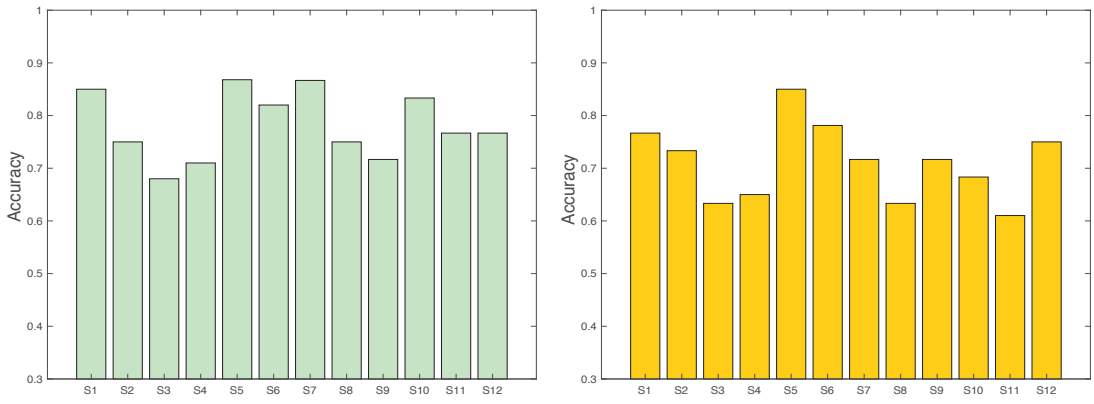
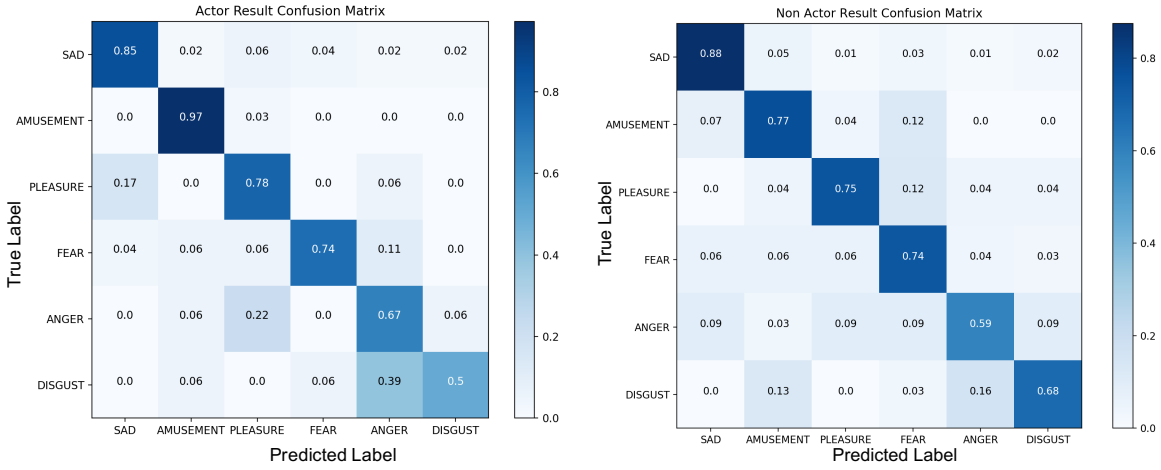


Fig. 15. Emotional expression recognition accuracy. (a) Intra-session performance. (b) Inter-session performance.

6.1.1 User-dependent Model. In this section, we evaluate and present the accuracy of the user-dependent classifier, which was trained and tested on the same user. Using the SonicFace system and the mounted camera, participants were asked to perform six different facial expressions and four emotional body gestures five times in each session and repeat for six separate sessions by following the experimental protocol in Section 5.1.3. In total, 300 samples were collected from each participant. We trained two types of classifiers: intra-session classifier and inter-session classifier.

1) Intra-Session Classification Result. To simulate the performance of recognizing the target facial expressions or emotional gestures, we performed the intra-session classification where we utilized four samples from each session as a training dataset and the remaining one sample in the same session as a testing dataset. Fig. 15(a) shows the recognition accuracy of six participants: the average accuracy is 78.6% with a standard deviation of 6.25%. The results demonstrate that the SonicFace system is able to detect different emotional expressions and thus provide immediate feedback of the user's instantaneous emotion with a reasonable accuracy. However, large variation of recognition accuracy indicates that the emotional expressions of the general population exhibit significant variance and it is less likely for participants (who are not proficient in acting) to have exactly the same and consistent emotional expressions even for the same type of stimuli. We expect to collect more data to enhance the recognition accuracy in the future.

2) Inter-session Classification Result. Maintaining a consistent performance across different sessions (e.g., re-sitting and re-playing at other times) has been a long-standing challenging problem for almost all ubiquitous sensing systems, because even a slight misalignment of these sensing systems can result in significant signal changes and performance drops [36]. In addition, as participants got more used to each stimulus, their emotional expressions would become less sensitive. To this end, we provide the performance of inter-session classification to investigate the effective performance. Here we utilize five sessions of one emotional expression from one subject as the training dataset, and the remaining one session of the same subject as the testing dataset. Fig. 15(b) shows the recognition accuracy of six participants for inter-session evaluations. The average accuracy is 71.2% with a standard deviation of 7.14%. As expected, average accuracy of the inter-session classification is lower than the one of the intra-session classification, due to the misalignment of acoustic signals across multiple sessions,



(a) (b)

Fig. 16. Confusion matrix for the recognition of six emotional classes: (a) acting participants, (b) non-acting participants.

possibly resulting from the re-sitting and re-positioning activities, as well as the changing familiarity level with the experimental stimuli.

3) Emotion Classes Recognition Result. In this section, we elaborate on recognition accuracy for different emotions (i.e., amusement, fear, pleasure, anger, disgust, sad) where the results are shown in Fig. 16. It is worth noting that emotion class "disgust" has the lowest accuracy for both the acting participants and non-acting participants (i.e., 50% for acting participants, 11.1% for non-acting participants). The predictions of "disgust" are easily mixed with the class of "angry." If we recall the emotional expression set in Fig. 13, we can observe that expressions of "disgust" and "angry" share a high similarity (i.e., lowering the cheek and winkling the nose). The main difference is the degree of mouth-opening. In addition to the high similarity, for the non-acting participants, it is challenging to precisely control facial muscles in a natural way, which would also lower the consistency of the same expression.

We also list more detailed evaluation metrics (e.g., precision, recall, F-1) for both the intra-session and inter-session assessment in Table 1. It is observed that the precision results are higher than recall for most emotions, which is aligned with our expectations. As an user engagement recognition system, precision is a better measure than recall, because false positives should be avoided as they will impact further user feedback analysis. Hence, it is more important to predict every recognized emotion correctly and precisely with the tolerance of some false negatives.

4) Valence-Arousal based Result. We show the user-dependent classification results of 6 participants with the valence and arousal scores as the coordinates, as shown in Fig. 17. The results show that our system can generally estimate the arousal and valence scales with the average accuracy of 82.22% for all participants. However, the average accuracy of acting participants (87.37%) is higher than non-acting participants (80.51%), which is consistent with our assumption that well-trained acting participants are better in controlling arousal and valence based emotion states.

5) Intra-class variations. As shown in Fig. 13, we provide an example of different facial expressions and emotion gestures. However, under our experimental protocols, participants were asked to watch film clips and perform

Table 1. Evaluation metrics over different emotional expressions

	Intra-session				Inter-session			
Emotions	Precision	Recall	F-1	Std.[2]	Precision	Recall	F-1	Std.[2]
Sad	**0.859**	0.822	0.840	0.166	**0.807**	0.778	0.792	0.200
Amusement	**0.778**	0.733	0.755	**0.223**	**0.786**	0.639	0.705	**0.238**
Pleasure	**0.814**	0.600	0.691	**0.256**	**0.556**	0.530	0.543	**0.247**
Fear	**0.786**	0.783	0.785	0.193	0.665	**0.733**	0.697	0.184
Anger	**0.657**	0.633	0.645	0.189	0.423	**0.433**	0.428	0.179
Disgust	**0.752**	0.700	0.725	**0.240**	**0.678**	0.667	0.673	**0.256**
Null[1]	0.962	**0.967**	0.965	0.036	0.962	**0.967**	0.965	0.036

[1] Null class result is estimated by event detection algorithm discussed in Section. 4.3.2.
[2] Std. represents the standard deviation of the precision scores.

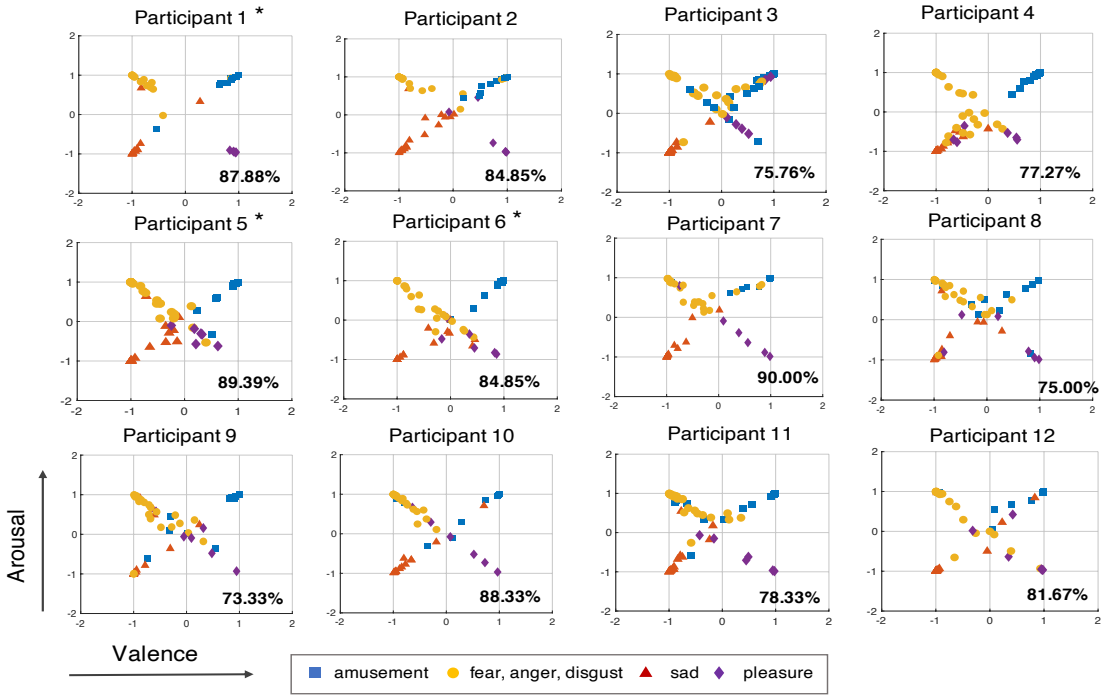


Fig. 17. Visualization of classification results for four quadrants in emotional valence and arousal 2D plane. Participants {1, 5, 6} have acting knowledge. Participants {2, 3, 4, 7, 8, 9, 10, 11, 12} have no acting background.

emotions naturally following their own styles. Given the fact that two thirds of our participants are people with no acting background, to evaluate the system performance on nuanced expressions for the same class (intra-class variations), we calculate the standard deviation of precision among different subjects in Table 1. It can be seen that, the "amusement," "pleasure," and "disgust" classes have relatively higher standard deviations among different subjects (i.e., for the same class, some users' emotions can be easily recognized while the others can not). It indicates that these three emotional expression classes have more significant inter-individual variability and inconsistency due to the users' behavioral uniqueness.

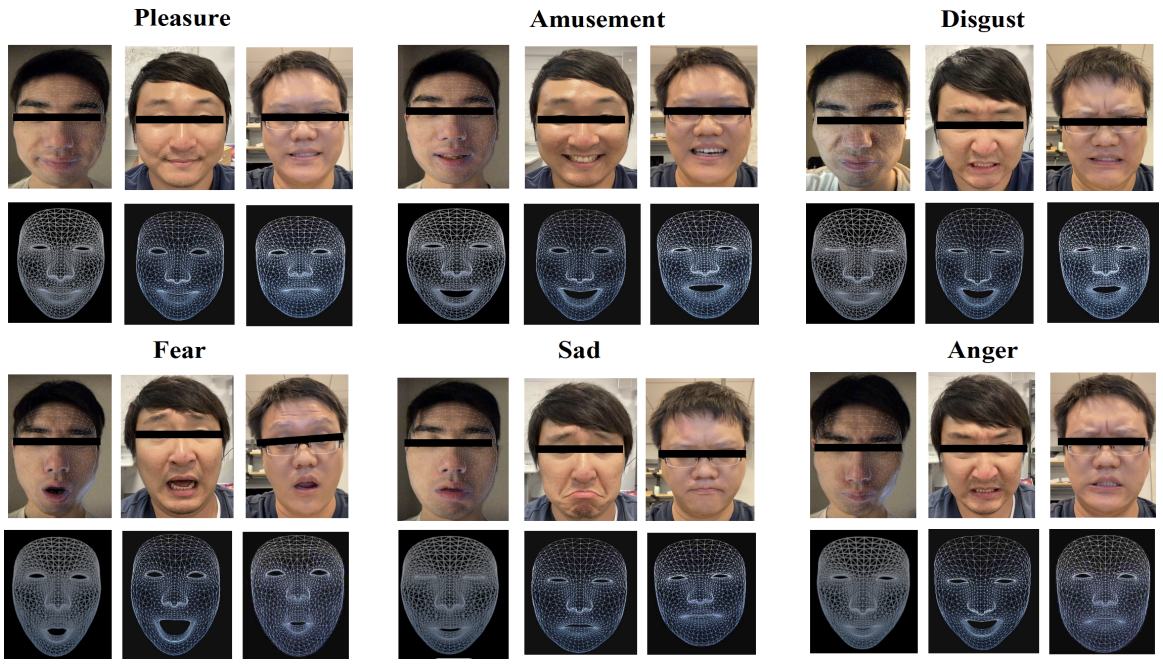


Fig. 18. An example of facial expression variations among different subjects. The top row shows the raw RGB facial images, and the bottom row shows the 3D face mesh models.

6.1.2 User-independent Model with Calibration. Given the individual differences (e.g., face geometries and behavior uniqueness) in emotional expressions as shown in Fig. 18, even for the same class, different individuals might generate different feature patterns. Thus, to investigate the generalizability of our system, we evaluate the user-independent performance with calibration by adopting a small amount of the new user's data for user calibration and fine-tuning the model. To obtain the baseline model for the user-independent analysis with no calibration process, we implement 12-fold cross-validation and train a model 12 times separately. Each time, we train the model based on the data of 11 participants and test this model on the data of the remaining one participant. Then, to further analyze how the calibration would improve the recognition performance, we repeat the leave-one-user-out evaluation by gradually adding different amounts of a new user's data into the model for training and tuning. Fig. 19 reports the results when adding 6, 12, 18, and 24 samples per class from the new user into the baseline model with no calibration. Given the large physical and behavioral variations among each individual and the limited data for training the baseline model, the recognition accuracy is really low (39%) when only 6 samples per class are collected for calibration. However, with the increasing number of calibration samples, we can observe a considerable performance increase to 54%, 64%, 72%, respectively.

6.2 Effectiveness of the Coarse-grained Motion Filter

As discussed in Section 3.2.3, the acoustic signal in the proposed SonicFace system is made up of the 20 kHz pure tone and 16-19 kHz FMCW signals. Its corresponding feature extraction pipelines can be divided into two parts: the fine-grained branch (pure tone) for facial expression recognition and the coarse-grained branch (FMCW) for irrelevant motion removal, as shown in Fig. 5. To understand the effectiveness of the proposed combination scheme and the FMCW-based motion removal strategy, we conducted a validation study in which we asked participants (N=2) to perform four different emotional hand-to-face gestures as shown in Fig. 13, some

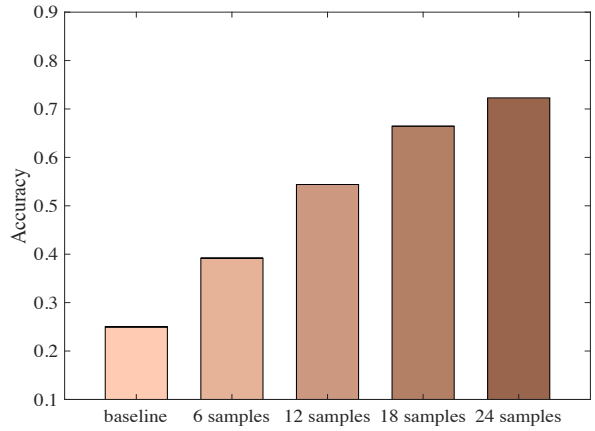


Fig. 19. The accuracy over increasing samples for calibration.

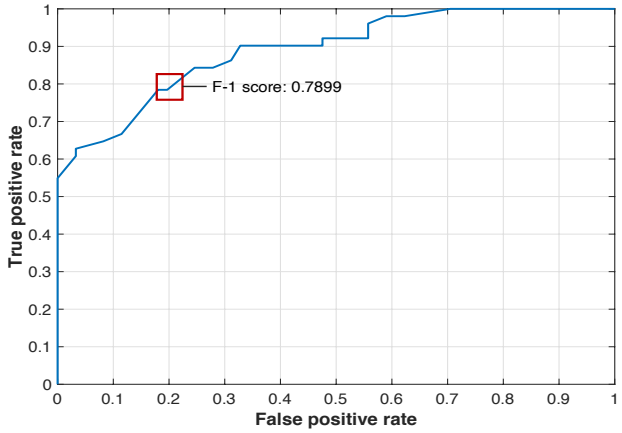


Fig. 20. ROC curve for emotional gesture detection.

common irrelevant body motions (e.g., leaning backward and forward), and hand movements (e.g., moving the mouse) in three sessions. In each session, participants were instructed to randomly and repeatedly perform the aforementioned body gestures or hand motions at a comfort level. To better segment the motions and mimic the real-world scenario of watching the film, participants were allowed to wait for 3-5 seconds between each motion. In total, we collected 186 common irrelevant motions (i.e., 54 lean-backward, 54 lean-forward, 78 move-mouse) and 112 emotional gestures (i.e., 42 hand-touch-head, 22 hand-cover-eyes, 24 hand-cover-mouth, 24 wipe-tear) for two participants. To validate the FMCW-based motion removal performance, we plot the receiver operating characteristic (ROC) curve in Fig. 20 by calculating the false positive rate (i.e., irrelevant motions were wrongly recognized as emotional hand-to-face gestures) and true positive rate (i.e., emotional gestures were correctly detected). We changed the amplitude threshold and duration of the distance for motion filtering. To achieve a balanced performance, we selected the optimized filter threshold with the highest F-1 score (0.7899).

Fig. 21. Heatmap of SonicFace's recognition accuracy with respect to different placements of the SonicFace.

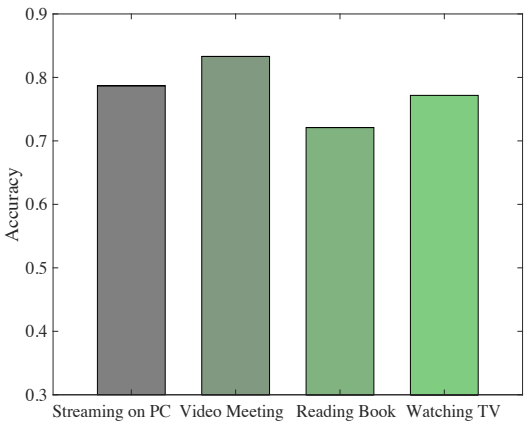
6.3 Robustness Quantification

To investigate the robustness of SonicFace in real-life application scenarios, we conducted a set of evaluations of performance with respect to different device placements and different environmental settings.

6.3.1 Impact of Device Placement. As the sound wave intensity decreases in inverse proportion to the squared distance between two objects, as well as due to multi-path channel fading, the reflected signal quality may vary depending on the specific locations. In previous evaluations with the default experimental setting, we asked participants to stay 40 cm away from the device. To investigate the impacts of the device placement and evaluate the corresponding recognition performance, we implemented a set of robustness tests for different positions in the hot zone (i.e., locations where the SonicFace will be placed) on the table. In the experiment, we divided the hot zone into six sub-zones corresponding to different distances (40 cm and 60 cm) and orientations ($\theta = 0°, -45°, 45°$). We placed the SonicFace in each sub-zone respectively and asked the participants (N=2) to follow the same procedure in Section 5.1.3 to make different facial expressions and emotional gestures. For all of the tests under different positions, we asked participants to face towards the monitor ($\theta = 0°$). We plot the heatmap of average recognition accuracy for all emotion categories in 6 different sub-zones, which are illustrated in Fig. 21. The results demonstrate that the highest accuracy locates in zone 5 ($d = 40cm, \theta = 0°$) and the rest zones show relatively lower performance degraded by $4\% \sim 10\%$ of accuracy. This observation is consistent with the hypothesis that, a higher recognition accuracy is achieved when the SonicFace is placed right in front of the user, facing towards the user's face and upper body region, at a smaller distance. The performance will slightly drop as the distance increases due to the loss of the signal intensity. Moreover, the acoustic signal's covered area of the face varies resulting from different aiming angles of the device which may also lead to performance degradation.

6.3.2 Impact of Application Scenarios. Multi-path interference caused by reflections from unrelated objects is a major challenge for acoustic sensing, especially at home or in the workplace. In this section, we aim to evaluate the effectiveness of SonicFace in recognizing emotional expressions under different real-world environmental settings. We conduct experiments in four common scenarios for daily usage of smart speaker, as shown in Fig. 22(a): 1) the user is watching the video streaming on PC in a room (8 meters wide × 10 meters long) at a normal ambient noise level (i.e., 40 dB); 2) the user is participating in a video conference with a laptop in a room (4 meters wide × 6 meters long); 3) the user is reading a book on the table with the downward face while SonicFace is placed nearby and playing the music; and 4) the user is watching TV on the sofa in a living room (4 meters

(a) Environment settings under different scenarios



(b) Accuracy in different scenarios.

Fig. 22. The impact of different application scenarios and environment settings.

wide × 7 meters long) with SonicFace placed on the coffee table. We recruited two participants to perform the emotional expressions (i.e., six facial expressions and four emotional body gestures) with 30 times repetition in four scenario settings separately. We evaluated the intra-session performance and presented the average accuracy of two subjects in different environmental settings in Fig. 22(b). The "reading book" scenario has the lowest accuracy of 72% because of the subtle head movements during reading and limited acoustic coverage of the downward face. Overall, it is observed that the accuracy in all scenarios reaches above 70%, which indicates the robustness of SonicFace system and its potential to be deployed in common daily scenarios for smart speaker users.

6.4 Long-term Case Study

In Section 6.1, we evaluated the recognition accuracy for session-based (1-2 minutes per session) emotional expression recognition. Those expressions and emotional states were detected and analyzed individually. To further evaluate the practicability of SonicFace, we conducted an extended experiment with one participant by watching a 20-minute Youtube video titled “Harry Potter and The Chamber of Secrets - Best Moments” [86]. We asked the participant to sit and watch the entire video in a comfortable position (e.g., leaning backward, keeping his face neutral, etc.). Then we used the model, which was trained using the data collected from this participant previously, to predict emotions in successive time series. The overall recognition accuracy for six emotions is 77.42%. The predicted emotions and their corresponding arousal scales are shown in Fig. 23. The white blocks indicate the time-step with neutral faces or predicted emotions with low confidence scores, as well as undesired body motions. It shows that the estimated emotions are aligned with every key film scene and match well with different scene types (e.g., fear matches the shocking scene; pleasure and amusement match the emotional scene).

6.5 Usability Analysis

6.5.1 Generalizability on Circular Microphone Array. In this study, we designed the prototype by utilizing dual speakers and a uniform rectangular microphone array (MiniDSP UMA-16) with 16 MEMS microphones (four of which in a line were used as the default setting), as shown in Fig. 14(b). As discussed in Section. 3.3, UMA-8 is another widely used commercial microphone array which is similar to Amazon Echo devices, it consists of 7 MEMS microphones, with 1 mic at the center and 6 mics uniformly spaced at the circumference of a circle with

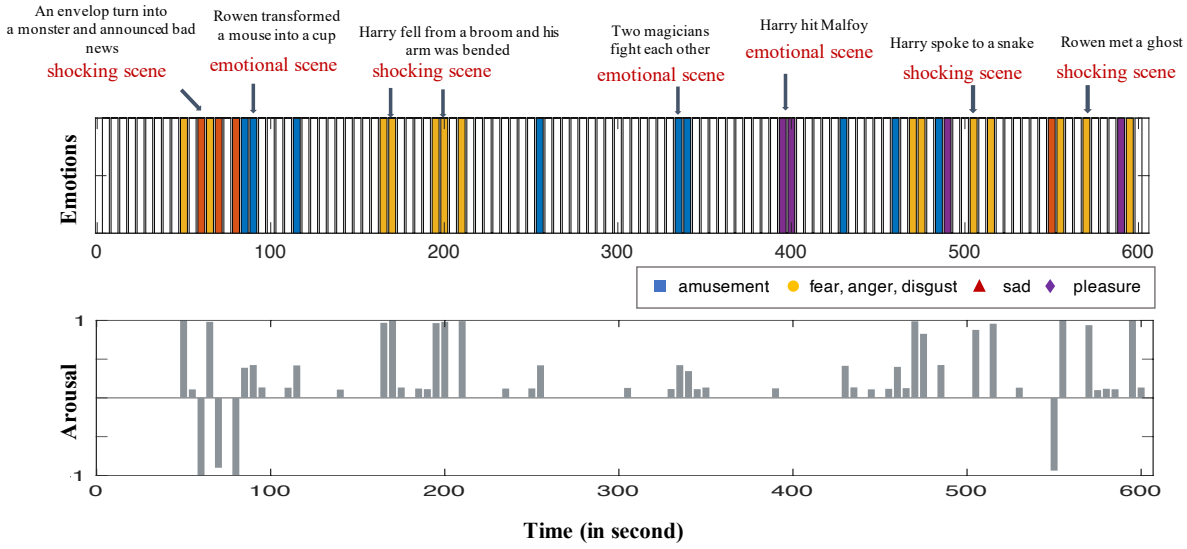
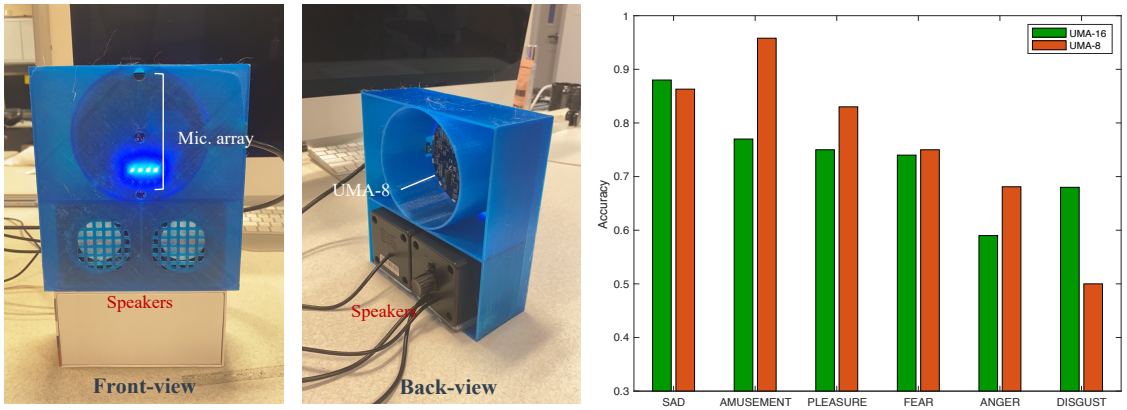


Fig. 23. Emotion recognition results and corresponding arousal scales for a video clip of 10 minutes.



(a) (b)

Fig. 24. a) Front and back views of the SonicFace prototype using the UMA-8 microphone array; (b) comparison of emotions recognition accuracy between the UMA-16 and UMA-8 based prototypes.

radius 42 mm. To better investigate the generalizability of the proposed system with different hardware settings, we also explored the SonicFace system by using the uniform circular microphone array (MiniDSP UMA-8), which was also widely used in off-the-shelf smart speakers. We asked two participants to perform facial expressions and hand-to-mouth gestures following the same protocol in Section 5.1.3. Fig. 24(b) shows the performance over six emotions using the UMA-8 microphone array in comparison with the result of the UMA-16 array. It is shown that the performance of using the uniform rectangular microphone array is approximately comparable with the circular microphone array, which indicates that the SonicFace system has good generalizability on popular microphone array configurations.

Table 2. Audio quality loss between the original audio and the audio mixed with near-ultrasound

	Video Clip[1]		**Video Clip**[2]		**Video Clip**[3]		**Video Clip**[4]	
	Original	Mixed	Original	Mixed	Original	Mixed	Original	Mixed
THD[5](dB)	-17.37	-16.39	-22.41	-18.98	-25.29	-23.33	-22.87	-19.86
SNR[6](dB)	13.39	12.94	21.21	17.53	21.01	20.68	18.33	17.97

[1] Source film - *Mission: Impossible (Brian De Palma 1996)* in Table A1.
[2] Source film - *Gladiator (Ridley Scott 2000)* in Table A1.
[3] Source film - *Saving Private Ryan (Steven Spielberg 1998)* in Table A1.
[4] Source film - *Harry Potter (Chris Columbus 2001)* in Table A1.
[5 & 6] Lower THD, higher SNR, better audio quality.

6.5.2 Audio Quality Measurement after Mixing Near-ultrasound. Theoretically, the mixed FMCW and pure tone signal have no frequency conflict with the background sound in movies and videos (which is typically less than 5 kHz [11]). However, as the audience engagement assessment solution for entertainment, it is necessary to investigate the effect of mixed near-ultrasound signals on the audio quality. In this section, we measured the audio quality loss of the mixed near-ultrasound signal and background sound based on the Total Harmonic Distortion (THD) and Signal-to-Noise-Ratio (SNR). THD measures the total amount of distortion produced by the speaker as the sound is produced from the original signal to audible sound. SNR is another common evaluation metric of the fidelity of audio to noise. We chose four video clips from Table A1 and extracted the background sounds as original references. We implemented the experiment by playing the original reference sound and the near-ultrasound mixed sound via the same speaker. A nearby microphone (distance = 0.5 m) recording the sound was simulated as the audience. As shown in Table 2, we listed THDs and SNRs for three different types of video clip background sounds. It can be observed that there is a slight audio quality loss of the mixed sound, ranging $0.98 \sim 3.43$ dB and $0.33 \sim 3.68$ dB for THD and SNR, respectively. This quality drop can be potentially improved by exploring different audio encoding and mixer techniques.

7 LIMITATIONS AND FUTURE WORK

7.1 Expanding Emotional Gestures and Postures Set

In this work, as discussed in Section 5.1.1, we selected six basic facial expressions and four hand-to-face gestures, which only cover a small portion of facial emotional gestures and body postures. To improve the SonicFace with better usability for in-the-wild studies, we will further collect data covering more different kinds of emotional gestures.

7.2 Exploring Fine-grained Resolution with More Microphones

Considering the typical design configurations of modern off-the-shelf smart speakers, we only utilized a microphone array with 4 mics in a line to capture the AoA-distance profile of the reflected echoes. However, with the increasing number of microphones in the array, it would help increase the spatial resolution of the pseudospectrum, which makes it possible to extract finger-level movements and even reconstruct hand gestures. We aim to keep improving the system by optimizing the number of involved microphones in the array with the minimum size as future work.

7.3 Enhancing Learner Engagement in Online Learning

As online learning has shown significant growth in recent years, the privacy issue has become one of the key concerns in online learning [14]. Compared with the camera-based online monitoring methods, this study can be

potentially used to assess the learner's engagement during the educational activities, which will significantly improve productivity and learning outcomes without sacrificing the user's privacy. However, compared with the six basic emotions investigated in this study, the learner engagement is more complicated to model [20], which includes facial expressions, gestures and postures, and eye movements. In the future study, instead of extracting coarse-grained features of body motions from FMCW signals, we plan to reconstruct hand gestures and body postures for a fine-grained learner's engagement analysis.

8 CONCLUSION

In this study, we propose the SonicFace, a ubiquitous emotional expressions recognition system for audience engagement analysis. The proposed approach leverages inaudible sonic wave mixed with background video sound to capture and recognize the fine-grained facial expressions and hand-to-mouth gestures. We conducted a conprehensive study with six common facial expressions and four emotional gestures, based on 12 participants. Our system can achieve the intra-session accuracy of 78.6% and the valence-arousal-based 4-class recognition accuracy of 82.22%. To evaluate the robustness and generalizability of SonicFace in real-world scenarios, we experimented with various configurations (e.g., device positions, environmental settings, and microphone array layouts.) and examined a real-life case study for online user emotion recognition with an extended period of movie watching. It is expected that this work could provide an exploratory foundation for future research on audience engagement and emotion recognition through privacy-preserving ubiquitous sensing.

REFERENCES

[1] 2018. https://www.apple.com/homepod-2018/. [Online; accessed 20-July-2021].
[2] 2020. https://www.nielsen.com/us/en/solutions/measurement/television/. [Online; accessed 23-April-2021].
[3] 2020. https://www.rottentomatoes.com/. [Online; accessed 23-April-2021].
[4] 2020. https://www.imdb.com/. [Online; accessed 23-April-2021].
[5] 2020. https://www.techhive.com/article/3516312/amazon-echo-studio-review.html. [Online; accessed 20-July-2021].
[6] 2020. https://github.com/WIKI2020/FacePose_pytorch. [Online; accessed 10-May-2021].
[7] 2021. https://go.affectiva.com/affdex-for-market-research. [Online; accessed 19-Oct-2021].
[8] 2021. https://imotions.com/blog/how-facial-expressions-analysis-fea-can-be-done-remotely/. [Online; accessed 19-Oct-2021].
[9] 2021. https://www.digitaltrends.com/smart-home-reviews/nest-mini-review-2/. [Online; accessed 20-July-2021].
[10] Anup Agarwal, Mohit Jain, Pratyush Kumar, and Shwetak Patel. 2018. Opportunistic sensing with MIC arrays on smart speakers for distal interaction and exercise tracking. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6403–6407.
[11] Abdullah I Al-Shoshan. 2006. Speech and music classification and separation: a review. *Journal of King Saud University-Engineering Sciences* 19, 1 (2006), 95–132.
[12] Kamran Ali, Mohammed Alloulah, Fahim Kawsar, and Alex X. Liu. 2021. On Goodness of WiFi based Monitoring of Sleep Vital Signs in the Wild. *IEEE Transactions on Mobile Computing* 01 (May 2021), 1–1. https://doi.org/10.1109/TMC.2021.3077533
[13] Mason Bretan, Guy Hoffman, and Gil Weinberg. 2015. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human-Computer Studies* 78 (2015), 1–16.
[14] Bo Chang. 2021. Student privacy issues in online learning environments. *Distance Education* 42, 1 (2021), 55–69. https://doi.org/10.1080/01587919.2020.1869527
[15] Chen Chen, Ke Sun, and Xinyu Zhang. 2021. ExGSense: Toward Facial Gesture Sensing with a Sparse Near-Eye Sensor Array. In *Proceedings of the International Conference on Information Processing in Sensor Networks (IPSN '21)*. 1–13.
[16] Yanjiao Chen, Runmin Ou, Zhiyang Li, and Kaishun Wu. 2020. WiFace: Facial Expression Recognition Using Wi-Fi Signals. *IEEE Transactions on Mobile Computing* (2020).
[17] Elizabeth A. Clark, J'Nai Kessinger, Susan E. Duncan, Martha Ann Bell, Jacob Lahne, Daniel L. Gallagher, and Sean F. O'Keefe. 2020. The Facial Action Coding System for Characterization of Human Affective Response to Consumer Product-Based Stimuli: A Systematic Review. *Frontiers in Psychology* 11 (2020), 920. https://doi.org/10.3389/fpsyg.2020.00920
[18] C. Corneanu, F. Noroozi, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari. 5555. Survey on Emotional Body Gesture Recognition. *IEEE Transactions on Affective Computing* 01 (Oct 5555), 1–1. https://doi.org/10.1109/TAFFC.2018.2874986
[19] Zhiwei Deng, Rajitha Navarathna, Peter Carr, Stephan Mandt, Yisong Yue, Iain Matthews, and Greg Mori. 2017. Factorized Variational Autoencoders for Modeling Audience Reactions to Movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR '17). IEEE, Honolulu, HI, USA, 2577–2586. https://doi.org/10.1109/CVPR.2017.637

[20] M Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. 2019. Engagement detection in online learning: a review. *Smart Learning Environments* 6, 1 (2019), 1–20.

[21] Paul Ekman. 1982. Methods for measuring facial action. *Handbook of methods in nonverbal behavior research* (1982), 45–90.

[22] Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion* 98, 45-60 (1999), 16.

[23] Paul Ekman, Wallace V. Friesen, and Phoebe Ellsworth. 1972. CHAPTER XIII - What Emotion Categories Can Observers Judge from Facial Behavior? In *Emotion in the Human Face: Guidelines for Research and An Integration of Findings*, Paul Ekman, Wallace V. Friesen, and Phoebe Ellsworth (Eds.). Pergamon General Psychology Series, Vol. 11. Pergamon, 57–65. https://doi.org/10.1016/B978-0-08-016643-8.50024-0

[24] Julien Fleureau, Philippe Guillotel, and Izabela Orlac. 2013. Affective benchmarking of movies based on the physiological responses of a real audience. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 73–78.

[25] Wallace V Friesen. 1973. *Cultural differences in facial expressions in a social situation: An experimental test on the concept of display rules*. Ph.D. Dissertation. ProQuest Information & Learning.

[26] Crystal A Gabert-Quillen, Ellen E Bartolini, Benjamin T Abravanel, and Charles A Sanislow. 2015. Ratings for emotion film clips. *Behavior Research Methods* 47, 3 (2015), 773–787.

[27] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-Based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 80 (Sept. 2020), 27 pages. https://doi.org/10.1145/3411830

[28] Arindam Ghosh and Giuseppe Riccardi. 2014. Recognizing Human Activities from Smartphone Sensor Signals. In *Proceedings of the 22nd ACM International Conference on Multimedia* (Orlando, Florida, USA) *(MM '14)*. ACM, New York, NY, USA, 865–868. https://doi.org/10.1145/2647868.2655034

[29] Guillaume Gibert, Martin Pruzinec, Tanja Schultz, and Catherine Stevens. 2009. Enhancement of Human Computer Interaction with Facial Electromyographic Sensors. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7* (Melbourne, Australia) *(OZCHI '09)*. ACM, New York, NY, USA, 421–424. https://doi.org/10.1145/1738826.1738914

[30] T Lee Gilman, Razan Shaheen, K Maria Nylocks, Danielle Halachoff, Jessica Chapman, Jessica J Flynn, Lindsey M Matt, and Karin G Coifman. 2017. A film set for the elicitation of emotion in research: A comprehensive catalog derived from four decades of investigation. *Behavior Research Methods* 49, 6 (2017), 2061–2082.

[31] Xiaojie Guo, Siyuan Li, Jinke Yu, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. 2019. PFLD: A Practical Facial Landmark Detector. arXiv:1902.10859 [cs.CV]

[32] Kunal Gupta, Jovana Lazarevic, Yun Suen Pai, and Mark Billinghurst. 2020. AffectivelyVR: Towards VR Personalized Emotion Recognition. In *26th ACM Symposium on Virtual Reality Software and Technology* (Virtual Event, Canada) *(VRST '20)*. ACM, New York, NY, USA, Article 36, 3 pages. https://doi.org/10.1145/3385956.3422122

[33] Søren Hauberg, Oren Freifeld, Anders Boesen Lindbo Larsen, John Fisher, and Lars Hansen. 2016. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *Artificial Intelligence and Statistics*. PMLR, 342–350.

[34] Eran Hof, Amichai Sanderovich, Mohammad Salama, and Evyatar Hemo. 2020. Face Verification Using mmWave Radar Sensor. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 320–324.

[35] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2019. BeamBand: Hand Gesture Sensing with Ultrasonic Beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) *(CHI '19)*. ACM, New York, NY, USA, 1–10. https://doi.org/10.1145/3290605.3300245

[36] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture Sensing Using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) *(CHI '19)*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300506

[37] Carroll E Izard. 1994. Innate and universal facial expressions: evidence from developmental and cross-cultural research. (1994).

[38] Carroll E Izard. 2013. *Human emotions*. Springer Science & Business Media.

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

[40] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D Human Pose Construction Using WiFi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (London, United Kingdom) *(MobiCom '20)*. ACM, New York, NY, USA, Article 23, 14 pages. https://doi.org/10.1145/3372224.3380900

[41] Crescent Jicol, Chun Hin Wan, Benjamin Doling, Caitlin H Illingworth, Jinha Yoon, Charlotte Headey, Christof Lutteroth, Michael J Proulx, Karin Petrini, and Eamonn O'Neill. 2021. Effects of Emotion and Agency on Presence in Virtual Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 529, 13 pages.

https://doi.org/10.1145/3411764.3445588
[42] Hideo Joho, Jacopo Staiano, Nicu Sebe, and Joemon M Jose. 2011. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications* 51, 2 (2011), 505–523.
[43] Soheil Khorram, Melvin G McInnis, and Emily Mower Provost. 2019. Trainable time warping: Aligning time-series in the continuous-time domain. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3502–3506.
[44] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2020. FM-Track: Pushing the Limits of Contactless Multi-Target Tracking Using Acoustic Signals. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems* (Virtual Event, Japan) *(SenSys '20)*. ACM, New York, NY, USA, 150–163. https://doi.org/10.1145/3384419.3430780
[45] Shan Li and Weihong Deng. 2020. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* 01 (Mar 2020), 1–1. https://doi.org/10.1109/TAFFC.2020.2981446
[46] Ting Li, Yoann Baveye, Christel Chamaret, Emmanuel Dellandréa, and Liming Chen. 2015. Continuous arousal self-assessments validation using real-time physiological responses. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*. 39–44.
[47] Dawei Liang and Edison Thomaz. 2019. Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 17 (Mar 2019), 18 pages. https://doi.org/10.1145/3314404
[48] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. 2010. EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering* 57, 7 (2010), 1798–1806.
[49] Helen Coster Lisa Richwine. 2020. Analysis: Fewer movies in theaters? Big Media turns focus to streaming video. https://www.reuters.com/article/walt-disney-restructuring-streaming/analysis-fewer-movies-in-theaters-big-media-turns-focus-to-streaming-video-idUSKBN26Z09E. [Online; accessed 10-May-2021].
[50] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kunpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. 2020. Real-Time Arm Gesture Recognition in Smart Home Scenarios via Millimeter Wave Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 140 (Dec 2020), 28 pages. https://doi.org/10.1145/3432235
[51] Jialin Liu, Lei Wang, Jian Fang, Linlin Guo, Bingxian Lu, and Lei Shu. 2018. Multi-Target Intense Human Motion Analysis and Detection Using Channel State Information. *Sensors* 18, 10 (2018). https://doi.org/10.3390/s18103379
[52] Yongsen Ma, Sheheryar Arshad, Swetha Muniraju, Eric Torkildson, Enrico Rantala, Klaus Doppler, and Gang Zhou. 2021. Location- and Person-Independent Activity Recognition with WiFi, Deep Neural Networks, and Reinforcement Learning. *ACM Trans. Internet Things* 2, 1, Article 3 (Jan 2021), 25 pages. https://doi.org/10.1145/3424739
[53] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. 2019. RNN-Based Room Scale Hand Motion Tracking. In *The 25th Annual International Conference on Mobile Computing and Networking* (Los Cabos, Mexico) *(MobiCom '19)*. ACM, New York, NY, USA, Article 38, 16 pages. https://doi.org/10.1145/3300061.3345439
[54] Daniel McDuff, Rana El Kaliouby, Jeffrey F Cohn, and Rosalind W Picard. 2014. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing* 6, 3 (2014), 223–235.
[55] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083* (2010).
[56] Walter Murch. 2001. *In the Blink of an Eye*. Vol. 995. Silman-James Press Los Angeles.
[57] Michal Muszynski, Leimin Tian, Catherine Lai, Johanna D. Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel. 2021. Recognizing Induced Emotions of Movie Audiences from Multimodal Information. *IEEE Transactions on Affective Computing* 12, 01 (jan 2021), 36–52. https://doi.org/10.1109/TAFFC.2019.2902091
[58] Rajitha Navarathna, Peter Carr, Patrick Lucey, and Iain Matthews. 2017. Estimating audience engagement to predict movie ratings. *IEEE Transactions on Affective Computing* 10, 1 (2017), 48–59. https://doi.org/10.1109/TAFFC.2017.2723011
[59] Rajitha Navarathna, Patrick Lucey, Peter Carr, Elizabeth Carter, Sridha Sridharan, and Iain Matthews. 2014. Predicting movie ratings from audience behaviors. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1058–1065.
[60] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning* (Bellevue, Washington, USA) *(ICML'11)*. Omnipress, Madison, WI, USA, 689–696.
[61] Jingping Nie, Yigong Hu, Yuanyuting Wang, Stephen Xia, and Xiaofan Jiang. 2020. SPIDERS: Low-cost wireless glasses for continuous in-situ bio-signal acquisition and emotion recognition. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 27–39.
[62] Tin Lay Nwe, Foo Say Wei, and Liyanage C De Silva. 2001. Speech based emotion classification. In *Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. TENCON 2001 (Cat. No. 01CH37239)*, Vol. 1. IEEE, 297–301.
[63] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D. Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2018. Multimodal Deep Learning for Activity and Context Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 157 (Jan 2018), 27 pages. https://doi.org/10.1145/3161174

[64] Javier Ramırez, José C Segura, Carmen Benıtez, Angel De La Torre, and Antonio Rubio. 2004. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication* 42, 3-4 (2004), 271–287. https://doi.org/10.1016/j.specom.2003.10.002
[65] Rebecca D Ray and James J Gross. 2007. Emotion elicitation using films. *Handbook of Emotion Elicitation and Assessment* 9 (2007).
[66] James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161.
[67] James A Russell and Lisa Feldman Barrett. 1999. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology* 76, 5 (1999), 805.
[68] Suman Saha, Rajitha Navarathna, Leonhard Helminger, and Romann M Weber. 2018. Unsupervised deep representations for learning audience facial behaviors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1132–1137.
[69] Panneer Selvam Santhalingam, Al Amin Hosain, Ding Zhang, Parth Pathak, Huzefa Rangwala, and Raja Kushalnagar. 2020. MmASL: Environment-Independent ASL Gesture Recognition Using 60 GHz Millimeter-Wave Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 26 (Mar 2020), 30 pages. https://doi.org/10.1145/3381010
[70] Tatsuya Shibata and Yohei Kijima. 2012. Emotion recognition modeling of sitting postures by using pressure sensors and accelerometers. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 1124–1127.
[71] Siddharth Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski. 2019. Impact of Affective Multimedia Content on the Electroencephalogram and Facial Expressions. *Scientific Reports* 9 (2019). https://doi.org/10.1038/s41598-019-52891-2
[72] Fernando Silveira, Brian Eriksson, Anmol Sheth, and Adam Sheppard. 2013. Predicting Audience Responses to Movie Content from Electro-Dermal Activity Signals. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) *(UbiComp '13)*. ACM, New York, NY, USA, 707–716. https://doi.org/10.1145/2493432.2493508
[73] Gabrielle Simcock, Larisa T McLoughlin, Tamara De Regt, Kathryn M Broadhouse, Denise Beaudequin, Jim Lagopoulos, and Daniel F Hermens. 2020. Associations between facial emotion recognition and mental health in early adolescence. *International Journal of Environmental Research and Public Health* 17, 1 (2020), 330.
[74] Bharath Sudharsan, Peter Corcoran, and Muhammad Intizar Ali. 2019. Smart Speaker Design and Implementation with Biometric Authentication and Advanced Voice Interaction Capability.. In *AICS*. 305–316.
[75] Bharath Sudharsan, Sree Prem Kumar, and Rakesh Dhakshinamurthy. 2019. Ai vision: Smart speaker design and implementation with object detection custom skill and advanced voice interaction capability. In *2019 11th International Conference on Advanced Computing (ICoAC)*. IEEE, 97–102.
[76] Thales Teixeira, Michel Wedel, and Rik Pieters. 2012. Emotion-induced engagement in internet video advertisements. *Journal of Marketing Research* 49, 2 (2012), 144–159.
[77] Meike K. Uhrig, Nadine Trautmann, Ulf Baumgärtner, Rolf-Detlef Treede, Florian Henrich, Wolfgang Hiller, and Susanne Marschall. 2016. Emotion Elicitation: A Comparison of Pictures and Films. *Frontiers in Psychology* 7 (2016), 180. https://doi.org/10.3389/fpsyg.2016.00180
[78] Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data Augmentation of Wearable Sensor Data for Parkinson's Disease Monitoring Using Convolutional Neural Networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, UK) *(ICMI '17)*. ACM, New York, NY, USA, 216–220. https://doi.org/10.1145/3136755.3136817
[79] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–20. https://doi.org/10.1145/3161188
[80] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the Limit of Acoustic Gesture Recognition. *IEEE Transactions on Mobile Computing* 01 (Oct 2020), 1–1. https://doi.org/10.1109/TMC.2020.3032278
[81] Mati Wax, Tie-Jun Shan, and Thomas Kailath. 1984. Spatio-temporal spectral analysis by eigenstructure methods. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 4 (1984), 817–827.
[82] Dan Wu, Ruiyang Gao, Youwei Zeng, Jinyi Liu, Leye Wang, Tao Gu, and Daqing Zhang. 2020. FingerDraw: Sub-Wavelength Level Finger Motion Tracking with WiFi Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 31 (Mar 2020), 27 pages. https://doi.org/10.1145/3380981
[83] Qingxin Xia, Joseph Korpela, Yasuo Namioka, and Takuya Maekawa. 2020. Robust Unsupervised Factory Activity Recognition with Body-Worn Accelerometer Using Temporal Structure of Multiple Sensor Data Motifs. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 97 (Sep 2020), 30 pages. https://doi.org/10.1145/3411836
[84] Wei Xu, Zhiwen Yu, Zhu Wang, Bin Guo, and Qi Han. 2019. AcousticID: Gait-Based Human Identification Using Acoustic Signal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 115 (Sep 2019), 25 pages. https://doi.org/10.1145/3351273
[85] Tong Xue, Abdallah El Ali, Tianyi Zhang, Gangyi Ding, and Pablo Cesar. 2021. RCEA-360VR: Real-Time, Continuous Emotion Annotation in 360° VR Videos for Collecting Precise Viewport-Dependent Ground Truth Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 513, 15 pages. https://doi.org/10.1145/3411764.3445487
[86] Youtube. 2016. Harry Potter and The Chamber of Secrets - Best/Funny Moments. https://www.youtube.com/watch?v=d69uAdbrprY. [Online; accessed 10-May-2021].

[87] Fusang Zhang, Kai Niu, Jie Xiong, Beihong Jin, Tao Gu, Yuhang Jiang, and Daqing Zhang. 2019. Towards a Diffraction-Based Sensing Approach on Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 33 (Mar 2019), 25 pages. https://doi.org/10.1145/3314420

[88] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling Word and Sentence-Level Lip Interaction for Smart Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 43 (March 2021), 28 pages. https://doi.org/10.1145/3448087

[89] Tianyi Zhang, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. 2020. RCEA: Real-Time, Continuous Emotion Annotation for Collecting Precise Mobile Video Ground Truth Labels. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376808

[90] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2016. Emotion Recognition Using Wireless Signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking* (New York City, New York) *(MobiCom '16)*. ACM, New York, NY, USA, 95–108. https://doi.org/10.1145/2973750.2973762

[91] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2018. Emotion Recognition Using Wireless Signals. *Commun. ACM* 61, 9 (Aug. 2018), 91–100. https://doi.org/10.1145/3236621

[92] Bing Zhou, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2018. EchoPrint: Two-Factor Authentication Using Acoustics and Vision on Smartphones. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking* (New Delhi, India) *(MobiCom '18)*. ACM, New York, NY, USA, 321–336. https://doi.org/10.1145/3241539.3241575

A APPENDIX

A.1 Valence-arousal 2D plane

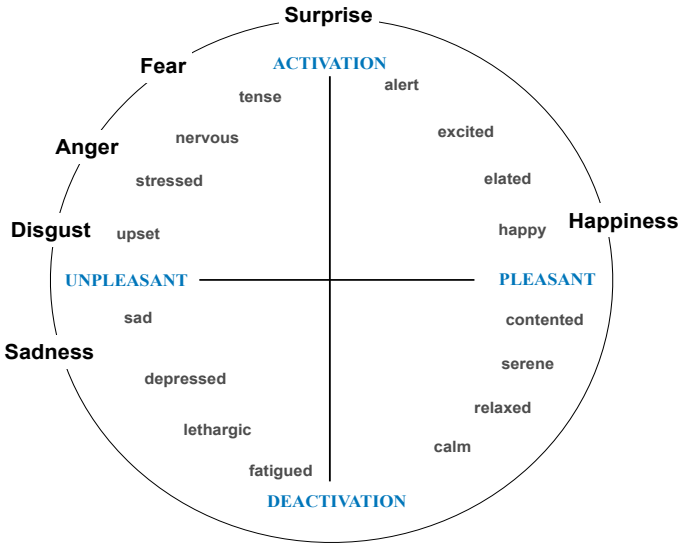


Fig. A1. Valence-Arousal 2D dimension plane [66].

A.2 Elicitation film clips

Table A1. Description of elicitation film clips [26, 30, 65, 77]

No.	Emotion	Source film	Clip description
1	Disgust	*Saving Private Ryan (Steven Spielberg 1998)*	Medium shot of a soldier lying on the floor, bleeding heavily from an abdominal wound.
2	Fear	*Scream (Wes Craven 1996)*	A young girl is running over a lawn fleeing from a masked man wielding a knife.
3	Pleasure	*Harry Potter (Chris Columbus 2001)*	A young boy is standing in the courtyard of a snow-covered school and releases a white owl into the air.
4	Fear	*Silence of the Lambs (Jonathan Demme 1991)*	A top shot shows an anxious woman praying. "I want my mommy." Followed by a cut to a man who looks down on her smiling and amused.
5	Disgust	*Silence of the lambs (Jonathan Demme 1991)*	An extreme close-up shows the open mouth of a man who has a larva removed with tweezers.
6	Pleasure	*Mission: Impossible (Brian De Palma 1996)*	A landscape shot shows a sunset behind a mountain backdrop with the rhythmic music.
7	Amusement	*Bruce Almighty (Tom Shadyac 2003)*	A man stands on the railing of his balcony and draws the moon closer with an invisible rope.
8	Sadness	*Gladiator (Ridley Scott 2000)*	A close-up shows the frightened face of a man, followed by a close-up of an alien filmed from below.
9	Anger	*Schindler's List (Steven Spielberg 1993)*	A man chooses a woman to be the housekeeper, then orders the execution of a woman in charge of the construction.
10	Sadness	*Finding Neverland (Marc Forster 2004)*	A small boy sits on a park bench next to a man dressed in black. The sad face of the boy is shown close-up.
11	Anger	*Straight Outta Compton (F. Gary Gray 2015)*	A man was harassed by the police for "looking like gang members"
12	Amusement	*Pirates of the Caribbean (Gore Verbinski 2003)*	A pirate is standing on the mast pole of his ship staring toward the port. Only through the backward movement of the camera, we can see that the ship has already sunk.
13	Disgust	*Joan of Arc (Christian Duguay 1999)*	Extreme close-up of an arrow sticking into an abdominal wound. Then a close-up of the pain contorted face of a female knight, followed by another close-up of the bleeding wound. Someone is trying to pull the arrow out of the wound. In the background, loud moaning of the injured woman can be heard.
14	Disgust	*Blade (Stephen Norrington 1998)*	In a crowded room that is completely smothered in red paint, a darkly dressed man shoots at a vampire, who then crumbles with a gurgling noise.
15	Fear	*Kill Bill II (Quentin Tarantino 2004)*	A woman is lying in a dark wooden box, only a flashlight is lighting her desperate face. Only her breath can be heard and the noise of the earth that is falling on the box in which she is trapped.
16	Fear	*Monster (Patty Jenkins 2003)*	Extreme close-up of bound hands, then a shot of a woman lying face down in a car. A man is kicking her, cursing and demanding that she screams. The woman screams and moans, contorted in pain.
17	Amusement	*What Women Want (Nancy Meyers 2000)*	Medium shot of a man. He is trying to get his legs into a woman's stockings. He has an anti-pimple plaster on his nose.
18	Amusement	*Finding Neverland (Marc Forster 2004)*	A man in a suit and bow tie sits at a dinner table in fine company and shows the children who are present a magic-trick.
19	Sadness	*Coach Carter (Thomas Carter 2005)*	Close-up of a basketball-player's sad face. Another close-up shows his disappointed team-mates. Blower's music accompanied the scene.
20	Sadness	*Moulin Rouge (Baz Luhrmann 2001)*	A top shot shows a loudly weeping young man who kneels on a bed of flowers and holds his dead lover in his arms. Slow orchestral music accompanies the scene.
21	Anger	*Crash (Paul Haggis 2004)*	Scene starts in a diner with a man talking on the phone (conversation starts with work "Look") (sets up context for later racism). A cop pulls over a black couple and sexually assaults the wife in front of her husband. Ends with the woman getting back in the car and closing door.
22	Pleasure	*Remember the Titans (Boaz Yakin 2000)*	Scene starts with coach saying "listen up, this is out time". A team wins its final football game and celebrates. End right before the music changes and the voiceover begins.
23	Pleasure	*Wall-E (Andrew Stanton 2008)*	Starts as a white robot flies forward. Two robots dance in outer space and fail in love as people in the spaceship watch and music plays. Ends when the two robots fly away together (before shot of big spaceship).