UFace: Your Smartphone Can "Hear" Your Facial Expression!

SHUNING WANG[*] and LINGHUI ZHONG[*], School of Computer Science and Engineering, Central South University, China
YONGJIAN FU, School of Computer Science and Engineering, Central South University, China
LILI CHEN, Department of Computer Science and Technology, Tsinghua University, China
JU REN[†], Department of Computer Science and Technology, BNRist, Tsinghua University, China and Zhongguancun Laboratory, China
YAOXUE ZHANG, Department of Computer Science and Technology, BNRist, Tsinghua University, China and Zhongguancun Laboratory, China

Facial expression recognition (FER) is a crucial task for human-computer interaction and a multitude of multimedia applications that typically call for friendly, unobtrusive, ubiquitous, and even long-term monitoring. Achieving such a FER system meeting these multi-requirements faces critical challenges, mainly including the tiny irregular non-periodic deformation of emotion movements, high variability in facial positions and severe self-interference caused by users' own other behavior. In this work, we present UFace, a long-term, unobtrusive and reliable FER system for daily life using acoustic signals generated by a portable smartphone. We design an innovative network model with dual-stream input based on the attention mechanism, which can leverage distance-time profile features from various viewpoints to extract fine-grained emotion-related signal changes, thus enabling accurate identification of many kinds of expressions. Meanwhile, we propose effective mechanisms to deal with a series of interference issues during actual use. We implement UFace prototype with a daily-used smartphone and conduct extensive experiments in various real-world environments. The results demonstrate that UFace can successfully recognize 7 typical facial expressions with an average accuracy of 87.8% across 20 participants. Besides, the evaluation of different distances, angles, and interferences proves the great potential of the proposed system to be employed in practical scenarios.

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

Additional Key Words and Phrases: Facial expression recognition, Acoustic sensing, Deep learning, Smartphone

ACM Reference Format:
Shuning Wang, Linghui Zhong, Yongjian Fu, Lili Chen, Ju Ren, and Yaoxue Zhang. 2024. UFace: Your Smartphone Can "Hear" Your Facial Expression!. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 22 (March 2024), 27 pages. https://doi.org/10.1145/3643546

[*] Both authors contributed equally to this research.
[†] Corresponding author.

Authors' addresses: Shuning Wang, shuning.wang@csu.edu.cn; Linghui Zhong, zlh2021@csu.edu.cn, School of Computer Science and Engineering, Central South University, ChangSha, China; Yongjian Fu, School of Computer Science and Engineering, Central South University, ChangSha, China, fuyongjian@csu.edu.cn; Lili Chen, Department of Computer Science and Technology, Tsinghua University, Beijing, China, lilichen@tsinghua.edu.cn; Ju Ren, Department of Computer Science and Technology, BNRist, Tsinghua University, Beijing, China and Zhongguancun Laboratory, Beijing, China, renju@tsinghua.edu.cn; Yaoxue Zhang, Department of Computer Science and Technology, BNRist, Tsinghua University, Beijing, China and Zhongguancun Laboratory, Beijing, China, zhangyx@tsinghua.edu.cn.

ACM 2474-9567/2024/3-ART22
https://doi.org/10.1145/3643546

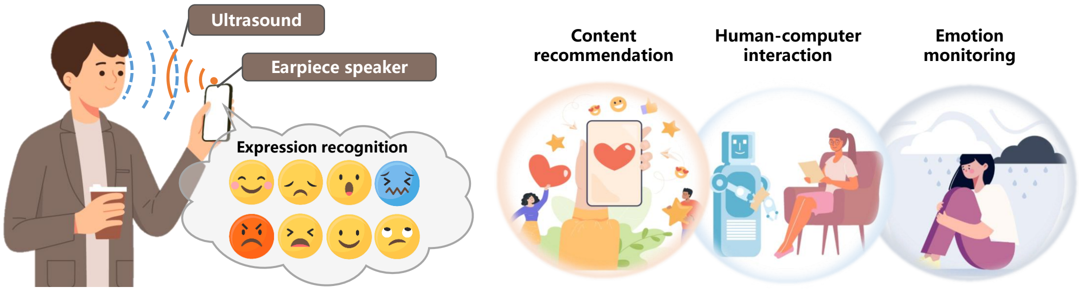


Fig. 1. Motivation examples and operating principle of UFace. The figure shows that a long-term and unobtrusive facial expression recognition system drives a variety of applications such as content recommendation, virtual reality and emotion monitoring. UFace emits ultrasound signals through the earpiece speaker on the smartphone and then extracts the expression-related signal variations from the reflected echoes received by the microphones.

1 INTRODUCTION

Affective computing has garnered wide attention from both industry and academia due to its potential to imbue computers with human-like emotions. Facial expressions, as the most natural and direct way to convey emotions, are exceptional indicators of individual psychological and physiological changes, explicitly reflecting people's emotions and states. Therefore, it is desirable to enable accurate facial expression recognition (FER) suitable for long-term monitoring, which underpins a variety of real-world applications, such as healthcare, education, entertainment, and human-computer interaction. For instance, smart homes can intelligently adjust lights and music according to the homeowner's mood for a more personalized experience. Smartphone applications can recommend content that aligns better with users' preferences based on their facial expressions. Similarly, video creators or film producers can use audiences' emotional feedback to shape their content in reverse. In education, FER offers teachers an effective way to gauge student engagement in online learning. Additionally, expression recognition can aid users in tracking their mood changes over time, which could play an essential role in the early diagnosis and intervention of mental disorders such as depression and bipolar disorder.

Existing facial expression recognition systems are achieved roughly by three categories of techniques: wearables, cameras and wireless signals. Wearables (e.g. Electromyography (EMG) [14], Electrical Field Sensing (EFS) [24], Photoplethysmography (PPG) [8]) based systems can achieve high recognition accuracy, however, they require users to wear additional devices for a long time, making them intrusive and intolerable in long-term recognition. For the purpose of user-friendliness, many researchers exploit cameras to capture face images or videos for non-contact and accurate monitoring [31, 40, 46], while the camera-based systems are sensitive to lighting conditions and may induce privacy concerns. To alleviate the aforementioned issues raised by wearables and cameras, recent work has demonstrated the feasibility of using wireless signals (e.g., WIFI [6], millimeter wave (mmWave) [43] and ultrasound [12]) to sense facial expressions, but these systems rely on extra hardwares (e.g., WIFI directional antennas, mmWave radars or smart speakers) that are not available to everyone in daily life. FacER [38] is the closest work to ours, which first explores the possibility of FER based on acoustic sensing on smartphones. However, it does not account for issues that may arise in actual use, such as the position of the phone, interference from fingers, etc., resulting in a significant limitation of their usability in real-world settings. To meet the application needs of users in real life, we envision a user-friendly, unobtrusive and ubiquitous technique that can detect facial expressions with desired accuracy for a long term.

In this paper, we propose UFace, an unobtrusive and reliable facial expression recognition system based on acoustic signals generated by commercial mobile devices (e.g., smartphones). As depicted in Figure 1, UFace utilizes the earpiece speaker on the smartphone to emit inaudible ultrasound signals toward the user's face and then captures the reflected echoes through the microphones. By analyzing the features of facial muscle changes embedded in the echoes, UFace enables ubiquitous facial expression detection without disturbing the user's original behavior, such as browsing news or watching videos on the smartphone. Although promising advances in acoustic sensing have been witnessed in various fields, such as breath detection [45], lip recognition [11, 13], target tracking [4, 7, 25], etc., recognizing facial expressions with smartphones remains a challenging task due to the following reasons:

(1) *Facial expression is a tiny irregular non-periodic deformation.* Unlike respiration and heartbeat, which are also recognized as subtle movements, facial expressions are a composite of multiple facial muscles working together in tiny movements with no apparent periodic pattern. This implies that there is a lack of opportunity to further assist signal regularization and feature extraction through prior regularity, leading to a higher probability of drowning in noise. Unlike smart speakers with linear microphone arrays, the number and positional limitations of microphones and speakers further restrict the sensing capabilities of smartphones, such as the inability to achieve 2D AoA-distance estimation as in [12], thus losing the opportunity to capture finer-grained information. Furthermore, compared to wearable device-based approaches, which have the natural advantage of being closer to the face (only a few centimeters, or contact with the skin) and relatively stationary with the face, using a smartphone in a more flexible way at a greater distance to extract such tiny non-periodic movement changes will encounter greater challenges.

(2) *High flexibility between face and smartphone.* Mobile devices have the tremendous benefit of being readily available and conveniently utilized in the way that is most comfortable for the user. However, this also introduces extreme flexibility simultaneously. When a user is using a smartphone, the relative position between his or her face and the smartphone is very variable. For example, they may be located at different distances and angles, or even the users may turn their heads to talk to others without facing the smartphone. Therefore, it is a non-trivial task to enable accurate and reliable facial expression detection under this flexible position relationship, which is also a key and unique challenge for smartphone scenarios.

(3) *Severe self-interference.* The third challenge comes from the significant interference caused by the user's own behavior, which is particularly evident in long-term monitoring scenarios. First, a non-negligible disturbance is finger-swiping, which is a natural behavior that often occurs while using a smartphone in order to obtain more content on the page. Because of its advantageous location (i.e., within 2 cm of the smartphone screen) and apparent amplitude of motion, it will dominate the effect on the acoustic signal, drowning the more distant and smaller-amplitude facial expressions in noise. In addition, other events related to the user's facial region, such as eating and talking, can also interfere with the detection of facial expressions, accordingly reducing the sensing performance of the system.

UFace is designed to tackle these challenges. To track fine-grained and non-periodic facial expressions, we first employ two microphones at the top and bottom of the smartphone, respectively, as dual receivers of echoes. This configuration provides dual viewpoints to capture more natural and comprehensive emotional movements (both face region and chest region). Based on the key insight that muscle movements of distinct facial expressions will lead to a combination of echoes from different distances, UFace exploits the feature of the distance variation pattern over time between the face and the smartphone to identify diverse facial expressions. Furthermore, aiming for a more thorough and abundant characterization, we integrate the In-phase/Quadrature (I/Q) data of the distance-time profile as an input, which comprises both amplitude and phase information. Regarding this, we design a specialized facial expression recognition network, DFNet, with dual-stream input based on the attention mechanism to maximize the utilization of the information from different viewpoints. To overcome the second issue arising from high flexibility, we design a head orientation targeting module whose essential role is to constrain the flexibility between the smartphone and the user. The key insight is that the facial expressions we desire represent the user's feedback on the smartphone's content, which means that a non-frontal face sample (i.e., the user is not looking at the phone) is not among our choices. To this end, we leverage blinking to localize the face profile, which is fed into the Support Vector Machine (SVM) algorithm to confirm the front face state. In addition, we employ a variety of data augmentation techniques to further enhance the generalization performance of the recognition system for distance and angle. Finally, to mitigate the effects of self-interference, we first develop a novel finger swipe elimination module. Specifically, we seize the opportunity behind the interference brought by finger swiping and accordingly design an algorithm based on the idea of self-elimination to implement finger swipe removal. Besides, we distinguish facial expressions from other irrelevant facial events based on the peak detection algorithm.

To summarize, our main contributions are as follows:

- We present UFace, a long-term, unobtrusive, and reliable acoustic sensing-based system for fine-grained facial expression recognition using smartphones that is capable of identifying 7 facial expressions. To the best of our knowledge, this is the first robust solution for the acoustic-based FER system via smartphones.
- We design an expression recognition model with dual-stream input based on the attention mechanism, which can leverage distance-time profiles from various viewpoints to facilitate more accurate identification.
- We design a novel reliability detection mechanism to safeguard the system from high flexibility and self-interference issues, advancing UFace one step closer to practical real-life adoption.
- We conduct extensive experiments on the commercial smartphone to demonstrate the performance of our system. The results show that UFace can recognize 7 typical facial expressions with an average accuracy of 87.8% under 20 participants. Meanwhile, UFace maintains high accuracy in a diversity of real-world scenarios, such as at various distances, angles, ambient noise, and the presence of finger swipes.

2 RELATED WORK

In this section, we review the existing related work, including acoustic sensing and facial expression sensing.

2.1 Facial Expression Sensing

FER has been widely researched because of its benefits for human-computer interaction, emotional inference, virtual reality, and content recommendation. Ekman proposed seven basic facial expressions in 1999 [9], including anger, disgust, fear, happiness, sadness, surprise, and contempt. In the past decades, computer vision has been masterfully applied to the recognition of facial expressions [41, 47]. Ryan [31] et al. proposed an automated facial expression recognition system to categorize seven expressions. Hickson [16] et al. designed an algorithm for VR headsets that can infer expressions from merely the image of the user's eyes.

With cameras requiring line-of-sight conditions and raising privacy concerns, some works have begun to utilize wearable devices to recognize expressions. NeckFace [5] is a wearable necklace that employs an infrared camera to continuously monitor expressions. A wearable device called EarIO [19] is designed to constantly detect facial expressions, allowing for low-power detection and great robustness when sitting, moving about, and reattaching the device. Electromyography (EMG) is capable of sensing skin deformation through electrodes and thus can be used to identify expressions [14, 24, 28]. Rostaminia [28] et al. proposed a system for sensing upper facial actions using commercially available Electrooculography-based glasses. It can detect five facial action units and continuously monitor pain through expression. Each of the aforementioned methods calls for users to wear specialized devices, especially the EMG-based methods, which demand that the electrodes are consistently in contact with the skin. For the majority of individuals, it is relatively unpleasant and uncomfortable to wear them for a long time.

Also for smart glasses, Xie [44] et al. leverage acoustic signals to capture facial action units. They designed an OFDM-based CSI scheme that mitigates the impacts of acoustic frequency selective fading and is able to classify six facial actions in the upper half of the face with an accuracy of 92%. Smart glasses have the potential to offer a novel approach to VR and human-computer interaction, but their view of the lower face is incredibly constrained. Smart headphones have also been shown to have the ability to infer expressions as they are embedded with microphones, speakers, IMUs, and a variety of other sensors. Takashi [1] et al. and CanalSense [2] took advantage of the physical deformation of the ear canal to identify expressions. ExpressEar [37] modified the IMU of commercial earbuds to capture subtle facial muscle movements, reaching an average recognition accuracy of 89.9% for 32 facial action units. FaceListener [32] transforms ordinary commercial headphones into acoustic sensors that recognize over 80% of expressions through knowledge distillation. However, many users don't always wear headphones, especially when browsing content.

Several studies have been presented to capture expressions by using contactless sensing. WiFace [6] proposes using the CSI of WiFi to recognize six expressions. SonicFace [12] implements contactless expression recognition on a smart speaker, capable of recognizing six expressions and four head movements by microphone arrays. mm3DFace [43] employs millimeter waves to enable a continuous 3D facial reconstruction process encompassing dynamic facial expressions. In contrast to the works mentioned above, we develop an expression recognition system for the scenario where the user is using a mobile phone without the help of any external additional devices.

2.2 Acoustic Sensing Based on Smartphone

Existing smartphones typically carry a pair of microphones and speakers located at the top and bottom of the device, which provides opportunities for acoustic sensing. Echoprint [51] exploits visual and ultrasonic signals reflected on the human face as dual authentication to verify identities on smartphones. There are many efforts to explore the acoustic tracking of smartphones [4, 7, 18, 21, 25]. FingerIO [25] achieves millimeter-level finger tracking via the phone's bottom microphone and speaker, whereas PDF [7] achieves sub-millimeter tracking. Even though they can follow a single object, like a finger, with millimeter-level accuracy, they do not apply to tracking facial expressions, which are the combined result of multiple muscle movements of the face. Qian [27] et al. extracted the heartbeat from the phase of the acoustic FMCW signal. To determine the status of the driver, Xu [45] et al. employed energy spectral density to monitor respiration. BlinkListener [22] is a blink recognition solution capable of observing blink changes in I-Q vector space that is deployed on commercial acoustic devices such as smartphones and Bela platforms. The works previously described are to extract features with periodic regularity. Facial expressions, on the other hand, are random and irregular, making them challenging to obtain using these methods. Some work also uses acoustics for lip movement recognition, which is engaged for speech enhancement [34] or silent speech recognition [11, 13] at a distance limited to a few centimeters. In contrast, we search for a method that can recognize expressions when a user is usually using a phone at a distance of roughly 25 to 50 cm. Ling [21] et al. proposed a finger motion sensing and recognition system based on the channel impulse response of the ultrasound, capable of recognizing 12 types of gestures. The modifications caused by expressions, however, are much more subtle. Therefore, it is very challenging to capture the user's expressions while browsing the phone normally (at a distance of about 25-50 cm) using only the acoustic device of the smartphone. The work most closely resembling ours is FacER [38], which similarly leverages the smartphone's microphone and speaker for expression recognition. The differentiating factor lies in our consideration of grip distance and angle challenges arising from the phone's flexibility, as well as significant self-interference from users under long-term monitoring, thereby advancing acoustic expression recognition on smartphones a significant step closer to practical implementation.

3 SYSTEM OVERVIEW

Fig. 2 gives an overview of UFace. Building and using UFace involves four steps, described as follows:

Acoustic Sensing. We implement the acoustic sensing module on the smartphone. Specifically, We leverage the earpiece speaker to emit inaudible ultrasound signals, and capture the reflected echoes using both top and

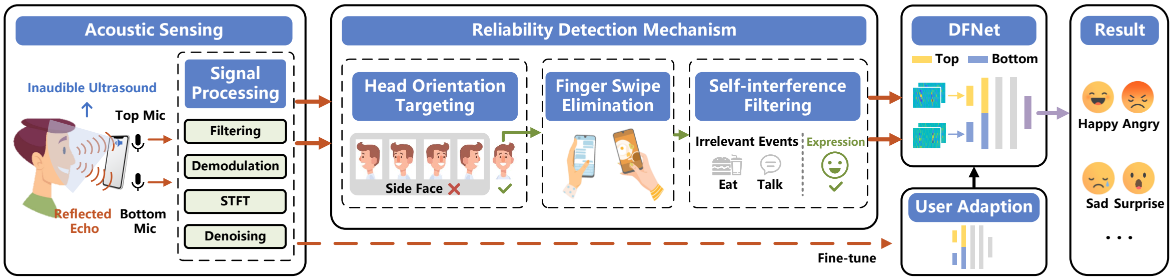


Fig. 2. The system overview of UFace.

bottom microphones. Then, specially designed signal processing algorithms are used to derive the expression pattern from the received signal. More details are in Section 4.

Reliability Detection Mechanism. Considering that strong interference from the target user himself/herself (e.g., face sideways to the smartphone screen, and other irrelevant activities such as eating or talking) may severely affect recognition performance, we develop a novel reliability detection mechanism for the data stream to facilitate credible classification. It mainly contains three key sub-modules, namely head orientation targeting, finger swipe elimination and self-interference filtering (Section 5).

DFNet. We design a tailored facial expression recognition model with dual-stream input, named DFNet, which can maximize the fusion of facial dynamic information extracted from two microphones to effectively predict expression categories. This will be discussed in detail in Section 6.

User Adaption. To personalize to new users, we apply transfer learning to fine-tune the network, which can assist the model to quickly capture individual characteristics while maintaining the learned knowledge of expressions, requiring only a small number of samples from new users (Section 6.4).

4 ACOUSTIC SENSING

In this section, we introduce the rationale for selecting signal features, the design of Frequency-Modulated Continuous Wave (FMCW) signals, and our signal processing methods.

4.1 Why Acoustic Signals?

UFace is dedicated to empowering unobtrusive, long-term facial expression monitoring on commercial mobile devices (e.g., smartphones) to enable the benefits of mood tracking and emotion understanding. To achieve this objective, we choose acoustic signals as the basis for UFace. The reasons can be summarized as follows: *(i) Low deployment costs.* Microphones and speakers that can be used to transmit and receive ultrasound are widely deployed in commercial mobile devices, resulting in no need for additional sensing equipment. *(ii) Fine-grained sensing.* Sound waves travel relatively slowly (about 340 m/s in the air). At the popular sampling rate of 48 kHz in smartphones, the resolution can reach 0.71cm, which is suitable for capturing subtle expression changes. In contrast, other commonly used wireless signals, such as WIFI and RFID-based solutions, are limited by indoor device deployment, while millimeter-wave radar is not widely utilized in mobile devices due to its high price. Consequently, a contactless and user-friendly acoustic-based system with no deployment requirements and no privacy risks is considered a more suitable choice for the target scenarios.

4.2 Speaker/Microphone Selection

A standard smartphone configuration typically has two speakers and two microphones [51], as shown in Figure 3. At the top of the smartphone, there is an earpiece speaker for answering calls and a noise-canceling microphone, while at the bottom, there is the main speaker and the main microphone.

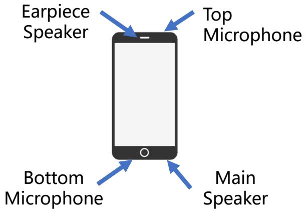


Fig. 3. Smartphone layout.

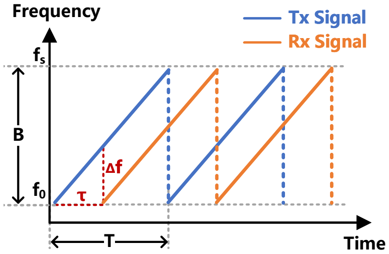


Fig. 4. Chirp signal.

Inspired by [51], we similarly choose the earpiece speaker as our FMCW signal emitter because of its advantageous location for "illuminating" the user's face. On the other hand, we observe that although the microphones on the phone are all omnidirectional [23], they possess distinct effective sensing areas. This is because signal variations induced by subtle motions are highly sensitive to propagation distance. If the signal is extremely weak when reaching the microphone due to severe attenuation, it may be buried in the noise, rendering it undetectable. Consequently, in order to capture facial expressions and accompanying unconscious body postures (such as tilting the body backward or chest heaving) [8] completely, we select a combination of top and bottom microphones to receive echoes together for a larger sensing coverage.

4.3 Signal Selection

Facial expressions are the result of the coordinated activation of multiple facial muscles, including corrugator supercilii, occipitofrontalis, orbicularis oculi, orbicularis oris, risorius, etc., and generally last from 0.4~ 5 s [10]. The following challenges exist in capturing facial expressions using acoustic sensing on smartphones. On the one hand, changes in the human face are subtle and non-periodic. On the other hand, swiping the screen while browsing the phone can interfere with the signal. Consequently, the signal features must satisfy several goals: *(i) The signal feature can fully detect the subtle and non-periodic facial muscle movements.* Expressions are primarily generated by three groups of muscle: orbital, nasal, and oral, with a variation of 1-5cm, accompanied by some mild movements of the shoulders and chest. For example, in the case of fear, in addition to lowered eyebrows, widened eyes, and slightly pulled-down mouth, there are also actions such as chin retraction and inhalation. Therefore the signal should need to be able to capture the expressions brought together by these parts. *(ii)The signal features need to be robust to interference.* Swiping the screen occurs frequently when people use their phones. Furthermore, it's unknown how far away the person is from the phone when surfing. As a result, we need the signal features to be only marginally influenced by these interferences.

We choose the FMCW signal as our transmitting signal, mainly for the following considerations: (i) FMCW is generally used in radar ranging. It can distinguish the movement of eyebrows, mouth, shoulders, and chest at distance. Besides, it can separate expressions and finger slides at distance. The distance between finger sliding and microphones is generally within 20cm, while it between the human face and microphones in browsing the phone is about 25-50cm. (ii) FMCW signals have a wide bandwidth and are hence not susceptible to multipath and frequency selective fading.

In addition, we consider several signal features, including Amplitude, Phase, Doppler shift, and Channel Impulse Response (CIR), but each of them has some limitations. The amplitude and phase of CW signals are vulnerable to multipath and frequency selective fading. Moreover, when there is finger sliding or surrounding object movement, the amplitude and phase changes brought by interfering objects and expressions will be blended together and difficult to decouple. The Doppler shift-based detection method is limited by the object's movement speed. The expression change speed is approximately 0.00125-0.1 m/s. According to the Doppler shift formula

$\Delta f = \frac{2\Delta v f_c}{c}$, the expression-induced frequency shift is about 0.14-11.56 Hz, where c is the speed of sound, Δv is the speed of facial expressions relative to the smartphone, and f_c is the signal frequency (taken here as 20 kHz). The minimum frequency resolution of Short-Time Fourier Transform (STFT) is given by $\Delta f_D = \frac{F_s}{W}$, where F_s is the sample rate and W is the window length. When the number of STFT points is set to 4096 and F_s is 48 KHz, the resulting Δf_D is 11.72 Hz. Therefore, the Doppler shift cannot fully capture the change in expression. Also, finger sliding will introduce Doppler shift and thus interference. CIR is better suited for single-target tracking and identification. In our case, it cannot differentiate between the interference caused by finger swiping and facial actions.

4.4 FMCW Signal Design

As introduced above, we choose FMCW as our transmit signal as shown in Figure 4, which is a series of chirp signals. The transmit chirp signal can be written as

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

where f_0 is the initial frequency, B is the bandwidth, and T is the duration of a chirp. Let's imagine the simplest case, where there is only one reflection path in the environment. The receiving signal can be written as

$$y(t) = \alpha cos\left(2\pi\big(f_0(t-\tau) + \frac{B}{2T}(t-\tau)^2\big)\right) \tag{2}$$

To obtain the I-part of the IF signal, we multiply the received signal by the transmitted signal before passing it through a low-pass filter. The I-part signal can be written as:

$$\begin{aligned} I(t) &= \frac{\alpha}{2}cos\left(2\pi(f_0\tau + \frac{B}{T}t\tau - \frac{B}{2T}\tau^2)\right) \\ &\approx \frac{\alpha}{2}cos\left(2\pi(f_0 + \frac{B}{T}t)\tau\right) \end{aligned} \tag{3}$$

Since the τ is very small, the last term of the equation $\frac{B}{2T}\tau^2$ can be ignored. Similarly, multiplying the received signal by a phase-rotated 90-degree transmit signal $s'(t) = sin\left(2\pi(f_0 t + \frac{B}{2T}t^2)\right)$ and passing it through a low-pass filter, we obtain the Q-part of the IF signal:

$$Q(t) \approx \frac{\alpha}{2}sin\left(2\pi(f_0 + \frac{B}{T}t)\tau\right) \tag{4}$$

Combining the I-part and the Q-part, we obtain the complete IF signal:

$$m(t) = \frac{\alpha}{2}e^{j2\pi(f_0+\frac{B}{T}t)\tau} \tag{5}$$

The distance of the reflected object can be given as

$$d = \frac{ct}{2} = \frac{cT\Delta f}{2B} \tag{6}$$

To determine the distance of the reflected object, we perform the Fast Fourier Transform (FFT) on the IF signal and thus obtain the Δf.

We set the transmit signal frequency in our system to 18−22 kHz with a bandwidth B of 4 kHz. To avoid sounding like noise to the user, the transmitted signal is placed in an ultrasonic band that is inaudible to the human ear. Given that popular smartphones' audio sampling rates F_s are 48 kHz or 44.1 kHz (for some Android phones), we set the highest frequency to 22 kHz according to Nyquist's sampling theorem. We set the number of samples in a chirp to 512 empirically.

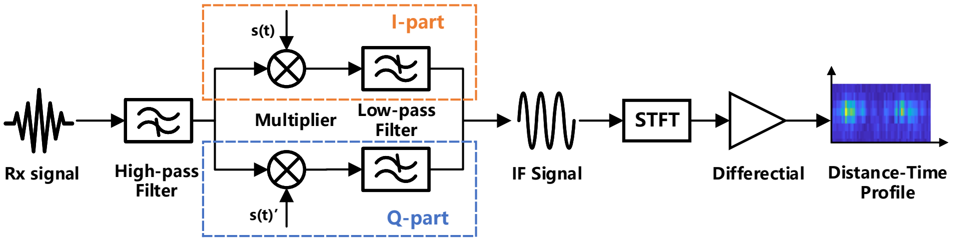
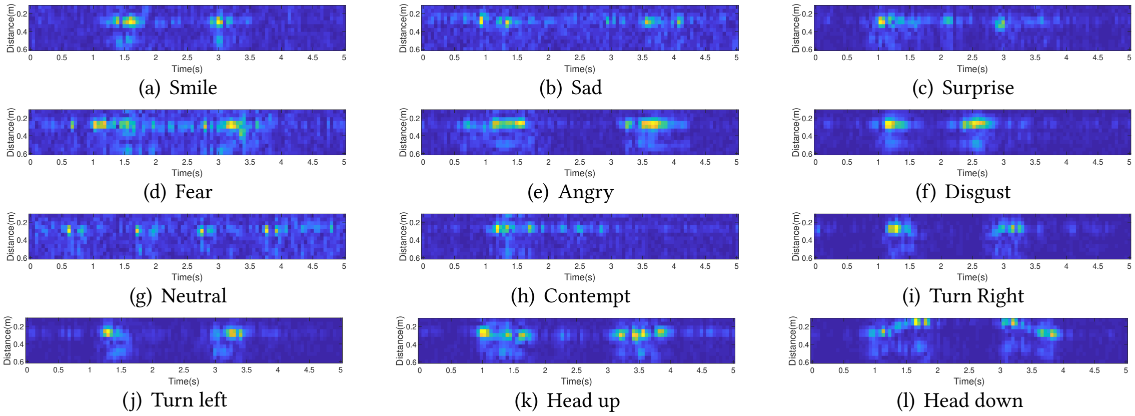


Fig. 5. The working pipeline of signal processing.



(a) Smile (b) Sad (c) Surprise

(d) Fear (e) Angry (f) Disgust

(g) Neutral (h) Contempt (i) Turn Right

(j) Turn left (k) Head up (l) Head down

Fig. 6. Feature profiles corresponding to different expressions and head movements. Note that for clearer visualization, we choose the amplitude values rather than the I/Q values for presentation, which applies to the following figures.

4.5 Signal Processing

Figure 5 details our signal processing process. In order to remove low-frequency noise, we first send the received signal through a high-pass filter. The IF signal is then obtained using the orthogonal demodulation approach mentioned above. After performing the short-time Fourier transform (STFT) on the IF signal, we are able to calculate the distance-time profile $S(f,t)$ using Equation 6. We set the window length of STFT to 512 based on the number of samples in a chirp, while oversampling the number of FFT points N_{FFT} to 1024 to extract more fine-grained information. This corresponds to a calculated bandwidth of $\Delta f = \frac{F_s}{N_{FFT}} = 46.875$ Hz for one bin. F_s is 48 kHz in our setting. Utilizing Equation 6, we can deduce a corresponding distance resolution of 2.125 cm. Given that sound travels faster through solids than through air and that the distance between the microphone and speaker is shorter than it is to the user's face, there is a significant direct interference path that is 1-2 orders of magnitude larger than the reflected path. To eliminate the interference of the direct path and better focus on the face, we chose 11-28 bins as face features, corresponding to a distance of 21.25-57.38 cm. To remove the static component, we employ the latter frame to subtract the former frame to get the $S_d(f,t)$. This differential method allows us to focus more precisely on the dynamic changes in expressions. Then, we make an average of 5 chirps to cancel ambient noise interference. The final feature sample rate is 18.75 Hz, which is sufficient for capturing expressions. After the aforementioned processing steps, we have the final feature, a distance-time profile with dimensions of 18 × 99. We extract the corresponding I/Q data of the distance-time profile as input to our network instead of merely amplitudes, since I/Q data provide a much more comprehensive characterization of expression movements, containing both amplitude and phase values.

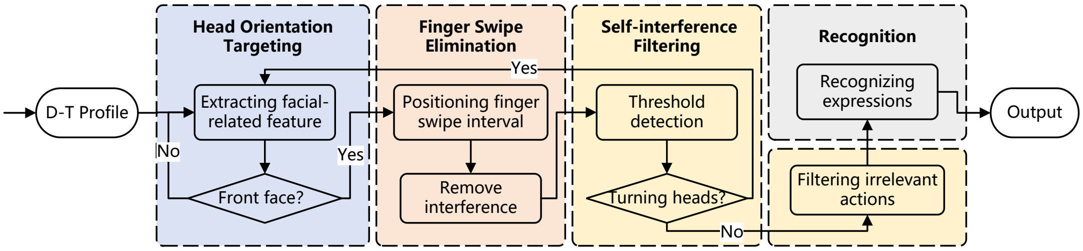


Fig. 7. The flow of the reliability detection mechanism for expression monitoring.

Algorithm 1 Head Orientation Targeting algorithm

```
Input: The Distance-Time profile $S(f,t)$, the differential Distance-Time profile $S_d(f,t)$, $N_{Th}$, $N_{In}$
Output: Head orientation
for $i = 3, 4, ..., f - 1$ do
    $Energy = sum(S_d(i-2 : i+1, :))$
    $pks = findpeaks(Energy)$
    for $j = 1, 2...length(pks) - 1$ do
        if $average(pks(j : j+1)) > N_{Th}$ and $interval(pks(j : j+1)) < N_{In}$ then $index = i$
        end if
    end for
end for
$FaceBin = S(index - 1 : index + 1, :)$
$HeadOrientation = SVM(FaceBin)$
```

The feature profiles of the various expressions are presented in Figure 6, demonstrating the possibility of differentiating between them. Notably, "smile" and "angry" are accompanied by distinguishable breathing features at a distance of 40 cm, while the movements associated with "sad" and "contempt" are more subtle.

5 RELIABILITY DETECTION MECHANISM

In this section, we delineate the structure of a reliable and long-term expression monitoring mechanism. Figure 7 shows the overall framework of the mechanism. Prior to sending the Distance-Time Profile (D-T Profile) to the network for expression recognition, the profile needs to be processed through algorithms, including head orientation targeting, finger swipe elimination, and self-interference filtering. Upon the detection of a head rotation by the threshold detection algorithm, the processing sequence undergoes a reset, reverting back to the stage of head orientation targeting. In the following, we will describe each part of the mechanism in detail.

5.1 Head Orientation Targeting

The mobility and convenience of smartphones have made them an essential part of our daily lives, but they also introduce distinct complexities for facial expression recognition.

Due to the flexibility of smartphone usage, where users hold the phone at varying distances and angles, we develop a head orientation targeting algorithm to identify the optimal viewing position for capturing facial expressions, which also serves as the appropriate moment to attempt sensing. Although we have roughly pinpointed the face's location by restricting the distance, it is not enough to accurately determine the face orientation. To overcome this problem, we need to locate the face region more precisely.

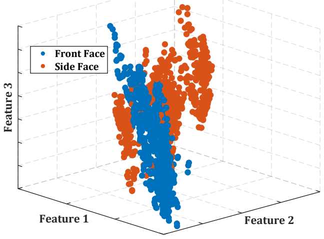


Fig. 8. Clustering result of front and side faces.

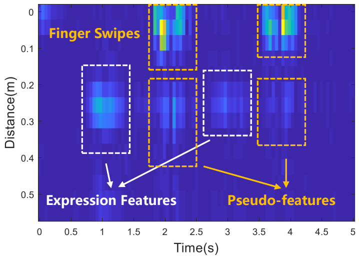
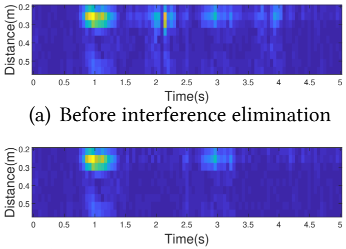


Fig. 9. Impact of finger swipe on distance-time feature profile.



(a) Before interference elimination

(b) After interference elimination

Fig. 10. Comparison of feature profiles with and w/o finger swipe elimination.

Algorithm 2 Finger Swipe Elimination algorithm

Input: The differential Distance-Time profile $S_d(f,t)$, neutral face profile $S_n(f,t)$

Output: $S_o(f,t)$

```
INITIALIZATION: $N_t$, $n$, $window$, $S_o(f,t) = S_d(f,t)$
for $i = 1 : t - n$ do
    $window = S_d(1:6, i:i+n)$
    if mean($window > N_t$) then
        $S_o(face\_region, i:i+n) = Gaussian\_noise \times S_n(face\_region, i:i+n)$
    end if
end for
```

Blinking, as a distinctive and persistent feature of the face, creates noticeable interference in the feature profile. Our key insight is that the interference caused by blinking can be transformed into a valuable feature for locating the frequency bin corresponding to the face. We observe that blinking produces two neighboring peaks in the feature profile because it consists of two distinct phases: closed and open eyes, which provides an opportunity to detect facial regions. Considering that a face's range can occupy roughly 4 bins, we employ a sliding window to combine every 4 bins into one. On the fused bins, a peak detection algorithm is applied to determine if there are two contiguous peaks larger than a threshold. Here, we set the threshold value to 5 and limit the peak interval to no more than 5. Regarding the filtered bins, we consider the bin with the highest peak to be the face region. With the specific location of the face, how do we distinguish a credible facial orientation? Due to the distinct facial geometry and structures between the front and side views, the matched original distance-time profiles will be different. The clustering results of the four bins in space are depicted in Figure 8, providing evidence that the front and side faces can be distinguished. We leverage the four bins mentioned above from the undifferentiated profile as key features. To create a more complete facial profile, we combine the features (both I and Q data) of the two microphones and obtain a 16-dimensional input. We use a Gaussian kernel Support Vector Machine (SVM) to classify the front and side faces of the user.

We consider the user's face to be inside the confidence interval and may activate expression recognition while more than 80% of the points in a frame are categorized as front faces. We are now prepared to detect expressions in this way. The specific process is summarized in Algorithm 1.

5.2 Finger Swipe Elimination

Even if the user's head is in an observable position, it is still difficult to monitor expressions while browsing the phone. This is because the user may operate the phone by swiping their finger across the screen, which can seriously interfere with the feature profile. The finger swipe drives a slight shaking of the phone, bringing a pseudo-feature on the profile that looks like a facial expression even though the face is still. Therefore, we must remove this pseudo-feature in order to avoid it from impairing expression recognition. We notice that in addition to the pseudo-features near the expression, finger swiping also introduced a prominent bump within 10 cm as shown in Figure 9, since the fingers are located at a distance of about 5 to 10 cm between the upper and lower microphones. This observation inspires us to extract the finger-swiping feature from the distance and thus eliminate its interference. We perform energy detection in the first 6 bins of the profile. We collect some finger sliding data and separate out the sliding part by the Local Extreme Value Detection (LEVD) algorithm. We calculate the average value of these features and set the threshold at 70% of that value. When energy is detected above the threshold, we assume that finger sliding has occurred. We take an 18×11 window at that place and replace it with a Gaussian distribution in a static state to remove the interference caused by the finger swiping. Figure 10 illustrates the feature profile before and after the elimination of finger swipe interference, and further validation will be shown in section 8.2.2. Algorithm 2 illustrates the detailed elimination steps.

5.3 Self-interference Filtering

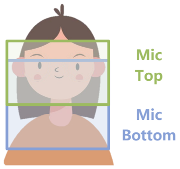
In addition to expressions, there are other facial actions may be misidentified as expressions by the DFNet, such as talking or chewing. To assure the performance of expression recognition, it is necessary to filter out these irrelevant actions. Our key observation is that expression is an irregular and low-frequency action. In contrast, chewing and speaking have a longer duration and a higher frequency of movement. Therefore, we employ peak detection algorithm to filter out irrelevant actions. Specifically, we utilize the variance on the time domain of the feature profile to represent the energy characteristics and detect peaks above the threshold. Based on the statistical analysis, we set the threshold to 0.8 of the average peak of the interfering motion. After excluding the peaks that are too small in interval, we scan the frames using a sliding window of 35. If the number of peaks within the window is greater than 7, we consider the frame to have a high-frequency motion of large amplitude and discard it. The remaining frames will be transmitted to DFNet for expression recognition. If we detect that the threshold is higher than the normal range for an expression, we can assume that the user made a pretty substantial action, such as moving their head or lifting their hand. The head orientation targeting algorithm will be restarted until we detect that the user's face is within the appropriate interval. At that time, expression monitoring will start.

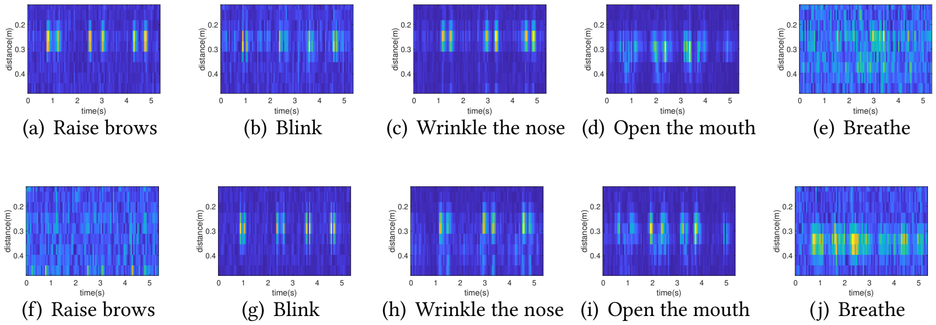
6 FACIAL EXPRESSION RECOGNITION NETWORK DESIGN

In this section, we introduce the detailed design of the proposed facial expression recognition network, DFNet.

6.1 Design Motivation

We target to achieve a reliable facial expression recognition system using acoustic signals obtained from two microphones. The crucial issue is how to fuse the information embedded in two signals to maximize the advantages of different viewpoints (i.e., top and bottom). As a widely used technique in array signal processing, the beamforming mechanism [36] seems to be the most intuitive option for multi-signal fusion. Based on the principle that a target beam from the desired direction can be formed by the weighted synthesis of the received signals, beamforming has already exhibited its unique advantages in silent speech recognition [13]. It enhances the reflected echoes from the mouth and suppresses those from other directions, leading to improved robustness of the system. However, it might not be suitable for detecting facial expressions. This is because expressions are generally the result of a combination of multiple facial muscles moving together, accordingly resulting in many reflected echoes from various directions. In this case, it is difficult to use beamforming techniques to locate a





(a) Raise brows (b) Blink (c) Wrinkle the nose (d) Open the mouth (e) Breathe

(f) Raise brows (g) Blink (h) Wrinkle the nose (i) Open the mouth (j) Breathe

Fig. 11. Sensing coverage of top and bottom microphones.

Fig. 12. Spectrograms corresponding to the five typical movements, where (a)-(e) are from the top microphone and (f)-(j) are from the bottom microphone.

specific location for signal enhancement, which may instead cause the loss of beneficial information and weaken the overall effect.

Therefore, to determine the appropriate fusion scheme, we conduct a benchmark experiment to further explore the sensing area characteristics of each of the two microphones. We select five typical movements associated with facial expressions: raising the eyebrows, blinking, wrinkling the nose, opening the mouth, and breathing, representing in turn the regions from top to bottom. The experiment is deployed as described in Section 7. Note that the participant is instructed to do each movement as independently as possible, such as refraining from blinking while breathing.

Figure 12 illustrates the sensibility of the top and bottom microphones for different parts. We can observe that the top microphone is more focused on the upper area, while the bottom microphone is more attentive to the comparatively lower area, as summarized in Figure 11. For example, raising the eyebrows three times can be clearly identified at about 0.3m in Figure 12(a); however, this is not a trivial task for the bottom microphone in Figure 12(f). When comparing Figure 12(e) and Figure 12(j), it can be seen that the bottom microphone provides a significantly better sensing ability of chest displacement than the top microphone, owing to its noticeable energy change at 0.3 m - 0.4 m. Additionally, for the middle "overlap" area, both successfully depict the movement's presence with similar but not identical variations of profiles. The key rationale for these differences is the disparity in the location and performance of the two microphones.

Aiming at acquiring more informative joint representations from multi-view data (e.g., various data modalities or data sources), multi-view learning has proven its effectiveness in computer vision, natural language processing, video analysis, etc [50]. Based on the above observations, our intention is to preserve consensus between the two microphones (i.e., corresponding to the same expression aligned in time), while leveraging the mutually complementary information from different viewpoints to gain deeper insights into subtle movements, which coincides with the core idea of multi-view learning. Therefore, inspired by multi-view learning, we design a novel dual-stream input facial expression recognition network, DFNet, which treats each distance-time profile as an image of a specific viewpoint and captures their intrinsic correlations to facilitate expression prediction.

6.2 Model Design

Our proposed DFNet model for facial expression recognition is illustrated in Figure 13, which is primarily based on the multi-view CNN architecture [33]. It consists of three major phases: feature extraction, feature fusion and expression prediction.

Define (X^T, X^B) as the input pair of DFNet, where $X \in \mathbb{R}^{D\times T}$ denotes the distance-time profile calculated from the received echoes (Section 4). Specifically, X is a stack of I data and Q data from the D-T profile. The blue part in Figure 13 demonstrates the basic structure of the feature extractor module. We first apply a dual-stream

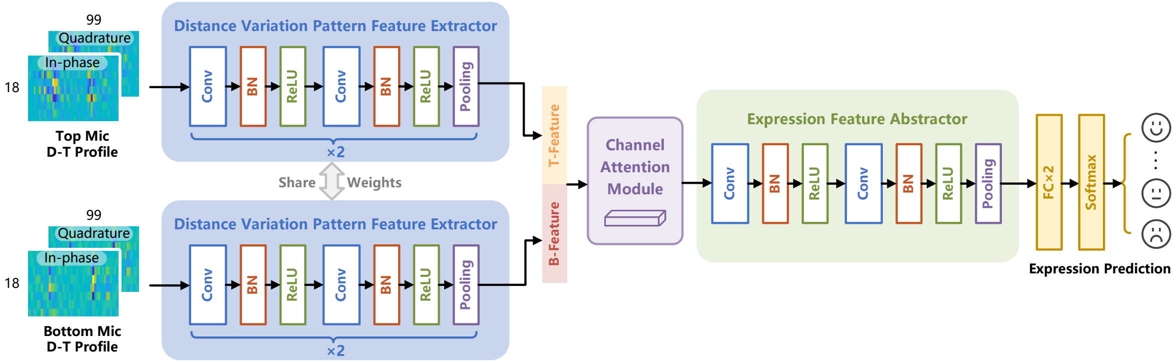


Fig. 13. DFNet model for facial expression prediction.

structure with 4 layer convolutional blocks to extract the respective distance variation pattern features of the top microphone and the bottom microphone from X^T and X^B, respectively. Notably, considering that X^T and X^B belong to the same data modality and have similar low-level spatial features, the extractor networks in the dual streams are designed to share weights. Let $F^T \in \mathbb{R}^{C\times H\times W}$ and $F^B \in \mathbb{R}^{C\times H\times W}$ denote the extracted feature representations for X^T and X^B, respectively, where C, H and W represent the channel dimension, height and width. F^T and F^B are as follows:

$$F^T = f_e(X^T;\theta_e) \tag{7}$$

$$F^B = f_e(X^B;\theta_e) \tag{8}$$

where f_e represents the feature extractor network and θ_e is the set of all its parameters.

By concatenating F^T and F^B, DFNet obtains the joint representation $F^J \in \mathbb{R}^{2C\times H\times W}$. In order to fully combine the advantages of the features F^T and F^B, we borrow the channel attention module from image processing [42], which enables adaptive re-tuning of the channel feature response by modeling the interdependencies between channels. Fortunately, this gives us an opportunity to further optimize the fusion of features, as different channels in F^J represent distinct information from the two viewpoints. The key insight is that a detectable movement implies a notable energy change, which should receive a higher weighting. Hence, by adjusting the weights given to the channels, we can focus on the more important feature areas. Figure 14 demonstrates the specific architecture of the channel attention module in DFNet, which essentially utilizes global pooling and a shared MLP network to compute the corresponding channel weights $M^C \in \mathbb{R}^{C\times 1\times 1}$ with the following formula:

$$M^C(F^J) = \sigma(MLP(AvgPool(F^J)) + MLP(MaxPool(F^J))) \tag{9}$$

where σ denotes the sigmoid function. F^C represents the final refined output of the channel attention module, which is computed as:

$$F^C = M^C(F^J) \otimes F^J \tag{10}$$

where $\otimes$ denotes element-wise multiplication.

Next, DFNet performs further abstraction of the fused features F^C by employing two-layer convolutional blocks to gain semantic information regarding the facial expressions. Followed by two fully connected layers and a Softmax layer, DFNet outputs the final expression prediction results.

The exact model specifications of DFNet are shown in Table 1. We utilize the AdamW optimizer with an L2 regularization term parameter of 0.1 to train DFNet. The other network hyperparameters are chosen empirically, such as learning rate, batch size, and number of epochs. Our code is publicly available [1] to encourage further research.

[1] https://github.com/Sulingy/UFace

Table 1. DFNet model specification.

Parameter	Input size	Conv. kernel	Pooling kernel	FC layer size	Output size	MLP ratio	Learning rate
Default	$2 \times 18 \times 99$	$256 \times 3 \times 3$	2×2	128	7	16	0.0001

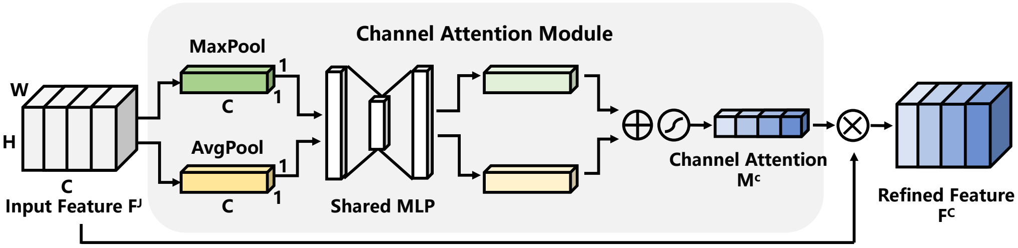
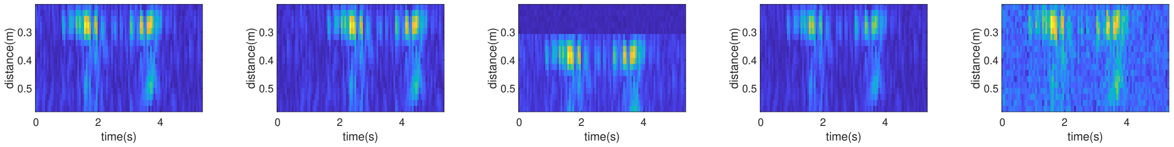


Fig. 14. Channel attention module.



(a) Original sample (b) Time-shift sample (c) Distance-shift sample (d) Scaled sample (e) Noise-added sample

Fig. 15. Examples of samples generated corresponding to the four data augmentation methods in Section 6.3.

6.3 Data Augmentation

To enhance the system's robustness in diverse real-world environments, we apply four augmentation methods to extend our dataset. Figure 15 exhibits the four augmentation samples derived from the original distance-time profile.

(1) **Time shifting**. To make facial expressions more widely distributed at any time position within a 5-second time window, we randomly shift the expression segments forward or backward along the time axis.

(2) **Distance shifting**. To improve the generalization performance of the system for different distances and angles, we exploit the augmentation technique of distance shifting. This is because distance-time profile features are sensitive to the relative position, such as distance and angle between the user and the smartphone in the real world. To assist the model training by providing more diverse training data, we shift the overall distance-time contour up or down by n bins. This generates more virtual samples at different distances by varying the location of the brightest energy row, which indicates the location of the face. Note that bin row gaps due to the offsets will be filled by randomly generated Gaussian noise to simulate the real environment.

(3) **Scaling**. To accommodate the complexity of real-world environments, the magnitude of the data in the distance-time profile is varied by multiplying by a series of scaling ratios, thus further enlarging the training dataset.

(4) **Noise adding**. To simulate possible interference in the real world, we add a series of scaled random Gaussian noises to the whole distance-time profile, artificially generating more noisy samples to further assist the model in improving its resistance to disturbances.

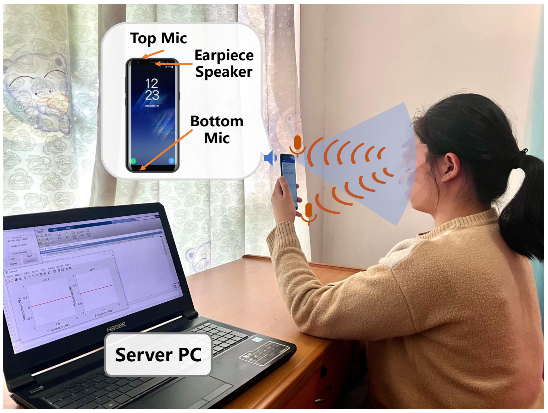


Fig. 16. The experimental setup.

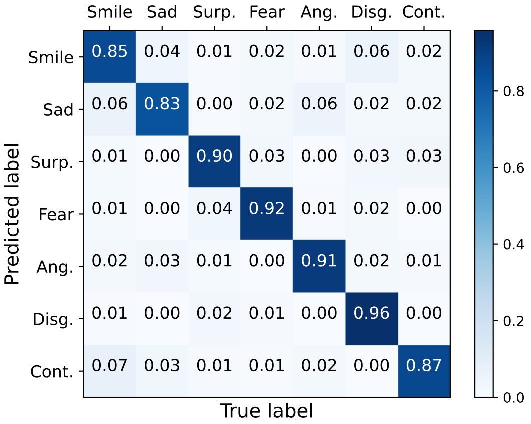


Fig. 17. Confusion matrix of 7 expressions.

6.4 User Adaption

Since each person has a unique facial shape, muscle characteristics and movement habits, facial expressions have obvious individual differences. Even if different people make the same expression, the features may not be exactly the same. Thus, it is challenging to build a model that generalizes directly to all users. To address the issue of user dependency, we adopt the widely used transfer learning [52] to assist in fine-tuning the model to fit new users. Specifically, we first train a basic model utilizing the collected dataset. When there is a new user, the model is retrained with a small amount of expression data collected from the new user. During the training process, the parameters of the previous convolutional layers used for feature extraction are frozen and only the last two fully connected layers are fine-tuned. This approach allows us to efficiently adapt the model to new users while retaining the knowledge of expressions learned from the previous dataset.

7 IMPLEMENTATION

7.1 Experimental Setup

We build a prototype of UFace using a commercial smartphone (i.e., Samsung S8), which consists of a speaker and two microphones (one is on the top and the other is on the bottom of the smartphone). Figure 16 shows the experiment setup. Specifically, we adopt the existing acoustic framework LibAS [35] to transmit and receive carefully designed acoustic signals. In each experiment, we use the earpiece speaker of the Samsung S8 to emit inaudible ultrasound signals, while the top and bottom microphones simultaneously receive the echoes. The distance between the smartphone and user is set at 30 cm (default setup). During the data collection phase, the received signal will be sent to a laptop for signal processing using Matlab software. Our DFNet is implemented in Python and runs on an NVIDIA GeForce RTX 2080 Ti. Note that the signal processing module and the model inference module can technically be run independently on a smartphone during the later practical deployment without additional laptop and server support.

7.2 Data Collection

In our experiment, we recruit 20 volunteers (8 males and 12 females aged 20-28 years), and choose 7 typical facial expressions (*smile, sad, surprise, fear, angry, disgust, contempt*) to evaluate system performance. Our data collection environments include a conference room, a study, and two offices with various layouts. In each environment, we instruct each volunteer to hold a smartphone with a distance of 30 cm to face and focus the smartphone screen in the normal browsing position.

In the basic evaluation, we show the participants pictures and text descriptions of seven typical facial expressions commonly used in computer vision in advance, and offer them some time to imitate and practice. During data collecting, each volunteer is then instructed to perform the assigned facial expression within 5 seconds at a time. Finally, we collect 2800 data samples (20 individuals × 7 actions × 20 times) with a total data duration of 233 minutes. To obtain more data, we augment the data using the method described in Section 6.3 and obtain 148,400 data samples with a total duration of about 206 hours.

To test the effect of head orientation targeting algorithm, we gather 40s non-frontal face data from each of the 20 users (10s for top, bottom, left, and right side faces). In addition, we gather a total of 575 interference data, consisting of eating, talking, waving and sliding fingers while performing expressions. These measurements are used to validate the performance of the method for eliminating self-interference.

7.3 Evaluation Metrics

We select four common metrics in classification models to evaluate the performance of our system. These metrics are *Accuracy, Precision, Recall, F1-score.* Accuracy is the proportion of accurately predicted samples relative to the total number of samples, i.e. $Accuracy = \frac{TP+TN}{FP+FN+TP+TN}$. Precision refers to the proportion of predicted accurate samples that are actually correct, i.e. $Precision = \frac{TP}{FP+TP}$. Recall refers to the proportion of correct samples that are actually selected, i.e. $Recall = \frac{TP}{TP+FN}$. F1-Score represents the reconciled average of precision and recall, i.e. $F1\text{-}Score = \frac{2\times Recall\times Precision}{Recall+Precision}$.

8 EVALUATION

In this section, we present the specific design of experiments and demonstrate the performance of our system.

8.1 Performance of Expression Recognition

8.1.1 Overall Performance. We employ a five-fold cross-validation method to evaluate the basic performance of our system. Note that the cross-validation is performed at the instance level. The average accuracy of DFNet is 87.8% and the average F1-Score is 87.78%. Figure 17 shows the confusion matrix of 7 facial expressions. Among the seven expressions, 'surprise', 'fear', 'angry', and 'disgust' have relatively high recognition accuracy of 90% and above. This is due to the fact that these facial expressions require the mouth and eyebrows to have a large amplitude of movement. 'Smile' and 'contempt' are two perplexing facial expressions because they share a similar movement pattern, with mild movements of the eyebrows and eyes and an upturning of the corners of the mouth. The difference between a 'contempt' and a 'smile' is that 'contempt' involves the corners of one side of the mouth rising upward, whereas 'smile' involves the corners of both sides of the mouth lifting upward. However, the movements on the left and right sides of the horizontally centered microphone and speaker on our smartphones are symmetrical and therefore readily confused. With even the lowest identification accuracy for 'sad' at 83%, we are confident that UFace exhibits high performance in expression recognition.

8.1.2 The advantages of dual-stream signals. To validate the effect of dual views (mic 1&2) in UFace, we compare the results using only the top microphone view (mic 1), only the bottom microphone view (mic 2) and dual views. Figure 18 shows the average recognition results of the different views for 7 expressions. We can observe that the top microphone produces better results than the bottom microphone because it is situated in a more favorable place to "hear" facial movements. Due to the broader detection range, the dual views exhibit considerably stronger recognition capabilities than the single view, with approximately 10% to 13% higher accuracy. The experimental results suggest that enabling dual views can indeed enhance facial expression recognition.

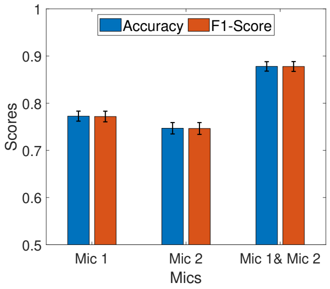


Fig. 18. Comparisons between using one Mic and two Mics.

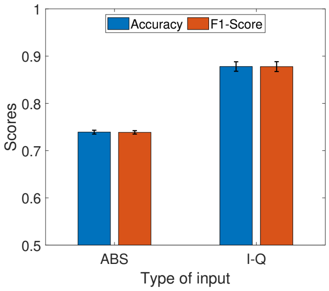


Fig. 19. Comparison of IQ data and absolute value data.

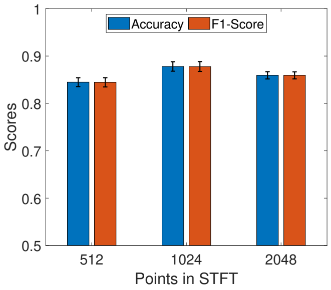


Fig. 20. Comparison of STFT Points.

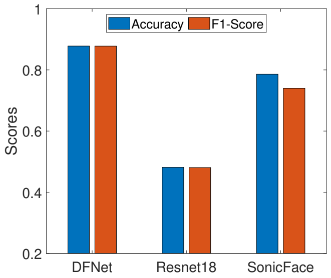


Fig. 21. Model comparisons.

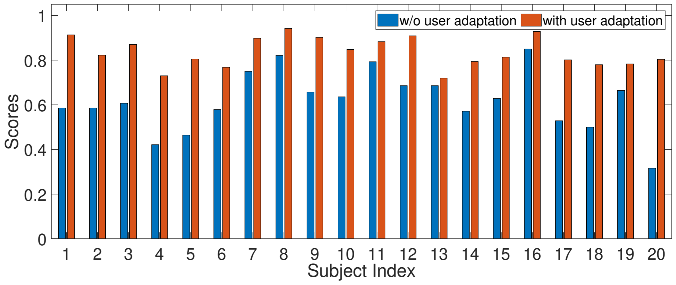


Fig. 22. FER with diverse targets with/without user adaptation.

8.1.3 The advantages of IQ data. To demonstrate the distinct advantage of opting for IQ data of distance-time profile as inputs to DFNet, an experiment is conducted to compare the impact of input absolute and IQ data on the performance of the recognition model using a fixed number of FFT points (1024). The outcomes of this experiment are depicted in Figure 19. It is unsurprising that the accuracy of the IQ data is about 14% higher than that of the Abs data due to the inclusion of a richer representation (containing both amplitude and phase), which fully highlights the necessity of adopting IQ as the model input.

8.1.4 Comparison of different STFT points. In order to determine the most appropriate number of FFT points in STFT, we evaluate the recognition performance of DFNet at different numbers of FFT points (512, 1024 and 2048). As depicted in Figure 20, the observed relatively low accuracy (84.48%) at 512 points is not unexpected, considering the restricted quantity of information that can be derived from a limited number of points. Furthermore, it's noteworthy that both 1024 and 2048 points yield comparable accuracy (the difference is about 1%). Taking into account the inherent resource constraints of the mobile phone, we choose 1024 as the number of FFT points in our system, which has relatively small overhead of signal processing and model inference. Also, it is used in the subsequent experiments.

8.1.5 Comparison between different models. To confirm the advanced recognition performance of our designed model, we compare our DFNet with two other baseline networks: the ResNet-18 [15] and SonicFace [12]. We employ the standard ResNet-18 model after pre-training, where data from two microphones are stacked into four channels fed together into the network. Due to the significant differences in the utilized signals (FMCW and CW signals), we directly report the best performance of SonicFace for comparison. Figure 21 shows the recognition

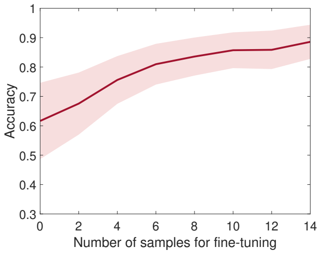


Fig. 23. Impact of different sample amounts on the fine-tuning effect.

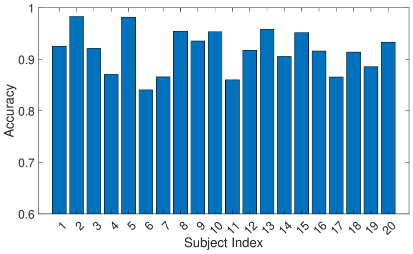


Fig. 24. Accuracy of head orientation recognition.

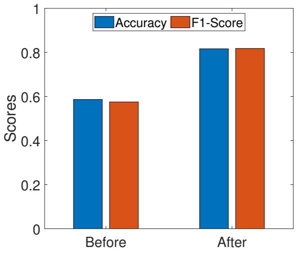


Fig. 25. Accuracy before and after finger swipe elimination

results of the three models. The experimental results demonstrate that DFNet has relatively superior recognition performance (with 87.8% accuracy). On the contrary, ResNet-18 does not attain the anticipated performance level, which we attribute to its design focusing on feature extraction from intricate images. With our smaller task (2 × 18 × 99 with only two paths), it is deep into the hazards of overfitting.

8.1.6 Adaptability to new users. We perform leave-one-user-out cross-validation tests to assess the model's performance on unseen users during training, demonstrating its ability to generalize effectively across diverse individuals. We use the data of 19 users for training and one other user for testing at a time. We compare the results with and without(w/o) fine-tuning as presented in Section 6.4. The average accuracy in the leave-one-user-out experiment without fine-tuning is 61.65%, a considerable reduction from the basic experiment (accuracy of 87.8%). Each individual's facial anatomy is unique, thus the amplitude of muscle movements required to create the same emotion varies. In addition, even for the same expression, each individual may have a distinct muscle movement pattern. Therefore, the features of the same expression might vary significantly across individuals, necessitating fine-tuning of the model. Figure 22 shows the experimental results of with/without fine-tuning. After fine-tuning the model with limited data per expression, the model's average accuracy increased to 83.57%. In order to determine the optimal amount of data for fine-tuning, we conduct experiments to evaluate the impact of varying amounts of data on the performance of the model. Figure 23 illustrates that the model can be fine-tuned to achieve an accuracy of over 80% with as few as six samples. This means that new users only need to collect approximately 30 seconds for each expression before implementing UFace to gain improved results, which we consider to be totally acceptable. Furthermore, as the number of samples used for fine-tuning increases, the accuracy of the fine-tuned model shows a gradual increase.

8.2 Performance of Reliability Detection Mechanism

8.2.1 Head orientation targeting performance. In this section, we test the algorithm's ability to recognize face regions that is proposed in Section 5.1. We collect an additional 40s of side neutral face data (10s each for the top and bottom left and right faces) and extract the frontal face data from each participant's neutral face data. We extract 760 segments each about front and side faces, respectively, of which 30% are used for training and 70% for testing. Figure 24 shows the experimental results of head orientation targeting. The average accuracy for recognizing front and side faces is 91.69%. Among these, fifteen subjects achieved an accuracy exceeding 90%, while the lowest one surpassed 84%, showing that our algorithm can accurately identify the user's face orientation with only 10 seconds of training data.

8.2.2 Validate finger swipe elimination. We collect a total of 245 expression data samples with finger swipes and eliminate the interference using the algorithm proposed in Section 5.2. As shown in Figure 25, the average accuracy is 58.6% before removing interference, highlighting the negative impact of finger swiping on expression

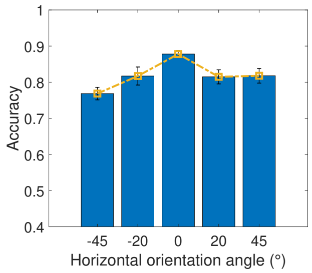


Fig. 26. Impact of horizontal angle of the user's face.

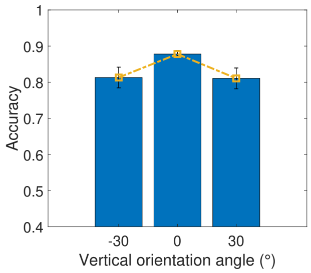


Fig. 27. Impact of vertical angle of the user's face.

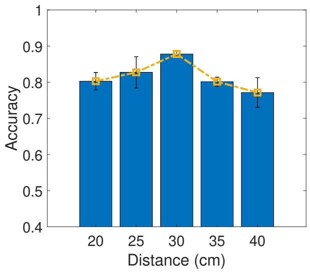


Fig. 28. Varying distance between the user's face and phone.

recognition. After eliminating finger sliding, accuracy improves to 81.6%. Despite a slight reduction in average accuracy after eliminating finger interference compared to the basic experiment, there is a notable improvement in accuracy before elimination. This demonstrates the efficacy of our devised finger interference elimination algorithm, offering the potential for 'opportunistic' capture of feedback expressions during mobile phone content browsing.

8.2.3 Self-interference filter performance. We collected 330 instances of interference data, including chewing, talking, and waving. We send these interference data together with 980 expression data samples to the self-interference elimination module for filtering. Of these data, 88.48% of the interference data were successfully filtered, and 91.2% of the expression data were successfully passed, proving the usability of the module.

8.3 Robustness Analysis

In this section, we evaluate the effects of the relative position of the device, the user's body posture, the speed of expression generation, and the noisy environment on the system's performance. There are five volunteers in total (among the previous 20 people) who participated in the robustness experiments. Each expression is collected 50 times under each experimental setting (e.g., 40 cm counts as one setting). Finally, we collect a total of 5950 data samples (50 times × 7 expressions × 17 settings) for testing.

8.3.1 The effect of the relative position of the devices. In this experiment, we investigate the recognition performance when the smartphone locates at different relative positions, i.e., the horizontal angle, the vertical angle, and the distance between the use's face and smartphone. Note that for each experiment, we keep the other factors unchanged.

Horizontal angle. We set the distance between the smartphone and the user's face as 30 cm. The vertical angle of the smartphone is 0°. Then, we vary the horizontal angle of the smartphone from -45° to 45°. Note that we use the right hand to move the smartphone along the positive angle, and use the left hand to move along the negative angle. The results are shown in Figure 26. We can see that 1) within the range of $[0°, 45°]$, UFace enables over 81.8% recognition accuracy; 2) The prediction accuracy of positive angles is higher than the negative angles. This is because the training datasets are collected by the right hand, which causes the distribution of testing data to be slightly different from the training data when we use the left hand to collect the testing data. Even so, the UFace can achieve 81.71% accuracy at $-20°$. We believe this accuracy is acceptable for real-world long-term expression monitoring.

Vertical angle. We set the distance between the smartphone and the user's face as 30 cm. The horizontal angle of the smartphone is 0°. Then, we vary the vertical angle of the smartphone from -30° to 30 °. From Figure 27, we can obtain that UFace can achieve over 81.06% recognition accuracy in $[-30°, 30°]$. This indicates that UFace is

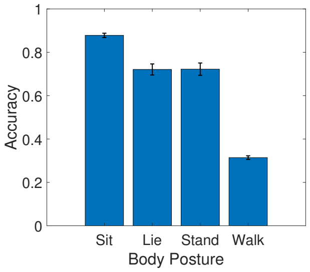


Fig. 29. Impact of different postures.

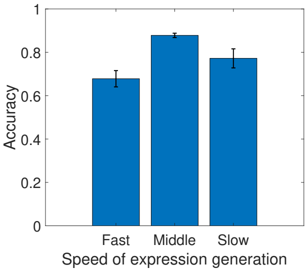


Fig. 30. Impact of the speed of expression generation.

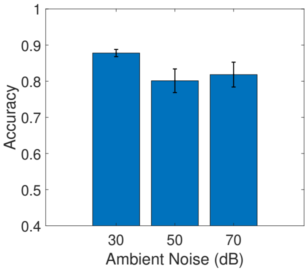


Fig. 31. Different degrees of ambient noise interference.

able to obtain satisfactory recognition results in the most natural position for using a smartphone, i.e., slightly looking down at the device.

Distance. We set the horizontal angle and vertical angle as 0° and vary the distance between the user's face and the smartphone from 20 cm to 40 cm with the step of 5 cm. Figure 28 illustrates the optimal recognition distance is $[25cm, 35cm]$.The recognition accuracy reached its highest level at 82.74% when the distance is set at 25cm. At a distance of 40 cm, the accuracy of recognition falls to 77.14%. This is because we ask participants to maintain the phone aligned with their face during data collection. At 40cm, participants are required to extend their arms far, which results in a significant tremor when they hold the phone, thereby interfering with recognition.

Overall, the effective field-of-view (FoV) of UFace is the horizontal angle within $[-20°, 45°]$, vertical angle within $[-30°, 30°]$, and the distance between user's face and smartphone within $[25cm, 35cm]$. This FoV demonstrates the consistent ability of UFace to accurately identify facial expressions in scenarios involving regular smartphone usage by covering the common positions people adopt when using their phones.

8.3.2 *Impact of user's body posture.*

In this experiment, we explore how the user's body posture impacts recognition performance. Specifically, we choose three states that represent common situations in normal life: sitting, standing, and lying. For each case, the center of the smartphone is parallel to the user's face and the distance between them is 30 cm. The results are shown in Figure 29. We can clearly see that when the user's state is sitting, the recognition accuracy can achieve over 87.8%. The accuracy of standing and lying down has decreased but still achieved 72.23% and 72.09% accuracy, respectively. The reason is that the body shakes slightly or the expression motion is limited when the user is standing or lying to hold onto the smartphone, which will cause the emitted and received signals to have deviations. The recognition accuracy during walking is seen to be merely 31.43%, mostly attributed to the substantial disruption caused by the vigorous motion associated with walking, which significantly affects the quality of our distance-time profile. Even so, the recognition accuracy under walking is still higher than the random result (14.29%). Experimental results demonstrate that UFace exhibits greater suitability for usage in conditions characterized by relative calmness. Conversely, the recognition rate of UFace is significantly diminished in scenarios where there is evident bodily shaking.

8.3.3 *Impact of the expression generation speed.*

To explore how the speed of expression generation impacts prediction accuracy, the participants perform expressions at three different speeds, i.e., fast, middle (normal), and slow. For each case, the center of the smartphone is parallel to the user's face and the distance between them is 30 cm. To be specific, the determination of "fast" and "slow" is established in relation to a baseline normal speed, whereby "fast" signifies that the expression is produced within about 0.1 to 0.2 seconds, and "slow" corresponds to a duration of around 0.5 seconds. It's essential to highlight that the duration discussed here refers to the process from the absence of the face to the completion of the expression (not yet recovered), i.e., the time of the occurrence of the expression, rather than the entire duration of the expression from the time it is produced to the return of the neutral face. This is because we do not limit the specific duration for which participants should perform each expression during data collection, as long as it is within 5 seconds. Consequently, for the latter, natural diversity is inherently embedded within the training dataset. Figure 30 depicts the recognition results. Under normal expression generation speed, the accuracy can achieve over 87.8%. Although the accuracy decreases if the speed is fast or slow, these situations are not common when we read the contents of smartphones, as it is a relatively unnatural behavior.

8.3.4 Impact of noisy environments. We now study how the noisy environment affects the system's performance. We run experiments in three environments by varying the noise level from 30 dB to 70 dB with the step of 20 dB. Specifically, the 30 dB noise level means a quiet room, the 50 dB noise level corresponds to the meeting room with discussion, and the 70 dB noise level belongs to playing loud music. The results are shown in Figure 31. We can see that the system achieves the best accuracy of 87.8% when the noise level is 30 dB. However, when the noise level increases, the accuracy decreases a bit but remains around 80-81%, which is within the normally fluctuating and acceptable range.

8.4 User Case Study

8.4.1 Potential applications. Facial expressions serve as the most straightforward indication of a user's emotions. Envision that many applications would benefit from enabling unobtrusive, long-term facial expression recognition on smartphones.

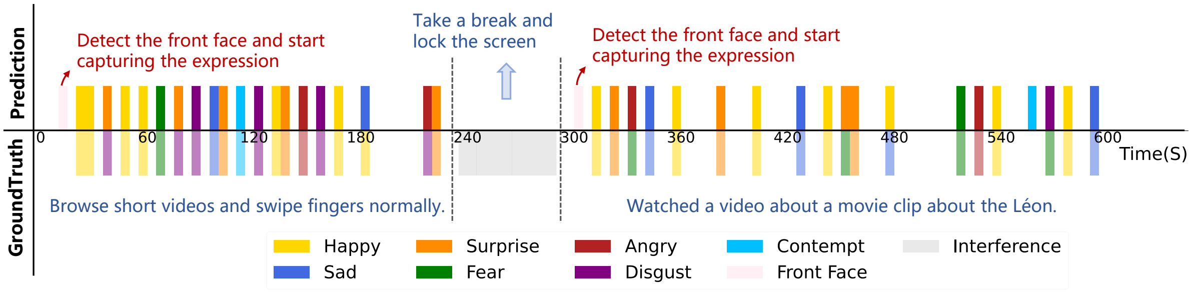
(1) *Long-term Emotion Monitoring*. Statistics indicate that almost half of people in the United States spend 5 to 6 hours on smartphones daily [29]. Considered a multi-purpose companion in real life, the smartphone witnesses many of the user's emotional states throughout the day. By analyzing facial expressions for the long term, mobile devices equipped with facial expression detection can track and monitor emotions for mental health purposes without any privacy risks, enabling users to understand and manage their well-being effectively.

(2) *Content Recommendation*. Smartphones with FER capabilities can provide valuable insights for content recommendation. For example, when users are browsing shopping platforms (e.g., Taobao, Amazon, etc.) or short-video platforms (e.g., TikTok), by detecting their expressions, the APPs can gauge users' responses to advertisements, products, or services, helping companies reach their target audiences with videos or products that match their preferences.

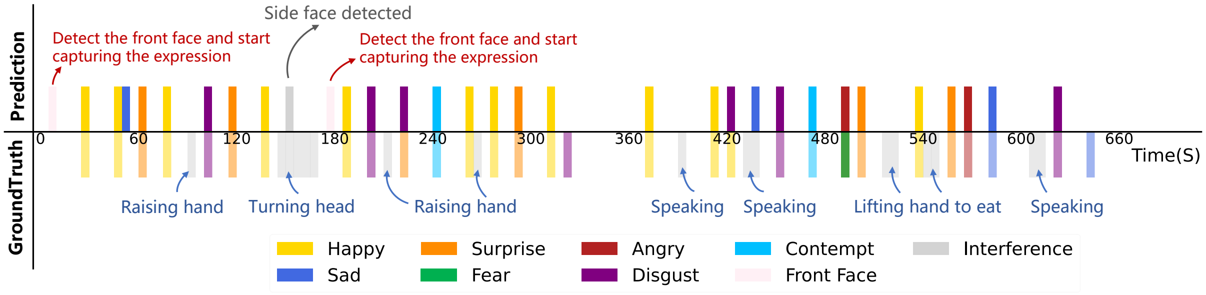
(3) *User Experience Enhancement*. FER can enhance the user experience of various mobile applications. For instance, it can enable hands-free control by allowing users to perform actions simply by making specific facial expressions. Additionally, personalized recommendations and adaptive interfaces can be offered based on user emotional states detected through facial expressions.

(4) *Video Feedback*. FER on smartphones can be employed to obtain real-time emotional feedback from viewers on different clips while watching videos or movies. Different from textual comments, it introduces a new dimension of feedback, which will further guide video and film creators in their content creation.

8.4.2 Case study. To explore the potential of our acoustic-based system in the above application scenarios, we conduct an extended user case study to evaluate its applicability in the real world. The case study consists of two 10-minute segments of continuous data with two participants, envisioning multiple scenarios of smartphone use in daily life. The first data segment is mainly oriented to scenarios of analyzing emotional feedback from watching videos, ranging from watching long videos (highlights of Léon: The Professional) and browsing short videos (TikTok). When browsing on TikTok, finger swipes occur approximately every 5-10 seconds, primarily used to play the next video, while there are no finger swipes for long videos. The second data segment focuses on daily emotion monitoring and content recommendation scenarios, covering situations where users browse



(a) Emotional feedback of watching videos.



(b) Emotion monitoring and content recommendation.

Fig. 32. Facial expression prediction results and corresponding ground truth in the user case study. There are two consecutive 10-minute segments: (a) mainly oriented to emotional feedback scenarios of watching videos, and (b) mainly oriented to daily emotional monitoring and content recommendation scenarios.

social and shopping platforms. It involves a variety of disruptive behaviors such as finger sliding (once every 2-5 seconds), talking, eating, waving, and head turning in addition to facial expressions, which better emulates the intricacies of the real world. Consistent with the experimental setup in Section 7.1, participants are instructed to sit while holding a cell phone 30 cm from their faces. We utilize a camera to record the user's video for the relevant clip as our ground truth.

Applying the carefully designed pipeline, which consists of signal processing, reliability detection mechanism, and trained DFNet model (from Section 8.1.1 involving 20 participants), to the received consecutive signal time series, the final prediction results are available, as shown in Figure 32. Note that we here utilize the majority voting mechanism in ensemble learning [26] to confirm the final prediction from the five trained models obtained from the five-fold cross-validation. The results show that our system is able to provide a valuable insight into facial expression detection. The overall accuracy for the seven facial expressions is 80.6%. Although there are some misrecognized expressions, such as some "fear" expressions being classified as "disgust" or "anger," they still fall under negative emotions, which remain in the same valence-arousal quadrant [12, 30]. Besides, we can observe from Figure 32 that UFace can successfully identify the optimal moment to attempt sensing, while effectively filtering out interfering movements to capture the occurrence of every subtle facial expression and accurately recognize its type. This further highlights the practicability of our system in the real world.

9 DISCUSSION

In this section, we will discuss the limitations of our system and the direction of future work.

Comparison with camera-based: There is no doubt that facial expression recognition based on the camera on a smartphone is the most direct and high-accuracy solution. Currently, the state-of-the-art methods based on static images or dynamic videos can mostly achieve more than 80% or 90% recognition accuracy with different datasets [17]. For example, the attention-based facial expression recognition network of wang et al. [39] can achieve 89.16% and 86.9% recognition accuracies under the FERPlus [3] and RAFDB [20] datasets, respectively. However, in our target scenario, long-term monitoring requires users to be exposed to the camera for a long time, posing serious privacy risks. In addition, accompanied by a high energy overhead, these camera-based methods are limited by lighting conditions. Consequently, we propose an acoustic sensing-based scheme conveniently implemented on smartphones that achieves comparable FER accuracy (87.8%), compensating for the shortcomings of camera-based schemes. It is noteworthy that our aim is not to replace cameras but rather to function as a complementary solution, providing an alternate method for recognizing facial expressions and assessing its viability.

Cross-device compatibility: Different smartphones have variations in the placement of their microphones and speakers, including differences in the number and spacing of these components. Transferring UFace directly to other devices may compromise its recognition accuracy. However, we are able to quickly adapt the model to a new device by collecting a small amount of data from the new device for fine-tuning. This approach has been demonstrated to be feasible in prior work [48].

Body motion: We also evaluate the performance of UFace during walking and find that the average accuracy is only 31.43%. This indicates that UFace does not currently have the ability to perform facial expression recognition in motion. The movement of the body frequently changes the relative position relationship between the phone and the face, which seriously disrupts the expression features. Based on our current understanding, it may be challenging to achieve accurate expression recognition in motion situations using only ultrasound on the smartphone. To address this issue, a possible feasible solution is to adopt a multi-modal approach to eliminate the effect of motion, e.g., with the help of IMU on the phone, which is the next step in our research.

System overhead: The energy consumption of our smartphones for transmitting and receiving ultrasonic signal is approximately 16.84 mAh[49], which is adequate for the current battery capacity of mobile devices to sustain the daily operation. The signal processing and detection mechanism have an average latency of 3.78s, while the model inference has an average latency of approximately 586ms, which is acceptable for our system with low real-time requirements. The signal processing, reliability detection mechanism, and model inference are currently executed on the server. We will explore how to implement a low-energy, lightweight mobile device-side expression recognition model in our future work.

Limited expression categories: In UFace, seven typical facial expressions commonly used in computer vision are employed for model training and performance evaluation to gauge the feasibility of the system. However, in real life, there are multiple ways of expressing facial expressions for the same emotion. Furthermore, even for the same individual, the muscle movement pattern of the same expression may vary over time (e.g., months later). These will lead to changes in signal characteristics, thereby significantly degrading the performance of the model. A potential solution is to utilize incremental learning techniques to facilitate the model's ability to absorb new facial expression information while retaining old FER knowledge. We will further explore its solutions in the future to generate personalized models for individuals and promote the practicability of the system.

Domain adaptation: The cross-domain problem remains an open challenge for wireless-based sensing systems. We have specifically designed a head orientation targeting module and a series of data augmentation schemes to significantly constrain and mitigate the impact of the high flexibility of mobile phones. However, from the results of the robustness experiments in Section 8.3, there is still a general decrease of $5\% - 10\%$ in accuracy under different distance, angle, and pose tests compared to the basic experiment (87.8%). This is due to the fact that the change in the location (i.e., domain) relationship between the person and the smartphone still has a slight impact on the data feature distribution. Leveraging a Generative Adversarial Network (GAN) or Diffusion models to generate data under different domains for more generalized augmentation or adopting domain adversarial learning methods to alleviate the effect of the domain may be a viable solution. Continuing optimizing this issue to make UFace more practical is one of the most important goals moving forward.

10 CONCLUSTION

In this work, we propose UFace, a FER system that performs long-term, unobtrusive, and reliable emotion monitoring by using acoustic signals generated by a smartphone. UFace achieves this through extracting high-quality features corresponding to tiny aperiodic emotion movements, and eliminating interference caused by users' facial position changes and other body motions. Experiment results show that UFace can recognize 7 typical facial expressions with an average accuracy of 87.8% across 20 participants. Extensive evaluations of UFace's prototype demonstrate the efficiency and robustness of our proposed schemes.

ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China under Grant No. 62122095, 62341201, 62072472 and 62202256, and by a grant from the Guoqiang Institute, Tsinghua University.

REFERENCES

[1] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial expression recognition using ear canal transfer function. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 1–9.

[2] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. Canalsense: Face-related movement recognition system based on sensing air pressure in ear canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 679–689.

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*. 279–283.

[4] Huijie Chen, Fan Li, and Yu Wang. 2017. EchoTrack: Acoustic device-free hand tracking on smart phones. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.

[5] Tuochao Chen, Yaxuan Li, Songyun Tao, Hyunchul Lim, Mose Sakashita, Ruidong Zhang, Francois Guimbretiere, and Cheng Zhang. 2021. NeckFace: Continuously Tracking Full Facial Expressions on Neck-mounted Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–31.

[6] Yanjiao Chen, Runmin Ou, Zhiyang Li, and Kaishun Wu. 2020. WiFace: facial expression recognition using Wi-Fi signals. *IEEE Transactions on Mobile Computing* 21, 1 (2020), 378–391.

[7] Haiming Cheng and Wei Lou. 2021. Push the limit of device-free acoustic sensing on commercial mobile devices. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.

[8] Seokmin Choi, Yang Gao, Yincheng Jin, Se jun Kim, Jiyang Li, Wenyao Xu, and Zhanpeng Jin. 2022. PPGface: Like What You Are Watching? Earphones Can" Feel" Your Facial Expressions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–32.

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

[10] Paul Ekman. 2004. Emotions revealed. *Bmj* 328, Suppl S5 (2004).

[11] Yongjian Fu, Shuning Wang, Linghui Zhong, Lili Chen, Ju Ren, and Yaoxue Zhang. 2022. SVoice: Enabling Voice Communication in Silence via Acoustic Sensing on Commodity Devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 622–636.

[12] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–33.

[13] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–27.

[14] Anna Gruebler and Kenji Suzuki. 2014. Design of a wearable device for reading positive expressions from facial emg signals. *IEEE Transactions on affective computing* 5, 3 (2014), 227–237.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. 2019. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1626–1635.

[17] Mohan Karnati, Ayan Seal, Debotosh Bhattacharjee, Anis Yazidi, and Ondrej Krejcar. 2023. Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–31. https://doi.org/10.1109/TIM.2023.3243661

[18] 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*. 150–163.

[19] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24.

[20] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2852–2861.

[21] Kang Ling, Haipeng Dai, Yuntang Liu, Alex X Liu, Wei Wang, and Qing Gu. 2020. Ultragesture: Fine-grained gesture sensing and recognition. *IEEE Transactions on Mobile Computing* (2020).

[22] Jialin Liu, Dong Li, Lei Wang, and Jie Xiong. 2021. BlinkListener: " Listen" to Your Eye Blink Using Your Smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–27.

[23] Wenguang Mao, Mei Wang, and Lili Qiu. 2018. Aim: Acoustic imaging on a mobile. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 468–481.

[24] Denys JC Matthies, Bernhard A Strecker, and Bodo Urban. 2017. Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1911–1922.

[25] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1515–1525.

[26] Robi Polikar. 2012. Ensemble learning. *Ensemble machine learning: Methods and applications* (2012), 1–34.

[27] Kun Qian, Chenshu Wu, Fu Xiao, Yue Zheng, Yi Zhang, Zheng Yang, and Yunhao Liu. 2018. Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. In *IEEE INFOCOM 2018-IEEE conference on computer communications*. IEEE, 1574–1582.

[28] Soha Rostaminia, Alexander Lamson, Subhransu Maji, Tauhidur Rahman, and Deepak Ganesan. 2019. W! nce: Unobtrusive sensing of upper facial action units with eog-based eyewear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–26.

[29] Daniel Ruby. 2023. Smartphone Usage Statistics 2023 (Facts & Data). https://www.demandsage.com/smartphone-usage-statistics/.

[30] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[31] Andrew Ryan, Jeffery F Cohn, Simon Lucey, Jason Saragih, Patrick Lucey, Fernando De la Torre, and Adam Rossi. 2009. Automated facial expression recognition system. In *43rd annual 2009 international Carnahan conference on security technology*. IEEE, 172–177.

[32] Xingzhe Song, Kai Huang, and Wei Gao. 2022. FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Commodity Headphones. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 145–157.

[33] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*. 945–953.

[34] Ke Sun and Xinyu Zhang. 2021. UltraSE: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th annual international conference on mobile computing and networking*. 160–173.

[35] Yu-Chih Tung, Duc Bui, and Kang G Shin. 2018. Cross-platform support for rapid development of mobile acoustic sensing applications. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 455–467.

[36] Barry D Van Veen and Kevin M Buckley. 1988. Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine* 5, 2 (1988), 4–24.

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

[38] Guangjing Wang, Qiben Yan, Shane Patrarungrong, Juexing Wang, and Huacheng Zeng. 2023. FacER: Contrastive Attention based Expression Recognition via Smartphone Earpiece Speaker. (2023).

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

[40] Shuliang Wang, Hehua Chi, Ziqiang Yuan, and Jing Geng. 2019. Emotion recognition using cloud model. *Chinese Journal of Electronics* 28, 3 (2019), 470–474.

[41] Shanmin WANG, Hui SHUAI, Lei ZHU, and Qingshan LIU. 2023. Expression Complementary Disentanglement Network for Facial Expression Recognition. *Chinese Journal of Electronics* 33 (2023), 1–11.

[42] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[43] Jiahong Xie, Hao Kong, Jiadi Yu, Yingying Chen, Linghe Kong, Yanmin Zhu, and Feilong Tang. 2023. mm3DFace: Nonintrusive 3D Facial Reconstruction Leveraging mmWave Signals. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 462–474.

[44] Wentao Xie, Qian Zhang, and Jin Zhang. 2021. Acoustic-based Upper Facial Action Recognition for Smart Eyewear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–28.

[45] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 54–66.

[46] Huiyuan Yang, Umur Ciftci, and Lijun Yin. 2018. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2168–2177.

[47] Nianyin Zeng, Hong Zhang, Baoye Song, Weibo Liu, Yurong Li, and Abdullah M Dobaie. 2018. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* 273 (2018), 643–649.

[48] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. Soundlip: Enabling word and sentence-level lip interaction for smart devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.

[49] Qian Zhang, Dong Wang, Run Zhao, Yinggang Yu, and Junjie Shen. 2021. Sensing to hear: Speech enhancement for mobile devices using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–30.

[50] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. 2017. Multi-view learning overview: Recent progress and new challenges. *Information Fusion* 38 (2017), 43–54.

[51] 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*. 321–336.

[52] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.