# C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-mounted Miniature Cameras

**Tuochao Chen** [1,2,3], **Benjamin Steeper** [1,2], **Kinan Alsheikh** [1,2], **Songyun Tao** [1,2]
**François Guimbretière** [2], **Cheng Zhang** [1,2]
[1] SciFi Lab, Cornell University, Ithaca, New York, United States
[2] Information Science, Cornell University, Ithaca, New York, United States
[3] EECS, Peking University, Beijing, China
chentuochao@pku.edu.cn, {bds238, kta33, st938, fvg3,chengzhang}@cornell.edu;

## ABSTRACT

C-Face (Contour-Face) is an ear-mounted wearable sensing technology that uses two miniature cameras to continuously reconstruct facial expressions by deep learning contours of the face. When facial muscles move, the contours of the face change from the point of view of the ear-mounted cameras. These subtle changes are fed into a deep learning model which continuously outputs 42 facial feature points representing the shapes and positions of the mouth, eyes and eyebrows. To evaluate C-Face, we embedded our technology into headphones and earphones. We conducted a user study with nine participants. In this study, we compared the output of our system to the feature points outputted by a state of the art computer vision library (Dlib[1]) from a font facing camera. We found that the mean error of all 42 feature points was 0.77 mm for earphones and 0.74 mm for headphones. The mean error for 20 major feature points capturing the most active areas of the face was 1.43 mm for earphones and 1.39 mm for headphones. The ability to continuously reconstruct facial expressions introduces new opportunities in a variety of applications. As a demonstration, we implemented and evaluated C-Face for two applications: facial expression detection (outputting emojis) and silent speech recognition. We further discuss the opportunities and challenges of deploying C-Face in real-world applications.

## Author Keywords

Wearable computing; Deep Learning; Computer Vision: Facial Expression Reconstruction and Tracking; Ear Sensing; Emoji Recognition; Silent Speech;

## CCS Concepts

**•Human-centered computing → Ubiquitous and mobile devices;**

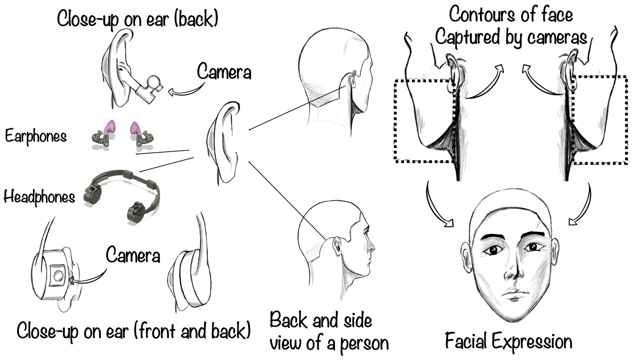

Close-up on ear (back)
Camera
Earphones
Headphones
Camera
Close-up on ear (front and back)
Back and side view of a person
Contours of face Captured by cameras
Facial Expression

**Figure 1. Overview of C-Face**

## INTRODUCTION

Humans use facial expressions as a natural mode of communication. The ability to continuously record and understand facial movements can improve interactions between humans and computers in a variety of applications. For example, reconstructing and recording facial movements in a learning environment can give instructors useful feedback into student engagement levels[55]. In virtual social environments, users can have their real-time facial expressions mirrored on avatars for more immersive social experiences. Tracking facial expressions can potentially improve daily communication experiences as well. For instance, in order to provide facial feedback on a mobile video call, a user must currently hold the phone in hand and point the camera towards the face. If the user's facial expression could be automatically recorded and presented in a hands-free fashion, communication experiences could be greatly improved in scenarios where the user is carrying groceries, doing the dishes, jogging, and more.

Traditional facial reconstruction methods require a camera positioned in front of the user's face at all times[5] and an entire view of the face without occlusions. These methods have

[1] dlib Library: http://dlib.net/

*UIST '20, October 20–23, 2020, Virtual Event, USA*
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7514-6/20/10 ...$15.00.
http://dx.doi.org/10.1145/3379337.3415879

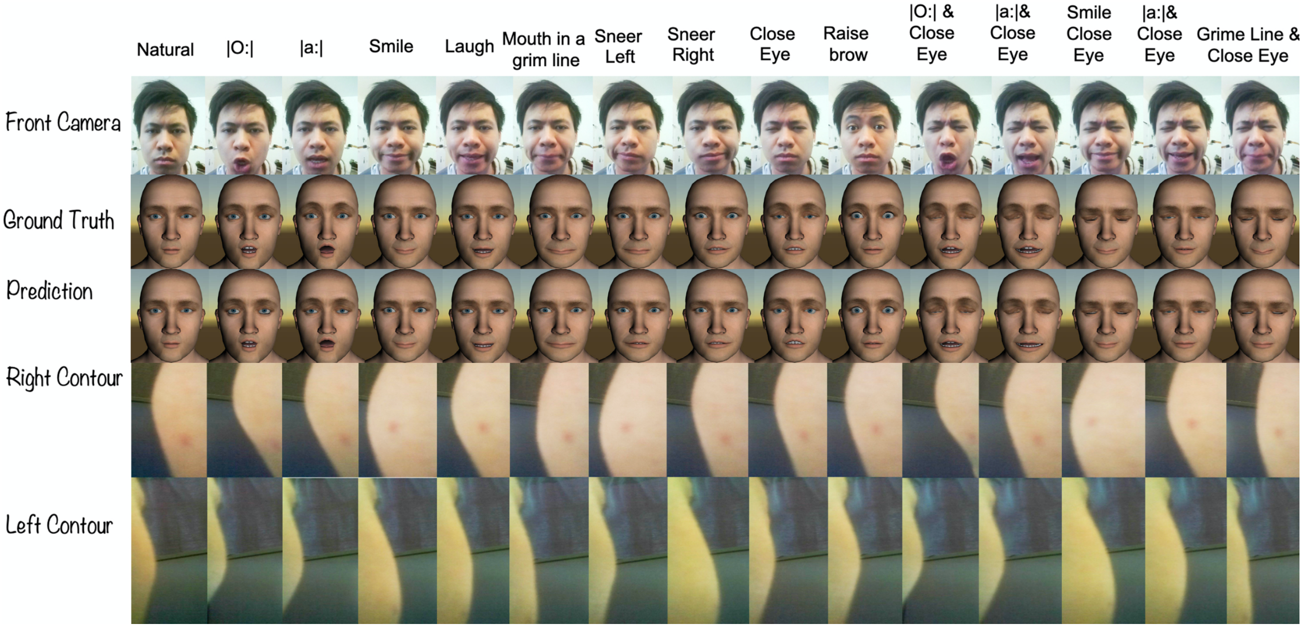


**Figure 2. The captured contour images, acquired ground-truth using Dlib, and predicted results of facial expressions used in the user study. The facial expressions are rendered using a 3D mesh model © TurboSquid.**

provided reliable facial tracking performance when certain criteria (e.g. position, view angle) are met. However, they have limitations in many use cases. They may not work well when 1) a user is in motion or in an environment where a camera can not be appropriately set up or 2) the user's face is partially occluded or is not fully visible due to camera positioning or angling relative to the user's face. The ability to track facial movements even when the face is partially covered could allow for facial reconstruction use in new settings. For instance, it could even be used to track facial expressions under the current COVID-19 pandemic when people wear face masks during their daily activities. Traditional computer vision based technology does not work well in these kinds of settings.

To overcome the above challenges, researchers have developed various wearable devices for facial expression recognition using sensing techniques such as acoustic interference, pressure sensing, electrical impedance tomography (EIT) and electromyography (EMG) [42, 43, 10]. Compared to the previously mentioned front-facing camera method, wearable devices are always mounted on the user. As a result, it does not require any pre-setup in the environment and has the potential to recognize facial expressions in mobile settings. However, these wearable technologies also present limitations. Many of them require heavy instrumentation on the user (e.g., covering the face), making it challenging for them to be adopted into common form factors. Also, most of these wearable devices can only recognize discrete facial expressions. We do not know of a prior wearable technology that can continuously reconstruct full facial expressions, capturing the shapes and positions of the mouth, eyes and eyebrows. There is a clear need for a practical wearable sensing technology that can continuously reconstruct full facial expressions under non-optimal scenarios(e.g., mobile setting, covered face).

We introduce C-Face, a novel wearable sensing device that continuously reconstructs facial expressions by learning contours of the face with ear-mounted miniature cameras. C-Face is designed based on the key observation that facial contours are highly informative of facial expressions. When we perform a facial expression, our facial muscles stretch and contract. They push and pull the skin and affect the tension of nearby facial muscles. This effect causes the outline of the cheeks (contours) to alter from the point of view of the ear. Based on this observation, we developed the key research question behind C-Face:

- *Is it possible to continuously reconstruct full facial movements by observing the contours of the face from the ear?*

To explore this research question, we developed our system, leveraging the latest advancements in wearable sensing, deep learning, and computer vision, while maintaining the mobility and minimal-obtrusiveness that a wearable device offers. To demonstrate the capability of C-Face, we integrated the system into two commonplace form factors: headphones and earphones. By embedding the cameras near the ear, our system is able to continuously reconstruct facial movements represented by 42 facial feature points without the need for a frontal view of the face. A user study with 9 participants (2 of which are authors) showed that the mean error for all 42 facial feature points was 0.77 mm (SD = 0.1 mm) using earphones and 0.74 mm (SD = 0.11 mm) using headphones. The mean error for 20 major feature points around the mouth and eye areas was 1.43 mm (SD = 0.21 mm) for earphones and 1.39 mm (SD = 0.28 mm) for headphones. This approach allows us to reconstruct the face even when it is partially occluded by a mask or a pair of glasses.

To the best of our knowledge, C-Face is the first wearable technology that is able to continuously reconstruct full facial expressions, capturing the positions and shapes of the mouth, eyes, and eyebrows without the need to view the entire face directly. The contributions of this paper are:

- We developed an ear-mounted wearable system, with two miniature RGB cameras to capture the contours of the face.
- We designed and implemented a deep neural networks to learn the subtle changes of the contour shapes of the face, which outputs 42 facial feature points representing the positions and shapes of the mouth, eyes and eyebrows.
- We conducted a user study with 9 participants to evaluate the performance of the system with two form factors, earphones and headphones, under different circumstances (after remounting and with face mask occlusion) and applications (silent speech and emoji recognition).
- We discuss the opportunities, challenges and limitations of C-Face for future work.

## RELATED WORK

In this paper, we evaluate C-Face's continuous facial reconstruction performance and explore its applications (i.e. silent speech, emoji input) as well. Therefore, we discuss related work in three sections: 1) traditional computer vision method for facial movement sensing 2) non-CV wearable facial movement sensing and 3) silent speech recognition.

### CV-based Facial Movement Sensing

Computer vision (CV) is one of the most popularly explored approaches on sensing the facial movements. Typically, research in this field involves placing an RGB camera [45], thermal camera [14], or depth camera [41] in front of the face and using the captured images for facial movement sensing. For facial movement recognition and reconstruction, there are two main types of approaches: pre-designed and learned [5]. As for the pre-designed method, researchers use conventional computer vision methods, such as appearance [52, 60] and geometry [29], in order to extract relevant information and determine facial expressions. As for the learned method, machine learning-based approaches are used to automatically learn the feature extraction and classification methods from the training data. Deep learning approaches in particular, namely CNN, DBN, RNN, and GAN, have demonstrated outstanding performances in recent years. Some researchers have built deep learning models focused on discrete classification such as [44, 46, 25] (CNN), [33, 11] (DBN), [6, 3] (RNN), and [30, 58, 56] (GAN). Others have succeeded in building models focused on continuous reconstruction and analysis such as [28] (RNN), [13] (DBLSTM) and [37] (Kernel Regression).

However, the key issues with traditional CV methods in facial movement detection are that 1) it requires a pre-set camera in front of the user with specific requirements for its angle and view and 2) the face of the user can not be blocked. In free living environments, these are significant limitations.

### Non-CV Wearable Facial Movement Sensing

A number of non-CV wearable methods have also been tested to reconstruct facial expressions. For example, Interferi [22] built a face mask with built-in ultrasonic transducers and used acoustic interferometry to detect facial movements. But the on-body acoustic sensing may suffer from performance decrease after remounting the device [22, 57]. Other researchers [42, 43, 10] secured electrodes to the human face, and used electromyography (EMG) or capacitive sensing to track the muscles that control the eyebrows and mouth corners. Researchers have also attached pressure sensors onto the face in order to sense skin deformation for facial expression recognition [48, 31]. However, all of the above methods require that the devices be attached directly onto the users' faces or body. This could block field of vision and interfere with normal daily activities like eating or socializing. Furthermore, most of them only recognize discrete facial expressions. Interferi [22] explored continuous reconstruction of the shape and position of the mouth when the user performs one expression (smiling) while covered by a mask.

Researchers have explored more unobtrusive methods and form factors to address this problem and reduce discomfort. For example, [36, 1] created an earpiece which fit inside the ear canal. The device measured contact impedance and air pressure from inside the canal to sense its deformation. They then mapped the ear canal deformation to facial movement. This paper [35] embedded infrared proximity sensors into eye glasses to detect facial expressions. However, despite the efforts and successes of these unobtrusive facial sensing methods, they are limited to discrete facial gesture classification. The lack of continuous facial gesture classification limits potential applications.

### Silent Speech Recognition

The idea of a 'silent speech interface' has emerged and shown promise in recent years as a method to aid the handicapped, or offer users text input capabilities in high-background-noise environments or settings where vocalizing speech is inappropriate [8]. Past approaches for silent speech detection can be divided into two main categories: contact and non-contact. In the contact approach, different sensors (magnetic [17, 4, 47], EMG [50, 49], EEG , ultrasonic [27, 24, 20, 7]) are attached directly onto the face or inside the mouth to detect the movement of articulators involved in speaking (i.e. lips, tongue, jaw) to recover speech content. However, these contact sensing methods may interfere with daily activities in real-world scenarios. Other researchers have focused on non-contact detection methods. The non-contact approach often involves positioning a camera in front of the face and capturing lip movement to recognize predetermined phrases [39, 53]. However, these systems still require users to position the camera by either holding it in front of their faces or securing it onto something. There is a need for a non-contact wearable approach that has the potential to detect silent speech phrases.

### Summary

Compared to these prior projects, C-Face is the first wearable sensing device using minimally invasive common form factors that can continuously reconstruct full facial expressions with the position and shapes of the mouth, eyes and eyebrows.

## THEORY OF OPERATION

In 1996, this paper [38] put forward one of the most influential facial motion models. In the model, the human face is composed of three main layers: skin, muscles, and skull. When humans make facial expressions, their facial muscles contract and their jaw rotates. The skin, being connected to both the facial muscles and bones, gets pushed and pulled as these articulators move. This deformation of the facial skin alters the contours of the face. C-Face is designed based on the above model and observation. FingerTrak [18] demonstrated the feasibility of using contours of the wrist to continuously reconstruct full hand posture using wrist-mounted miniature cameras. Thus, we hypothesize that facial contours are also informative to predict full facial expressions.

To validate the feasibility of our proposed theory, one of the researchers fixed a camera under each ear. The researcher then made different facial expressions and captured his facial contours with the cameras while doing so. As shown in Fig. 3, the contours of the user's face uniquely varied depending on the facial expression. We then trained a simple machine learning model to distinguish between discrete facial expressions. With encouraging results, this preliminary experiment verified that facial contours are informative on predicting facial expressions.

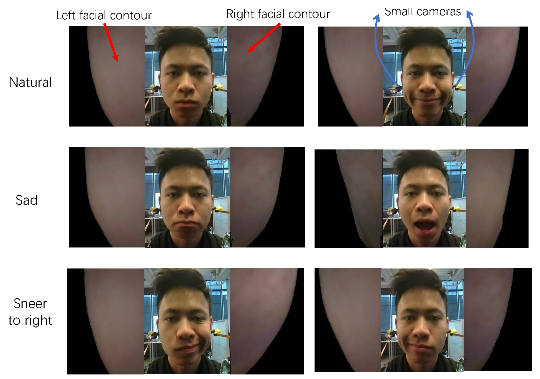


**Figure 3. Facial contour images captured in the preliminary feasibility test**

## HARDWARE AND FORM FACTOR DESIGN OF C-FACE

In this section, we discuss form factor design considerations, present our data acquisition system and an experiment to determine optimal camera positioning.

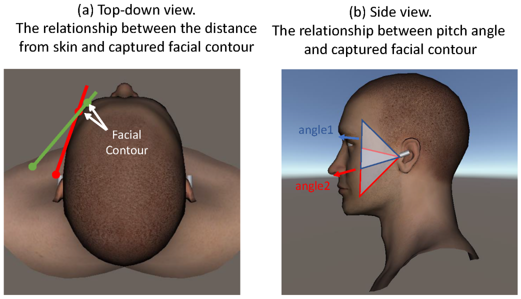
### Hardware

To capture facial contours, we had many options for sensing: depth cameras, an array of proximity sensors, thermal cameras, and RGB cameras. We intended to find a balance between practicality and capacity when choosing our sensing approach. Depth cameras are very informative, but too bulky to be attached to an earphone. The resolution of proximity sensor arrays and thermal cameras are too low. Only RGB cameras offer enough picture resolution while being compatible with form factors that are small in size and light enough to be worn on the ear. Admittedly, RGB cameras are susceptible to ambient light and background interference. However, our main focus of this paper is to demonstrate the feasibility of predicting full facial expressions from facial contours. Other sensing methods to capture facial contours can be tested in the future.

We built two types of form factors for our study: headphones and earphones as shown in figure 5. For the headphones, we embedded two 14x14mm ArduCAM cameras (OV5647 with adjustable focus, auto exposure and a view angle of 120x120 degrees) into the side earpieces. For the earphones, we secured two smaller 6x6mm cameras (OV5647 with fixed focus, auto exposure and the view angle of 64 x 48 degrees) to an attachment which slid onto the base of each earphone. Both cameras had an image resolution of 640x480 pixels and a frame rate of 30 fps. Each camera was connected to a Rasberry Pi board which read the images through a CSI interface. Next, we transmitted data from the Raspberry Pi boards to the server through WiFi or Ethernet for further data processing. Before data processing, the server synchronized the images from the right and left cameras.

### Form Factor Design and Camera Position Selection

We conducted an experiment to explore the influence of camera positioning on facial reconstruction performance.


(a) Top-down view. The relationship between the distance from skin and captured facial contour

(b) Side view. The relationship between pitch angle and captured facial contour

**Figure 4. The relationship between camera positions/angles and the facial contours captured, 3D mesh model© TurboSquid**

#### *Pre-experiment Considerations*

Fig 4(a) visualizes a top view of the 3D head model (by TurboSquid [2]) with two camera positions shown. The camera placed farther from the skin better captures skin deformation than the camera placed closer because it has a better line of sight to the facial contour (referred to by the and green lines in Fig.4(a)). Fig.4(b) visualizes a side view of the head with two camera angles shown. Naturally, when the camera tilts upward, it better captures the upper part of the face near the eye. When the camera tilts downward, it better captures the lower part of the face near the mouth. The closer the skin is to the mouth or eye, the more apparent its deformation. So, to maximize changes in facial contours captured by the cameras, it makes sense to keep both the eye and mouth in view.

#### *Experiment on the Camera Position*

To conduct our experiment, we designed 3D printed headphones with adjustable ear pieces to allow the camera's position and angle to be adjusted as shown in Fig. 5. We tested three camera distances from the skin (1 cm, 2 cm, 3 cm) and

[2] Turbosquid:https://www.turbosquid.com/Search/3D-Models.

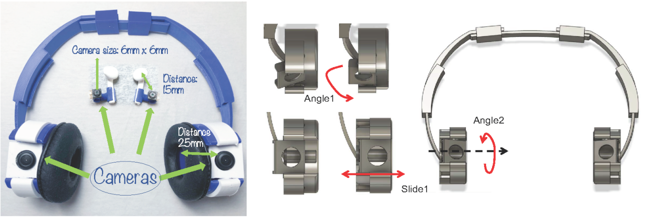


**Figure 5. Prototypes of headphones and earphones**

four angles (-10°, -20°, -30°, -40°). For each camera position and angle, a researcher was asked to make the same set of facial expressions. Images from the 12 different camera variations are shown in Fig. 6. The collected data from each variation was divided into training and testing datasets to train a deep learning model and evaluate its final performance. As the camera moves away from the skin, reconstruction performance improves according to the mean square distance (MSD) in pixels between our predictions and the ground truth. We determined that for our study, a distance of 1.5cm to 2.5cm would be an appropriate range. We therefore set the camera distance to 2.5cm from the skin for the headphones, and 1.5cm from the skin for the earphones. Intuitively, as the camera angle tilts downward, eye reconstruction worsens. As the camera angle tilts upward, mouth reconstruction worsens. We found that the total MSD reached its minimum value in the range of -20° -30°, where the camera images capture the skin around both the mouth and eye. We designed and built both form factors to allow their cameras to swivel upwards and downwards, and then set their angles within this range for the user study.

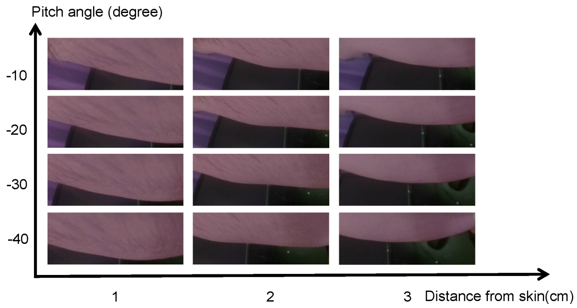


**Figure 6. Different facial contours from the perspective of the camera in different positions**

## CONTINUOUS FACIAL RECONSTRUCTION

With the captured contour images from the ear-mounted cameras, we put forward a CV-based pre-processing method and a deep learning model to reconstruct continuous facial expressions. Since our raw data is 2D images, we use a convolutional neural network (CNN) since they have shown exemplary performance in image classification, detection, and retrieval [26] compared to other traditional ML methods. In addition, previous works have applied CNN to human gesture reconstruction tasks (such as hand gestures [18], facial expressions [23], and body poses [2]) and demonstrated good performance. Therefore, we chose to use CNN to reconstruct facial expressions from facial contours. We first introduce the ground-truth acquisition method to our system for training and testing.

### Ground Truth Acquisition

*Definition of Ground Truth*

CV-based methods using a front-facing camera have achieved reliable performance tracking full facial expressions. Therefore, we set up a frontal camera and used a state-of-the-art computer vision library (Dlib library [28]) as our ground-truth acquisition method. Dlib library can extract 68 facial landmarks (feature points). When humans make different facial expressions, changes mainly occur in the mouth, eyes and eyebrows. So, we removed the less informative feature points, and selected 42 of the 68 landmarks (outlining the mouth, right eye, right eyebrow, left eye and left eyebrow) for our ground-truth for continuous reconstruction, as shown in Figure 7.

*Ground Truth Alignment*

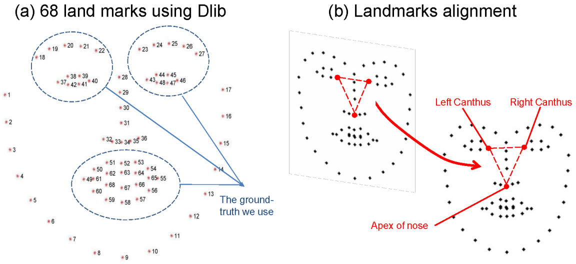


**Figure 7. Feature points that represent full facial expressions**

One common issue with CV-based methods is that the camera's view can influence results. During data collection, if a user slightly changes his or her facing direction, the acquired ground-truth for the same facial expressions can vary. To overcome this limitation, we first align the ground-truth before using these landmarks as output labels for training. Because changes in head position are relatively small, we regarded any resulting landmark transformations as affine transformations. To recover original landmark positions, we selected three landmarks whose relative positions change little when making facial expressions (right canthus, left canthus, and the apex of the nose as shown in Fig. 7). We then set these three landmarks fixed in the picture and used them to calculate the affine matrix for each ground-truth image. Using this method, we aligned the landmark positions to the same range and reduced the influence of head position change during ground-truth collection.

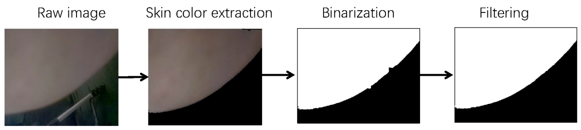


**Figure 8. The pre-processing procedure**

### Pre-processing Facial Contour Images

Pre-processing of the captured images is composed of four steps, as shown in Fig. 8. First, we convert the raw camera image color space from RGB to YCrCb, and apply the Otsu threshold [59] algorithm to extract the skin color from the background. Second, we seek out the max contour area (the facial contour) and remove other parts (background), which can help to decrease the influence of the background on our system's performance. Third, to reduce the influence of ambient light, we binarize the images after extraction. Another merit of binarization is that it can help to reduce the datasize of an image (from over 50KB to less than 5KB) during transmission. Lastly, we apply the morphological transformations (like erosion and dilation) and median filter to remove noisy points and smooth the image. These pre-processed images are sent to the deep learning model described below.

## Deep Learning Pipeline

As described in the preliminary experiments, to capture more detailed facial expressions on both sides of the face, we chose to use one camera on each side. We divided our ground-truth into two parts: left and right, as shown in Fig 9. We used the left and right facial contours to train the landmarks of the left and right side of the face, respectively. In our deep learning model, we implemented two identical deep convolutional neural networks respectively for estimating the facial landmarks on each side. This combination of two models performs better than using one model to predict landmarks on both sides.

### *Network Architecture*

The deep convolutional neural network has two parts: a backbone network and a regression network behind it, as shown in Fig. 9. The backbone network follows the same design as the 18-layer residual network [12] (ResNet-18), which has been proven to be highly effective for visual recognition tasks and less prone to over-fitting. A convolutional block in ResNet includes several convolution operations, each followed by batch normalization [21] and rectified linear unit (ReLU). A global average pooling is performed at the end of the backbone network to extract a vector representation of each image. Then, the extracted feature vector is inputted into the regression network, which consists of two fully connected layers with ReLU in between them and a dropout [51] ($p = 0.5$) before the last layer. The regression network outputs the landmarks of either side of the face. A matching module then concatenates the landmarks from the left and right by matching them to the same level using translation and scaling. It then outputs the final reconstruction result as 84 parameters (42 feature points x 2 dimensions for X and Y). Our deep learning model is built under the *Pytorch* framework.

### *Training Process*

**Weighted Loss Function.** Our model was trained with ground-truth facial landmarks using the weighted Huber loss function [19] to provide a robust regression. The weighted loss function is designed to tackle the common issue of imbalanced data in real-world data collection [40]. Since facial movement data is collected in continuous streams, the amount of background frames (frames of natural facial expressions) is significantly larger than foreground ones (frames of other facial expressions).

**Data Augmentation** When users remount form factors or the form factor has a small shifting, facial contours from the camera's POV can change with respect to translation, rotation and scaling. We apply data augmentations on the contour images in the training set to address this issue. Specifically, we set a probability of 0.5 to conduct certain transformations on the images before they are sent to the backbone network for training. The transformation can be shifting (range: -20 to 20 pixels), rotation (range: -10°to 10°), scaling (range: 0.8 to 1.2) or any combination of these. The distribution within each range following a Gaussian distribution ($\mathcal{N}(\mu = 0, \sigma^2 = 0.01)$).

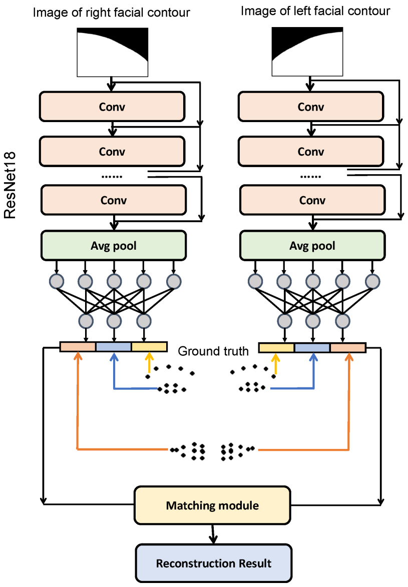


**Figure 9. The structure of our deep learning model**

**Training Parameter Setting.** We select the following training hyper-parameters based on the common practices established in previous CV research [15]: standard mini-batch stochastic gradient descent (SGD) with momentum (0.9), weight decay (1e-4), batch size 60, and a learning rate of 0.01 with Cosine learning rate annealing [34]. For all experiments, our model was trained for 50 epochs on the training set, with each epoch being a pass over the full training set with random shuffling. The trained model was further evaluated on a hold-out non-overlapping test set.

## USER STUDY

Due to the COVID-19 pandemic, recruiting participants for an in-person user study was extremely challenging. After discussing with the Institutional Review Board (IRB) in our university, we were allowed to recruit the roommates of two co-authors as participants in our study. We successfully recruited 7 roommates as study participants. To provide more information on C-Face's performance with different people amidst these guidelines, the two co-authors (P1 and P2) also participated in this study. They followed the same study procedure as the other participants. In total, we evaluated C-Face with 9 participants in our user study - 6 males and 3 females ranging from 21 to 25 years old.

### Setup

To begin the study, a researcher introduced the study protocol and answered any participant questions. Then, the participant sat down in a chair in front of a small RGB camera (Arducam cameras (OV5647) 30 Hz, 480 x 640) taped to a monitor for ground truth collection. The monitor would be used later to play pre-recorded instructional videos throughout the study. Next, the researcher adjusted the ground truth camera to a suitable position and angle to capture the user's facial feature landmark points. Next, the researcher helped the participant put on the earphones (the first form factor tested). Both the left and right cameras were adjusted to fit participants' face size and shape. We first adjust the camera positions to guarantee that each camera can capture the facial contour lines of the participant. Then, we adjust pitch angle (around -20°to -30°based on our camera positioning study) until the camera's frame can capture the participant's eye and mouth. For each form factor (headphones and earphones) the cameras were only adjusted one time before collecting data. Once data collection began, we did not adjust the cameras again. Before the study began, the participant was encouraged to practice making facial movements to ensure comfort.

### Data Collection

For data collection, we played a series of pre-recorded videos portraying faces for the participant to imitate. We first tested earphones followed by headphones. The earphone data collection procedure consisted of three categories: continuous reconstruction, emoji input detection, and silent speech detection. Data from the last two categories, emoji input detection and silent speech detection, were used to conduct a preliminary applications study which is outlined in the "Applications" section. For each category, training data and testing data were collected separately in different sessions (training data first, followed by testing data).

After collecting all training and testing data, the researcher reinserted the earphones and ran the continuous reconstruction testing session again to evaluate remounting performance. The researcher then removed the earphones and mounted the headphones on the participant's head. Thes headphone cameras were adjusted the same way as the earphones'. For headphones, we only tested continuous reconstruction. Similar to the earphones, we remounted them and ran the continuous reconstruction testing session again to evaluate remounting performance.

The continuous reconstruction training and testing videos prompted the participant to make the following facial expressions as Figure 2 shows: |o:| face (mouth "oh"), |a:| face (mouth "ah"), smile without teeth, smile with teeth, |sh| face (mouth "sh"), grimace with flat lips, right sneer, left sneer, right wink, left wink, raise eyebrows, frown and a combination of both eye motion and mouth motion. Figure 2 demonstrates the facial expressions rendered by a 3D human face mesh model provided by TurboSquid [3]. When mapping the facial expression data to the 3D mesh model, we adjusted the ratio of Dlib landmarks to fit with the model. Each expression was separated by a neutral, relaxed facial expression. We prompted each expression four times in the training video, and once in the testing video. The training video was about 6 minutes long and the testing video was 2 minutes. For each participant, since our sampling rate is 30 fps, we obtained an average of 6 x 60 x 30 = 10.8K samples for training and 2 x 60 x 30 = 3.6K samples for testing, leading to a total of (10.8K + 3.6K) x 9 = 129.6K samples. Each sample contains 2 images of both facial contours captured at the same time and one frontal image of the user's face for ground-truth labelling.

## CONTINUOUS RECONSTRUCTION RESULTS

### Evaluation Protocol

To evaluate the continuous reconstruction results, we used MSD (Mean Square Distance), calculating the mean of the square distances of each pair of landmarks between our prediction results and the ground truth. However, after adopting this method we found that many landmarks on more inactive parts of the face (like the lower eyelid) move very little (less than 1 mm) throughout the study. If we were to include all landmarks and weigh them equally, our results would be skewed since these inactive landmarks have a very low MSD regardless of facial expression. Thus, to more accurately represent our data quantitatively, we created a new variable called "MMAE (Major Mean Absolute Error)" comprised of the more meaningful landmarks. To find these landmarks, we first calculated the variance of each landmark using the collected ground-truth data as Fig. 10 shows. Higher variance signifies a higher degree of change for a given landmark during a facial expression. We then chose the landmarks with the highest variance capturing both the mouth and the eye to be considered major landmarks, as shown in Fig. 10. We derived MMAE from these major landmarks to better represent our results.

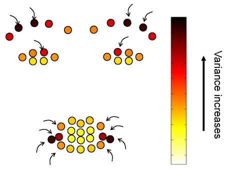


**Figure 10. The variance of landmarks during data collection. The arrows indicate chosen major landmarks.**

Since the landmarks are image points, the MMAE unit must be converted from pixels to a physical unit. We thus measured the positions of the right and left canthus relative to the nose apex and combined these values with the alignment to scale the landmark positions from pixels to millimeters.

### Continuous Reconstruction Evaluation

We evaluated our continuous facial reconstruction model on all nine participants with both form factors. Across all nine participants, our system achieves an MMAE of 1.43 mm (SD = 0.21 mm) and an MSD of 0.77mm (SD = 0.1 mm) for earphones, and an MMAE of 1.39 mm (SD = 0.28 mm) and an MSD of 0.74 mm for headphones (SD = 0.11 mm). Fig. 11 (a) (b) (c) shows the MMAE for mouth motion, eye motion and overall face motion for all nine participants with both form factors. To give an intuitive understanding of our reconstruction error, we calculate the MMAE for each facial expression

[3] Turbosquid:https://www.turbosquid.com/Search/3D-Models.

of participant 1 (p1) when wearing earphones and visualize the facial expressions with highest MMAE, median MMAE, and lowest MMAE as shown in Fig. 12. Additionally, we use a gradient to represent the error of each point. We can see most error occurs on the eyebrows, upper eyelids and the corners of the mouth. Fig. 2 provides a visualization for more reconstruction results. We also conducted a one-way ANOVA test on the MMAE results between earphones and headphones and found no significant effect (F(1,16) = 0.006 and p = 0.93).

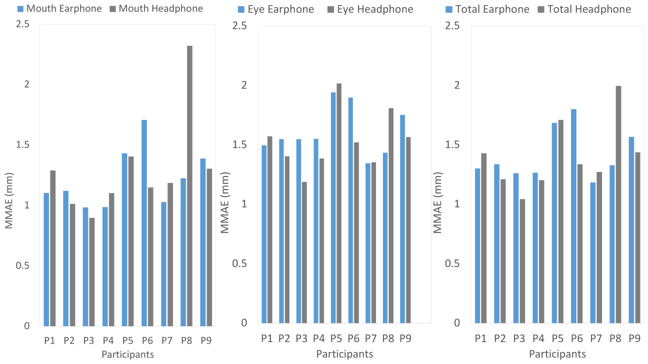


**Figure 11. The continuous reconstruction results for facial expressions**

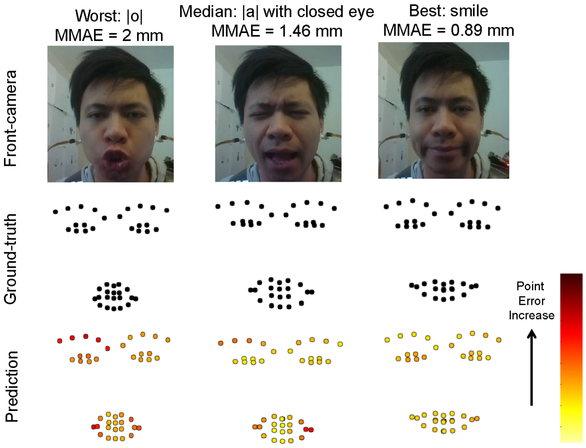


**Figure 12. The visualization for the expressions of highest, median and lowest MMAE of participant 1.**

### Form Factor Remounting Performance

During data collection, the device was remounted on the same participant to emulate a real-world scenario. Remounting the device can shift the camera positions and lead to adaptability challenges for our deep learning model. In the evaluation, we test the model after remounting both form factors. Additionally, to validate our data augmentation method, we train a model without data augmentation and test it with the new data after remounting.

The results for the models with and without data augmentation are summarized in Fig 13. For the earphones, with data augmentation, the MSD after remounting is 1.32 mm (SD = 0.37 mm) and the MMAE is 2.79 mm (SD = 1.1 mm). Without data augmentation, the MSD is 1.79 mm (SD = 0.77 mm) and the MMAE is 3.59 mm (SD = 1.5 mm). For the headphones, with data augmentation, the MSD after remounting is 0.867 mm (SD = 0.08 mm) and the MMAE is 1.72 mm (SD = 0.24 mm). Without data augmentation, the MSD rises to 1.10 mm (SD = 0.15 mm) and the MMAE to 2.14 mm (SD = 0.26 mm). Fig. 13 visualizes the remounting results (MMAE) among different participants and Fig. 14 displays the average ΔMMAE of remounting results with a 95% confidence interval. The results show that our data augmentation helped to decrease the remounting error by 26% for the earphones and 21.2% for the headphones. It demonstrated that our system can adapt to minor remounting variations for earphones and headphones and thus has potential for real world use cases. We ran a two-way ANOVA test to study possible effects between data augmentation and form factor. We found the main effect on data augmentation (F(1,32) = 3.9 and p = 0.05) and the main effect on form factor (F(1,32) = 14 and p = 0.001). There were no interactions between these two variables (F(1,32) = 0.35 and p = 0.55).

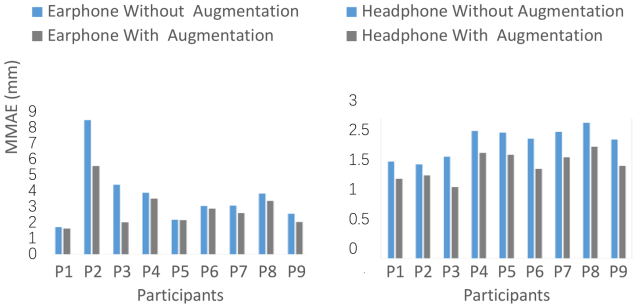


**Figure 13. The results of the remounting experiment among 9 participants.**

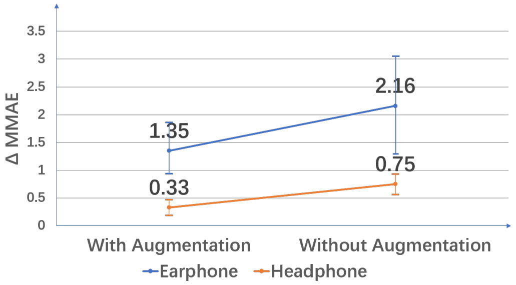


**Figure 14. The statistical results of the remounting experiment**

There are two elements of our remounting results which need further explanation. Firstly, even though the remounting results are encouraging, if the camera shifts too much (such as with participant 2), data augmentation may still not work, as the system may not be able to capture enough facial contour. Secondly, headphones performed better than earphones in the remounting experiment, and we think this was because earphones have a larger freedom of movement compared to headphones while being worn. A carefully designed earphone to constrain the freedom of movement can potentially alleviate this issue. Furthermore, a camera with a wider view could also help address this issue, as it could potentially capture all facial contours regardless of the camera's exact physical orientation.

## APPLICATIONS - EMOJIS AND SILENT SPEECH

To further evaluate the performance and potential applications of C-face, we selected two real-world scenarios using reconstructed facial expressions provided by C-Face: emoji input and silent speech input. Emoji input is an input technique whereby users can input emojis into a text message, online comment or post by imitating the desired emoji with a facial expression. Silent speech is a language interaction interface enabling speech-to-text input to take place without the need to voice an active acoustic signal. It has the potential to allow speech-to-text input in environments where speaking out loud may be difficult due to background noise or uncomfortable due to social norms.

### Data Collection Procedure

As a part of the user study, we collected data for emoji input and silent speech recognition from the same 9 participants. For emoji input recognition, we selected eight commonly used emojis to display in our emoji videos: natural, smile, laugh, angry, kissy-face, surprise, sneer and wink. For silent speech recognition, we chose eight commands designed to control a music player: "play", "stop", "next song", "previous song", "volume up", "volume down", "share", and "open lyric".

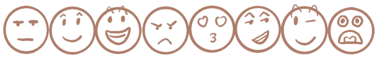


**Figure 15. The emojis prompted to the participants**

We prompted each emoji and each word 10 times in the training video, and five times in the testing video. For the silent speech portion of our study, participants were instructed to silently mouth commands that were displayed on the monitor. For each participant, we collected 15K samples on average for training and 6K samples for testing.

### Data processing pipeline

C-Face is able to predict 42 facial landmarks per frame. This allows us to create a data flow containing the landmark clusters. To recognize emoji and silent speech input gestures, we first segment the gestures in the data flow and then train a classifier to distinguish between various emoji faces or mouthed-words.

To segment the gestures, we begin by calculating the total difference between the predicted landmarks from C-Face and the landmarks from a natural facial expression each frame. Since facial landmarks change most significantly at the peak of each gesture (when making a face or mouthing a word), a peaking seeking algorithm is applied to find the peaks in the stream of landmark differences. We then sort out the primary peaks, which represent the gestures. These landmarks are software-segmented and inputted into the trained BLSTM described below.

Since the extracted events after segmentation are temporal series with variable lengths, we deploy a two-layer Bidirectional Long Short-Term Memory (BLSTM) model [16, 9] followed by a fully connected layer with a softmax function to classify them. The input of our classifier is a sequence of 42 landmark positions from the ResNet, and the output is a one-hot encoding of the facial event class. This BLSTM network was trained with batch size 30, stochastic gradient descent (SGD) with momentum (0.9), learning rate 0.001 and training epochs 50. A similar data augmentation method described in the previous section is applied on this labeled data and fed into the BLSTM to improve generalizability of the model.

### Evaluation

Once the data was collected, we used the data from the training set of each participant to train the deep learning model, which was evaluated using the testing set of that participant. The average emoji recognition accuracy is 88.6% (SD = 6%). and the average silent speech word accuracy is 84.7%(SD = 7.8%). The confusion matrix of emoji recognition and silent speech recognition results are shown in figure 16 and figure 17 respectively.

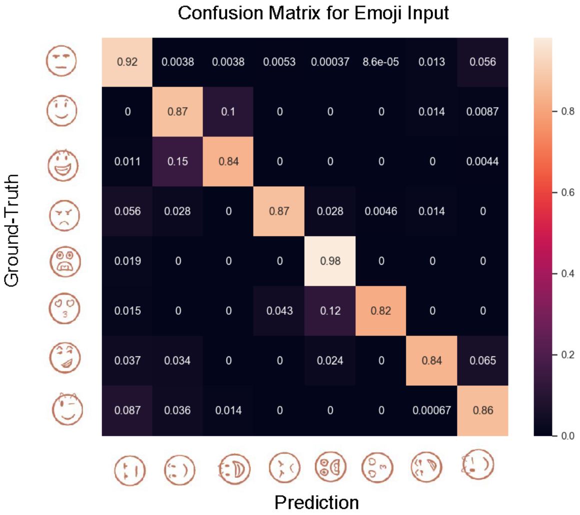


**Figure 16. The confusion matrix for emoji recognition**

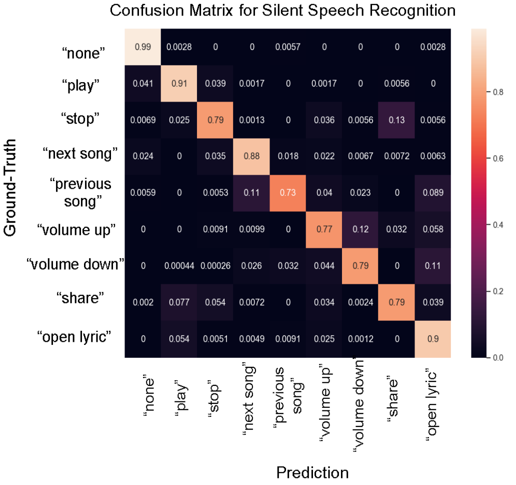


**Figure 17. The confusion matrix for silent speech recognition**

These two preliminary experiment results recognizing emojis and silent speech phrases validate the possibilities of putting

C-Face to use in real-world applications. However, the performance is not yet optimal. To better interpret these results, we first compared the results of feeding C-Face data into our gesture recognition protocol with ground truth data using the three lowest scoring participants (P2, P4, P6). We did this in order to gauge the extent to which inaccurate C-Face reconstructions were to blame for the sub-optimal performances. C-Face's emoji recognition accuracy for each of the three users was 89.76%, 77.05% and 78.9% respectively, while ground-truth results were 94.4%, 81.36%, and 83.88%. C-Face's silent speech recognition accuracy for each of the three users was 77.4%, 70.2% and 80.8% respectively, while ground-truth results were 77.8%, 67.45% and 73.8%. This indicates that gesture recognition performance using C-Face is comparable to direct CV-based methods, which is consistent with our continuous facial reconstruction results.

One possible factor that could have lowered performance is our small dataset size. We only required each participant to perform each utterance 15 times. If words were mouthed differently across utterances, our data may have lacked the diversity necessary for our model to accurately classify gestures. Also, given the small training data size, other ML models may have performed better than BLSTM. In the future we can focus on improving our classification results.

## DISCUSSION

### System Performance with Face Occlusion

Traditional CV methods using frontal cameras to reconstruct facial expressions are often unreliable if the face is partially blocked. For example, a camera pointed directly at the face relies on capturing images of the lips to determine the feature points of the mouth. If a user wears a face mask, their lips are blocked and the system fails to reconstruct their mouth. As face mask use increases worldwide amidst the COVID-19 pandemic, a system like C-Face could be especially practical and valuable going into the future.

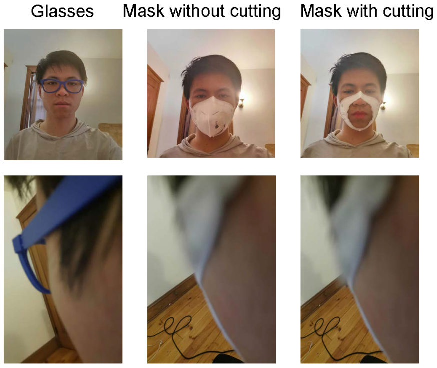


**Figure 18. The face and facial contours when wearing the mask and glasses**

We tested partial facial obstruction with two common wearables: eye glasses and face masks. For facial reconstruction data with eye glasses, we simply allowed the participants who naturally wore glasses to leave them on throughout the original study. After the study, we compared the average results from the users who wore glasses with users who did not.

To evaluate our system's performance while wearing a face mask, we conducted a short follow-up study on five participants in our original study. The study procedure went as follows: First, we cut a hole in the center of a face mask so that the ground truth camera could collect the facial feature points of the nose and mouth. Second, we secured the cut mask on the participant's face. Lastly, we put the earphones on the user, adjusted the cameras on them, and ran the continuous reconstruction data collection as described in the "User Study" section. We demonstrated that the camera views with the cut mask (for ground truth collection) and the regular uncut mask were the same by capturing pictures of both masks on one of the participants. Figure 18 shows that the cut hole had little to no effect on the captured image of the facial contours.

#### *Occlusion from Eye Glasses*

In our user study, we had four participants who naturally wore glasses: P5, P6, P7 and P9. For earphones, the average MSD of these four users is 0.824 mm (SD = 0.07 mm) and the average MMAE is 1.44 mm (SD = 0.19 mm), while the average MSD of the other 5 users without glasses is 0.743 mm (SD = 0.09 mm) and the MMAE is 1.392 mm (SD = 0.2 mm). For headphones, the MSD with glasses is 0.798 mm (SD = 0.09 mm) and the MMAE is 1.59 mm (SD = 0.26 mm), while the MSD without glasses is 0.703 mm (SD = 0.06 mm) and the MMAE is 1.24 mm (SD = 0.14 mm). Even though the MSD and MMAE of users with eye glasses is slightly higher, their overall performance is quite similar.

#### *Occlusion from a Face Mask*

In our mask blocking experiment, the mask training dataset and mask testing dataset of four participants (P1, P3, P5, P6) were fed into a deep learning model. Then, we compared the continuous reconstruction results of the participants when wearing and not wearing a mask. The four participants with a mask had an average MSD of 0.717 mm (SD = 0.07) and average MMAE of 1.36 mm (SD = 0.28 mm), compared with an average MSD of 0.801 mm (SD = 0.068 mm) and MMAE of 1.51 mm (SD = 0.24 mm) without a mask. The results show that the performance while wearing mask is even a little better than the results without a mask.

These results also demonstrate the strength of C-Face over other traditional CV methods. C-Face still provides reasonably accurate facial expression tracking even when part of the face is blocked.

### User-independent Models

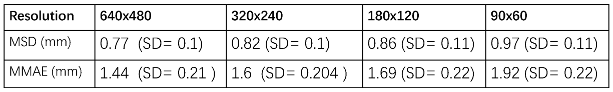
The results we presented so far use user-dependent models, where the training and testing data come from the same participant. We started with user-dependent models because contour shapes may vary between people. However, this model may introduce a burden as users need to provide training data before practical use of the system.

In order to understand how C-Face performs with user-independent models, we conducted a follow-up study using the data collected in the earphones user study with 9 participants for continuous facial reconstruction. We built user-independent models, where we used 8 participants' data to train the models, and one remaining participant's data as the testing set. We iterated this process over each participant. The average error across 9 participants of our 42 feature points (MSD) and 20 major feature points (MMAE) were 1.95 mm and 3.78 mm, respectively. The user-dependent model had smaller errors of 0.76 mm for MSD and 1.8 mm for MMAE. The results for user-independent models still show potential for C-Face to perform well on different participants if a larger training set can be obtained.

### Power Consumption

Improving C-Face's power consumption is a critical step towards real world deployment to conserve battery life. In our experiment, we use multiple Raspberry Pis which consume a relatively high amount of power (over $2W$ [32]). It prevents the current prototype from an immediate large-scale deployment. In future work, we can design the main control board ourselves with a low-power MCU, a wireless module, and other chips to reduce power use. For example, we could choose ESP, a low power MCU with a WiFi antenna ($210mW$), which could save around 89% of the power used by a Raspberry Pi.

Additionally, according to [54], lowering camera resolution could also reduce power consumption. However, lowering image resolutions may reduce facial reconstruction accuracy. To investigate this issue, we conducted an experiment where we tested four image resolutions (640x480, 320x240, 160x120 and 80x60) on the data set we collected from the 9 users wearing our earphones in the previous study. The results in Fig. 19 demonstrate that a lower resolution does lead to a slightly lower reconstruction accuracy. But with an image resolution of 90*60, C-Face can still reconstruct facial movements with a MSD of 0.97mm and MMAE of 1.02 mm. This indicates C-Face has potential to be a more energy efficient to be deployed in real-world devices.



| Resolution | 640x480 | 320x240 | 180x120 | 90x60 |
|---|---|---|---|---|
| MSD (mm) | 0.77 (SD= 0.1) | 0.82 (SD= 0.1) | 0.86 (SD= 0.11) | 0.97 (SD= 0.11) |
| MMAE (mm) | 1.44 (SD= 0.21 ) | 1.6 (SD= 0.204 ) | 1.69 (SD= 0.22) | 1.92 (SD= 0.22) |

**Figure 19. The reconstruction performance under different resolutions**

### Image Segmentation and Different Backgrounds

We conducted the user study indoors, but have yet to investigate how the system works in different environments. Varying light conditions and backgrounds can lead to poor image segmentation results, so we added noise in the segmentation process to address this issue during the training process in our experiment. However, this method may not work well when segmentation results are far from expected values. In the future, we plan to improve segmentation by training a deep model with a larger dataset dedicated to segment human skin from the environment. Also, we can add more sensing methods for segmentation, such as using a depth camera. Depth-cameras are large, but we believe they could become small enough to fit into an ear-mounted device in the future.

### Privacy Concerns

If C-Face is deployed, it may raise serious privacy concerns as the cameras can potentially capture information in private environments. There are ways to address this issue, like extracting features on the fly and not saving raw captured images. To fully address privacy concerns, we need a much longer discussion aside from our main paper topic. We plan to further explore this issue in the future.

### Improving Ground Truth Acquisition Method

Our current ground-truth acquisition method uses a state-of-the-art computer vision library to track facial expressions. This data acquisition method may not be the most accurate way to track facial expressions. It is an approximation of actual ground-truths, which limits the performance of C-Face. The ceiling of C-Face's performance is set by the performance of the CV library we chose. If there is a more accurate method for acquiring ground-truth data, C-Face's performance can potentially improve. Our CV-based ground-truth acquisition also limits the use cases of C-Face.

### Other Limitations

Like most CV-based systems, ambient light may influence C-Face's real-world implementation. The main procedure that may be affected is the facial contour extraction when preprocessing. If the ambient light becomes too bright or dim, the face segmentation algorithm may fail, affecting the final continuous reconstruction result. Also, longer hair can block C-Face's view of a user's facial contours. This issue could lead to inaccurate model predictions of landmark positions and in turn result in poor facial reconstruction.

## CONCLUSION

In this paper, we present C-Face, a minimally intrusive ear-mounted technology that continuously reconstructs full facial expressions by capturing the positions and shapes of the mouth, eyes and eyebrows. It uses two miniature cameras to capture the contours of the face, which are used to train a deep learning model to predict facial expressions. A user study with 9 participants demonstrated that C-Face can continuously reconstruct facial feature points with an MSD of 0.77mm using earphones and 0.74mm using headphones, and an MMAE of 1.43mm using earphones and 1.39mm using headphones. When the face is partially covered by a face mask or eye glass frame, C-Face can still reconstruct the facial feature points with an MSD of 0.717mm while wearing a mask, and 0.824mm while wearing glasses, and an MMAE of 1.36mm when wearing a mask and 1.44mm while wearing glasses.

## ACKNOWLEDGMENTS

This work is supported by Information Science Department at Cornell University. We thank all study participants for participating in the study, reviewers for constructive comments, and lab mates in Cornell SciFi Lab for providing feedback.

## REFERENCES

[1] 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.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008* (2018).

[3] George Caridakis, Lori Malatesta, Loic Kessous, Noam Amir, Amaryllis Raouzaiou, and Kostas Karpouzis. 2006. Modeling naturalistic affective states via facial and vocal expressions recognition. In *Proceedings of the 8th international conference on Multimodal interfaces*. 146–154.

[4] Lam Aun Cheah, James M Gilbert, José A González, Phil D Green, Stephen R Ell, Roger K Moore, and Ed Holdsworth. 2018. A Wearable Silent Speech Interface based on Magnetic Sensors with Motion-Artefact Removal.. In *BIODEVICES*. 56–62.

[5] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence* 38, 8 (2016), 1548–1568.

[6] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine* 18, 1 (2001), 32–80.

[7] Bruce Denby, Yacine Oussar, Gérard Dreyfus, and Maureen Stone. 2006. Prospects for a silent speech interface using ultrasound imaging. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. IEEE, I–I.

[8] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg. 2010. Silent speech interfaces. *Speech Communication* 52, 4 (2010), 270–287.

[9] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* 18, 5-6 (2005), 602–610.

[10] Anna Gruebler and Kenji Suzuki. 2010. Measurement of distal EMG signals using a wearable device for reading facial expressions. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 4594–4597.

[11] Jun He, Jianfeng Cai, Lingzhi Fang, and Z He. 2016. Facial expression recognition based on LBP/VAR and DBN model. *Appl. Res. Comput* 33 (2016), 453–461.

[12] 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.

[13] Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli. 2015. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. 73–80.

[14] Shan He, Shangfei Wang, Wuwei Lan, Huan Fu, and Qiang Ji. 2013. Facial expression recognition using deep Boltzmann machine from thermal infrared images. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 239–244.

[15] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2019. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 558–567.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[17] Robin Hofe, Stephen R Ell, Michael J Fagan, James M Gilbert, Phil D Green, Roger K Moore, and Sergey I Rybchenko. 2013. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Communication* 55, 1 (2013), 22–32.

[18] Fang Hu, Peng He, Songlin Xu, Yin Li, and Cheng Zhang. 2020. FingerTrak: Continuous 3D Hand Pose Tracking by Deep Learning Hand Silhouettes Captured by Miniature Thermal Cameras on Wrist. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–24.

[19] Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics*. Springer, 492–518.

[20] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone. 2010. Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips. *Speech Commun.* 52, 4 (April 2010), 288–300. DOI: `http://dx.doi.org/10.1016/j.specom.2009.11.004`

[21] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[22] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture Sensing Using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[23] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. 2017. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proceedings of the IEEE International Conference on Computer Vision*. 1031–1039.

[24] Eloi Moliner Juanpere and Tamás Gábor Csapó. 2019. Ultrasound-Based Silent Speech Interface Using Convolutional and Recurrent Neural Networks. *Acta Acustica united with Acustica* 105, 4 (2019), 587–590.

[25] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, and others. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 543–550.

[26] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. 2019. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review* (2019), 1–62.

[27] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: an ultrasound imaging-based silent speech interaction using deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[28] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.

[29] Irene Kotsia and Ioannis Pitas. 2006. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE transactions on image processing* 16, 1 (2006), 172–187.

[30] Ying-Hsiu Lai and Shang-Hong Lai. 2018. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 263–270.

[31] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–9.

[32] Robert LiKamWa, Bodhi Priyantha, Matthai Philipose, Lin Zhong, and Paramvir Bahl. 2013. Energy characterization and optimization of image sensing toward continuous mobile vision. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. 69–82.

[33] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. 2014. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1805–1812.

[34] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[35] Katsutoshi Masai, Yuta Sugiura, Masa Ogata, Kai Kunze, Masahiko Inami, and Maki Sugimoto. 2016. Facial expression recognition in daily life by embedded photo reflective sensors on smart eyewear. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 317–326.

[36] 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.

[37] Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. 2012. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. 501–508.

[38] Frederic I Parke and Keith Waters. 1996. *Computer facial animation*. CRC press.

[39] Stavros Petridis, Jie Shen, Doruk Cetin, and Maja Pantic. 2018. Visual-only recognition of normal, whispered and silent speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6219–6223.

[40] Huy Phan, Martin Krawczyk-Becker, Timo Gerkmann, and Alfred Mertins. 2017. DNN and CNN with weighted and multi-task loss functions for audio event detection. *arXiv preprint arXiv:1708.03211* (2017).

[41] Subramanian Ramanathan, Ashraf Kassim, YV Venkatesh, and Wu Sin Wah. 2006. Human facial expression recognition using a 3D morphable model. In *2006 International conference on image processing*. IEEE, 661–664.

[42] Ville Rantanen, Pekka-Henrik Niemenlehto, Jarmo Verho, and Jukka Lekkala. 2010. Capacitive facial movement detection for human–computer interaction to click by frowning and lifting eyebrows. *Medical & biological engineering & computing* 48, 1 (2010), 39–47.

[43] Ville Rantanen, Hanna Venesvirta, Oleg Spakov, Jarmo Verho, Akos Vetek, Veikko Surakka, and Jukka Lekkala. 2013. Capacitive measurement of facial activity intensity. *IEEE Sensors journal* 13, 11 (2013), 4329–4338.

[44] Marc'Aurelio Ranzato, Joshua Susskind, Volodymyr Mnih, and Geoffrey Hinton. 2011. On deep generative models with applications to recognition. In *CVPR 2011*. IEEE, 2857–2864.

[45] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. 2012a. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*. Springer, 808–822.

[46] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. 2012b. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*. Springer, 808–822.

[47] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014. The tongue and ear interface: a wearable system for silent speech recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers*. 47–54.

[48] Jocelyn Scheirer, Raul Fernandez, and Rosalind W Picard. 1999. Expression glasses: a wearable device for facial expression recognition. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*. 262–263.

[49] Tanja Schultz. 2010. ICCHP keynote: Recognizing silent and weak speech based on electromyography. In *International Conference on Computers for Handicapped Persons*. Springer, 595–604.

[50] Tanja Schultz and Michael Wand. 2010. Modeling coarticulation in EMG-based continuous speech recognition. *Speech Communication* 52, 4 (2010), 341–353.

[51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[52] Bo Sun, Liandong Li, Tian Zuo, Ying Chen, Guoyan Zhou, and Xuewen Wu. 2014. Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In *Proceedings of the 16th international conference on multimodal interaction*. 481–486.

[53] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 581–593.

[54] Anestis Terzis. 2016. *Handbook of camera monitor systems: The automotive mirror-replacement technology based on ISO 16505*. Vol. 5. Springer.

[55] Jacob Whitehill, Marian Bartlett, and Javier Movellan. 2008. Automatic facial expression recognition for intelligent tutoring systems. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1–6.

[56] Huiyuan Yang, Zheng Zhang, and Lijun Yin. 2018. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 294–301.

[57] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Ruichen Meng, Sumeet Jain, Yizeng Han, Xinyu Li, Kenneth Cunefare, Thomas Ploetz, Thad Starner, and others. 2018a. FingerPing: Recognizing fine-grained hand poses using active acoustic on-body sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–10.

[58] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. 2018b. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3359–3368.

[59] Minghua Zhao and Yonggang Zhao. 2010. Skin color segmentation based on improved 2D Otsu and YCgCr. In *2010 International Conference on Electrical and Control Engineering*. IEEE, 1954–1957.

[60] Ruicong Zhi, Markus Flierl, Qiuqi Ruan, and W Bastiaan Kleijn. 2010. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41, 1 (2010), 38–52.