CanalScan: Tongue-Jaw Movement Recognition via Ear Canal Deformation Sensing

Yetong Cao* Huijie Chen* Fan Li* Yu Wang†
* School of Computer Science and Technology, Beijing Institute of Technology, China
† Department of Computer and Information Sciences, Temple University, USA

***Abstract*—Human-machine interface based on tongue-jaw movements has recently become one of the major technological trends. However, existing schemes have several limitations, such as requiring dedicated hardware and are usually uncomfortable to wear. This paper presents *CanalScan*, a nonintrusive system for tongue-jaw movement recognition using only commodity speaker and microphone mounted on ubiquitous off-the-shelf devices (e.g., smartphones). The basic idea is to send an acoustic signal, then captures its reflections and derive unique patterns of ear canal deformation caused by tongue-jaw movements. A dynamic segmentation method with Support Vector Domain Description is used to segment tongue-jaw movements. To combat sensor position-sensitive deficiency and ear-canal-shape-sensitive deficiency in multi-path reflections, we first design algorithms to assist users in adjusting the acoustic sensors to the same valid zone. Then we propose a data transformation mechanism to reduce the impacts of diversities in ear canal shapes and relative positions between sensors and the ear canal. *CanalScan* explores twelve unique and consistent features and applies a Random Forest classifier to distinguish tongue-jaw movements. Extensive experiments with twenty participants demonstrate that *CanalScan* achieves promising recognition for six tongue-jaw movements, is robust against various usage scenarios, and can be generalized to new users without retraining and adaptation.**

I. Introduction

Being able to interact with the system naturally is becoming ever more important in the field of Human-Computer Interaction (HCI). In recent years, it has facilitated various types of HCI technology (e.g., speech recognition and gesture recognition). However, all of these manners are easy to be eavesdropped and only subject to healthy users.

In addition, tongue and jaw movement can present rich information with diverse motion combinations. Compared with the above interaction manner, it is good for privacy due to the hidden characteristic and allows interactions for those who have language barrier or poor finger coordination. Therefore, there is of great interest to develop a recognition algorithm for tongue-jaw movement to create an alternative human-computer interface (e.g., tongue-controlled wheelchairs [1], tongue-teeth typing systems [2], and silent speech output systems [3]).

Existing tongue-jaw movement recognition methods can be divided into three groups, which rely on cameras [4], oral cavity devices [1], [5], and wearable devices [2], [6], [7].

Fan Li is the corresponding author. The work of Fan Li is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No.62072040,61772077 and Beijing Natural Science Foundation under Grant No. 4192051.

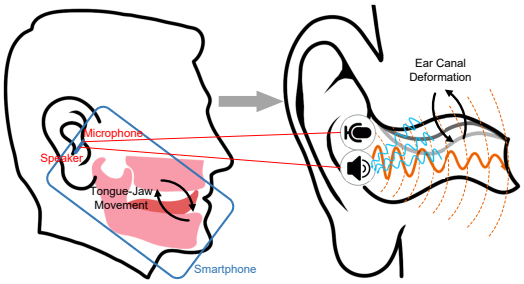

Fig. 1. Illustration of *CanalScan*.

These methods suffer from the following limitations: 1) for vision based methods, the advantage of hiding interactions is broken because only the out-of-mouth tongue movements can be recognized for interaction. Besides, these methods also face various issues on privacy, social awkwardness, and application scenarios (e.g., light condition, orientation). 2) for oral cavity device based methods, they not only suffer from obvious hygiene and intrusion disadvantages, but also may impair verbal communication and other oral functions. 3) for wearable device based methods, they require dedicated hardware with high cost, which makes them difficult to be adopted widely (especially in developing counties).

The above limitations motivate us to design a nonintrusive tongue-jaw movement recognition system, called *CanalScan*, which uses speaker and microphone integrated into ubiquitous off-the-shelf devices (e.g., smartphones) to detect tongue-jaw movements. Users can use the system simply by pressing the smartphone to their ears like making a phone call. The basic idea emerged from our finding that different tongue-jaw movements cause different amounts of movements of the ear canal wall in anterior-posterior, superior-inferior, and medial-lateral [8]. As illustrated in Fig. 1, the speaker and microphone serve as an active sonar that sends an acoustic signal into the ear canal and captures acoustic reflections. As the ear canal wall moves upon tongue-jaw movements, multi-path reflections interfere with each other, which leads to reflections with strength corresponding to the direction, speed, and intensity of the movement of the ear canal wall. These reflections are decoded for tongue-jaw movement recognition.

Despite its simple idea, three major challenges underlie the design of *CanalScan*:

1) Multi-path reflections are highly sensitive to ear canal shape and the relative position between the smartphone acoustic sensors and the ear canal, which makes it intractable to extract reliable features for recognition. We mainly through the effort of two sides to solve this. The first is to design a sensor position detection method to assist users in adjusting the smartphone acoustic sensors to the same valid zone every time they collect acoustic signals. The second is to design a data transformation mechanism to reduce the impacts of ear canal shape diversity and sensor position difference on the received signals.
2) The presence of extra movements between tow consecutive tongue-jaw movements, facial expressions, and head movements are common in real-world usage. They introduce jitter and pause similar to tongue-jaw movements in the received multi-path reflections [7], [9], which is challenging to distinguish. To address this, we segment movements based on dynamic threshold generated by a percentile measurement, and select tongue-jaw movements leveraging Support Vector Domain Description (SVDD) [10].
3) We observe that people exhibit different patterns for the same tongue-jaw movement and perform movement slightly differently from time to time. This makes it hard for the system to realize robust and user-independent recognition for tongue-jaw movements. To facilitate user-independent recognition, enhance robustness, and increase accuracy, we explore twelve kinds of features that are robust to user behavior diversity and movement inconsistency. Random Forest (RF) classifier is then adopted for tong-jaw movement recognition.

Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to develop a tongue-jaw movement-based human-computer interface in off-the-shelf devices. We use only commodity speaker and microphone to build an active sonar. By characterizing multi-path reflections induced by dynamic ear canal deformation, we investigate new measurement for tongue-jaw movement recognition.
- We design a set of novel techniques including a sensor position detection method and multi-path instability reduction method that overcome the ear-canal-shape-sensitive deficiency and sensor-position-sensitive deficiency in multi-path reflections, and a movement segmentation method that accurately segments and selects tongue-jaw movements from other interference movements. Also, we explore twelve kinds of features and adopt RF for final classification.
- We evaluate *CanalScan* with 20 participants extensively. The results show that *CanalScan* achieves 94.84% recall and 95.00% precision in tongue-jaw movement recognition. Results also show that *CanalScan* can generalize to new users without retraining or adaptation and is robust under various usage scenarios and environments.

II. Related Work

Previous studies on tongue and jaw movement recognition vary in sensing modalities and sensor placement. Tongible [4] leverages RGB camera to track tongue positions. However, only outside-mouth tongue movements can be detected, thus limits its application scope. Tongue drive [1] extracts rich tongue gestures by instrumenting the tongue with magnetic piercings. Sahni *et al.* [3] track tongue and jaw motion by tongue-mounted magnetic sensors with one headset mounted magnetometer, combined with a proximity sensor in the ear. TongueBoard [5] enables absolute position tracking of the tongue by placing 124 capacitive touch sensors on the roof of the mouth and holding a palate sensor in the mouth. However, the use of intraoral sensors is inconvenient, uncomfortable, and brings hygiene concerns. TYTH [2] uses the electroencephalography sensor, the electromyography sensor, and the miniature skin surface deformation sensor to identify tongue movement. Tongue-n-Cheek [6] captures tongue gestures using an array of radars integrated into helmets. TongueSee [11] realizes high-fidelity tongue gesture recognition using EMG signals from the surface of the skin. However, these techniques require sophisticated and dedicated hardware and users to wear noticeable sensors, which is socially awkward and might attract unnecessary attention in public.

The ear canal, which reflects the mouth-related activities, has drawn significant attention in recent years. Some prior contributions have been made in capturing ear pressure signals using barometers [7], [12] and microphones [9], [13], [14] embedded in earbuds to detect facial expressions, head movements, and tongue movements. However, measuring ear pressure changes requires to seal the ear canal, which can significantly affect hearing. Meanwhile, electrodes [15], infrared LEDs [16], and proximity sensors [17], which are placed inside the ear canal, have been exploited to recognize facial expressions and tongue movements. However, such dedicated hardware is not always available and is not compatible with off-the-shelf devices. Also, placing sensors inside the ear canal is uncomfortable and brings safety concerns. So far, tongue-jaw movement recognition through sensing in the ear canal still lacks highly accurate, robust, and nonintrusive solutions.

Another aspect of related researches focuses on using ear canal acoustic properties for authentication. These approaches use modified earphones with microphones. The basic principle is to send audible signals [18]–[21] or inaudible signals [22] and derive static characteristics of the ear canal shape. However, dynamic ear canal deformation is a relatively new area. Existing static-pattern-based works can not be applied directly because ear canal deformation always accompanies by the rotation of earphone prototypes they built. The position and direction change of the excitation signal will bring great interference, making it difficult for existing works to achieve high accuracy sensing. Moreover, [23] warns that these kinds of devices might result in the collapse of the external ear canal.

Over the past few years, many acoustic-based activity sensing systems have been developed on smartphones, such as

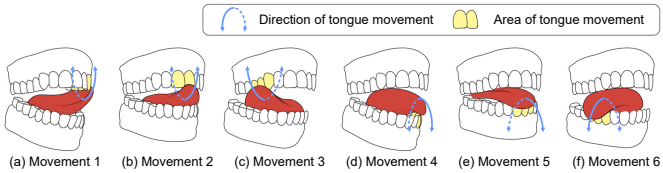

Fig. 2. Illustration of six movements involving the tongue and jaw.

hand tracking [24], [25], driving behavior sensing [26], [27], and breathing monitoring [28]. However, ear canal deformations are minute and have more subtle differences in acoustic properties, leading to more challenges for accurate sensing.

Compared with the previous efforts, *CanalScan* only relies on the built-in microphone and speaker on smartphones, does not require additional sensors and modification. While using *CanalScan*, a use hold the smartphone like making a phone call, which is unobtrusive, nonintrusive, and user-friendly. By analyzing dynamic acoustic properties of ear canal deformation, *CanalScan* achieves high accuracy in nonintrusive tongue-jaw movement recognition.

III. Observations

The ear canal is a roughly S-shape elliptical cylinder with a length of about 30 mm [29]. Ear canal shape and volume change upon tongue and jaw movements. When an acoustic signal is sent into the ear canal, ear canal deformations cause variations in acoustic reflections. Tongue and jaw reaching out to different areas cause different amounts of movements of ear canal wall in anterior-posterior, superior-inferior, and medial-lateral, which has been shown in studies such as [8], [15]–[17]. This motivates us to explore the feasibility of using acoustic reflections to characterize different tongue-jaw movements.

To understand the relationship between ear canal deformation and multi-path reflections, we conduct experiments on a smartphone that sends 16kHz continuous acoustic signals and continuously collects acoustic reflections at 48kHz. We recruit two volunteers to perform six tongue-jaw movements illustrated in Fig. 2, respectively. These tongue-jaw movements are performed in different areas of the oral cavity, and they are composed of two stages: (i) the tongue starts from the back of the teeth, lick over the teeth, reaches the front of the teeth, and the jaw moves naturally with the movement of the tongue. (ii) the tongue returns to the back of the teeth and jaw returns to its original position. During experiments, we ask volunteers to hold the smartphone like making a phone call and align the top microphone and earpiece speaker with the ear canal entrance. In particular, volunteer 1 rotates the smartphone counterclockwise around the sensor-to-ear axis by 135 degrees and 140 degrees, then collects continuous reflections twice, respectively. Volunteer 2 rotates the smartphone counterclockwise around the sensor-to-ear axis by 140 degrees and collects continuous reflections twice.

We extract the multi-path reflection envelope in the time window and illustrate examples of the vibration patterns of six tongue-jaw movements in Fig. 3.

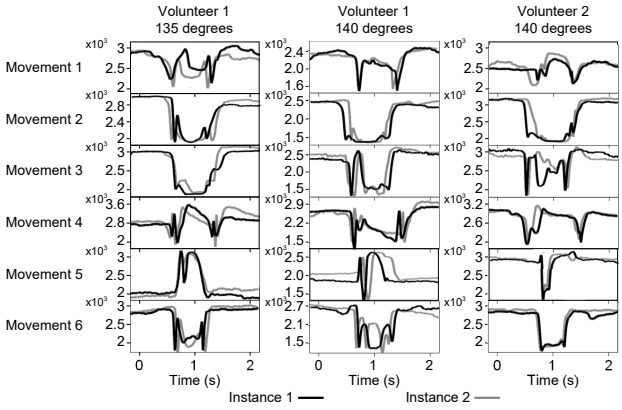

Fig. 3. The reflection envelope of movement 1-6 in three conditions.

Feasibility: We can easily observe that each kind of tongue-jaw movement has unique patterns in the reflection envelope, such as the same number of peaks, the same or near positions for peak, trough, and turning point. This demonstrates the feasibility of characterizing different tongue-jaw movements based on multi-path reflection from the ear canal.

Interference: Meanwhile, we can observe that two instances collected from the same movement in the same situation are slightly different in curve shape and signal amplitude. Also, when volunteer 1 rotates the smartphone acoustic sensor at different angles, envelopes from the same movement differ in curve shapes and signal amplitudes, such as movement 3 and 4. Moreover, when two volunteers rotate the sensor at 140 degrees, the same movement can have different curve shapes and signal amplitudes, such as movement 1, 3, and 5. The results demonstrate the impacts of movement inconsistency, acoustic sensor position difference, ear canal shape difference, and user behavior diversity.

According to our experiments, the presence of peak and trough is caused by changing movement directions of the ear canal wall. The curve shape and signal amplitude are related to ear canal shapes and sensor positions. Therefore, to address ear canal shape diversity and relative position difference between acoustic sensors and the ear canal, we need to modify information related to ear canal shape and sensor position (e.g., curve shape and peak amplitude) while keeping motional information (e.g., number of peaks and peak/trough position) unchanged.

IV. System Design

A. Overview

CanalScan utilizes off-the-shelf speaker and microphone integrated into smart devices (e.g., smartphone) for tongue-jaw movement recognition. Fig. 4 shows the overall design of *CanalScan*, which mainly comprised of four models: *Acoustic Signal Collection*, *Tongue-jaw Movement Segmentation*, *Multi-path Reflection Instability Reduction*, and *Tongue-jaw Movement Recognition*.

In *Acoustic Signal Collection*, the earpiece speaker and top microphone of a smartphone serve as an active sonar,

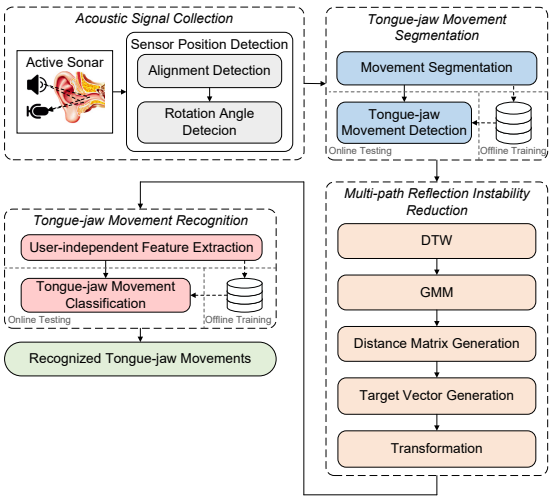

Fig. 4. *CanalScan* framework.

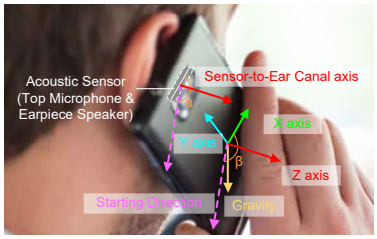

Fig. 5. Rotation angle and smartphone coordinate system.

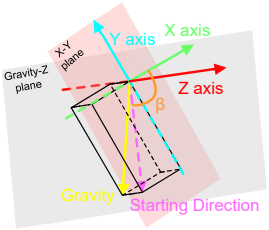

Fig. 6. Rotation angle and smartphone coordinate system.

which generates inaudible acoustic signals and collects their reflections. *Sensor Position Detection* is performed to monitor the relative position between acoustic sensors and the ear canal thus to assist users in placing the acoustic sensors in the same valid zone every time they use *CanalScan*.

In *Tongue-jaw Movement Segmentation*, we first segment all possible movement frames with a dynamic threshold. We then use a pre-trained Support Vector Domain Description (SVDD) [10] classifier to select real tongue-jaw movements from extra movements and non-tongue-jaw movements.

During *Multi-path Reflection Instability Reduction*, envelope segments of each tongue-jaw movement serve as input. We first apply Dynamic Time Warping (DTW) and Gaussian Mixture Model (GMM) to separate the input signal. We then leverage Kullback-Leibler (KL) divergence to generate a distance matrix that describes the similarity between Gaussian components from the input signal and envelope examples. Afterward, we select Gaussian components from examples that are most similar to Gaussian components of the input signal and generate a target vector. Finally, we transform the input signal into a new signal with characteristics of the target vector based on Minimum Mean Square Error (MMSE).

In *Tongue-jaw Movement Recognition*, *User-independent Feature Extraction* extracts twelve statistic features unique to each tongue-jaw movement and consistent across different users. A Random Forest (RF) classifier is used to obtain a prediction probability for each tongue-jaw movement. *CanalScan* takes prediction with the highest probability as the recognized tongue-jaw movement.

B. Acoustic Signal Collection

1) Acoustic Signal Selection: There are several considerations in selecting the excitation acoustic signal. It should be as inaudible as possible to avoid annoyance. Sounds above 16kHz are candidates because they are hard to hear for adults over 25 [30]. Most smartphones support a sampling rate of 48kHz, so the excitation acoustic signal is restricted to below 24kHz. However, speaker and microphone distortion at high frequencies narrows our choices to below 17kHz [31]. To enable *CanalScan* compatible with various smartphones, we send 16kHz sound to overcome the frequency selectivity of acoustic sensors and collect its reflection at 48kHz.

2) Sensor Position Detection: For reliable multi-path reflection collecting, two conditions need to be fulfilled. One is to allow the sensor to collect effective multi-path reflection, which has strength corresponding to the direction, speed, and intensity of the movement of the ear canal wall. The other is to minimize the relative position difference between the sensor and the ear canal entrance every time acoustic signals are collected. Note that the smartphone should be pressed on the ear to avoid interference from the surrounding environment. Thus, there needs no adjustment of the distance between the acoustic sensor and the ear canal entrance.

Alignment Detection Most readily available smartphones employ a slender earpiece speaker about 1cm long and mount a smaller top microphone inside the earpiece. The ear canal entrance of an adult is about the size of the speaker. Therefore, the acoustic sensor should be placed in a valid zone to collect effective multi-path reflection in the ear canal. In other words, the sensor should be aligned with the ear canal. However, it is very difficult to determine the relative position between the ear canal entrance and the acoustic sensor.

We solve this by a simple but efficient mechanism. We let users perform a pre-agreed tongue-jaw movement. If a unique pattern presents in the collected reflection signal, we consider that as aligned. Otherwise, we consider that as not aligned. Specifically, movement 4 that involves larger jaw and tongue motions is employed as the pre-agreed movement. We determine whether the acoustic sensors are aligned with the ear canal by checking whether the reflection envelope has more than two peaks or troughs with prominence higher than 30% of the maximum prominence of the highest peak and lowest though, which is observed through experiments.

Rotation Angle Detection: To measure the angle of smartphone rotating around an axis, coordinate system conversion is often required to address data variety caused by users facing different directions. However, data conversion between coordinate systems is time-consuming. Instead, we design a lightweight algorithm to work in different facing directions. Fig. 5 shows an example of the smartphone coordinate system,

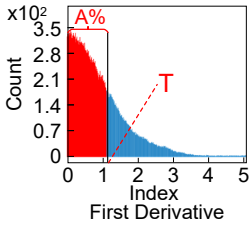

Fig. 7. Calculation of T based on intensity distribution.

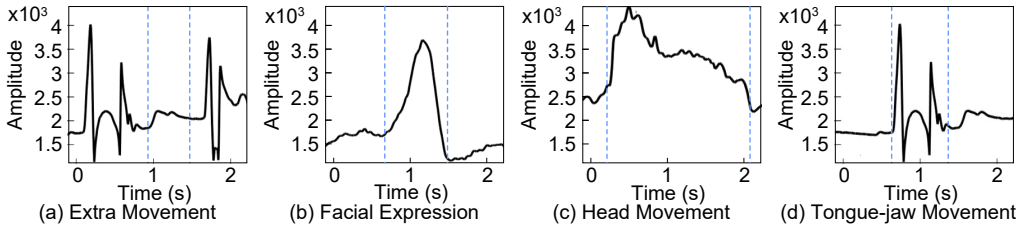

Fig. 8. Examples of extra movement, facial expression, head movement, and tongue-jaw movement.

sensor-to-ear-canal-axis, and rotation angles. We define the intersection line of the X-Y plane and gravity-Z plane along the smartphone's bottom as the start direction of rotation. We define the acoustic sensor rotating α degrees around the sensor-to-ear-canal-axis, and the smartphone rotates β degrees around its Z-axis. When the acoustic sensor is aligned with the ear canal, the sensor-to-ear-canal-axis is parallel or nearly parallel to the smartphone Z-axis. We can easily derive that α equals to β, which is the angle between the starting direction and the smartphone X-axis. Therefore, we now turn to the problem of obtaining β. Fortunately, inertial measurement unit mounted on modern smartphones provide easy access to such tilt angle:

$$\beta = acctan(\frac{g_x}{g_y}) + \frac{\pi}{2} \quad (1)$$

where g_x and g_y is the gravity component in X and Y axis. Gravity is typically derived from the accelerometer where the magnetometer and the gyroscope help remove the linear acceleration from the data.

According to our experiment with 50 people, a comfortable posture of holding the smartphone close to the ear canal (like making a phone call) is to make the smartphone rotates 130-140 degrees. By calculating the smartphone rotation angles, we guide the users to rotate the smartphone at the same or similar angle when collecting signals. Thus, we can minimize the relative position difference between the acoustic sensor and the ear canal during each collection and mitigate the impacts of various relative positions on multi-path reflections.

C. Tongue-Jaw Movement Segmentation

Tongue-jaw movement segmentation is a two-step process: the first step is to segment all candidate movements; the second step is to select tongue-jaw movements from other movements.

1) Movement Segmentation: The tongue and jaw pause for a very short while between two consecutive tongue-jaw movements to felicitate segmentation. Intuitively, we can segment movements by detecting a pause and a huge jitter in the envelope signal. We make use of the fact that the first derivatives of jitters are high, and the first derivatives of pauses are low and relatively stable. The first derivative that exceeds a certain threshold at a point is considered the start of a movement, and that is below a certain threshold for a while after a point is considered the end of a movement. The threshold T must be sufficiently small to capture all tongue-jaw movements but sufficiently large to avoid capturing random noise in the collected signal. However, finding the threshold suitable for everyone is extremely difficult due to diversity in movement amplitude range and noise uncertainty. Therefore, we determine a dynamic threshold by using a percentile measurement procedure.

Given the absolute value of the first derivative of the input signal, we first calculate its intensity distribution $I(a)$, which is weighted according to the scattering intensity of signal strength a [32]. Then, the threshold T is calculated as $\int_0^T I(a) = A\%$. Fig. 7 shows an example of calculating threshold T based on intensity distribution. We set A as 63 based on our experimental study.

2) Tongue-jaw Movement Detection: Extra movements of tongue and jaw are often required when switching between two consecutive tongue-jaw movements. In addition, facial expressions, head movements, and other movements are common in real-world use. To avoid high computational costs and misclassification, we only take real tongue-jaw movements for further process and recognition.

Fig. 8 illustrates the envelope of extra movement between two consecutive movements, facial expression, head movement, and tongue-jaw movement, respectively. The blue dash lines mark the start and end of each movement. A key observation is that tongue-jaw movements have more peaks, and the peaks are sharper. This motivates us to discriminate six tongue-jaw movements and other movements using a statistical-based method. We first extract features to represent each segmented movement, including *kurtosis, standard derivation, length*, and *the number of peaks*. Then, we use a classifier to select tongue-jaw movements. Since non-tongue-jaw movements are unpredictable and training the classifier with limited samples leads to limited accuracy, we employ a one-class classifier, SVDD. We take six tongue-jaw movements as a whole to train a tongue-jaw movement-class. SVDD determines the boundary of the tongue-jaw movement-class and assigns a sample to that class according to whether it falls within or outside the boundary. After that, facial expressions, head movements, extra movements, and other movements outside the boundary are discarded, and tongue-jaw movements are further processed and recognized by the following proposed techniques. Specifically, SVDD receives 93.88% recall and 91.93% precision, which is described in Section V-E.

D. Multi-path Reflection Instability Reduction

Multi-path reflections are highly sensitive to ear canal shape and the relative position between the smartphone acoustic sensor and the ear canal. To overcome the instability in multi-path reflection caused by these factors and facilitate robust tongue-jaw movement recognition, we propose a data transformation technique.

1) Design Guidelines: We aim to reduce pattern instability through a transform function. Such a transformation process involve three design guidelines:

- Data from the same tongue-jaw movement should be more similar after transformation.
- Data from different tongue-jaw movements should be distinct after transformation.

Based on the above goals and our discussion in Section III, we aim to modify information related to ear canal shape and sensor position (e.g., curve shape and peak amplitude) while keeping motional information (e.g., number of peaks and relative peak/trough position) unchanged. The basic idea is to generate a representative target vector for each type of tongue-jaw movement, then derive the statistical relations between target vector and the collected data and finally transform the collected signal into a new signal with characteristics of the target vector.

2) Data Transformation Process: Fig. 9 illustrates the process of data transformation. The envelope samples are random selections of representative envelopes. We consider the envelope of the newly collected data $\boldsymbol{x}$ and stored envelope examples of six tongue-jaw movements $\boldsymbol{y}_m, m = 1, 2...6$ are vectors with different lengths.

Step 1: We first adopt the DTW method to process them. After that, $\boldsymbol{x}$ and $\boldsymbol{y}_m$ are time-aligned.

Step 2: We then apply Gaussian Mixture Model (GMM) to represent them as the sum of K multivariate Gaussian function:

$$P_{\boldsymbol{x}} = \sum_{i=1}^{K} \alpha_i \mathcal{N}(\mu_i, \sigma_i), \tag{2}$$

$$P_{\boldsymbol{y}_m} = \sum_{j=1}^{K} \beta_j \mathcal{N}(\mu_j, \sigma_j), \tag{3}$$

where $\mathcal{N}$ is the normal distribution with the constraints that $\sum_{i=1}^{K} \alpha_i = 1, \alpha_i \geqslant 0$ and $\sum_{j=1}^{K} \beta_j = 1, \beta_j \geqslant 0$.

Step 3: Since we do not know what kind of tongue-jaw movement is performed, we introduce a distance matrix to find the most similar components in the stored templates. Specifically, we adopt the Kullback–Leibler (KL) divergence to measure the distance of two Gaussian components. Each entry $D_{i,j}$ of the distance matrix is calculated as:

$$D_{i,j} = \frac{1}{2}[KL(\mathcal{N}_{\mu_i,\sigma_i} || \mathcal{N}_{\mu_j,\sigma_j}) + KL(\mathcal{N}_{\mu_j,\sigma_j} || \mathcal{N}_{\mu_i,\sigma_i})], \tag{4}$$

where the KL divergence is defined as:

$$KL(\mathcal{N}_{\mu_i,\sigma_i} || \mathcal{N}_{\mu_j,\sigma_j}) = log\frac{\sigma_j}{\sigma_i} + \frac{(\mu_i - \mu_j)^2 + \sigma_i^2 - \sigma_j^2}{2\sigma_j^2}. \tag{5}$$

Step 4: Then, we search the distance matrix to find K components from the Gaussian distribution set that are most similar to the K components from the collected data. In our case, those with the minimum distance is considered as the most similar components. We add up K components in the form of GMM to obtain probability density of the representative target vector $\boldsymbol{y'}$:

$$P_{\boldsymbol{y'}} = \sum_{i=1}^{K} \gamma_i \{\mathcal{N}(\mu_j, \sigma_j) | arg\, min D_{i,j}\}. \tag{6}$$

By applying Bayes's rule, the weight of each component is defined as follows:

$$\gamma_i = \frac{\alpha_i \mathcal{N}(\mu_i, \sigma_i)}{\sum_{j=1}^{K} \alpha_j \mathcal{N}(\mu_j, \sigma_j)}. \tag{7}$$

Step 5: We now turn to the problem of finding a transformation function to transform the collected data $\boldsymbol{x}$ into the target vector $\boldsymbol{y'}$. Motivated by speech transformation [33], we introduce a transformation function $\mathcal{F}(x)$ assumed by the Minimum Mean Square Error (MMSE) estimation:

$$\begin{aligned} \mathcal{F}(\boldsymbol{x}) &= E(\boldsymbol{y'}|\boldsymbol{x}) \\ &= \int \boldsymbol{y'} \frac{P(\boldsymbol{x}, \boldsymbol{y'})}{P_x(\boldsymbol{x})} d\boldsymbol{y'}, \end{aligned} \tag{8}$$

where $P_x(\boldsymbol{x})$ is the probability density of $\boldsymbol{x}$, which is modeled by Equ (2). The joint probability density $P(\boldsymbol{x}, \boldsymbol{y'})$ should be modeled carefully to refine the description of the statistical distribution of $\mathbf{x}$ and $\mathbf{y}'$. Therefore, we apply GMM to model the joint vector $\boldsymbol{z} = [\boldsymbol{x}^T, \boldsymbol{y'}^T]^T$. The choice of GMM is based on its ability to provide a "*soft classification*", and the desired transformation relationship between the target vector and the collected data only relies on their time index. The two-dimensional joint probability density is defined by:

$$P_{\boldsymbol{z}} = \sum_{i=1}^{K} \omega_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \sum_{i=1}^{K} \omega_i = 1, \quad \omega_i \geqslant 0, \tag{9}$$

where mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ defined by:

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_i^{\boldsymbol{x}} \\ \mu_i^{\boldsymbol{y'}} \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} cov(\boldsymbol{x}, \boldsymbol{x}) & cov(\boldsymbol{x}, \boldsymbol{y'}) \\ cov(\boldsymbol{y'}, \boldsymbol{x}) & cov(\boldsymbol{y'}, \boldsymbol{y'}) \end{bmatrix}, \tag{10}$$

where cov is the covariance operator. To fit GMM with the weights, means, and covariance matrix, we adopt the Expectation Maximization (EM) algorithm.

Proceeding as before yields a transformation function from Equ (8) in the following:

$$\mathcal{F}(\boldsymbol{x}) = \sum_{i=1}^{M} P_{\boldsymbol{y}}(C_i|\boldsymbol{x})[\mu_y + \frac{cov(\boldsymbol{y}, \boldsymbol{x})}{cov(\boldsymbol{x}, \boldsymbol{x})}(\boldsymbol{x} - \mu_x)], \tag{11}$$

in which $P_{\boldsymbol{y}}(C_i|\boldsymbol{x})$ is the conditional probability that $\boldsymbol{x}$ belongs to component C_i. Through the application of Bayes's rule, it is easily derived that $P_{\boldsymbol{y}}(C_i|\boldsymbol{x})$ can by calculated using Equ (7). Using more GMM mixture components can better model the signal, but also cause high computational costs. In our case, 12 GMM components are used. After processed by data transformation, differences between data from the same tongue-jaw movement are effectively reduced, and data

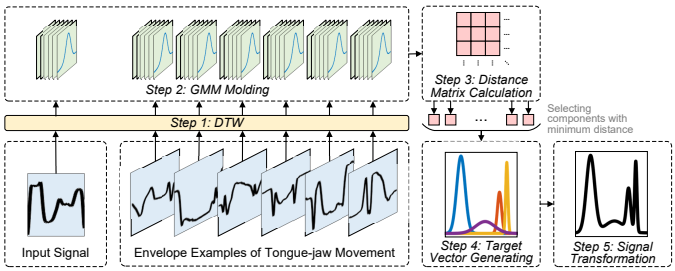

Fig. 9. Illustration of data transformation process.

from different tongue-jaw movements and non-tongue-jaw movements are still distinct.

Since this data transformation technology reduces the impacts of ear canal shape diversity and phone position difference on the reflection signal, it improves the average recall from 69.35% to 91.41%, and the average precision from 70.46% to 91.58%. Experiment details are described in Section V-E.

E. Tongue-Jaw Movement Recognition

1) Feature Extraction: Intuitively, we can recognize different tongue-jaw movements with similarity matching (e.g., DTW method). However, it is arduous to generate standard templates for each type of tongue-jaw movement because of the diversity of movements performed by different users. Instead, we extract unique and consistent statistic features of each type of tongue-jaw movement. The basic idea is to build a database with profiles of each type of tongue-jaw movements before classification, and use the database to train a classifier to infer the performed tongue-jaw movement.

We extensively explore plenty of features and apply RF classifier to rank these features by a feature importance feedback. Afterward, we select twelve kinds of features that contribute most to recognize various tongue-jaw movements and consistent across different users, including *variance, absolute energy, vectorized approximate entropy, autocorrelation, count above/below mean, the first location of maximum/minimum, linear least-squares regression, the mean over the absolute differences between subsequent time series values, mass center index*, and *energy ratio of ten chunks.*

2) Tongue-Jaw Movement Classification: We employ Random Forest (RF) to train a six-class classier to recognize different types of tongue-jaw movements. We feed twelve kinds of features extracted from reflection envelopes into the RF classifier and obtain prediction probabilities for the input data. Then we take prediction with the highest probability as the recognized tongue-jaw movement. Although several classifiers such as decision tree, support vector machine, and k-nearest neighbor perform well in related works, we chose RF because it has the best performance in our experimentally study, which is presented in Section V-E.

V. Evaluation

A. Implementation

We implement *CanalScan* to verify its performance in recognizing tongue-jaw movements. In our proof-of-concept implementation, we use LIBAS [34] to send acoustic signals at 16kHz and receive the reflections at a sampling rate of 48kHz. LIBAS is a cross-platform framework, which simplifies the development of acoustic-based applications. We pair the sensing smartphone with Matlab by using LIBAS's server-client remote mode.

B. Experimental Setup

We recruit 20 adult participants (10 male and 10 female) for evaluation. This study is conducted with the approval of our institute's IRB. All participants are healthy, right-handed, and cleaned their ears before collecting experimental data. During data collection, participants align the top microphone and earpiece speaker with their ear canals and press the smartphone tightly. To accommodate slight sensor position differences, we encourage participants to rotate the smartphone 130-140 degrees. Participants are asked to respectively perform the six tongue-jaw movements for 5 sessions, each session includes 10 rounds, and each round lasts 2-4 minutes. Between sessions, every participant take a five minutes break. To understand the performance of *CanalScan* against various issues, we ask participants to collect data with various sensor rotation angles and different devices. After the first phase of data collection, we collect data from all participants one month later to validate the long-term performance. To evaluate *CanalScan*'s performance under different usage scenarios, we ask participants to collect data in three common usage conditions: standing, sitting in a moving car, and standing on a moving bus. The start and end of each tongue-jaw movement are indicated by clicking a computer mouse using their left hands. We use the following metrics to evaluate *CanalScan*:

Confusion Matrix: Each row and column of the matrix represent the ground truth and the predicted results, respectively. Each entry $c_{i,j}$ is the ratio of instances belonging to the i^{th} class predicted as the j^{th} class to all instances belong to the i^{th} class.

Precision: the ratio of the instances correctly classified as label A to all instances predicted as label A.

Recall: the ratio of the instances correctly classified as label A to all instances belong to label A.

C. Overall Performance of Tongue-jaw Movement Recognition

We conduct five-fold cross-validation to evaluate the tongue-jaw movement recognition performance of *CanalScan*. The average recall and precision of *CanalScan* are 94.84% and 95.00%, respectively. The results demonstrate that *CanalScan* achieves accurate recognition of tongue-jaw movements. Fig. 10 shows the confusion matrix of the recognition results of *CanalScan*. Each entry is the average result of five sessions across 20 participants. The entries on the diagonal show the average accuracy of recognizing each tongue-jaw movement, which reaches 94.06%, 93.23%, 94.99%, 96.90%, 95.08%, and 94.78%, respectively. We observe that movement 4 and 5 receive higher recall and precision. A possible reason is that they involve more significant movements of the lower jaw,

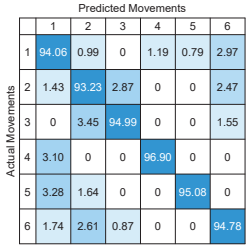

Fig. 10. Overall recognition performance.

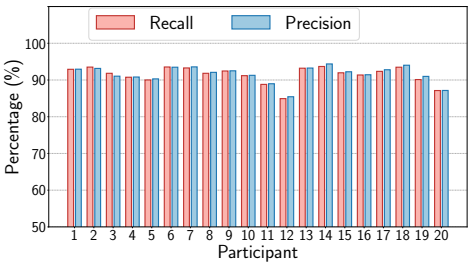

Fig. 11. Performance of leave-one-person-out-validation.

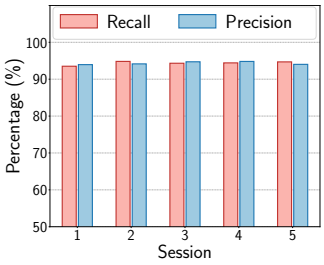

Fig. 12. Performance of leave-one-session-out-validation.

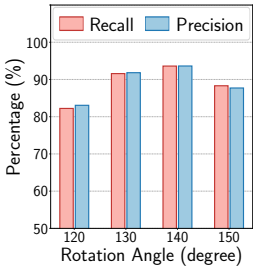

Fig. 13. Impact of rotation angle.

making the ear canal deformation reflections more distinguishable than other tongue-jaw movements.

D. Use Issue Study

We study the performance of *CanalScan* from many aspects, including universality across users, stability against movement inconsistency, the impacts of smartphone sensor rotation angles, the result of various devices, and a long-term study.

1) Universality: To understand whether *CanalScan* can generalize to new users without retaining or adaptation, we conduct leave-one-person-out-validation. We use data from one participant for testing and data from nineteen participants for training. Evaluations of all combinations are shown in Fig. 11. We observe that 17 participants have recall higher than 90% and precision higher than 90%. And the average recall and precision are 91.41%, and 91.58%, respectively. These excellent results suggest *CanalScan* can effectively work across different users. Participate 12 has relatively low performance. We carefully check the recognition results of different tongue-jaw movements from participate 12 and find that movement 1 contributes the most to error. The study of this special case is left as future work.

2) Stability: To evaluate the stability of *CanalScan* against movement inconsistency, we conduct leave-one-session-out-validation, where data from one session are used for testing and the remaining data for training. As shown in Fig. 12, the results reach 94.35% average recall and 94.33% average precision across sessions. The leave-one-session-out-validation results show good agreement with the cross-validation results, confirm that *CanalScan* works effectively against movement inconsistency.

3) Impact of Sensor Rotation Angle: We use the pre-trained classifier described in Section V-C to evaluate *CanalScan*'s robustness against different sensor rotation angles. Four angles are tested, including 120 degrees, 130 degrees, 140 degrees, and 150 degrees. Fig. 13 shows the recognition results under these four conditions. The recall results of four cases are 82.22%, 91.58%, 93.60%, and 88.31%. The precision of four cases are 83.06%, 91.82%, 93.62%, and 87.71%, respectively. *CanalScan* receives the highest recall and precision at 140 degrees. As participants place the smartphone outside the valid zone of 130-140 degrees, the multi-path reflection in the ear canal changes significantly, resulting in decreases of recall and precision.

4) Impact of Device: CanalScan's performance is related to the hardware of smart devices. Therefore, we conduct cross-validation experiments on data collected from four different devices. Specifically, we implement LIBAS to collect acoustic signals with iPhone X, iPhone 8, HUAWEI Mate 9, and HUAWEI Mate 9pro. These devices differ in size and audio hardware. Fig. 14 shows the comparison of the recall and precision across four devices. The results show that *CanalScan* is highly effective with all devices. There is no noticeable difference in their tongue-jaw movement recognition results. This indicates that our system is compatible with different mobile phone modules.

5) Long-Term Performance: Existing related approaches that send and receive acoustic signals in the ear canal do not support long-term use. They focus on the static characteristics of the ear canal shape. However, ear wax, naturally produced by the human body, can greatly affect the static characteristics. Our proposed system focuses on dynamic characteristics: the direction, speed, and amount of the ear canal wall movement. We conduct a long-term experiment, where data collected in the first data collection phase are used for training, and data collected one month later for testing. Also, we conduct a five-fold-cross-validation with data collected from two data collection phases. When using data collected one month later for testing, the average recall is 92.26%, and the average precision is 92.18%. Meanwhile, the cross-validation recall and precision of data collected from two data collection phases show good performance, reaching 94.06% and 93.64%, respectively. The results suggest a regular update of the training data set of *CanalScan* enables high accurate tongue-jaw movement recognition.

E. Key Algorithm Study

We evaluate the performance of movement segmentation, tongue-jaw movement detection, data transformation, and various classifiers.

1) Performance of Movement Segmentation: We segment the movement between the start and end points indicated by the computer mouse and compare them with the segmentation results based on the dynamic threshold. Experiment results show that 90% of the time difference between segments

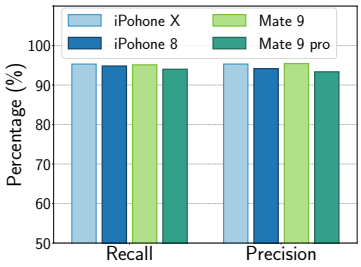

Fig. 14. Performance under four different smartphones.

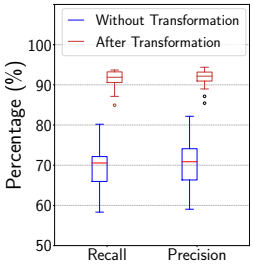

Fig. 15. Performance of multi-path reflection instability reduction.

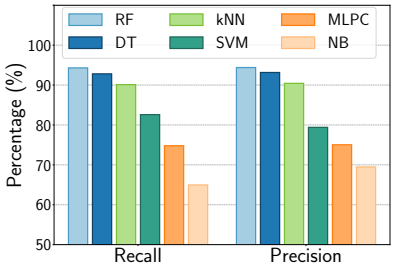

Fig. 16. Performance of different classifiers.

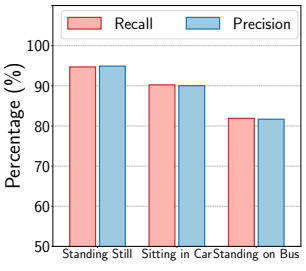

Fig. 17. Performance in three different usage scenarios.

and ground truth is less than 0.1s, which demonstrates the effectiveness of the proposed method.

2) Performance of Tongue-jaw Movement Detection: By carefully checking the results of SVDD classifier, we found that 93.88% of the tongue-jaw movement is correctly detected. While in the segments classified as tongue-jaw movements, 91.93% of them truly belongs to the tongue-jaw movement class, which shows that *CanalScan* can effectively detect tongue-jaw movement. This result can be improved by fusing other sensory data, which is part of our future work.

3) Performance of Multi-path Reflection Instability Reduction: The proposed data transformation technique provides an efficient mechanism for *CanalScan* to reduce the impacts of ear canal shape diversity and sensor position difference on the received signals. Fig. 15 compares the leave-one-person-out validation results without and after data transformation. When we do not perform data transformation, the average recognition recall is 69.35%, and the average precision is 70.46%. The recall and precision of participants with the worst results are both lower than 60%. After data transformation, we observe a significant increase in recognition results, which reach 91.41% average recall and 91.58% average precision. Form the results, data transformation shows high efficiency and is the key to realize accurate tongue-jaw movement recognition.

4) Impacts of Training Data Size and Classifier: We first compare the impact of training data size by varying the training data size from 25% to 85%. We find that with more training data, the system could receive higher recall and precision. Then, we compare the performance of several highly used classifier including Random Forest (RF), Decision Tree (DT), k-Nearest Neighbor(kNN), radial basis function kernel Support Vector Machine (SVM), Multi-layer Perceptron Classifier (MLPC), and Naive Bayes (NB), All classifiers are implemented with dealt values. Fig. 16 shows the tongue-jaw movement recognition performance of different classifiers. We can observe that RF and DT have better performance than other classifiers. When using 85% data for training, RF achieves its best recall of 94.30% and best precision of 94.39%. Since RF classifier outperforms than the other classifiers, we employ RF to recognize different tongue-jaw movements.

F. Usage Scenarios

To validate if *CanalScan* can work well in various usage scenarios, we ask the participants to collect data in three conditions: standing still, standing on a moving bus, and sitting in a moving car. The classifier is trained as described in Section V-C. The recognition results of six tongue-jaw movements are shown in Fig. 17. Standing still yields the highest recall and precision, which are 94.71% and 94.91%, respectively. The performance of sitting in a moving car is slightly worse. The recall decreases to 90.24%, and precision decreases to 90.03%, which is acceptable in real environments. In terms of standing on a moving bus, body vibration obfuscates acoustic reflections in the ear canal, resulting in 81.90% recall and 81.68% precision.

VI. Conclusion

In this paper, we propose a nonintrusive tongue-jaw movement recognition system, *CanalScan*. Our system only relies on commodity speaker and microphone mounted on ubiquitous off-the-shelf devices (e.g., smartphones), which sends an inaudible acoustic signal to the ear canal, then capture its multi-path reflections. By deriving unique patterns of ear canal deformation caused by tongue-jaw movements, *CanalScan* is capable of recognizing six tongue-jaw movements. *CanalScan* adopts a set of novel signal processing techniques. Extensive experiments with twenty participants demonstrate that *CanalScan* reaches the goal of accurate, robust, and user-independent recognition of six tongue-jaw movements. However, the general methods proposed in this work can be extended to other tongue-jaw movements easily.

CanalScan still has limitations and spaces to improve. Firstly, *CanalScan* currently focuses on six tongue-jaw movements. To support more tongue-jaw movements for more complex applications, extensions of proposed methods (such as different classification methods or advanced features) can be further explored. Secondly, the proposed data transformation method is the most time-consuming step in *CanalScan*. Further investigation to reduce its computational cost can improve the overall latency of recognition. Thirdly, *CanalScan* is currently implemented on smartphones. We will further evaluate its performance with other wearable devices such as a headset, which may expand the applications to wider ranges.

REFERENCES

[1] G. Krishnamurthy and M. Ghovanloo, "Tongue drive: a tongue operated magnetic sensor based wireless assistive technology for people with severe disabilities," in *Proc. IEEE ISCAS'06*, 2006, pp. 5551–5554.

[2] P. Nguyen, N. Bui, A. Nguyen, H. Truong, A. Suresh, M. Whitlock, D. Pham, T. Dinh, and T. Vu, "TYTH-Typing on your teeth: Tongue-teeth localization for human-computer interface," in *Proc. ACM MobiSys'18*, 2018, pp. 269–282.

[3] H. Sahni, A. Bedri, G. Reyes, P. Thukral, Z. Guo, T. Starner, and M. Ghovanloo, "The tongue and ear interface: A wearable system for silent speech recognition," in *Proc. ACM ISWC'14*, 2014, pp. 47–54.

[4] L. Liu, S. Niu, J. Ren, and J. Zhang, "Tongible: A non-contact tongue-based interaction technique," in *Proc. ACM ASSETS'12*, 2012, pp. 233–234.

[5] R. Li, J. Wu, and T. Starner, "Tongueboard: An oral interface for subtle input," in *Proc. ACM AH'19*, 2019, pp. 1–9.

[6] Z. Li, R. Robucci, N. Banerjee, and C. Patel, "Tongue-n-cheek: Non-contact tongue gesture recognition," in *Proc. IPSN'15*, 2015, pp. 95–105.

[7] T. Ando, Y. Kubo, B. Shizuki, and S. Takahashi, "Canalsense: Face-related movement recognition system based on sensing air pressure in ear canals," in *Proc. ACM UIST'17*, 2017, pp. 679–689.

[8] J. Grenness, M.J.and Osborn and W. Weller, "Mapping ear canal movement using area-based surface matching," *Sensors*, vol. 111, no. 3, pp. 960–971, 2002.

[9] T. Amesaka, H. Watanabe, and M. Sugimoto, "Facial expression recognition using ear canal transfer function," in *Proc. ACM ISWC'19*, 2019, pp. 1–9.

[10] D. M. J. Tax and R. P. W. Duin, "Support vector domain description," *Pattern Recogn. Lett.*, vol. 20, pp. 1191–1199, Nov. 1999.

[11] Q. Zhang, S. Gollakota, B. Taskar, and R. P. Rao, "Non-intrusive tongue machine interface," in *Proc. ACM CHI'14*, 2014, pp. 2555–2558.

[12] B. Maag, Z. Zhou, O. Saukh, and L. Thiele, "Barton: Low power tongue movement sensing with in-ear barometers," in *Proc. IEEE ICPADS'17*, 12 2017, pp. 9–16.

[13] R. Vaidyanathan and C. J. James, "Independent component analysis for extraction of critical features from tongue movement ear pressure signals," in *Proc.IEEE EMBS'17*, 2007, pp. 5481–5484.

[14] M. Mace, K. A. Mamun, S. Wang, L. Gupta, and R. Vaidyanathan, "Ensemble classification for robust discrimination of multi-channel, multi-class tongue-movement ear pressure signals," in *Proc. IEEE EMBS'11*, 2011, pp. 1733–1736.

[15] D. J. C. Matthies, B. A. Strecker, and B. Urban, "Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions," in *Proc. ACM CHI'17*, 2017, pp. 1911–1922.

[16] K. Taniguchi, H. Kondo, M. Kurosawa, and A. Nishikawa, "Earable tempo: A novel, hands-free input device that uses the movement of the tongue measured with a wearable ear sensor," *Sensors*, vol. 18, no. 3, p. 733, 2018.

[17] A. Bedri, D. Byrd, P. Presti, H. Sahni, Z. Gue, and T. Starner, "Stick it in your ear: Building an in-ear jaw movement sensor," in *Proc. ACM UbiComp/ISWC'15 Adjunct*, 2015, pp. 1333–1338.

[18] T. H. M. Akkermans, T. A. M. Kevenaar, and D. W. E. Schobben, "Acoustic ear recognition," in *Proc. ICB'06*, 2006, pp. 697–705.

[19] A. H. M. Akkermans, T. A. M. Kevenaar, and D. W. E. Schobben, "Acoustic ear recognition for person identification," in *Proc. IEEE AUTOID'05*, 2005, pp. 219–223.

[20] T. Arakawa, T. Koshinaka, S. Yano, H. Irisawa, R. Miyahara, and H. Imaoka, "Fast and accurate personal authentication using ear acoustics," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2016, pp. 1–4.

[21] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin, "EarEcho: Using ear canal echo for wearable authentication," *Proc. ACM IMWUT.*, vol. 3, no. 3, Sep. 2019.

[22] S. Mahto, T. Arakawa, and T. Koshinak, "Ear acoustic biometrics using inaudible signals and its application to continuous user authentication," in *Proc. IEEE EUSIPCO'18*, 2018, pp. 1407–1411.

[23] I. M. Ventry, J. B. Chaiklin, and W. F. Boyle, "Collapse of the ear canal during audiometry," *Archives of Otolaryngology-head and Neck Surgery*, vol. 73, no. 6, pp. 727–731, 1961.

[24] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys'17, 2017, pp. 15–28.

[25] H. Chen, F. Li, and Y. Wang, "Echotrack: Acoustic device-free hand tracking on smart phones," in *Proc.IEEE INFOCOM'17*, 2017, pp. 1–9.

[26] X. Xu, J. Yu, Y. Chen, Y. Zhu, and M. Li, "Steertrack: Acoustic-based device-free steering tracking leveraging smartphones," in *Proc. IEEE SECON'18*, 2018, pp. 1–9.

[27] Y. Xie, F. Li, Y. Wu, S. Yang, and Y. Wang, "D3-guard: Acoustic-based drowsy driving detection using smartphones," in *Proc. IEEE INFOCOM'19*, 2019, pp. 1225–1233.

[28] K. Qian, C. Wu, F. Xiao, Y. Zheng, Y. Zhang, Z. Yang, and Y. Liu, "Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices," in *Proc. IEEE INFOCOM'18*, 2018, pp. 1574–1582.

[29] R. Oliveira and G. Hoeker, "Ear canal anatomy and activity," *Semin Hear*, vol. 24, pp. 265–275, 11 2003.

[30] M. Jilek, D. Suta, and J. Syka, "Reference hearing thresholds in an extended frequency range as a function of age," *Journal of the Acoustical Society of America*, vol. 136, no. 4, p. 1821, 2014.

[31] H. Lee, T. H. Kim, J. W. Choi, and S. Choi, "Chirp signal-based aerial acoustic communication for smart devices," in *Proc. IEEE INFOCOM'15*, 2015, pp. 2407–2415.

[32] F. Yan, H. Zhang, and C. R. Kube, "A multistage adaptive thresholding method," *Pattern Recognition Letters*, vol. 26, no. 8, pp. 1183–1191, 2005.

[33] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[34] Y. C. Tung, D. Bui, and K. G. Shin, "Cross-platform support for rapid development of mobile acoustic sensing applications," in *Proc. ACM MobiSys'18*, 2018, pp. 455–467.