# Tongue-Jaw Movement Recognition Through Acoustic Sensing on Smartphones

Yetong Cao, *Student Member, IEEE,* Fan Li, *Member, IEEE,* Huijie Chen, Xiaochen Liu, and Yu Wang, *Fellow, IEEE*

**Abstract**—Past tongue-jaw movement interaction systems typically require dedicated hardware and are uncomfortable to use, limiting their scalability and generalizability. This paper introduces *CanalScan*, the first system that recognizes tongue-jaw movements using commodity speakers and microphones mounted on ubiquitous off-the-shelf devices (e.g., smartphones). What inspires us is that tongue-jaw movements always cause ear canal deformations, and we find that for different tongue-jaw movements, dynamic features of ear canal deformations present unique patterns on acoustic reflections in the ear canal. Specifically, *CanalScan* first sends an acoustic signal to the ear canal, then parses the reflection signals for tongue-jaw movements recognition. To eliminate the impacts of body movements, we develop a body movement noise filtering method and a dynamic segmentation method to identify and separate the tongue-jaw movements-associated ear canal deformations from other types of body movements. We further propose a sensor position detection method and 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 validate the generalizability, effectiveness, robustness, and high accuracy of *CanalScan*.

**Index Terms**—Human-computer interaction, tongue-jaw movement, multi-path reflection, random forest.

## 1 Introduction

In recent years, tongue-jaw movement-based interactions have gained particular attention due to various benefits: (i) As one of the most natural movements, tongue-jaw movements are easy to perform and can present rich information with diverse motion combinations, and (ii) Compared to traditional interaction manners (e.g., speech recognition and gesture recognition), tongue-jaw movements are good for privacy due to the hidden characteristic and allow interactions for those who have language barrier or poor finger coordination. Tongue-jaw movement recognition systems thus have gained particular attention to create an alternative human-computer interface (e.g., tongue-controlled wheelchairs [1], tongue-teeth typing systems [2], and silent speech input systems [3]). As shown in Fig. 1, changes in tongue-jaw movements are converted into user control commands that communicate to the targeted devices in the user's surrounding environment.

There have been active research efforts for recognizing tongue-jaw movements. The computer-vision-based methods [4], [5] can track tongue positions using cameras. However, it requires a camera positioned in front of the user's face and the users to stick their tongues out, thereby preventing them from being applied to many scenarios (e.g., dim lighting environment and face covered by a mask) and impairing hiding interactions. Besides, these methods also raise privacy concerns. Alternatively, there exist approaches employing oral cavity devices to recognize tongue and jaw movements [1], [6]. However, all these approaches suffer from obvious hygiene and intrusion drawbacks and impair verbal communication and other oral functions. Additionally, various wearable-sensor-based methods have been developed to monitor tongue and jaw movements [2], [7], [8]. However, the requirement of dedicated hardware with high cost hinders them from being adopted widely.

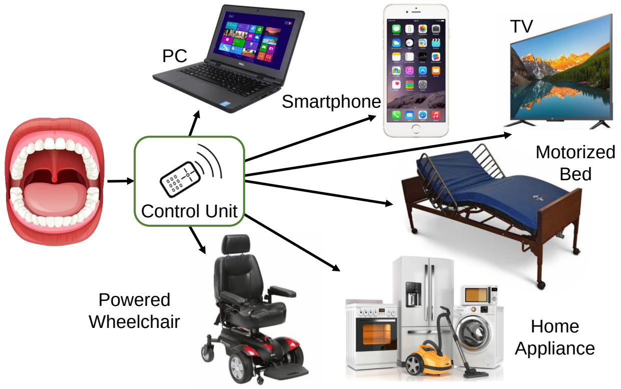


Fig. 1. Interaction applications based on tongue-jaw movements.

To circumvent the limitations of prior works, this paper proposes *CanalScan*, a novel approach that uses speaker and microphone integrated into ubiquitous commodity devices (e.g., smartphones) to support accurate tongue-jaw movement recognition in real-life environments. As shown in

- *Y. Cao, F. Li, and X. Liu are with School of Computer Science, Beijing Institute of Technology, Beijing, 100081, R.P.China.*
  *E-mail: {yetongcao, fli, xiaochenliu}@bit.edu.cn*
- *H. Chen is with School of Computer Science, Beijing University of Technology, Beijing, 100124, R.P.China.*
  *E-mail: chenhuijie@bjut.edu.cn*
- *Y. Wang is with Department of Computer and Information Sciences, Temple University, Philadelphia, Pennsylvania 19122, USA.*
  *E-mail: wangyu@temple.edu*
- *F. Li is the corresponding author.*

*Manuscript received MONTH DAY, YEAR; revised MONTH DAY, YEAR.*

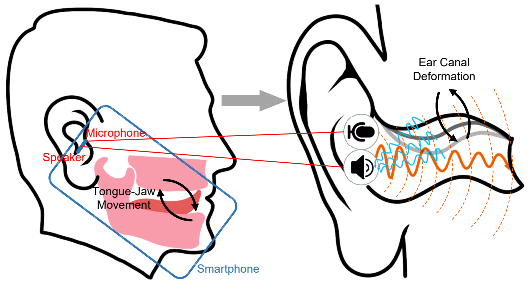


Fig. 2. An illustration of *CanalScan*.

Fig. 2, our idea emerges 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, which is also supported by existing researches [8], [9]. *CanalScan* employ the speaker and microphone on the smartphone to measure such ear canal wall movements (we refer to them as ear canal deformations) and interpret them into tongue-jaw movements.

Realizing an accurate system that captures tongue-jaw movements in real-life environments is very challenging. Our design should consider the following aspects:

1) We profile the ear canal deformation by multi-path reflections in the ear canal to decode tongue-jaw movements. However, the multi-path reflections are sensitive to ear canal shape and the relative position between the smartphone acoustic sensors and the ear canal entrance, making it intractable to profile the ear canal deformation reliably. To handle this, we design a sensor position detection method to ensure that the smartphone acoustic sensors are placed in the same valid zone every time the users collect acoustic signals. Furthermore, we design a data transformation mechanism to reduce the instability of sensor measures.
2) The presence of extra movements between two 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 [8], [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) To ensure reliable measures, prior works mostly require users to have minimal body motions when collecting sensory data. However, activities of daily living (e.g., walking) and passive movements from vehicle moving are inevitable in real-life environments. These body movements usually distort the reflection patterns and impair recognition of tongue-jaw movements. Therefore, we propose a body movement noise filtering method to detect the presence of noise in the collected signals and remove the corresponding fragments.
4) We profile the tongue-jaw movements from novel features of the acoustic signal reflected by the ear canal wall, which has not been explored previously. Although initial works show that different face-related movements can produce unique ear canal deformation consistent over multiple users [8], [9], the relationship between the acoustic measurements and tongue-jaw movements remains unclear. To facilitate user-independent recognition, enhance robustness, and increase accuracy, we explore twelve significant features that are robust to user behavior diversity and movement inconsistency. Random Forest (RF) classifier is then adopted for tongue-jaw movement recognition.

In summary, our contributions are 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 the commodity speaker and microphone to build an active sonar. By characterizing multi-path reflections induced by dynamic ear canal deformation, we investigate new measurements for tongue-jaw movement recognition.
- We design a set of novel techniques to eliminate the impact of ear canal shape diversity and sensor position variations on multi-path reflections. Specifically, we propose a sensor position detection method that monitors the relative position between acoustic sensors and the ear canal. Moreover, we propose a multi-path instability reduction method that selects the most significant movement examples and converts the collected signal into a new signal conducive to user-independent tongue-jaw movement recognition.
- We propose novel algorithms to realize accurate and robust tongue-jaw movement recognition in real-life environments, including a body movement noise filtering method that detects the presence of noise in acoustic reflections and removes the polluted fragments, 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 movements recognition. Results also show that *CanalScan* can generalize to new users without retraining or adaptation and is robust under various usage scenarios and environments.

The rest of the paper is organized as follows. We first review several related works in Section 2. Then we show the preliminary in Section 3. Section 4 presents the details of system design of *CanalScan*. The evaluation of the system is presented in Section 5, followed by the discussion of future works in Section 6. Finally, Section 7 concludes the paper.

## 2 Related Work

There have been many studies on recognizing tongue and jaw movements due to their good privacy and capability to present rich information. These methods vary in sensing modalities and sensor placement. Among them, Tongible [4] leverages RGB camera to track tongue positions. However, it can only detect the outside-mouth tongue movements, which limits its application scope. In addition, some methods use intraoral sensors to track tongue and jawbone positions. For example, Tongue drive [1] utilizes magnetic piercing devices instrumented inside the mouth that monitors rich tongue gestures. Sahni *et al.* [3] design a tongue and jaw motion tracking system combining magnetic sensors mounted on the tongue, proximity sensors inside the ear, and a headset mounted magnetometer. TongueBoard [6] is a tongue position tracking system that uses 124 capacitive touch sensors on the roof of the mouth and a palate sensor holding in the mouth. However, the use of intraoral sensors is inconvenient, uncomfortable, and brings hygiene concerns. To avoid the above drawbacks, some attempts have been made that leverage wearable devices for permitting convenience and comfort. TYTH [2] uses the electroencephalography sensor, the electromyography sensor, and the miniature skin surface deformation sensor to identify tongue movement. Tongue-n-Cheek [7] 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, the requirement of expensive and dedicated hardware prevents them from being adopted widely.

Alternatively, the ear canal has drawn significant attention due to the close correlation between tongue-jaw movements and ear canal wall motions. Some prior contributions have been made in capturing ear pressure signals using barometers and microphones embedded in earbuds to detect facial expressions [8], [9] and tongue movements [12]–[14]. These approaches achieve good performance over multiple users and at different times, demonstrating the feasibility of capturing the ear canal deformation to recognize tongue-jaw movements. However, measuring ear pressure changes requires sealing 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 non-intrusive solutions.

Another aspect of related work focuses on acoustic sensing in ears, such as using the in-ear microphones to capture body sounds for vital sign tracking [18], [19], human-commuter interactions [20], [21], and extracting unique biometrics [22]. In our initial study, we attempt to leverage the in-ear microphones to capture the sounds induced by tongue and jaw movements. However, ear canal deformation induced by tongue-jaw movements often pushes the in-ear microphones out of the ear canal, making them unreliable for tracking the tongue-jaw movement sounds. Moreover, [23] warns that moving the ear canal wall while wearing earplugs can lead to collapse of the external ear canal.

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 user holds the smartphone, like making a phone call, which is unobtrusive, non-intrusive, and user-friendly. By analyzing dynamic acoustic properties of ear canal deformation, *CanalScan* achieves high accuracy and robust tongue-jaw movement recognition in real-life environments.

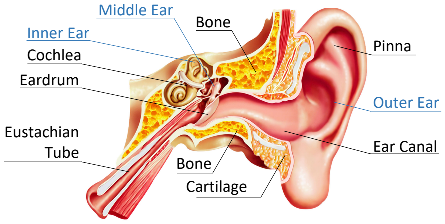


Fig. 3. Structure of human ear.

## 3 Preliminary

In this section, we first describe the basics of the ear structure. Then, we introduce the sensing principle of ear canal deformation. Afterward, we present the basics of multi-path propagation. Finally, we show the observations which validate the feasibility of using ear canal deformation-related acoustic signals for tongue-jaw movement recognition.

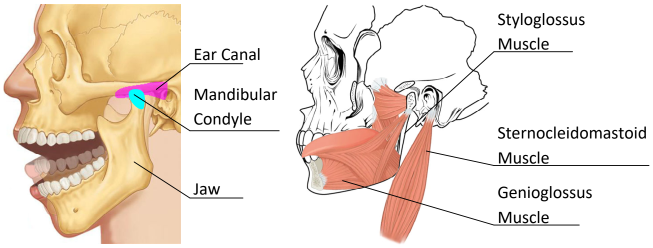
### 3.1 Basic of Ear Structure

The ear can be divided into three parts, outer ear, middle ear, and inner ear (as shown in Fig. 3). The outer ear is the part of the ear that can be seen from the outside of the human body, which consists of the pinna and the ear canal. The ear canal is a roughly S-shape passage comprised of bone and skin leading to the eardrum, which mainly has a length of about 30 mm [24]. Its primary function is to transmit sound to the tympanic membrane and protects the middle and inner ears. Under the skin of the ear canal is the cartilage laterally and bone medially. Although the shape and size of the cartilaginous portion and bony portion of the ear canal vary among individuals, the skeletal structure is essentially the same for all individuals.

### 3.2 Sensing Principle of Ear Canal Deformation

Promoting tongue and jaw movements requires the active participation of masticatory muscles bilaterally and both temporomandibular joints. Fig. 4 shows the musculoskeletal system related to the movement of the tongue and jaw. When the user performs a tongue-jaw movement, the musculoskeletal system changes, which affects the shape of the ear canal. Particularly, the positional relationship between the ear canal and the mandibular condyle changes when the jaw is moved, which causes the ear canal deformation. The styloglossus muscle starts from the styloid process and ends at the junction of the hyoid body and the greater horn of the hyoid bone. The main function is to retract and elevate the tongue, which will relax or contract according to the tongue and thus change the shape of the ear canal. The genioglossus muscle is responsible for protracting the tongue and has the same principle as the styloglossus. In addition, the sternocleidomastoid muscle, which connects the back of the ear to the clavicle, either relaxes or contracts when the face or head is moved. This also changes the shape of the ear canal by expanding or compressing the ear canal.

When performing different tongue and jaw movements, different changes in the ear canal shape are introduced due to the above factors. Subsequently, the acoustic signal reflected by the ear canal wall implies the generated tongue-jaw movement's characteristics. Therefore, we can recognize



(a) Tongue-jaw movements related bones (b) Tongue-jaw movements related muscles

Fig. 4. Musculoskeletal system related to the movement of the tongue and jaw.

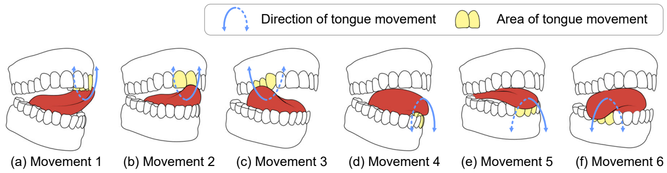


Fig. 5. An illustration of six movements involving the tongue and jaw.

tongue-jaw movements by characterizing changes in the multi-path reflections caused by ear canal deformation.

### 3.3 Basic of Multi-Path Propagation

An acoustic signal reflected by different surfaces could produce multiple reflections with different propagation directions, amplitudes, and phases. Such reflections are called multi-path reflections and have a long and rich research history in capturing the geometry of certain surfaces [25]–[27]. Given an acoustic signal $R(t) = sin(2\pi ft + \phi)$ with the frequency $f$ and the initial phase $\phi$, the received signal at time $t$ can be defined as:

$$R(t) = \sum_{i \in \chi} A_i sin(2\pi ft + \phi_i), \tag{1}$$

where $\chi$ is the set of acoustic signals of all paths, $\phi_i$ is the phase change coefficient. $A_i$ depicts the amplitude reduction and $A_i \propto 1/d_i$ with $d_i$ representing the corresponding propagation distance.

When emitting an acoustic signal to the ear canal, the ear canal deformation can induce multi-path reflections while performing tongue-jaw movements. Different tongue-jaw movements cause the ear canal wall to move differently, thus presenting unique information in the multi-path reflections. We are motivated to exploit the uniqueness of the acoustic reflections in the ear canal for tongue-jaw movement recognition.

### 3.4 Observations

According to [15]–[17], [28], 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. In other words, 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.

To demonstrate the feasibility of using acoustic reflections to characterize different tongue-jaw movements, we conduct experiments on a smartphone that continuously sends 16kHz acoustic signals and collects acoustic reflections at 48kHz. While the design space of tongue-jaw movements is large, we focus on six tongue-jaw movements performed in different areas of the oral cavity, as shown in Fig. 5. These tongue-jaw movements are composed of two stages: (i) the tongue starts from the back of the teeth, licks 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 the jaw returns to its original position. The selected six movements have wide user acceptance and are natural to perform. We ask two volunteers to hold the smartphone as if they are on the phone, align the top microphone and earpiece speaker with the ear canal entrance, and synchronously perform the six tongue-jaw movements, respectively. 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 then extract the multi-path reflection envelope in the time window and illustrate examples of the movement patterns of six tongue-jaw movements in Fig. 6.

**Feasibility**: The reflection envelope shows that each kind of tongue-jaw movement has unique patterns, 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

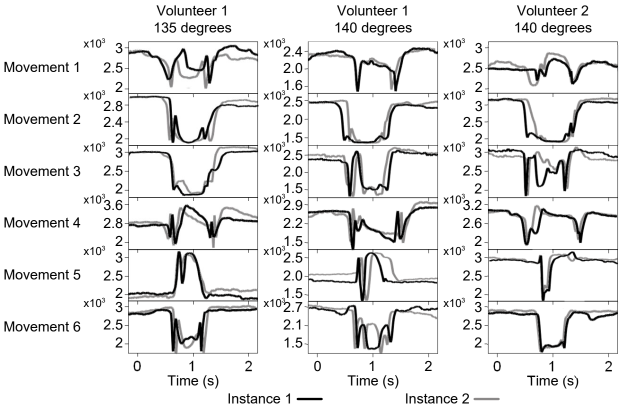


Fig. 6. The reflection envelope of movement 1-6 in three conditions.

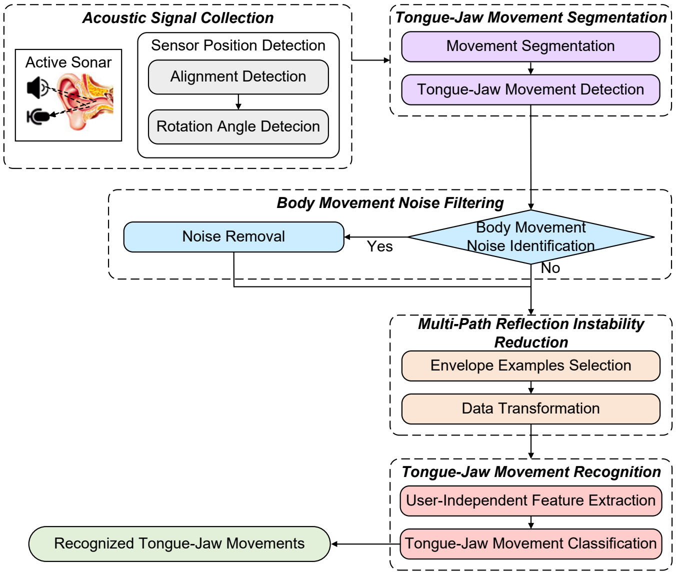

Acoustic Signal Collection
Active Sonar
Sensor Position Detection
Alignment Detection
Rotation Angle Detecion
Tongue-Jaw Movement Segmentation
Movement Segmentation
Tongue-Jaw Movement Detection
Body Movement Noise Filtering
Noise Removal
Yes
Body Movement Noise Identification
No
Multi-Path Reflection Instability Reduction
Envelope Examples Selection
Data Transformation
Tongue-Jaw Movement Recognition
User-Independent Feature Extraction
Tongue-Jaw Movement Classification
Recognized Tongue-Jaw Movements

Fig. 7. Overall framework of *CanalScan*.

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.

## 4 System Design

In this section, we present the design of *CanalScan*, which recognizes tongue-jaw movements through acoustic sensing and captures unique patterns of ear canal deformation.

### 4.1 Overview

*CanalScan* utilizes the off-the-shelf speaker and microphone integrated into smart devices (e.g., smartphones) for tongue-jaw movement recognition. Fig. 7 shows the overall design of *CanalScan*, which is mainly comprised of five models: *Acoustic Signal Collection*, *Tongue-Jaw Movement Segmentation*, *Body Movement Noise Filtering*, *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, 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 and 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.

After *Tongue-Jaw Movement Segmentation*, the body movements between tongue-jaw movements are eliminated, but those overlapping the tongue-jaw movements are retained. Therefore, we perform *Body Movement Noise Filtering* to eliminate the signals contaminated by body movement noise. Specifically, we develop a body movement noise identification method using an extreme learning machine (ELM) and optimize its parameters using a particle swarm optimization (PSO) algorithm. If body movement noise is detected, we further filter out the corresponding frames to eliminate the impact of body movement noise.

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). To find the most significant envelop examples, we perform *Envelope Examples Selection* through modeling of the within-class distance and between-class distance of each candidate envelope example.

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.

Fig. 8 shows the walkthrough of the detection system. Firstly, we check the relative position between the smartphone and the ear canal. If the smartphone is aligned and the rotation angle is appropriate, the system will continue to collect and process data. Otherwise, the system will prompt the user to adjust the smartphone position. The acoustic signals induced by tongue-jaw movement are then segmented and transformed to facilitate the subsequent training. If the tongue-jaw movement segments are distorted by body movements, we eliminate them, and we only process those that are not contaminated by body movements. In the tongue-jaw movement recognition phase, we extract features, and then make the prediction decision according to the pre-learned knowledge.

### 4.2 Acoustic Signal Collection

#### 4.2.1 *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 [29]. Most smartphones support a sampling rate of 48kHz, so the excitation acoustic signal is restricted to below 24kHz.

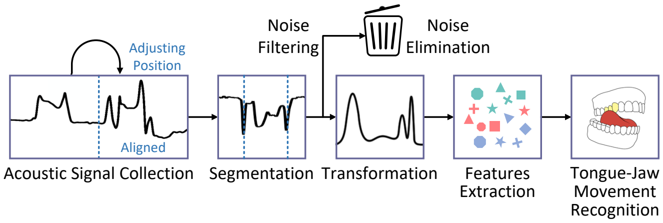


Fig. 8. An illustration of the major steps of *CanalScan*.

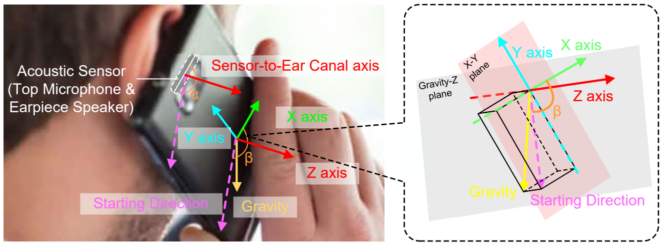


Fig. 9. Rotation angle and smartphone coordinate system.

However, speaker and microphone distortion at high frequencies narrows our choices to below 17kHz [30]. 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.

### 4.2.2 *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 with 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. 9 shows an example of the smartphone coordinate system, 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 $\alpha$ degrees around the sensor-to-ear-canal-axis, and the smartphone rotates $\beta$ 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 $\alpha$ equals $\beta$, which is the angle between the starting direction and the smartphone X-axis. Therefore, we now turn to the problem of obtaining $\beta$. Fortunately, inertial measurement unit mounted on modern smartphones provide easy access to such tilt angle:

$$\beta = arctan(\frac{g_x}{g_y}) + \frac{\pi}{2}, \tag{2}$$

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.

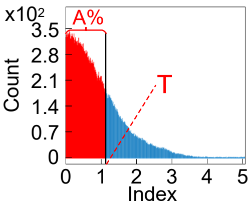
According to our experiment with $50$ people, a comfortable posture of holding the smartphone close to the ear canal (like making a phone call) is to make the smartphone rotates $130$-$140$ degrees. By calculating the smartphone rotation angles, we guide the users to rotate the smartphone at the same 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.

## 4.3 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.

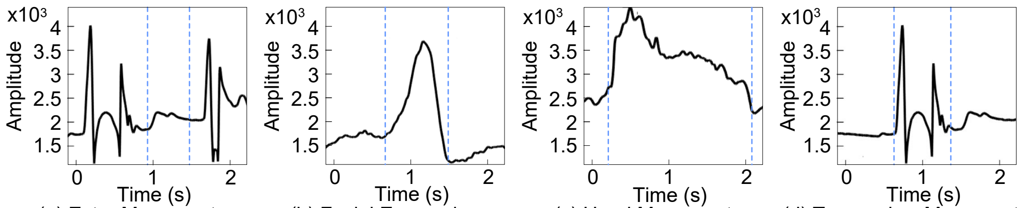
### 4.3.1 *Movement Segmentation*

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 derivative of jitters is high, and the first derivative of pauses is 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



First Derivative

Fig. 10. Calculation of $T$ based on intensity distribution.



(a) Extra Movement (b) Facial Expression (c) Head Movement (d) Tongue-jaw Movement

Fig. 11. Examples of extra movement, facial expression, head movement, and tongue-jaw movement.

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 $a$ [31]. Then, the threshold $T$ is calculated as $\int_0^T I(a)\mathrm{d}a = A\%$. Fig. 10 shows an example of calculating threshold $T$ based on intensity distribution. We set A as 63 based on our experimental study.

#### 4.3.2 Tongue-Jaw Movement Detection

Extra movements of the 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. 11 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 between 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 5.5.

### 4.4 Body Movement Noise Filtering

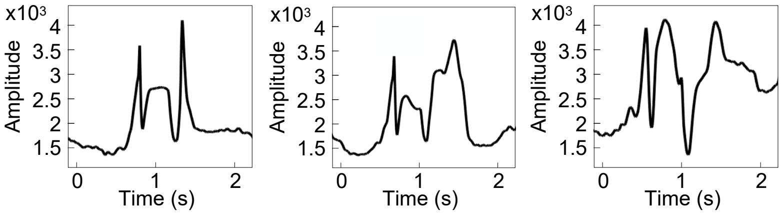
In Section 3.4, we demonstrate the feasibility of characterizing tongue-jaw movement based on multi-path reflection from the ear canal. In the experiment, participants are required to maintain stationary while collecting acoustic signals. However, users inevitably involve body movements in real-life environments, making the collected multi-path reflections unreliable. After interviewing many volunteers, we find that the common usage scenarios of tongue-jaw movement-based interactions include standing still (static state), sitting in a car (containing environment noise), and walking (containing body movement noise). Fig. 12 illustrates the envelope of acoustic reflection caused by performing movement 4 under the three conditions. We can observe that noise significantly distorts the shape of the acoustic reflection envelope more than that under the static state. Characterizing the acoustic reflection envelope contaminated by noise would lead to poor movement recognition performance. Therefore, it is necessary to identify the presence of noise before recognizing tongue-jaw movements.

After careful research with twenty participants, we find that the relative position between the smartphone acoustic sensor and the ear canal changes when the human body is moving. Hence, we analyze the rotation angle of the smartphone for body movement noise identification. Fig. 13 shows examples of smartphone angles of static state and moving state. Different instances are illustrated in different colors. We can observe that the body movements lead to time-varying rotation angles, and the variation ranges are different across instances. However, we can also observe that the rotation angle variation pattern in the moving state (e.g., Instance 1) could be very similar to that in the stationary state. Therefore, identification of body movement noise should be designed carefully to ensure efficiency.

#### 4.4.1 Body Movement Noise Identification

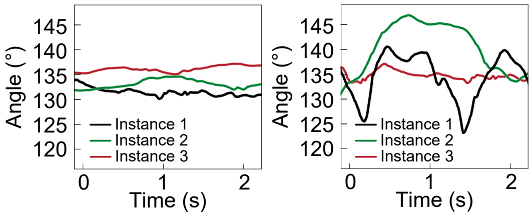
Noise detection is essential to ensure signal quality. To this end, we develop an extreme learning machine (ELM)-based body movement noise identifier. Compared with other noise identification solutions, such as Convolutional Neural Network (CNN), ELM has a low computational cost and can handle the variation of rotation angles well. Moreover, its performance is less subjected to user-specified parameters. Although ELM has been previously applied to human action recognition [32]–[35], to the best of our knowledge never for our application scope.

We apply a 2 seconds sliding window with 50% overlap to process the rotation angle data. In each window, we extract *mean, max, min, standard deviations, auto-regression coefficients, entropy,* and *energy* to build the data represen-



(a) Stand still (b) Sit in a moving car (c) Walk

Fig. 12. The reflection envelope of movement 4 under three conditions.



(a) Stationary state (b) Moving state

Fig. 13. Rotation angle of different body state.

tation to input to the ELM. Given the input $H_i$ and the corresponding label $T_i$, the output of ELM is modeled by

$$f(H_i) = \sum \lambda \cdot G(\mathbf{W} \cdot H_i + \mathbf{b}_i), \tag{3}$$

where $G$ is the activation function. $\mathbf{W}$, $\mathbf{b}_i$, and $\lambda$ denote the input weights, biases to the hidden layer, and output wrights, respectively. We rewrite Eqn (3) as $f(H_i) = \mathbf{K}\lambda$, where $\mathbf{K}$ is the hidden layer output matrix defined as:

$$\mathbf{K} = \begin{bmatrix} G(\mathbf{W}_1 H_1 + \mathbf{b}_1) & \cdots & G(\mathbf{W}_L H_1 + \mathbf{b}_L) \\ \vdots & \ddots & \vdots \\ G(\mathbf{W}_1 H_N + \mathbf{b}_1) & \cdots & G(\mathbf{W}_L H_1 + \mathbf{b}_L) \end{bmatrix}, \tag{4}$$

where $N$ is the sample length and $L$ is the number of hidden notes. The ELM classifier is trained to minimize the difference between output labels and the true labels of the input samples. Specifically, we first assign random values to $\mathbf{W}$ and $\mathbf{b}_i$ and optimizes $\lambda$ via least squares.

To eliminate the sensitivity of ELM to the number of hidden neurons, input weights, and bias values, we use the particle swarm optimization (PSO) algorithm [36] to optimize the three parameters. Given a swarm of particles at position $p = \{p_1, ..., p_d\}$ with velocity $v = \{v_i, ...v_d\}$, PSO firstly assigns $p$ and $v$ arbitrarily, then iteratively updates their values according to

$$\begin{aligned} v_i(t+1) &= \rho \cdot v_i(t) + c_1 \cdot r_1(p_i^*(t) - p_i(t)) \\ &\quad + c_2 \cdot r_2(p_g^*(t) - p_i(t)), \\ p_i(t+1) &= p_i(t) + v_i(t+1), \end{aligned} \tag{5}$$

where $r_1$, $r_2$ are within [0,1] range to maintain the diversity of the population. $c_1$ and $c_2$ are the positive coefficients of the self-recognition component and the social component, respectively. $p_i^*$ denotes the best position of the particle $i$ at each iteration, while $p_g^*$ denotes the best position of optimal particle in the swarm at each iteration. $\rho$ is the inertia factor defined as,

$$\rho = \rho_{max} - \frac{\rho_{max} - \rho_{min}}{I} \times t, \tag{6}$$

where $\rho_{max}$ and $\rho_{min}$ represent the upper and lower bounds of $\rho$. $I$ is the allowed iteration number, and $t$ is the iteration count.

We consider an ELM as a particle. The variable $p$ can be expressed as

$$p = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1N} & b_1 \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2N} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{L1} & \rho_{L2} & \cdots & \rho_{LN} & b_L \end{bmatrix}, \tag{7}$$

where $b_i$ is the biases of $i^{th}$ hidden neurons. By treating $\rho$ as the input weighs, the optimization is implemented of finding the globally optimal particle. We evaluate the performance of the developed method in Section 5.5.5. The results demonstrate that the recall and precision reach the best performance of 93.23% and 93.07%, respectively, when using 55 hidden nodes. Therefore, we use 55 hidden nodes for body movement noise identification.

#### 4.4.2 Body Movement Noise Removal

The impacts of body movement noise on the acoustic reflection envelope of tongue-jaw movements are complex and unpredictable. The signal envelope is significantly distorted when body movements are detected in the data window. Building noise elimination algorithms require substantial computational effort and can only work in limited scenarios. Instead, we consider the periods with noise unrecoverable. We remove data periods affected by body movements and only perform classification during non-body movement periods. Such a strategy is adopted by many well-recognized existing systems [37], which gives a delayed response instead of an error response.

We also find that the acoustic measurements contaminated by noise may not be continuous. The interval between two adjacent noisy periods can be much shorter than a complete tongue-jaw movement, which does not allow *CanalScan* to extract enough features for tongue-jaw movement recognition. Hence, we remove the interval of less than $\tau$ seconds between two adjacent noise periods. We set $\tau$ to be 1 experimentally.

### 4.5 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 and propose an effective method to select essential envelope examples.

#### 4.5.1 Design Guidelines

We aim to reduce pattern instability through a transform function. Such a transformation process involve two 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.

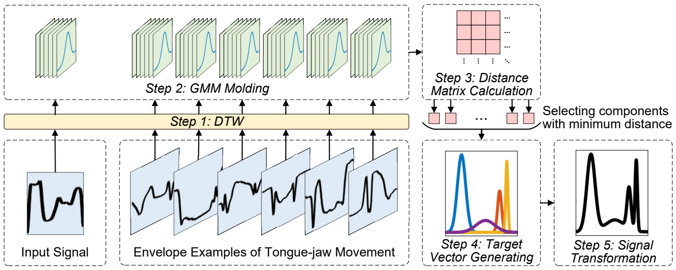

Step 2: GMM Molding
Step 3: Distance Matrix Calculation
Step 1: DTW
Selecting components with minimum distance
Input Signal
Envelope Examples of Tongue-jaw Movement
Step 4: Target Vector Generating
Step 5: Signal Transformation

Fig. 14. An illustration of the data transformation process.

Based on the above goals and our discussion in Section 3.4, 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 the target vector and the collected data, and finally transform the collected signal into a new signal with characteristics of the target vector.

### 4.5.2 Data Transformation Process

Data transformation techniques have been investigated for voice generation [38] and data augmentation [39]. However, they usually build user-specific models and demand a large amount of reference data to learn the mapping between a reference signal and the pre-transformation signal, which makes them unsuitable for our system. Therefore, we propose a user-independent data transformation method based on only six envelope examples.

Fig. 14 illustrates the process of data transformation. The envelope examples 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:

$$P_{\boldsymbol{x}} = \textstyle\sum_{i=1}^{K} \alpha_i \mathcal{N}(\mu_i, \sigma_i), \tag{8}$$

$$P_{\boldsymbol{y}_m} = \textstyle\sum_{j=1}^{K} \beta_j \mathcal{N}(\mu_j, \sigma_j), \tag{9}$$

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:

$$D_{i,j} = \frac{1}{2}[KL(\mathcal{N}_{\mu_i,\sigma_i}||\mathcal{N}_{\mu_j,\sigma_j}) + KL(\mathcal{N}_{\mu_j,\sigma_j}||\mathcal{N}_{\mu_i,\sigma_i})], \tag{10}$$

where the KL divergence is defined as:

$$KL(\mathcal{N}_{\mu_i,\sigma_i}||\mathcal{N}_{\mu_j,\sigma_j}) = log\frac{\sigma_j}{\sigma_i} + \frac{(\mu_i - \mu_j)^2 + \sigma_i^2 - \sigma_j^2}{2\sigma_j^2}. \tag{11}$$

**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 are considered the most similar components. We add up $K$ components in the form of GMM to obtain the probability density of the representative target vector $\boldsymbol{y'}$:

$$P_{\boldsymbol{y'}} = \textstyle\sum_{i=1}^{K} \gamma_i \{\mathcal{N}(\mu_j, \sigma_j)|arg\,min D_{i,j}\}. \tag{12}$$

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

$$\gamma_i = \frac{\alpha_i \mathcal{N}(\mu_i, \sigma_i)}{\sum_{j=1}^{K} \alpha_j \mathcal{N}(\mu_j, \sigma_j)}. \tag{13}$$

**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 [40], we introduce a transformation function $\mathcal{F}(x)$ assumed by the Minimum Mean Square Error (MMSE) estimation:

$$\begin{aligned} \mathcal{F}(\boldsymbol{x}) &= E(\boldsymbol{y'}|\boldsymbol{x}) \\ &= \int \boldsymbol{y'} \frac{P(\boldsymbol{x}, \boldsymbol{y'})}{P_x(\boldsymbol{x})} d\boldsymbol{y'}, \end{aligned} \tag{14}$$

where $P_x(\boldsymbol{x})$ is the probability density of $\boldsymbol{x}$, which is modeled by Equ (8). 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:

$$P_{\boldsymbol{z}} = \sum_{i=1}^{K} \omega_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \sum_{i=1}^{K} \omega_i = 1, \quad \omega_i \geqslant 0, \tag{15}$$

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

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_i^{\boldsymbol{x}} \\ \mu_i^{\boldsymbol{y'}} \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} cov(\boldsymbol{x}, \boldsymbol{x}) & cov(\boldsymbol{x}, \boldsymbol{y'}) \\ cov(\boldsymbol{y'}, \boldsymbol{x}) & cov(\boldsymbol{y'}, \boldsymbol{y'}) \end{bmatrix}, \tag{16}$$

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 (14) in the following:

$$\mathcal{F}(\boldsymbol{x}) = \sum_{i=1}^{M} P_{\boldsymbol{y}}(C_i|\boldsymbol{x})[\mu_y + \frac{cov(\boldsymbol{y}, \boldsymbol{x})}{cov(\boldsymbol{x}, \boldsymbol{x})}(\boldsymbol{x} - \mu_x)], \tag{17}$$

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 be calculated using Equ (13). 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 being processed by data transformation, differences between data from the same tongue-jaw movement are effectively reduced, and data 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 5.5.

### *4.5.3 Envelope Example Selection*

The transformation process converts the input data to a new vector based on the similarity between the input data and envelope examples that represent six tongue-jaw movements. It is essential to select envelope examples to provide the basics of data transformation. There are two selection criteria:

- Because an envelope example represents a tongue-jaw movement, it should exhibit general features, i.e., the selected envelope example should be very similar to those belonging to the same tongue-jaw movement.
- The ultimate goal of data transformation is to distinguish different tongue-jaw movements. Envelope examples of different tongue-jaw movements should be distinct to facilitate recognition.

We collect ear canal reflections from multiple participants and form a dataset of tongue-jaw movement envelope examples $Y = \{y_{i,i=1,...,N}\}$. Selecting of the most significant example of each tongue-jaw movement is to solve an optimal subset $Y' = \{y_1, y_2, y_3, y_4, y_5, y_6\}$, where subscripts $1$ to $6$ represent six tongue-jaw movements. We propose to model the within-class and between-class distances and find examples with minimal within-class and maximal between-class distances. The six tongue-jaw movements are denoted as six classes $C_{l,l=1,2,..,6}$. The subset selection problem is defined as:

$$arg\min_{y_i} \frac{\sum_{i,j} d(y_i, y_j)\nu_{i,j}}{\sum_{i,j} d(y_i, y_j)\eta_{i,j}}, \tag{18}$$

where $d(y_i, y_j)$ measures the Euclidean distance between normalized examples $y_i$ and $y_j$. Two-dimensional weight matrices $\nu_{i,j}$ and $\eta_{i,j}$ describe the relationship between $y_i$ and $y_j$. If $y_i$ and $y_j$ belong to the same class $C_l$, $\nu_{i,j} = 1$, otherwise, $\nu_{i,j} = 0$. Thus, $\sum_{i,j} d(y_i, y_j)\nu_{i,j}$ models the total within-class distance between $y_i$ and $C_l$. Our objective is a small within-class distance. To obtain $\eta_{i,j}$, we first determine $K_d$-nearest neighbors of $y_i$. If $y_j$ is one of the $K_d$-nearest neighbors of $y_j$, and $y_i$ does not belong to the same class, we set $\eta_{i,j}$ and $\eta_{j,i}$ to be 1, otherwise, to be 0. When $y_i$ has small between-classe distances, $\sum_{i,j} d(y_i, y_j)\eta_{i,j}$ should get a smaller value.

Fig. 15 shows the selection process. The candidate envelope examples collected from multiple users and different sessions are firstly normalized and input to the weight metrics generator. The weight matrices generator is the core of the proposed method. It groups the candidate examples into six classes and determines the $K_d$-nearest neighbors of each example. After that, $\nu_{i,j}$ and $\eta_{i,j}$ of each example are generated. Finally, the envelope example selector models the within-class distance and between-class distance of each candidate example based on Equ (18). By finding the examples with the minimum within-class distance and the maximum between-class distance, we determine the most significant examples for data transformation.

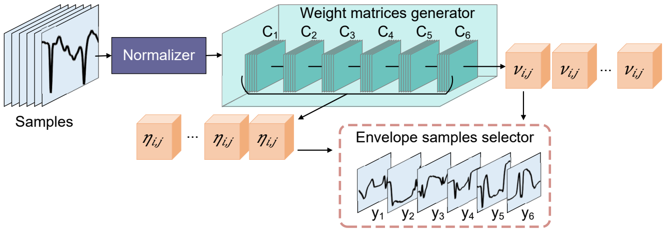


Fig. 15. An illustration of envelope examples selection process.

## 4.6 Tongue-Jaw Movement Recognition

### *4.6.1 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 movement before classification, and use it to train a classifier to infer the performed tongue-jaw movement.

We select over fifty candidate features by extensively exploring the features suggested by plenty of related systems and applying the feature extraction toolbox (e.g., *tsfresh* [41]). These features characterize the acoustic reflection envelope in both the time and frequency domains. But using redundant features would degrade the classification performance. Therefore, we apply an RF classifier to rank these features by feature importance feedback and evaluate the importance of features on the movement classification task [42]. We pick twelve kinds of features that contribute most to recognizing various tongue-jaw movements, 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.* Fig. 16 shows the t-SNE (t-distributed stochastic neighbor embedding) projections of features from five volunteers, with each user performing six tongue-jaw movements six times. We can observe that the extracted features are consistent for the same tongue-jaw movement and are unique across different tongue-jaw movements.

### *4.6.2 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

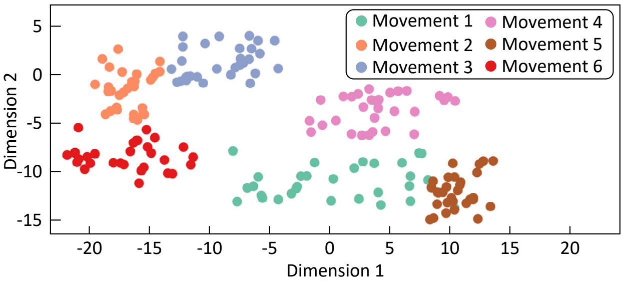


Fig. 16. The t-SNE visualization of features.

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 choose RF because it has the best performance in our experimentally study, which is presented in Section 5.5.

# 5 Evaluation

## 5.1 Implementation

We implement *CanalScan* to verify its performance in recognizing tongue-jaw movements. In our proof-of-concept implementation, we use LIBAS [43] to send acoustic signals at 16kHz and receive the reflections at a sampling rate of 48kHz. LIBAS is a cross-platform framework that simplifies the development of acoustic-based applications. Using the server-client remote mode of LIBAS, we transfer the acoustic measurements to a laptop (i.e., Intel Core i7-11800H, 16GB RAM, and Nvidia GeForce GTX 3050Ti graphics card) and perform data processing in MATLAB.

## 5.2 Experimental Setup

There exists no state-of-the-art dataset of acrostic reflection from the ear canal. Therefore, we collect our dataset in real-life environments to evaluate *CanalScan*. We recruit 20 adult participants (10 males and 10 females) from colleagues and students at the institution. All participants are healthy, right-handed, and cleaned their ears before collecting experimental data. This study is conducted with the approval of our institute's IRB. During the data collection phase, we ask participants to align the top microphone and earpiece speaker with their ear canals and simultaneously press the smartphone tightly. To accommodate slight sensor position differences, we encourage participants to rotate the smartphone 130-140 degrees. Participants are asked to 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 takes a five minutes break. The start and end of each tongue-jaw movement are indicated by clicking a computer mouse.

To evaluate the key algorithms and explore the robustness of *CanalScan* against various issues, we ask participants to collect data with various sensor rotation angles, different devices, and diverse body motions. In addition, we evaluate *CanalScan* in long-term study. Moreover, we conduct experiments in common and representative real-life scenarios covering stationary (noise-free), slight noise, and intense noise conditions.

To evaluate *CanalScan*, we define several metrics as follows:

**Confusion Matrix**: Each row of the matrix represents the ground truth while each column represents the predicted results. Each entry $c_{i,j}$ of the matrix shows the percentage of instances belonging to the $i^{th}$ class predicted as the $j^{th}$ class to all instances belonging 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 belonging to label A.

## 5.3 Overall Performance of Tongue-jaw Movement Recognition

We first evaluate the tongue-jaw movement recognition performance of *CanalScan* through conducting five-fold cross-validation. Fig. 17 shows the confusion matrix of the recognition results. 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. Overall, the average recall and precision are 94.84% and 95.00% respectively which demonstrate that *CanalScan* achieves accurate recognition of tongue-jaw movements. We find that movements 4 and 5 receive higher recall and precision than other tongue-jaw movements. A possible reason is that they involve more significant lower jaw movements, making the ear canal deformation reflections more distinguishable. After carefully interviewing participants, we find that some participants like to clean their ears after bathing or swimming, which could affect the reflection properties and consequently cause recognition errors.

## 5.4 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.

### *5.4.1 Universality*

For a movement recognition system, the ability to generalize to new users without retraining or adaptation can provide a satisfactory user experience. To evaluate the universality of *CanalScan*, we conduct leave-one-person-out-validation. That is, we use data from nineteen participants for training and data from one participant for testing. Fig. 18 illustrates the behavior of the proposed system of all combinations. We observe that 17 participants have recall higher than 90% and precision higher than 90%. *CanalScan* achieves the average recall and precision of 91.41% and 91.58%, respectively. These excellent results suggest *CanalScan* can effectively work across different users. Furthermore, 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.

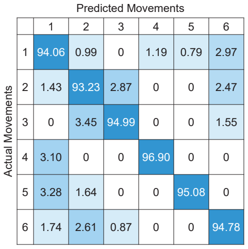


Fig. 17. Overall recognition performance.

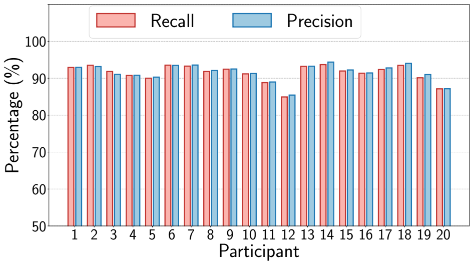


Fig. 18. Performance of leave-one-person-out-validation.

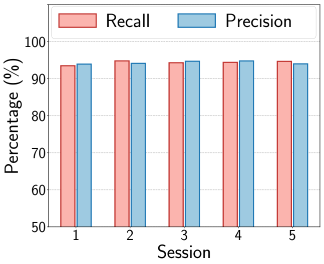


Fig. 19. Performance of leave-one-session-out-validation.

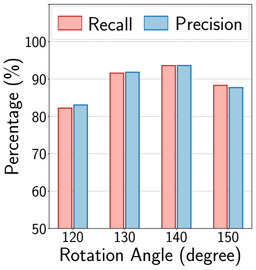


Fig. 20. Impact of rotation angle.

### 5.4.2 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. 19, 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, which confirms that *CanalScan* works effectively against movement inconsistency.

### 5.4.3 Impact of Sensor Rotation Angle

We use the pre-trained classifier described in Section 5.3 to evaluate *CanalScan*'s robustness against different sensor rotation angles. Whereas daily usage purposes, we test four angles, i.e., 120 degrees, 130 degrees, 140 degrees, and 150 degrees. Fig. 20 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. Furthermore, it can be observed that when the angle is 140 degrees, *CanalScan* receives the highest recall and precision. 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 in recall and precision.

### 5.4.4 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. We implement LIBAS to collect acoustic signals with iPhone X, iPhone 8, HUAWEI Mate 9, and HUAWEI Mate 9pro. Specifically, these smartphones are different in size and audio hardware. Then, we compare the recall and precision across four devices and show the result in Fig. 21. The results demonstrate 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.

### 5.4.5 Long-Term Performance

Existing related approaches that send and receive acoustic signals in the ear canal mainly focus on the static characteristics of the ear canal shape. However, the static characteristics can be greatly affected by the ear wax which is naturally produced by the human body. Thus, these approaches do not support long-term use. 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 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 that a regular update of the training data set of *CanalScan* enables high accurate tongue-jaw movement recognition.

## 5.5 Key Algorithm Study

We evaluate the performance of movement segmentation, tongue-jaw movement detection, data transformation, various classifiers, body movement noise identifier, and envelope example selection.

### 5.5.1 Performance of Movement Segmentation

Under the direction of the computer mouse, we segment the movement between the start and end points as the ground truth. Then, we compare them with the segmentation results based on the dynamic threshold. Experiment results show that 90% of the time difference between segments and ground truth is less than 0.1s, which demonstrates the effectiveness of the proposed method.

### 5.5.2 Performance of Tongue-jaw Movement Detection

By carefully checking the results of the 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.

### 5.5.3 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

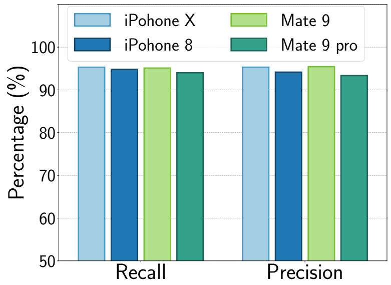


Fig. 21. Performance under four different smartphones.

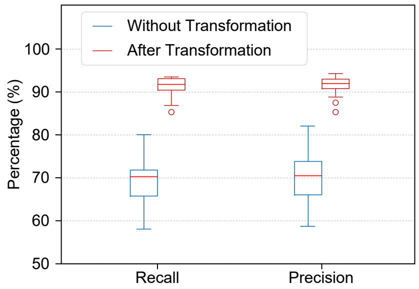


Fig. 22. Performance of multi-path reflection instability reduction.

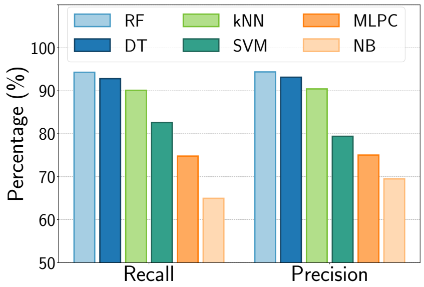


Fig. 23. Performance of different classifiers.

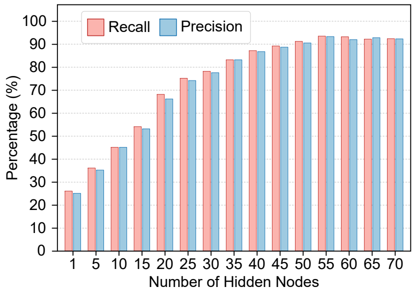


Fig. 24. Performance with different number of hidden nodes.

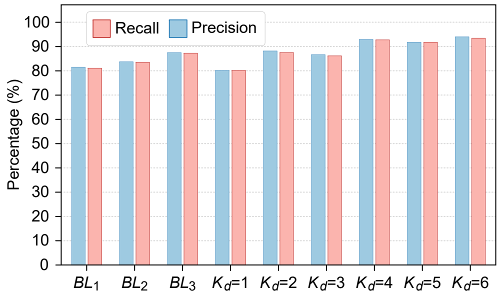


Fig. 25. Performance of baseline and with different $K_d$.

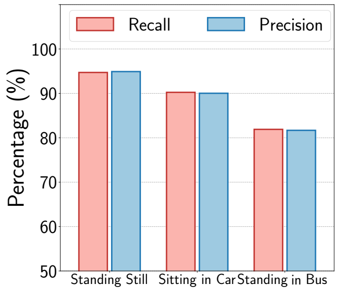


Fig. 26. Performance in three different usage scenarios.

the received signals. We compare the leave-one-person-out validation results without and after data transformation. As shown in Fig. 22, 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, there is a significant increase in recognition results with 91.41% average recall and 91.58% average precision. The results demonstrate that data transformation shows a high efficiency, which is the key to realize accurate tongue-jaw movement recognition.

### 5.5.4 Impacts of Training Data Size and Classifier

We evaluate the performance of *CanalScan* under different training data sizes by varying the training data size from 25% to 85%. The results show that the system could receive higher recall and precision when the training data is more. Then, we compare the performance of several highly used classifiers, 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 default values. As shown in Fig. 23, RF and DT have better performance than other classifiers. When using 85% data for training, RF achieves its best recall and precision of 94.30% and 94.39%, respectively, and is adopted in this work.

### 5.5.5 Performance of Body Movement Noise Identification

Body movement noise identification serves as a layer of defense against envelope distortions. To evaluate the performance of the body movement noise identifier, we ask participants to record the acoustic reflections of tongue-jaw movements in both stationary and moving states. After carefully interviewed all participants, we summarize the common usage scenarios of *CanalScan* in Table 1, which covers different intensity levels of noise. Each activity is repeated by multiple participants and data are collected from different days to include the slight changes in participants' behaviors. We group the data of different activities into stationary state and moving state. Five-fold cross-validation is conducted through the proposed ELM-based method. As our focus is mainly to learn the behavior of body movement noise identification in detail, we compare its performance with the different number of hidden nodes. Fig. 24 shows the recall and precision results. We can observe that, with a small number of nodes, the recall and precision are low. As the number of nodes increases from 40 to 55, recall and precision keep improving. Specifically, the recall and precision reach 93.23% and 93.07% when using 55 hidden nodes. Adding more hidden nodes does not help to further boost the performance as the number of hidden nodes goes beyond 55. This supports our choice of 55 neurons for ELM in the experiments.

TABLE 1
Body state category and activities.

| State Category | Activity |
|---|---|
| Stationary | Sit still, stand still |
| Moving | Sit in a car, stand in a bus, stand in the subway, walk, eat, deep breath, ride a stationary bike. |

### 5.5.6 Performance of Envelope Example Selection

We conduct five-fold cross-validation to study the impact of parameter $K_d$. Moreover, we randomly select three sets of envelope examples as the baseline (denoted as $BL_1$, $BL_2$, and $BL_3$) and compare their performance in recognizing tongue and jaw movements with the template selected by the proposed method. As shown in Fig. 25, three sets of randomly selected examples obtain different recall and precision results, and the recall and precision are between 80% to 90%. This confirms that envelop examples do influence the recognition of tongue-jaw movements. Besides, when $K_d$ is set from 1 to 6, we can observe an interesting pattern of alternating increase and decrease. As it is trivial that the running time of the algorithm is directly proportional to the value of $K_d$, we found $K_d = 4$ to be a good tradeoff between the performance of the algorithm and its running time. Nevertheless, when having sufficient computing power, we suggest using $K_d = 6$ to get the best recognition results.

### 5.6 Usage Study in Real-Life Environments

In real-world usage, the user can deliberately avoid body activities but cannot avoid the noise caused by the environment. For example, when a user rides in a vehicle that is moving may introduce noise into the collected signal. To understand how well *CanalScan* can work in real life, we ask 6 participants to perform six tongue-jaw movements and collect data in three conditions: standing still, standing in a moving bus, and sitting in a moving car. These scenarios are common and representative of scenarios containing environmental noise. The classifier is trained as described in Section 5.3. We report the detailed recognition results of *CanalScan* under three different environments in Fig. 26. It can be seen that 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, with the recall decreases to 90.24%, and precision decreases to 90.03%, demonstrating that *CanalScan* can bare slight motions. In terms of standing in a moving bus, body movement obfuscates acoustic reflections in the ear canal, resulting in 81.90% recall and 81.68% precision. The results are acceptable in real-life environments, but *CanalScan* would better be used under static and slight motion states.

### 5.7 User Study

We are interested in the user experience of *CanalScan*. We ask the participants to fill a System Usability Scale (SUS) [44] questionnaire to gather feedback, which ranks from 1 (strongly disagree) to 5 (strongly agree). The SUS questionnaire has ten questions and proved to be a valuable evaluation tool, being robust and reliable. Table 2 summarizes the questions and the average scores gathered from each participant. We can notice that positive statements (question 1,3,5,7, and 9) have high scores around 4 and 5. Beside, the negative statements (question 2,4,6,8, and 10) have low scores around 1 and 2. The results demonstrate that *CanalScan* offers good user experience and great practical usability.

TABLE 2
Results of user study questionnaire.

| Item | Question | Score |
|---|---|---|
| 1 | I think that I would like to use this system frequently. | 4.65 |
| 2 | I find the system unnecessarily complex. | 1.1 |
| 3 | I think the system is easy to use. | 4.55 |
| 4 | I think that I would need the support of a technical person to be able to use this system. | 1.7 |
| 5 | I find the various functions in this system are well integrated. | 4.3 |
| 6 | I think there is too much inconsistency in this system. | 1.3 |
| 7 | I would imagine that most people would learn to use this system very quickly. | 4 |
| 8 | I find the system very cumbersome to use. | 1.85 |
| 9 | I feel very confident using the system. | 4.3 |
| 10 | I need to learn a lot of things before I could get going with this system. | 1.1 |

## 6 Discussion and Future Work

As a new technology, *CanalScan* certainly leaves a number of limitations to explore further. First, the design space of tongue-jaw movements is large. In addition to the six movements studied in this work, we also study many other movements, such as jaw protrusion, jaw retrusion, tongue moving along a clockwise path, and tongue moving along a counterclockwise path. We find that *CanalScan* can achieve precision and recall rates of over 90%, and the more complex the trajectory of the tongue-jaw movements, the better the system performs. We focus on six predefined tongue-jaw movements so far as they are representative, common, and well-accepted by users. Expanding to a larger movement set or supporting user-defined movements is more attractive and we recommend those with complex tongue and jaw movement trajectories when choosing additional tongue-jaw movements. Second, we only implement *CanalScan* using smartphones. Today's earphones usually integrate in-ear microphones and speakers for a superior listening experience. Such acoustic front-ends are very promising to implement *CanalScan*. We are planning to implement *CanalScan* with Active Noise Cancellation (ANC) earphones to conduct more experiments. Last but not least, we have not seriously evaluated *CanalScan*'s performance under various noise conditions. It is impractical to exhaust all usage scenarios, and we conduct experiments at various noise intensity levels instead. In particular, our experiments have been conducted during intensive noise (i.e., walking, eating, riding a stationary bike) and slight noise (i.e., sitting in a car, standing in a bus, standing in the subway, deep breathing). The experiments demonstrate that *CanalScan* is promising to cope with a wide range of noise conditions. We will extensively conduct experiments with more body movements to gather more evaluation.

## 7 Conclusion

In this paper, we propose a non-intrusive tongue-jaw movement recognition system, *CanalScan*. Our system only relies on the 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 captures 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. Specifically, to mitigate the impacts of various relative positions and individual's ear canal shape on multi-path reflections, a sensor position detection method and a data transformation method with movement examples selection algorithm are introduced. Then, to remove noise and segment tongue-jaw movements, a body movement noise filtering method and a dynamic segmentation method are developed. Afterward, we extract twelve unique and consistent features and adopt an RF-based classifier for recognition. 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.

## ACKNOWLEDGMENTS

The work of Fan Li is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62072040. The work of Huijie Chen is supported by NSFC under Grant No. 62202019, and China Postdoctoral Science Foundation under Grant No. 2021M700302.

## 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. of IEEE ISCAS 2006*, 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. of ACM MobiSys 2018*, 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. of ACM ISWC 2014*, 2014, pp. 47–54.

[4] L. Liu, S. Niu, J. Ren, and J. Zhang, "Tongible: A Non-Contact Tongue-Based Interaction Technique," in *Proc. of ACM ASSETS 2012*, 2012, pp. 233–234.

[5] B. Leung and T. Chau, "A Multiple Camera Tongue Switch for a Child With Severe Spastic Quadriplegic Cerebral Palsy," *Disability and Rehabilitation: Assistive Technology*, vol. 5, no. 1, pp. 58–68, 2010.

[6] R. Li, J. Wu, and T. Starner, "TongueBoard: An Oral Interface for Subtle Input," in *Proc. of ACM AH 2019*, 2019, pp. 1–9.

[7] Z. Li, R. Robucci, N. Banerjee, and C. Patel, "Tongue-n-Cheek: Non-Contact Tongue Gesture Recognition," in *Proc. of IPSN 2015*, 2015, pp. 95–105.

[8] 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. of ACM UIST 2017*, 2017, pp. 679–689.

[9] T. Amesaka, H. Watanabe, and M. Sugimoto, "Facial Expression Recognition Using Ear Canal Transfer Function," in *Proc. of ACM ISWC 2019*, 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. Rao, "Non-intrusive Tongue Machine Interface," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 04 2014, pp. 2555–2558.

[12] R. Vaidyanathan and C. J. James, "Independent Component Analysis for Extraction of Critical Features from Tongue Movement Ear Pressure Signals," in *Proc. of IEEE EMBS 2017*, 2007, pp. 5481–5484.

[13] 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.

[14] B. Maag, Z. Zhou, O. Saukh, and L. Thiele, "BARTON: Low Power Tongue Movement Sensing with In-ear Barometers," in *Proc. of IEEE ICPADS 2017*, 12 2017, pp. 9–16.

[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. of ACM CHI 2017*, 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, pp. 733–744, 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. of ACM UbiComp/ISWC 2015 Adjunct*, 2015, pp. 1333–1338.

[18] A. Martin and J. Voix, "In-Ear Audio Wearable: Measurement of Heart and Breathing Rates for Health and Safety Monitoring," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 6, pp. 1256–1263, 2017.

[19] S. Nirjon, R. F. Dickerson, Q. Li, P. Asare, J. A. Stankovic, D. Hong, B. Zhang, X. Jiang, G. Shen, and F. Zhao, "MusicalHeart: A Hearty Way of Listening to Music," in *Proc. of the 10th ACM SenSys*, 2012, pp. 43–56.

[20] X. Xu, H. Shi, X. Yi, W. Liu, Y. Yan, Y. Shi, A. Mariakakis, J. Mankoff, and A. K. Dey, "Earbuddy: Enabling On-Face Interaction via Wireless Earbuds," in *Proc. of ACM CHI 2020*, 2020, pp. 1–14.

[21] J. Prakash, Z. Yang, Y.-L. Wei, H. Hassanieh, and R. R. Choudhury, "EarSense: Earphones as a Teeth Activity Sensor," in *Proc. of the 26th ACM MobiCom*, 2020, pp. 1–13.

[22] Y. Xie, F. Li, Y. Wu, H. Chen, Z. Zhao, and Y. Wang, "TeethPass: Dental Occlusion-based User Authentication via In-ear Acoustic Sensing," in *Proc. of IEEE INFOCOM 2022*. IEEE, 2022, pp. 1789–1798.

[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] R. Oliveira and G. Hoeker, "Ear Canal Anatomy and Activity," *Semin Hear*, vol. 24, pp. 265–275, 11 2003.

[25] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "EchoPrint: Two-Factor Authentication Using Acoustics and Vision on Smartphones," in *Proc. of the 24th ACM MobiCom*, 2018, pp. 321–336.

[26] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin, "EarEcho: Using Ear Canal Echo for Wearable Authentication," *Proc. of ACM IMWUT 2019.*, vol. 3, no. 3, pp. 1–24, 2019.

[27] Z. Wang, S. Tan, L. Zhang, Y. Ren, Z. Wang, and J. Yang, "EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables," *Proc. of the ACM IMWUT 2021*, vol. 5, no. 1, pp. 1–27, 2021.

[28] 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.

[29] 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.

[30] 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.

[31] F. Yan, H. Zhang, and C. R. Kube, "A Multistage Adaptive Thresholding Method," *Pattern Recognition Letters*, vol. 26, no. 8, pp. 1183–1191, 2005.

[32] Z. Zhao, Z. Chen, Y. Chen, S. Wang, and H. Wang, "A Class Incremental Extreme Learning Machine for Activity Recognition," *Cognitive Computation*, vol. 6, no. 3, pp. 423–431, 2014.

[33] Y. A. Jeroudi, M. A. Ali, M. Latief, and R. Akmeliawati, "Online Sequential Extreme Learning Machine Algorithm Based Human Activity Recognition Using Inertial Data," in *2015 10th Asian Control Conference (ASCC)*, 2015, pp. 1–6.

[34] A. Budiman and M. I. Fanany, "Pose-Based 3D Human Motion Analysis Using Extreme Learning Machine," in *2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE)*, 2013, pp. 3–7.

[35] X. Ma, H. Wang, B. Xue, M. Zhou, B. Ji, and Y. Li, "Depth-Based Human Fall Detection via Shape Features and Improved Extreme Learning Machine," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1915–1922, 2014.

[36] ——, "Depth-Based Human Fall Detection via Shape Features and Improved Extreme Learning Machine," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1915–1922, 2014.

[37] T. Zhao, Y. Wang, J. Liu, Y. Chen, J. Cheng, and J. Yu, "Trueheart: Continuous Authentication on Wrist-Worn Wearables Using PPG-Based Biometrics," in *Proc. of IEEE INFOCOM 2020*, 2020, pp. 30–39.

[38] Y. Stylianou, "Voice Transformation: A Survey," in *Proc. of IEEE ICASSP 2009*, 2009, pp. 3585–3588.

[39] B. Liu, X. Wang, M. Dixit, R. Kwitt, and N. Vasconcelos, "Feature space transfer for data augmentation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9090–9098.

[40] 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.

[41] "tsfresh," 2022. [Online]. Available: https://tsfresh.readthedocs.io/en/latest/

[42] G. Zhang, V. Davoodnia, A. Sepas-Moghaddam, Y. Zhang, and A. Etemad, "Classification of Hand Movements From EEG Using a Deep Attention-Based Lstm Network," *IEEE Sensors Journal*, vol. 20, no. 6, pp. 3113–3122, 2019.

[43] Y. C. Tung, D. Bui, and K. G. Shin, "Cross-Platform Support for Rapid Development of Mobile Acoustic Sensing Applications," in *Proc. of the 16th ACM MobiSys*, 2018, pp. 455–467.

[44] J. Brooke *et al.*, "SUS-A Quick and Dirty Usability Scale," *Usability Evaluation in Industry*, vol. 189, no. 194, pp. 4–7, 1996.

**Yetong Cao** received the BEng degree in computer science and technology from Shandong University, Shandong, China, in 2017. She is now working toward the PhD degree in the School of Computer Science and Technology, Beijing Institute of Technology. Her research interests include mobile computing and ubiquitous computing.

**Fan Li** received the PhD degree in computer science from the University of North Carolina at Charlotte in 2008, MEng degree in electrical engineering from the University of Delaware in 2004, MEng and BEng degrees in communications and information system from Huazhong University of Science and Technology, China in 2001 and 1998, respectively. She is currently a professor at School of Computer Science in Beijing Institute of Technology, China. Her current research focuses on wireless networks, ad hoc and sensor networks, and mobile computing. Her papers won Best Paper Awards from IEEE MASS (2013), IEEE IPCCC (2013), ACM MobiHoc (2014), and Tsinghua Science and Technology (2015). She is a member of ACM and IEEE.

**Huijie Chen** received the B.Eng degree from the School of Computer Science, Henan University of Economics and Law, Zhengzhou, China in 2010, the M.Seng degree from the School of Computer Science, Taiyuan University of Science and Technology, Taiyuan, China in 2013, and Ph.D. degree in computer since from the Beijing Institute of Technology, Beijing, China in 2020. He currently works in the school of computer science, Beijing University of Technology, Beijing, China. His research interests include Smart Sensing, crowdsensing, and mobile computing.

**Xiaochen Liu** received the BEng degree in Internet of Things from China University of Petroleum, Shandong, China, in 2020. She is now working toward the MEng degree in the School of Computer Science and Technology, Beijing Institute of Technology. Her research interests include computer networks and the IoT.

**Yu Wang** is currently a Professor in the Department of Computer and Information Sciences at Temple University. He holds a Ph.D. from Illinois Institute of Technology, an MEng and a BEng from Tsinghua University, all in Computer Science. His research interest includes wireless networks, smart sensing, and mobile computing. He has published over 200 papers in peer reviewed journals and conferences. He has served as general chair, program chair, program committee member, etc. for many international conferences (such as IEEE IPCCC, ACM MobiHoc, IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC), and served as Editorial Board Member for several international journals, including IEEE Transactions on Parallel and Distributed Systems. He is a recipient of Ralph E. Powe Junior Faculty Enhancement Awards from Oak Ridge Associated Universities (2006), Outstanding Faculty Research Award from College of Computing and Informatics at the University of North Carolina at Charlotte (2008), Fellow of IEEE (2018), and ACM Distinguished Member (2020).