Jawthenticate: Microphone-free Speech-based Authentication using Jaw Motion and Facial Vibrations

Tanmay Srivastava
tsrivastava@cs.stonybrook.edu
Stony Brook University
New York, USA

Shijia Pan
span24@ucmerced.edu
University of California, Merced
Merced, USA

Phuc Nguyen
vp.nguyen@cs.umass.edu
University of Massachusetts Amherst
Amherst, USA

Shubham Jain
jain@cs.stonybrook.edu
Stony Brook University
New York, USA

ABSTRACT

In this paper, we present *Jawthenticate*, an earable system that authenticates a user using audible or inaudible speech without using a microphone. This system can overcome the shortcomings of traditional voice-based authentication systems like unreliability in noisy conditions and spoofing using microphone-based replay attacks. *Jawthenticate* derives distinctive speech-related features from the jaw motion and associated facial vibrations. This combination of features makes *Jawthenticate* resilient to vocal imitations as well as camera-based spoofing. We use these features to train a two-class SVM classifier for each user. Our system is invariant to the content and language of speech. In a study conducted with 41 subjects, who speak different native languages, *Jawthenticate* achieves a Balanced Accuracy (BAC) of 97.07%, True Positive Rate (TPR) of 97.75%, and True Negative Rate (TNR) of 96.4% with just 3 seconds of speech data.

CCS CONCEPTS

• **Security and privacy → Biometrics**; • **Human-centered computing → Ubiquitous and mobile computing**.

KEYWORDS

Speech Authentication, Biometrics, IMU Sensing, Signal Processing

ACM Reference Format:
Tanmay Srivastava, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2023. *Jawthenticate*: Microphone-free Speech-based Authentication using Jaw Motion and Facial Vibrations. In *ACM Conference on Embedded Networked Sensor Systems (SenSys '23), November 12–17, 2023, Istanbul, Turkiye.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3625687.3625813

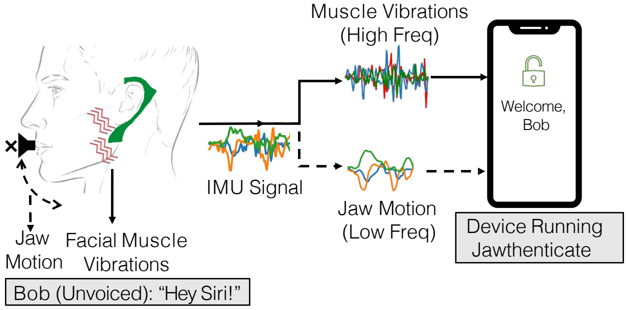

Figure 1: *Jawthenticate* concept.

1 INTRODUCTION

As earables are garnering greater attention and driving the wearables market [63], this paper proposes *Jawthenticate*, which fuses speech-based (behavioral) and biometrics-based (physical) features for continuous authentication on earables. Existing user authentication methods for smart devices often rely on audible speech [14, 15, 60, 76], which can be prohibitive in public environments due to privacy concerns, or biometrics such as fingerprint [35, 69] or Face ID [29, 84]. Unlike fingerprint and Face ID authentication, earables have the potential to offer a completely hands-free approach, which is essential for emerging Augmented Reality (AR) applications and even day-to-day activities, where the user's hands may be occupied. This form of authentication is also inclusive for users with verbal or physical disabilities.

Recent work on hands-free authentication using earables manifests two key limitations. First, speech-based works such as MandiPass [49], Face-Mic [74], and VAuth [21] authenticate a user by capturing vibrations produced by *audible* speech, requiring the user to speak audibly, thereby rendering these systems less secure in public places, compromising user privacy, and opening them to audio-based spoof attacks [16, 65]. Second, many systems require the use of an active audible or inaudible probe. To facilitate biometrics based authentication using earables, EarEcho [25] continuously plays audible sound into users' ears to capture unique ear canal geometry. However, this system is not ideal for scenarios where continuous authentication is required for a seamless user experience (e.g., AR game or interactive movies). To address this, EarDyamic [89] and the work by Mahto et al. [52] use inaudible signals for continuous authentication; however, these high-frequency inaudible signals can potentially negatively impact the ear. Studies

SenSys '23, November 12–17, 2023, Istanbul, Turkiye

ACM ISBN 979-8-4007-0414-7/23/11…$15.00
https://doi.org/10.1145/3625687.3625813

have shown that long-term exposure to high-frequency sounds can cause discomfort, difficulty in concentrating, and pain in the ears [20, 23, 44, 45].

In this paper, we develop *Jawthenticate*, a novel hands-free user authentication technique that unifies the advantages of speech-based (continuous) recognition with biometrics-based authentication (secure). *Jawthenticate* is the first system to explore and exploit the distinctiveness of the user's jaw movement for authentication using an around-the-ear design. It computes the *intonation*, *randomness*, *phase variations*, and *rhythm* of user speech, and *facial muscle vibrations*, using inertial motion sensors (no microphones) to recognize users. It ensures safe and comfortable authentication without external probe sounds or the requirement for a user to produce audible speech. *Jawthenticate* removes the impact of environmental noise and motion artifacts via a twin-sensor setup, allowing for robust continuous authentication.

In contrast to previous works, *Jawthenticate* offers several significant advancements. It is the first system that effectively captures both audible and inaudible speech articulations, relying solely on jaw motion data and does not require the user to produce any sound. This provides an inclusive biometric-based authentication solution. Furthermore, our system stands out in its ability to support language- and content-agnostic continuous authentication. This feature enables broader applicability across diverse linguistic backgrounds and speech patterns. Combining efficient feature extraction and language/content-agnostic continuous authentication, makes *Jawthenticate* well-suited for various real-world scenarios, ranging from voice-controlled applications to secure access control systems, where reliable and versatile authentication is required.

However, realizing *Jawthenticate* encounters the following three challenges: (1) *Microphone-free speech feature learning*: While speech has been shown to be sufficiently distinctive to identify users [86], learning these speech-based features from jaw motion is extremely challenging due to the indirect sensing of a secondary speech articulator - the jaw. (2) *Extracting signatures from inaudible and audible speech:* Many authentication systems require a passcode or passphrase. However, passphrases cannot be used with audible speech-based systems due to privacy concerns, particularly in public environments or when other people are within hearing distance. This necessitates the ability to understand and extract speech-based features from inaudible speech. (3) *Lightweight authentication:* There is a need for a system that is robust yet lightweight, without requiring heavy-weight machine-learning solutions to enable authentication in real-time.

To summarize, we make the following contributions:

- We develop *Jawthenticate*, a novel earable-based authentication system. Our system authenticates the user based on their jaw motions and facial vibration while they are talking (*audible and inaudible*).
- We derive features that are representative of the user's speech mannerisms, and invariant to the content (conversation, phrase, or numerical passcode) and language (English, Hindi, Greek, etc.) of the speech.
- We conduct various real-world impersonation attacks to demonstrate the robustness to observation-based and advanced video-based mimic attacks.

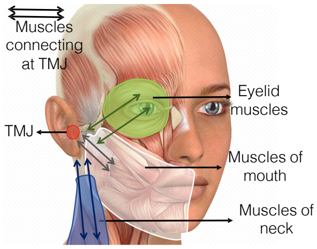

Figure 2: Muscle groups involved in producing facial vibrations during speech-related jaw motion.

- In an IRB-approved study, we collect data with 41 participants and evaluate the performance of our system under various scenarios, achieving up to 97.07% BAC.

To the best of our knowledge, *Jawthenticate* is the first authentication system to learn speech-based features from jaw motion and facial muscle vibration *without requiring voiced speech*. This enables us to accommodate both voiced and unvoiced speech for authentication, making it reliable, resilient to audio and video-based spoofing attacks, and privacy-preserving.

2 BACKGROUND

We present a primer on human speech articulation and how we leverage those principles in *Jawthenticate*.

2.1 Speech Articulation

Given their active participation in speech production, lips, teeth, tongue, alveolar ridge, hard and soft palate are termed *primary articulators*. On the other hand, the jaw is only responsible for facilitating the lower lip and is hence termed a *secondary articulator* [58]. The temporomandibular joints (TMJ), located at the junction of the lower jaw and skull, allow the lower jaw to move up and down [19].

Not only articulators but facial muscles also perform important tasks during speech production [53]. Figure 2 demonstrates three major muscle groups related to speech production. The muscles around the mouth (e.g., obicularis oris) control the shape and movements of the mouth and lips. The muscles around the eyes (e.g., obicularis oculi) contract and pull the skin of the forehead and cheek towards the nose, indirectly participating in speech production.

These muscles are tightly connected and contraction or relaxation of one muscle group can be sensed in another. The movement of these muscles due to speech-related jaw motion produces subtle vibrations, called *facial vibrations* [53]. There is another class of vibrations present during speech articulation: bone-borne vibrations [55], that are generated at the vocal cords during audible speech articulation, propagating through bone and muscles around the face to the TMJ. The movement of these muscle groups along with jaw motion dictates speech characteristics like the speech rate variation, intonations, rhythm, the magnitude of jaw opening, etc., which is the unique speech mannerisms of an individual.

2.2 Experimental Validation

TMJ is sensitive to capturing both the jaw and facial muscles' information related to speech production. To understand different types of signals generated during speech articulation, we place an

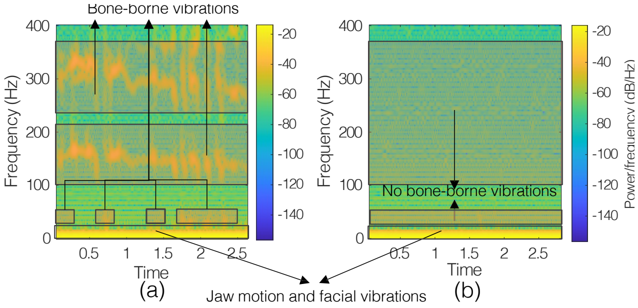

Figure 3: Spectrogram when a user articulates a phrase (a) audibly (voiced) (b) inaudibly (unvoiced). We can see the presence of bone-borne vibrations in (a), which are absent in (b). Jaw motion and facial vibrations, marked by dashed boxes are present in both audible and inaudible speech.

IMU on the right TMJ. The gyroscope captures rotational jaw movements during speech articulation. In addition, the accelerometer captures subtle low-magnitude vibrations generated by bone-borne vibrations and facial vibrations. To validate that these vibrations are caused due to speech articulation, we conduct the following experiments with the accelerometer signals.

Characterizing bone-borne vibrations. When users articulate audibly, their sounds could conduct through the mandible bone and reach the TMJ. To understand the characteristics of these bone-borne vibrations, we ask the user to articulate a phrase of their choice, once audibly and once inaudibly. In these experiments, we sample data from the IMU at 800 Hz since previous works [39, 74] have shown that data sampled at this frequency contains bone-borne vibrations. Figure 3 shows the spectrogram for the same phrase articulated audibly and inaudibly. In Figure 3(a) we can see the presence of a high-frequency (>25Hz) signal, which is absent in Figure 3(b). This higher frequency band appearing in the audible articulation is the bone-borne vibrations.

Additionally, we observe a low-frequency band, as marked by the dashed boxes, in both (a) and (b). This band is associated with two types of vibrations: jaw motion and facial vibrations. Since we expect jaw motion and facial vibrations to manifest in both audible and inaudible speech articulation (as they involve rotation of the jaw around TMJ and associated muscles) the observed low-frequency band is consistent with our expectation. In this work, we aim to achieve authentication for both audible and inaudible speech articulation, focusing on their shared frequency – [1,20] Hz.

Isolating jaw motion and facial vibrations. The jaw moves differently for different words [39, 79], therefore it is closely related to distinctive behavioral features of an individual's speech (the rhythm of speech, variation in speech rate, etc.). On the other hand, facial vibrations are caused as the user's facial muscles contract and relax, which is closely related to the person's physical characteristics. To verify that we can capture both, jaw motion and facial vibrations induced by the contraction/relaxation of the muscles, we conduct experiments that characterize these signals within [1,20] Hz. To isolate these signals, we record a video, along with IMU data collected from the user's TMJ. We ask the user to articulate phrases inaudibly to omit the bone-borne vibrations. We use the video feed as ground truth to track the jaw motion by calculating the displacement of a marker on the IMU, and then computing acceleration in m/s^2. We sample data from the IMU and camera at 60 Hz as this is sufficient to capture jaw motion and facial vibrations. Figure 4 shows the frequency spectrum plot for the same phrase, for the y-axis of the accelerometer (blue) and camera (red). We observe a similar magnitude in the 1-5 Hz range in the frequency spectrum in both the camera and IMU. However, the IMU captures the 5-20 Hz facial vibration signals (absent in the camera data). We verified this phenomenon for different users articulating different words/phrases. Our insights in characterizing jaw motion and facial vibrations are significant to defend against advanced attacks simulating IMU data from videos (§ 6.3).

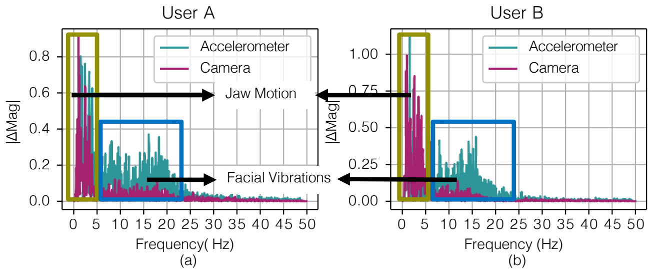

Figure 4: (a) Comparison of the frequency spectrum from accelerometer and camera when the user articulates a phrase inaudibly. The only attribute captured by the camera is jaw motion (0-5 Hz), and not the facial vibrations. (b) The same observation for another user.

3 SYSTEM OVERVIEW

Jawthenticate is capable of authenticating users whether they speak (audibly or inaudibly) a predetermined passphrase or a numerical passcode, or have a regular conversation. Figure 5 shows the overview of our system. There are 4 main modules that constitute *Jawthenticate*: (1) *Data Pre-processing*: This module is responsible for removing body motion, gravity, and DC noise from the accelerometer data, segregating jaw motion and facial vibrations from the signal, and distinguishing between speech and non-speech signals. Speech signals are passed on to the next module. (2) *Jaw motion feature extraction*: With the isolated jaw motion signal, *Jawthenticate* extracts intonations and randomness in the time domain and rhythm and phase in the frequency domain that are representative of the distinctive jaw motion of the user. Features such as intonation and rhythm have been successfully employed in speech-based authentication [64, 77]. We extract similar features from jaw motion to encode equivalent knowledge. (3) *Facial vibrations feature extraction*: *Jawthenticate* extracts frequency domain features from the facial vibrations that encode the distinctive skin/muscle vibration trend for each user. (4) *User Identification:* Once all the features are extracted, *Jawthenticate* uses an SVM-based classifier to authenticate the user.

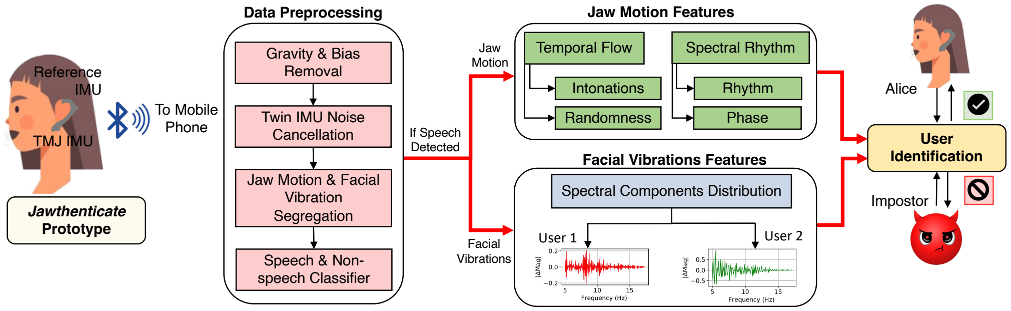

Figure 5: Overview of *Jawthenticate*.

3.1 Potential Applications

Jawthenticate aims for both traditional speech-based and secure unvoiced authentication applications, such as one-time authentication in public settings (privacy-preserving), continuous authentication to a voice-enabled device (e.g., continuous VR game, one-time login/payments), and novel authentication modality for post-laryngectomy patients unable to produce sound. We discuss two potential applications:

Authentication in public settings: Voice-based authentication is popular and convenient [4, 33], however, it is an open channel that is susceptible to privacy leakage and impersonation attacks, and can be affected by ambient noise. In contrast, *Jawthenticate* enables users to authenticate themselves by using unvoiced speech, which helps protect their privacy and prevent impersonation attacks. For example, it can allow users to inaudibly convey sensitive information, like SSNs to bank operators, where voice authentication can breach privacy. In addition, *Jawthenticate* can be used in conjunction with silent speech recognition systems, such as MuteIt [80], to provide an additional layer of security by using the user's articulated password or other confidential information as the primary authentication factor, while their speech mannerisms serve as a biometric that helps verify their identity. This can enable a secure and private authentication experience in scenarios where traditional voice based authentication might fall short. In addition to enabling hands-free payments in crowded cafes, *Jawthenticate* can also be used by medical staff to access patient records and medication securely in a busy environment where vocal passwords may breach confidentiality, touch-based systems are unhygienic, and face might be covered with masks and glasses.

Authentication for VR/AR and other Head Mounted Devices (HMD): HMDs lack input devices, such as keyboards, making it difficult for users to authenticate themselves (when logging in or making purchases). As a result, users often have to use a secondary device, such as a mobile phone or tablet, which can negatively impact the usability and experience, as they have to take off the HMD to use the secondary device [3]. *Jawthenticate* can enable users to log in to the system without using a secondary device. Most HMDs already consist of a head-mounted IMU which can act as the reference sensor. With an additional retro-fitted IMU placed at the TMJ, *Jawthenticate* can be used for authentication without requiring users to take their device off, thereby improving their overall experience.

3.2 Threat Models

We identify the following threat models for *Jawthenticate*.

No knowledge attack. A no-knowledge attack can occur when an impostor tries to breach the system without any knowledge of how the system works. However, as they have no insights from observation or other sources, it is very hard to break the system. Due to its low probability of success, we do not evaluate the system against this attack.

System aware attack. In this attack, the impostor gets access to the user's wearable device without their knowledge and permission. Through observation, the impostor knows about the working of the system and articulates some phrases. But, due to distinctive jaw motion as well as facial vibrations for each user, *Jawthenticate* will be able to deal with such attacks by rejecting impostors whose facial vibrations and jaw motion do not match.

Mimic attack. In this attack, the impostor, after observation and with insights about the data used by the system, tries to mimic a legitimate user's speaking style to break the system. We asked 5 users to mimic other users' speech mannerisms while they articulate a few phrases. We evaluate the system's performance against this attack in § 6.3.

Advanced mimic attack. In this attack, the impostor tries to mimic the users' jaw motion by using existing techniques that use automated computer vision algorithms to reconstruct the IMU data stream [42, 51]. However, since cameras are not capable of capturing subtle facial vibrations (as shown in §2.2), this attack can lead to simulating jaw motion only. *Jawthenticate* extracts distinctive facial vibrations in addition to jaw motion, making it resilient to such attacks that can only mimic jaw motion. We evaluate *Jawthenticate*'s defensive performance against this attack in § 6.3.

4 SYSTEM DETAILS

In this section, we delve into the intricate workings of *Jawthenticate* and its functionality.

4.1 Pre-processing IMU data

Jawthenticate adopts the twin-IMU noise cancellation design presented in MuteIt [79]. In the twin-IMU setup, a reference IMU is placed at a location such that it captures head and body motion but not the jaw motion, and another IMU is placed at the TMJ such that it captures the jaw motion, but is corrupted by head and body movements. This enables the system to remove body motion noise (e.g., head nodding) from the jaw motion signal. However, instead

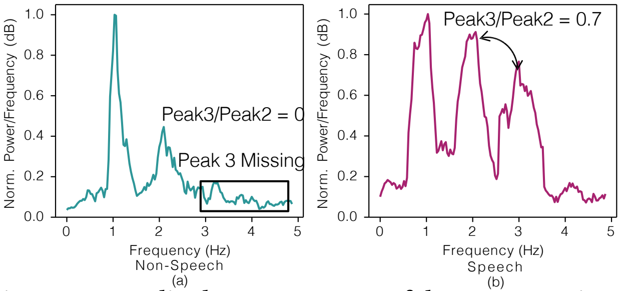

Figure 6: Normalized power spectrum of the gyroscope signal for (a) non-speech-related jaw motion (eating); (b) speech-related jaw motion. In (a) there is no Peak 3 and most of the power is distributed between the first two peaks, unlike (b) which has 3 high power peaks.

of subtracting the two signals as in [79], we use an adaptive finite impulse response (FIR) filter that converges an input signal to the desired signal [31]. This can model complex and non-linear noise characteristics more effectively, and provide better cancellation performance [8]. Next, we remove the effect of gravity and any DC bias from the accelerometer data. After applying noise cancellation [79] to remove body motion from jaw motion signals, we use a third-order band-pass filter to isolate jaw motion (0-5 Hz) and facial vibrations (5-20 Hz). After this preprocessing, we obtain segregated jaw motion and facial vibration signals.

Jawthenticate is convenient and can potentially be used continuously. To achieve this, we discard any non-speech-related jaw motion, the most common of which is eating and drinking. The jaw motion related to eating has a more regular motion of opening and closing compared to speech whereas there can be irregular pauses for better articulation. In the frequency spectrum of eating-related IMU signals, this reflects as one or two frequencies comprising most of the energy. Figure 6(a) shows the normalized power spectrum for the z-axis of the gyroscope when a user is eating, and (b) talking. We observe that the non-speech jaw motion signal has one or two frequencies dominating the distribution while speech has a spread-out power distribution with multiple peaks (at least 3). We use this property to identify a window as speech or non-speech jaw motion. Specifically, for every 0.5 seconds of jaw motion data, we find peaks (whose magnitude is greater than the empirically determined threshold of 0.4 and are at least 0.5 Hz apart) in the normalized power spectrum signal. We compute the inter-peak ratio between the 3^{rd} and 2^{nd} peaks, and if this ratio is found to be less than 0.2 (threshold determined from an ROC curve), this window is detected as non-speech and discarded, else it is marked as speech and passed on to the next module.

4.2 Features from Jaw Motion

Next, we extract speech-based features from jaw motion. The complex movements of articulators (tongue, lips, jaw, etc.) are responsible for speech mannerisms like speech rate, the rhythm of talking, accents, etc. One way to represent these features would be to encode them via statistical features, like maximum/minimum and velocity of jaw motion, which are bound to change with content and emotions. With *Jawthenticate* we delve into the intricate process of human speech articulation and employ a novel methodology to extract features from both jaw motion and skin deformation. This approach effectively encodes the unique speech mannerisms exhibited by each individual user. Notably, this method offers advantages over deep learning approaches that rely on complex layers for feature extraction– it requires less training data, making it a more efficient alternative (Section 6.7). Also, the use of speech mannerism extraction in our system ensures its invariance to speech content, language, and even breathing patterns, given that speech mannerisms tend to exhibit relatively stable characteristics Section (6.2;6.4). We extract the following speech-specific features:

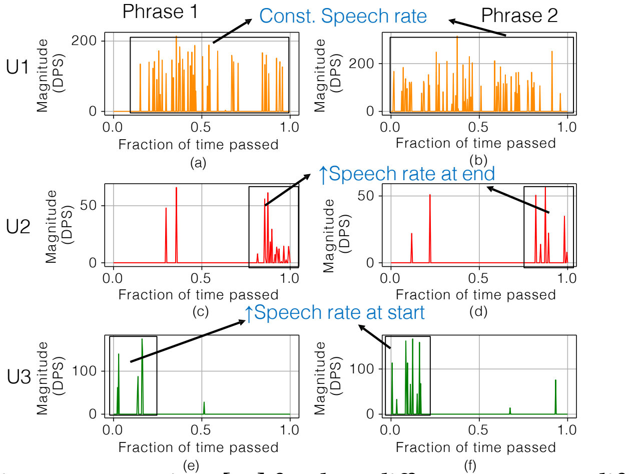

Figure 7: Intonations [F1] for three different users, two different phrases. The rows represent a user and the columns a phrase. (a), (b) show similar values (location and magnitude), like (c), (d), and (e), (f). For the same phrase (each column), intonations show different trends for different users.

■ **Feature 1 (F1):** ***Intonations in speech.*** Different people have different speaking rates, which can be estimated by the angular and linear velocity of the jaw. Speech rate is a distinctive speech-related feature as it is dependent upon social, physical, and psychological bio-markers which vary across users [36]. However, speech rate can change with content. Therefore, instead of speech rate, we look at the intonation and stress in one's speech. Intonation is the rise and fall of voice while talking, and stress is a property determined by the volume and pitch of the vowels. Although we can not measure these directly from a secondary articulator without audio data, we extract features that encode these properties. We observe that some users tend to have higher speech rates at the start of their phrase than in the middle and the end, while some users tend to have a monotonous speech rate. Figure 7 shows the intonations for 3 different users. Each row represents a user, and each column represents a phrase[1].

To leverage intonations, we look for repeatability within each row, and variation within each column. For the same phrase, a user exhibits similar patterns for F1, which is different than when other users say the same phrase. With this insight, we find the locations of the intonation points where the rate changes, i.e. where the signal crosses a certain threshold. We find those samples when the signal transitions outside ±20% of the mean. We save the magnitude of the signal at these locations, forming a two-value tuple of [magnitude,

[1] Phrase 1: "How is the weather?" and Phrase 2: "Have a nice day." for Figure 7- 9

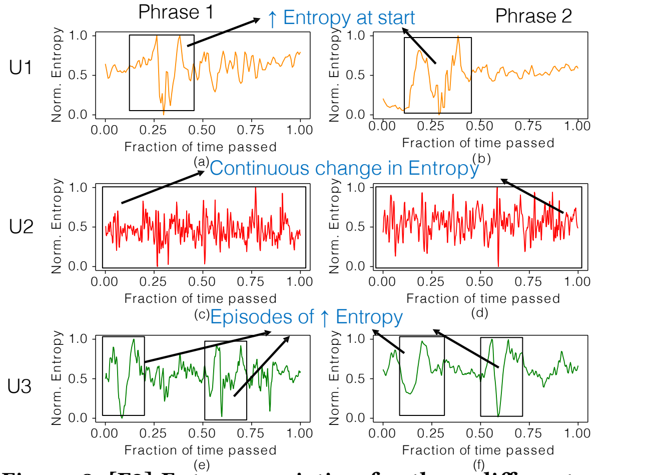

Figure 8: [F2] Entropy variation for three different users, two different phrases. The y-axis is the Normalized Shannon Entropy for the z-axis of the accelerometer. Each row represents a user and each column is a phrase. We can see similarity within different phrases articulated by a user, that is dissimilar from others.

location]. Though this is somewhat similar to the mean crossing rate that can be a measure of speech rate, it is noteworthy that people might have similar speech rates but the variation in speech rate is distinctive, which is captured by this 2-value tuple.

■ **Feature 2 (F2) -** ***Randomness in speech.*** Another feature in speech articulation is the randomness or the change in speech rate over time. To this end, we measure the entropy of the jerk of the linear acceleration and angular velocity. The intuition is that when there will be a change in the speech rate, the randomness of the signal will increase at that time with respect to signal windows before and after. Specifically, we measure the Shannon entropy [73] for each 0.1 second window. We leverage these thresholds to learn speech-based features from jaw motion since they have been previously used in the audio domain [11, 56]. We form a tuple with all the values of entropy in each window. Figure 8 shows the jerk of angular velocity for two phrases spoken by three users. We can see that the entropy for different phrases for the same user has similar characteristics, while the entropy or randomness patterns are different across users.

■ **Feature 3 (F3) -** ***Phase variations over time.*** With this feature, we aim to extract spectral properties of the jaw motion signal. However, Discrete Fourier Transform (DFT) cannot detect the temporal distribution of different spectral components, and therefore Short Time Fourier Transform (STFT) is usually employed. The drawback with STFT is that the temporal information of the frequency spectrum is not fine-grained and we cannot obtain high resolution both in the time and frequency domain [81]. We want a time-frequency domain transform that encodes how the phase and amplitude of each frequency of interest varies over time for low-frequency jaw motion signals (0-5 Hz). We, therefore, use S-Transform [82], to produce a frequency-time representation of the time series data, as it has been shown to provide high resolution for low-frequency signals [81]. Figure 9 shows the output of the S-transform for two phrases for three users. Similar to Figure 7 we see repeatability in each row (same user) and variation within a column (different users). We calculate the phase for 1 Hz, 2 Hz, 3 Hz, 4 Hz, and 5 Hz signals by calculating the location of the peak with respect to the length of the phrase.

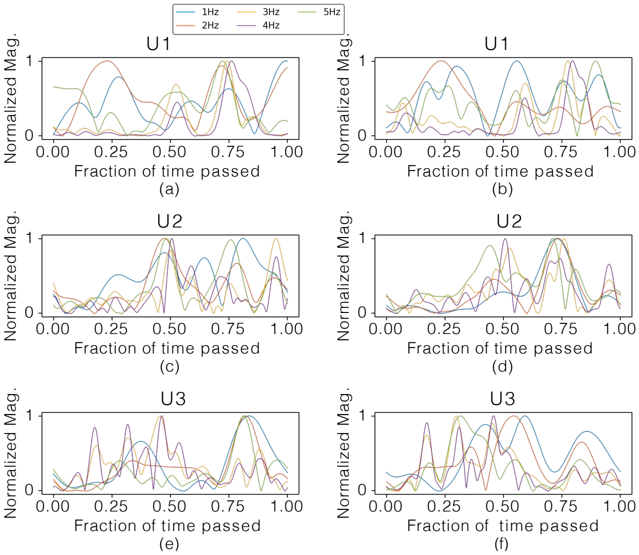

Figure 9: [F3] Variation of frequencies 1-5 Hz over time, for three users, two phrases. The phase, especially at the start of the phrase, is different for different users, while repeatable across phrases for a user.

■ **Feature 4 (F4):** ***Speech rhythm.*** In linguistics, rhythm is one of the aspects of prosody [62]. It is the beat of one's speech and recent discovery from the linguistic literature shows that speech rhythm has a temporal structure of high regularity [27, 46]. We extract this regularity by using the Fourier Transform and finding the frequency bin with maximum amplitude. An assumption of the Fourier Transform is that the signal can be approximated well as a weighted sum of sinusoidal basis elements. However, a complex signal cannot always be broken down into these basic elements and be approximated as sinusoidal elements. Also, like the Fourier Transform, Wavelet Transform does not look for regularities directly in a signal, but represents the signal as a sum of the mother wavelet, like Mexican hat [5]. To overcome this, we use Periodicity Transform (PT), to find the period of a signal such that for a signal *x(k)* with a period *p*, *x(k+p) = x(k).* The PT breaks the signal into a sum of periodic sequences by projecting onto a set of "periodic subspaces" and hence this decomposition is accomplished directly in terms of periodic sequences and not in terms of frequency [71]. PT has been used to find rhythm in different domains of signal analysis [7, 67, 72]. Given a signal, there are multiple ways to decompose it into different sub-periodic signals by projecting them into periodic subspaces. We use the M-Best Algorithm [72], which finds M-best periodicities; the best is based on the amount of energy lost from the signal once that periodicity has been removed. We select the top M components that capture at least 90% of the power as that would imply that the projections made by these components have captured a higher order of periodicity. We experiment with different values of M and empirically select M=4 based on the dominant periodicities method [70].

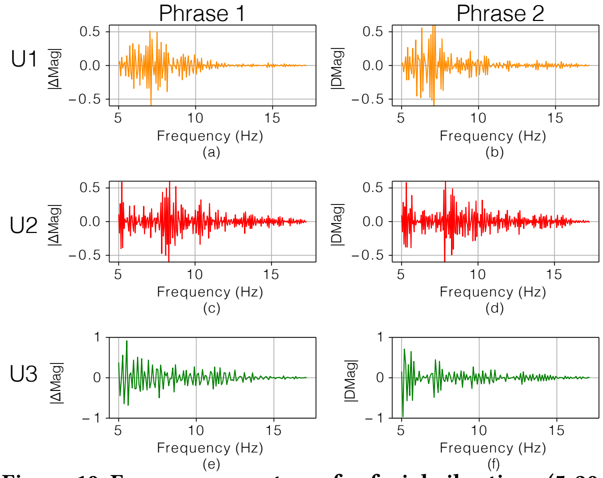

Figure 10: Frequency spectrum for facial vibrations (5-20 Hz). (a), (b) show similar trends, like (c), (d), and (e), (f). For the same phrase (each column), the frequency spectrum shows different trends.

4.3 Features from Facial Vibrations

Facial vibrations are obtained from the preprocessing step. They are generated based on how different facial muscle groups interact and the facial structure of the user [53]. The distinctive movements of the articulators dictate how facial muscle groups will interact [59]. Hence, our speech-related features are not only indicative of the user's speech mannerisms but also their facial structures and facial muscle interactions, known to be distinctive for a large population [54, 61]. There are limited insights from previous work on how facial skin and muscles contract and relax during speech articulation, and therefore it is hard to extract speech-based features from the facial vibrations. However, we observe that these movements and subtle vibrations vary in accordance with one's speech mannerisms and facial structure. To capture this, we first calculate 100 point FFT for the three axes of the accelerometer in the frequency range of 5-20 Hz, with a sliding window of 0.1 seconds and 80% overlap. A 100-point FFT gives us a frequency resolution of 1 Hz at the given sampling rate. Figure 10 shows the frequency spectrum of facial vibrations for three users for two phrases of different lengths. The frequency spectrum for each user is consistent across different phrases. To represent this feature, we form a tuple with phase and magnitude for each frequency bin.

4.4 User Identification

Next, we use a Support Vector Machine (SVM) to classify a user as a legitimate user or an impostor. We collect data from users in different scenarios explained in Section 5. We chose SVMs as they are known to be robust to overfitting, can handle noisy data [13], and have been extensively used by previous authentication systems [21, 22, 25]. We follow different data splitting techniques (balanced and imbalanced) to gauge the system's capability to deal with data imbalance. To evaluate how the system will perform when negative data points are not present during training of the model, we train a one-class SVM. We also performed other splitting schemes as in previous studies [22, 49, 93] to comprehensively evaluate *Jawthenticate*. During training the SVM, we use 30% of the data as the validation set to find the best parameters for the SVM. We perform the following splitting schemes for all users.

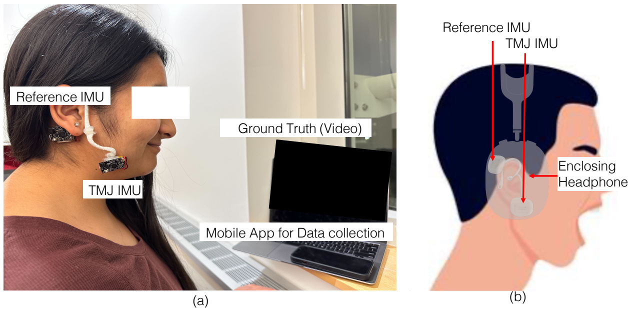

Figure 11: (a) Data collection setup. (b) *Jawthenticate* can be integrated with commercial head-worn devices. Example superposition of our current prototype with Apple AirPods Max for scale.

- *Legitimate user-only classifier (LOC)* - This is an imbalanced data split scheme, wherein we train a one-class SVM for each user. For each legitimate user U, the training data consists of 50% of the data from U, and the testing data consists of the remaining 50% of the data from U and 100% of the data from all other subjects.
- *Legitimate user-Intruder Classifier (LIC)* - This is an imbalanced data split scheme, wherein we take 50% of the data from legitimate user U and 50% of the data from all other subjects in the training set. The remaining data from U and other subjects is used for testing. We train a two-class SVM for each user with U as the legitimate user and the remaining subjects as impostors.
- *Legitimate user-Intruder Proportional Classifier (LIPC)* - This is a balanced data split scheme where we train a two-class SVM for each user. We have the same number of training samples in legitimate and impostor classes, with the number of training samples in the impostor class equal to 50% of the total samples in the legitimate class (user U). We chose impostor samples stratified for equal representation from all subjects.

5 EXPERIMENT SETUP

In this section, we describe the design and implementation of our prototype, and the data collection procedure.

5.1 Implementation

Inspired by [79], we build a custom earable prototype with two IMUs [43]: one placed on TMJ and another placed on the temporal bone, connected via a silicon ear hook [1] as shown in Figure 11 (a). While our current prototype design targets feasibility, we envision that it can be integrated into open-ear and around-the-ear earables allowing for optimal sensor placement [75, 78]. A to-scale example of potential integration with Apple AirPods Max is shown in Figure 11(b). Most existing earables/headphones/HMDs have at least one IMU which can be used as a reference IMU; our system would need only one additional IMU to be placed on the TMJ. The current prototype weighs 20 mg with plastic sensor casings (36mmx27mmx10mm) and the ear hook. This twin-IMU prototype has been shown to work effectively for the removal of body motion artifacts [79]. We keep the IMUs in contact with the skin using a gentle medical-grade adhesive. Besides skin-adhered research prototypes [37, 87] and commercial products [85], FDA-approved [18] adhesive-backed wearables are widely used over long-term (e.g. in glucose [83] and nicotine patches [57]) and are known to be comfortable [41]. Data is collected at 100 Hz; we log timestamp, 3-axis accelerometer (± 2g), and 3-axis gyroscope (250 DPS) data. We stream to an Android Device running the authentication pipeline.

5.2 Data Collection Setup

We evaluate *Jawthenticate* with 41 users (29 Male; 12 Female) in the age range of 16-38 years, enrolled in an IRB-approved study. We collect video as ground truth for labeling. Our participants speak 9 different native languages, Hindi (10), English (8), Greek (4), Persian (4), Tamil (3), Chinese (3), Korean (3), Italian (3), and Kannada (3). We ask the users to talk as they normally do and not restrict their body motion.

A summary of our data collection is presented in Table 1. We collect conversational data as well as ask the user to read some phrases. Conversations are spurred by asking questions, such as "What is your daily schedule?", responses to which are of varying lengths and content. Phrase data is used to determine if our system can distinguish between users when they articulate the same phrase. We selected 15 phrases used in daily life, with each phrase having 15-20 syllables. For example: "How are you doing?" For multilingual data collection, users were asked to create their own passphrases in their native language. We ask the users to not produce any sound for inaudible speech and since we do not use any signal from the microphone, mumbled sounds will not affect the system. For data containing body motion noise, participants were instructed to walk and jog at their regular speeds in a hallway or an empty parking lot. When experimenting with acoustic noise, we ask users to wear AirPods and play top Billboard songs 2021 at three levels – 30dB, 45dB, and 60dB – and converse with the users.

- *Same user; different phrases (Audible)*: We collect conversational data (20 samples (phrases) × 3 sessions) to capture users' natural style of articulation. Conversations are spurred by asking questions, such as "What is your daily schedule?", responses to which are of varying lengths and content.
- *Different users; same phrase (Audible and Inaudible)*: This data simulates a passphrase scenario and determines if our system can distinguish between users when they articulate the same phrase. We selected 15 phrases used in daily life, with each phrase having 15-20 syllables. For example - "How are you doing?" We collect 90 samples ((15 audible + 15 inaudible) × 3 sessions) per user.
- *Same user; different language (Audible and Inaudible)*: We evaluate if the features extracted are invariant to the language. We ask non-native users (33 users) to create 5 passphrases in their native language and articulate them audibly and inaudibly. We have 30 samples from each user ((5 audible + 5 inaudible) × 3 sessions).
- *Numerical data (Audible and Inaudible)*: We asked 10 users to form five 6-digit passcodes (e.g. when a user articulates their passcodes/SSN). Users repeated each passcode 5 times. We collected 50 samples ((5 audible + 5 inaudible) × 5 numerical passcodes) from each user.
- *Presence of motion noise*: We gathered data from 10 users (7 males and 3 females) while they engaged in physical activities, specifically walking and jogging. Participants were instructed to walk and jog at their regular speeds in a hallway or an empty parking lot. They were asked to engage in a 5-minute conversation, both audibly and inaudibly, during walking and jogging. Additionally, we collected 10 minutes of data (5 minutes conversational and 5 minutes audible phrases) after participants finished their physical activities to assess the influence of changes in breathing patterns. On average, we obtained 48 samples per user during walking/jogging and 44 samples post-activity.
- *Presence of acoustic noise*: We collect data from 12 users with music playing. We ask users to wear AirPods and play top Billboard songs 2021 at three levels (30dB, 45dB, and 60dB) and converse with the users for 5 minutes audibly, and 5 minutes in an inaudible manner. We collect a total of 68 samples in this setting.

6 EVALUATION

In this section, we show that: (1) *Jawthenticate* achieves 97.07% Balanced Accuracy. (2) It can authenticate users even when they speak multiple languages. (3) *Jawthenticate* is robust to external attacks. (4) In exit survey, 90% users reported the system to be comfortable.

We identify the following metrics based on their extensive use to evaluate authentication systems [22, 25, 88, 89, 93]: (1) True Positive Rate (TPR; how well the system identifies legitimate users). (2) False Rejection Rate (FRR; how often the system incorrectly classifies legitimate users as an impostor). $FRR = 1 - TPR$. (3) True Negative Rate (TNR; how well the system can block impostors). (4) False Acceptance Rate (FAR; how often the system incorrectly labels an impostor as a legitimate user). $FAR = 1 - TNR$ (5) Balanced Accuracy (BAC). $BAC = (TPR + TNR)/2$.

6.1 Overall Performance

We discuss the overall performance of the three data splitting schemes (Section 4.4). Figure 12 (a) shows the mean value of TPR, TNR, and BAC for the LOC, LIC, and LIPC data splits, which are all greater than 92%. This shows that *Jawthenticate* can adapt to different proportions of data from legitimate users and impostors. The mean values for one-class SVM (LOC) are lower than the other two schemes. This is understandable because this model has no explicit information about the impostor distribution. *Jawthenticate* achieves less than 7% mean FAR and FRR. The mean values of all users' BAC are greater than 90%, indicating robustness across different native languages and accents. For the rest of the evaluations, we report the performance of the LIPC data splitting scheme, unless otherwise mentioned.

6.2 Impact of Spoken Language

Language-invariant authentication is a significant contribution of *Jawthenticate* rendering it more usable, convenient, and practical than other voiced authentication systems [49, 74]. To demonstrate this, we leverage data from the bilingual users in our sample population. We use the data points in the English language as the training set and data from the native language as the test set for each user. Figure 12 (b) shows the mean value of TPR, TNR, and BAC, which are all greater than 92%, indicating that *Jawthenticate* is able to extract and learn speech-related features that are specific to the user and not dependent on the language. The highest mean values are reported for native Hindi and Tamil users, who were proficient in the English language and hence their speech rates and

Setup	Evaluates for	Data	Audible SPS	Inaudible SPS	#Users	#Sessions
Same user; Different phrase	Content invariance (Figure 14)	C	20	-	41	3
Different user; Same phrase	Speech mannerism (Figure 14)	P	15	15	41	3
Same user; Different language	Language invariance (Figure 12(b))	P	5	5	33	3
Numerical passcode	Password entry (Figure 14)	D	5	5	10	1
Walking and Jogging	Body motion noise (Figure 15)	C	52	44	10	1
Post running/jogging	Breathing variations (Figure 15)	C	48	40	10	1
Music on earphones	Acoustic noise (Figure 15)	C	82	52	12	1

Table 1: We collect data in diverse conditions to evaluate *Jawthenticate*'s performance. C: Conversational, P: Phrases, D: 6-digit numerical passcode, SPS: samples per session per user.

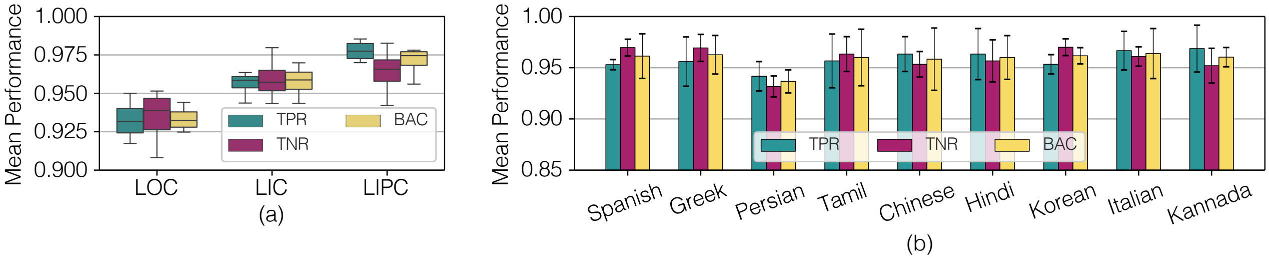

Figure 12: (a) The mean of TPR, TNR, and BAC is >90% for all training schemes. (b) Performance of *Jawthenticate* when the training set is in the English language and the testing set is in the user's native language.

User	U1	U2	U3	U4	U5
FAR	0.067	0.055	0.071	0.059	0.047

Table 2: Impersonation-based mimic attack.

accent were similar in both languages. We do not expect the performance to be affected by users' language proficiency unless their speech mannerisms vary significantly across training and testing languages.

6.3 Attacking *Jawthenticate*

We now evaluate how well *Jawthenticate* can defend against mimic attacks mentioned in §3.2.

Mimic attack (impersonation-based). We collect audible data with 5 users to simulate a mimic attack. Two users (U1 and U2) are native English speakers, and three are native Hindi speakers (U3-U5). Each English speaker mimics another English, and Hindi speakers mimic other Hindi users. Every user articulates 5 phrases, 15 times each, while the other user(s) observe. After each phrase, the other user(s) makes 20 attempts to mimic. We report the FAR in Table 2. The highest FAR is 7.1%, showing that *Jawthenticate* is robust to impersonation-based mimic attacks with TNR>92%.

Advanced mimic attack (video-based). We evaluate *Jawthenticate* with 15 users for the advanced video-based mimic attack, where the impostor has access to a video of the user talking and intends to extract jaw motion from the video. We use video for mimic attacks due to their popularity for behavioral biometric systems [9, 51, 92]. We employ a pre-trained object tracking model [6] to track the TMJ sensor, thereby acquiring the linear acceleration and angular velocity necessary to simulate jaw motion of the legitimate user. To synchronize the IMU data with the video, we manually mark the initiation of jaw motion in the video and align it with the corresponding timestamp in the IMU data stream for each phrase. Given that *Jawthenticate* also requires facial vibration features along with jaw motion, we incorporate the facial vibration features of the impostors. We train a LOC for each user. To attack the system, we use the feature set extracted from the simulated jaw motion using the video of the legitimate user and the facial vibration feature set of the impostor. For each user, we make 700 attempts to break the system. The mean FAR is 11.8% and the mean TPR is 89.4%. This shows that *Jawthenticate* is robust to advanced video-based mimic attacks. It also substantiates that even if an attacker has access to precise jaw motion data, facial vibrations are distinctive and can prevent such attacks.

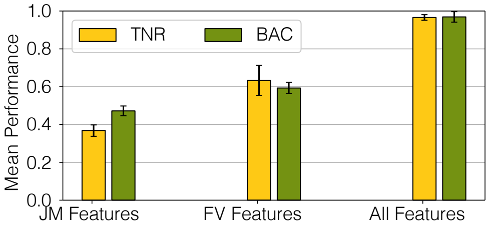

Figure 13: BAC and TNR exhibit a sharp drop when only Jaw Motion (JM) or Facial Vibrations (FV) features are used, compared to all features.

6.4 Impact of Different Settings

In this section, we evaluate the effects of different settings on the performance of *Jawthenticate*.

■ **Impact of different features** We evaluate the impact of jaw motion features and facial vibration features on the overall system performance. We report mean values of TNR and BAC in Figure 13, when the system is trained on individual feature sets

Metric	TPR	TNR	BAC
Mean	0.952 ±3.6%	0.946 ±2.6%	0.949 ±3.1%

Table 3: Change in mean BAC is ≈ 3% for leave-one-session-out evaluation, showing that *Jawthenticate* is robust to variations in sensor placement.

	Phrase	BAC
Ph1	Hey Siri.	0.89
Ph2	How was your day?	0.91
Ph3	How are you doing?	0.92
Ph4	Hope you have a great day.	0.92
Ph5	I think I should get going.	0.93
Ph6	I can not wait for the world cup	0.97
Ph7	Spring break is so short.	0.97
Ph8	I would prefer no meat.	0.98
Ph9	Every cloud has a silver lining.	0.99
Ph10	Let us do apple picking for the weekend.	0.99

Table 4: Short and commonly used phrases have lower BAC (Ph1-Ph5) compared to longer ones (Ph6-Ph10).

(jaw motion and facial vibrations), and a combination of both. The performance for the given metrics exhibit a sharp drop when using only jaw motion or facial vibration features, as compared to using all the features. Specifically, the TNR and BAC drop by 44.3% and 42% respectively. This means that by combining jaw motion and facial vibration features, *Jawthenticate* provides higher accuracy. Additionally, we assessed the impact of sequentially omitting one feature at a time to determine the necessity of each extracted feature. By executing all possible permutations, we observed at least 10% decrease in BAC for each omitted feature, indicating the significance of every feature in maintaining the system's robust performance.

■ **Impact of variation in sensor placement.** We evaluate *Jawthenticate*'s robustness to variations in sensor placement by collecting data over 3 sessions for all users and conducting leave-one-session-out evaluation. In the first session, users were instructed on where and how to place the sensors, and in the other two sessions, they placed sensors on their own. Hence, the position is never exactly the same from one session to the next. The sessions were on different days (separated by > 1 week). For the leave-one-session-out evaluation, we train on data collected from two sessions and test on the third session. As seen in Table 3, the mean value of TPR, TNR, and BAC is more than 90% with a change in mean < 3% over the sessions. These results show that system performs reliably over different sessions despite minor variations in sensor placement.

■ **Impact of phrase selection.** We assess the impact of both, phrase content and length on *Jawthenticate*. We analyze the phrases with the lowest and highest BAC and report BAC for ten phrases across all users in Table 4. The first five phrases (Ph1-Ph5) exhibit the lowest BAC. We speculate that this is because these phrases are short and commonly used, and are therefore articulated similarly by users. These phrases also last less than 3 seconds. The last five phrases (Ph6-Ph10) are longer and less commonly used phrases; these phrases have the highest BAC. To further analyze the impact of phrase length, we input testing phrase samples ranging from one second to nine seconds, and report mean BAC in Figure 17. We observe that with just 3 seconds of data acquisition, *Jawthenticate* can achieve a BAC>92%. This is much faster than recent earable-based systems that require 10 seconds of data to authenticate [22].

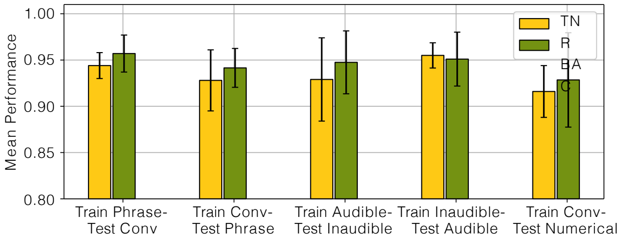

Figure 14: *Jawthenticate* is robust to variations in speech mannerisms.

■ **Impact of speech mannerism.** To evaluate *Jawthenticate*'s robustness to variations in speech mannerism when users speak audibly/inaudibly, or when they speak a predetermined phrase vs. have a conversation, we train and test it on different types of data. We identify 3 cases: (1) Conversational (audible) vs. phrase data (audible+inaudible), (2) Audible vs. inaudible using phrase data, and (3) Conversational (audible) vs. numerical passcodes (audible+inaudible). Figure 14 (first two groups) shows the results for case 1. We can see that *Jawthenticate* can successfully recognize a user whether they use a phrase or have a conversation (mean BAC ≈ 95%). Figure 14 (last two groups) shows the results for case 2, demonstrating that our system learns fundamental jaw motion features that can authenticate users for unvoiced speech even when trained on voiced speech, and vice versa. For case 3, when using audible and inaudible numerical passcodes for testing, *Jawthenticate* achieves 91.7% BAC. The slight drop is due to less variations in how different people say numbers compared to phrases.

■ **Impact of body motion and variation in breathing.** We investigate the impact of motion noise and variations in breathing. Our evaluation involves training the model on data collected in a noise-free environment (sitting at a desk with no large body movements) and testing it on data obtained during walking/jogging and immediately after walking/jogging. The first set of bars in Figure 15 presents the mean values of TNR and BAC for walking/jogging. The BAC>0.9 indicates that motion noise resulting from body motion has minimal effect on *Jawthenticate*'s performance. Additionally, we examine *Jawthenticate*'s performance when there are variations in breathing as it can influence speech characteristics. Figure 15 (second set of bars) illustrates the TNR and BAC after walking/jogging. The BAC is >0.9 in both scenarios indicating that changes in breathing have no significant impact on the system's performance.

■ **Impact of acoustic noise.** We evaluate *Jawthenticate*'s performance when users listen to music on earphones. We trained the system with data collected in noise noise-free environment (no music playing) and tested it on data collected with music playing on earphones. Figure 15 shows that *Jawthenticate* is not affected by music playing on the earphones. This can be because the vibrations generated by music playing are above 20Hz. From this, we conclude that *Jawthenticate* can be used with earphones, without compromising their functionality.

6.5 Longitudinal Study

To evaluate the performance of *Jawthenticate* in capturing the speech mannerisms over the long term, we further conducted a 4-week longitudinal study involving 7 users. We use the data described in Section 5.2 to train a model. Then, we use data collected

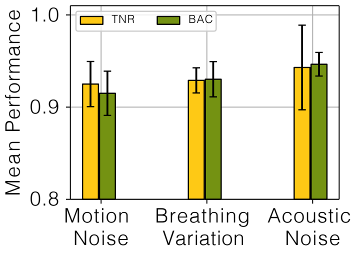

Figure 15: *Jawthenticate* is robust to various external noises.

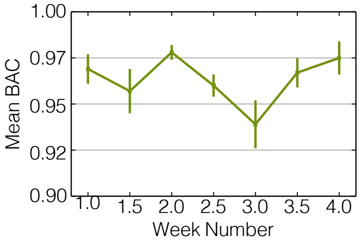

Figure 16: Consistent performance for 7 users over a month (3 sessions/week).

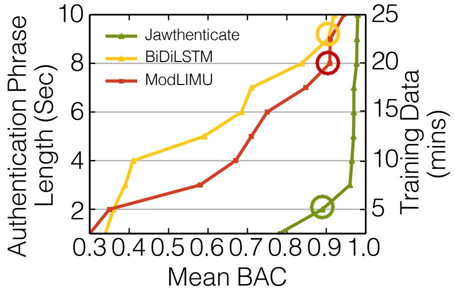

Figure 17: Comparison with baselines.

from the same users over the following 4 weeks to serve as the testing set. In each session, users were asked to audibly articulate 20 phrases (15 phrases provided by us, 5 passphrases created by the user) and engage in a 5-minute conversation, both audibly and inaudibly, and repeat this study over a month. On average, one user finishes 12 sessions in 4 weeks, with a total of 520 samples from each user. Figure 16 demonstrates a BAC exceeding 0.92, confirming our system's robustness over time.

6.6 Identifying Speech-related Jaw Motion

To separate speech from non-speech-related jaw motion, we conduct conversations with 5 users while they eat. They were asked to chew and talk as they felt comfortable. On average, we had 21 minutes of eating and 13 minutes of talking for each user. We manually annotate the boundaries for speech and non-speech data from the video. If the detected boundary is within 0.1 seconds from the ground truth we label it as correctly identified speech-related motion. We achieve a true positive rate of 99% with 2% false positives, allowing our system to successfully discard non-speech-related jaw motions in real-world settings.

6.7 Comparison with DL Baselines

We compare *Jawthenticate* with two deep learning (DL) models as baselines. We use a Bidirectional LSTM (BiDiLSTM) and LIMU-BERT [90] with a classifier head (ModLIMU). In BiDiLSTM we have one Bidirectional LSTM layer, a fully connected layer, and a classifier layer. For the ModLIMU baseline, we use the pre-trained model [90] to extract representations, followed by a bidirectional LSTM as a classifier head. We try different inputs to the two baselines: (1) a 3D array of the orientation, (2) a 1D array created by flattening the orientation, (3) 1D array created by flattening the 6 axis IMU data. We report the evaluation results for input with the best BAC. Figure 17 shows the amount of training data and authentication data for each system, along with the achieved BAC. The baselines achieve comparable mean BAC as *Jawthenticate*. However, they need a considerably longer sequence of training data and test data (authentication phrase) to achieve a mean BAC greater than 90%. *Jawthenticate* achieves a BAC of 90% with 5 minutes of training data and only 3 seconds of test data, while BiDiLSTM and ModLIMU need more than 20 minutes of training data and 8-9 seconds of data for authentication to achieve the same BAC.

System	Unique Biometric	Speech-based Features	Validated with Multiple Languages	Supports Inaudible Speech	Audio/ Replay Attack Resistant	Performance
Vocal Resonance [41]	Vocal Resonance	✗	✗	✗	✗	96% (BAC)
EchoVib [2]	Phonatory Vibrations	✗	✗	✗	✓	90% (BAC)
EarDynamic [78]	Ear Canal Geometry	✗	✗	✓	✗	97.38% (Recall)
MandiPass [40]	Bone Borne Vibrations	✗	✗	✗	✓	1.28% (FRR)
VAuth [17]	Bone Borne Vibrations	✗	✓	✗	✓	97% (BAC)
Face-Mic [63]	Bone Borne and Facial Muscle Vibrations	✗	✗	✗	✓	97% (TPR)
Jawthenticate	Jaw Motion and Facial Muscle Vibrations	✓	✓	✓	✓	97.07% (BAC), 2.25% (FRR), 97.75% (TPR), 98.1% (Recall)

Table 5: Comparing *Jawthenticate* with related earable speech-based authentication systems.

6.8 Implementation and Study Results

■ **Latency and Power Consumption.** We implement *Jawthenticate* on Android Smartphone Pixel 3XL running Android 12.0. We use TensorFlow lite to run the SVM on mobile device. We run the complete pipeline for 150 phrase instances. *Jawthenticate* takes 733ms and consumes 268mJ for running the end-to-end pipeline to authenticate a user. This includes streaming data over Bluetooth from the prototype, preprocessing, anti-noise filtering, extracting the features, and running LIPC two-class SVM. These latency numbers are comparable with other recent earable-based authentication systems [22, 49] and can further be improved by optimizing the code.

■ **Exit Survey.** We conduct an exit survey with all participants, using an anonymous Google Form, designed after the SUS survey [47] to validate the usability and practicality of *Jawthenticate*. The form contains several questions on a Likert scale [38] to gauge the users' perception of *Jawthenticate* on comfort level, usability, and willingness to use it continuously over long periods. 90.2% users strongly agree that *Jawthenticate* is comfortable, and 92% users think it is intuitive to use. When asked if they would be willing to use *Jawthenticate* as a second-factor authentication device, 87% of the users strongly agree, and 88% of users report being willing to wear it for more than 3 hours. These responses are encouraging and show affinity to accept *Jawthenticate*.

7 RELATED WORK

Jawthenticate is the first work to use jaw motion as a biometric for authentication using voiced and unvoiced speech for multiple languages. Table 5 presents a summary of comparison with existing work on speech-based authentication earable systems, where we compare the capability to derive a user's speech mannerism (speech-based features), generalizability over multiple languages, support of inaudible speech, and resistance to audio/replay attacks.

While EarEcho [25] and EarDynamic [89] exploit changes in a user's ear canal when they speak by transmitting active probe signals into the ear, Vocal Resonance [50], EarPrint [24], and MandiPass [49] capture voice-induced body sound transmission. ToothSonic [88] leverages the sonic effect produced when a user performs teeth gestures for earable authentication that are captured via an inward-facing microphone, but its functionality is hindered when music plays on the earable. Hu Et al. [34] estimate the user-specific occluded ear canal transfer function by capturing the difference between the sounds inside and outside the ear. Like ToothSonic, [34] cannot perform authentication when music is playing on the earable and requires strong ambient sounds that can be captured by the in-ear microphone. Unlike both these systems, *Jawthenticate* does not require a microphone or strong ambient sounds, and can be used while music is playing on the earable. Face-Mic [74] uses the speech-associated facial vibrations (require audible speech) captured via IMUs in VRs to identify a user. EchoVib [2], Accuth [30] (sampling IMU at 500 Hz), and VocalPrint [48] (directs mmWave signals at user's throat) also leverage vocal cord vibrations generated during audible speech articulation. VAuth [21] uses a microphone and IMU for continuous authentication by matching the speech-associated body-surface vibrations. All these systems either require the user to speak audibly, which may not always be practical due to privacy and accessibility constraints, sample at a high frequency making real-world use impractical, or use external setups constraining them to specific environments.

Bone-conduction-based systems like SkullConduct [68] do not require a user to produce audible sound. However, they play white noise in the earphones which might be unpleasant to the user, susceptible to external sounds, and prone to sound/voice injection-based attacks. EarGate [22] requires that a user walk every time to authenticate themselves, which can be impractical in many real-world scenarios. Similarly, EarID [93] captures behavioral characteristics of the user via earphone IMU. In contrast to our work, none of the systems have been demonstrated to work with both audible and inaudible speech. By using jaw motion and facial vibrations as the signature and validating it with a larger population/languages than other works, we establish the unique contributions of *Jawthenticate*.

8 DISCUSSION

Jawthenticate is an early authentication prototype and a first of its kind that leverages jaw motion and facial vibrations to perform microphone-free speech-based authentication.

Limitations. Currently, we have evaluated *Jawthenticate* with 41 users. It is likely that a larger population size will affect the result. To address this, we will conduct a larger study and employ techniques [10, 12, 40] that could adjust based on the specific user population. We also plan to evaluate the system at high speeds (in a bus or car) and the impact of variations in speech mannerisms due to the influence of alcohol, cold, cough, caffeine, and emotions.

Comparison with standard biometric authentication systems: Our system has lower performance compared to top-3 standard biometric authentication systems [66], fingerprint [91] (TPR1%), facial recognition [26] (99% BAC), and iris recognition [17] (>99.9% TPR). However, it is important to note that *Jawthenticate* is the first of its kind and is still in its early stages. We believe that with further research, *Jawthenticate* can be as accurate as the standard biometric authentication systems and can even be used as a secondary authentication medium.

Frequency of retraining, model extraction, and physical attacks: The speech mannerisms of users can change over time due to various factors like age, health, and emotions. We plan to study the frequency of retraining the system to keep it up-to-date. This can be done via a closed-loop system where we ask users to provide some data for training and developing a system with adaptive learning capability. We also plan to make *Jawthenticate* robust against model extraction attacks using techniques like limiting the number of unsuccessful trials, adding noise to the model's output [32], and differential privacy during training [28].

9 CONCLUSION

Jawthenticate learns speech-based features without using a microphone and is invariant to changes in the content (conversational, passphrase, or numerical passcode) or the spoken language. It performs well for both, audible and inaudible speech. Experiments with 41 users show that *Jawthenticate* achieves a balanced accuracy of 97.07%. We envision that *Jawthenticate* can enable private hands-free voice-free authentication in noisy environments (e.g. factory or train), in quiet spaces (e.g. library or theatre), or for patients who cannot produce sound but have their jaw movements intact (e.g. tracheostomy). *Jawthenticate* can also be used as a secondary authentication factor to enhance the security of existing systems.

ACKNOWLEDGMENTS

We thank the shepherd and the anonymous reviewers for their insightful comments. This material is based in part upon work supported by the National Science Foundation under Award Numbers 2238553, 2110193, and 2152357.

REFERENCES

[1] Amazon. 2021. Ear hooks. shorturl.at/IPR35

[2] S Abhishek Anand, Jian Liu, Chen Wang, Maliheh Shirvanian, Nitesh Saxena, and Yingying Chen. 2021. Echovib: Exploring voice authentication via unique non-linear vibrations of short replayed speech. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*. 67–81.

[3] auth0. 2022. Authenticating Users in Your VR Apps. https://auth0.com/blog/authenticating-users-in-your-vr-apps/

[4] Barcalys. 2016. Barclays rolls out voice biometrics for phone banking. https://www.finextra.com/newsarticle/29245/barclays-rolls-out-voice-biometrics-for-phone-banking

[5] Paul M Bentley and JTE McDonnell. 1994. Wavelet transforms: an introduction. *Electronics & communication engineering journal* 6, 4 (1994), 175–186.

[6] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*. Springer, 850–865.

[7] Andrzej K Brodzik and Olivia Peters. 2005. Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 5. IEEE, v–373.

[8] DS Broomhead, JP Huke, and MR Muldoon. 1992. Linear filters and non-linear systems. *Journal of the Royal Statistical Society: Series B (Methodological)* 54, 2 (1992), 373–382.

[9] Arne Bruesch, Ngu Nguyen, Dominik Schürmann, Stephan Sigg, and Lars Wolf. 2019. Security properties of gait for mobile device pairing. *IEEE Transactions on Mobile Computing* 19, 3 (2019), 697–710.

[10] Mario Parreño Centeno, Aad van Moorsel, and Stefano Castruccio. 2017. Smartphone continuous authentication using deep learning autoencoders. In *2017 15th annual conference on privacy, security and trust (pst)*. IEEE, 147–1478.

[11] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4960–4964. https://doi.org/10.1109/ICASSP.2016.7472621

[12] Xue-wen Chen and Jong Cheol Jeong. 2007. Enhanced recursive feature elimination. In *Sixth international conference on machine learning and applications (ICMLA 2007)*. IEEE, 429–435.

[13] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20 (1995), 273–297.

[14] Haipeng Dai, Wei Wang, Alex X Liu, Kang Ling, and Jiajun Sun. 2019. Speech based human authentication on smartphones. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[15] Rohan Kumar Das, Sarfaraz Jelil, and SR Mahadeva Prasanna. 2017. Development of multi-level speech based person authentication system. *Journal of Signal Processing Systems* 88, 3 (2017), 259–271.

[16] Rohan Kumar Das, Xiaohai Tian, Tomi Kinnunen, and Haizhou Li. 2020. The attacker's perspective on automatic speaker verification: An overview. *arXiv preprint arXiv:2004.08849* (2020).

[17] John Daugman. 2009. How iris recognition works. In *The essential guide to image processing*. Elsevier, 715–739.

[18] DexCom. 2022. Dexcom G7 Receives FDA Clearance. https://investors.dexcom.com/news/news-details/2022/Dexcom-G7-Receives-FDA-Clearance-The-Most-Accurate-Continuous-Glucose-Monitoring-System-Cleared-in-the-U.S/default.aspx#:~:text=(NASDAQ%3ADXCM)%2C%20the,ages%20two%20years%20and%20older.

[19] Richard P Di Fabio. 1998. Physical therapy for patients with TMD: a descriptive study of treatment, disability, and health status. *Journal of orofacial pain* 12, 2 (1998).

[20] H Ebelthite. 2016. Could your health be ruined by noises you can't hear? Some gadgets emit silent ultra-high whines that may hurt you. *The Daily Mail. Available from: http://www. dailymail. co. uk/femail/article-3527060/Could-health-ruined-noises-t-hear-gadgets-emit-silent-ultra-high-whines-hurt-you. html [Last viewed 2 August 2017]* (2016).

[21] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 343–355.

[22] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 337–349.

[23] Mark D Fletcher, Sian Lloyd Jones, Paul R White, Craig N Dolder, Timothy G Leighton, and Benjamin Lineton. 2018. Effects of very high-frequency sound and ultrasound on humans. Part I: Adverse symptoms after exposure to audible very-high frequency sound. *The journal of the acoustical society of America* 144, 4 (2018), 2511–2520.

[24] Yang Gao, Yincheng Jin, Jagmohan Chauhan, Seokmin Choi, Jiyang Li, and Zhanpeng Jin. 2021. Voice In Ear: Spoofing-Resistant and Passphrase-Independent Body Sound Authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 12 (mar 2021), 25 pages. https://doi.org/10.1145/3448113

[25] Yang Gao, Wei Wang, Vir V. Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using Ear Canal Echo for Wearable Authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 81 (sep 2019), 24 pages. https://doi.org/10.1145/3351239

[26] Patrick J Grother, Mei L Ngan, and Kayee K Hanaoka. 2018. Ongoing face recognition vendor test (frvt) part 2: Identification. (2018).

[27] Frank H Guenther. 2016. *Neural control of speech*. Mit Press.

[28] Trung Ha, Tran Khanh Dang, Tran Tri Dang, Tuan Anh Truong, and Manh Tuan Nguyen. 2019. Differential privacy in deep learning: an overview. In *2019 International Conference on Advanced Computing and Applications (ACOMP)*. IEEE, 97–102.

[29] Abdenour Hadid. 2014. Face biometrics under spoofing attacks: Vulnerabilities, countermeasures, open issues, and research directions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 113–118.

[30] Feiyu Han, Panlong Yang, Haohua Du, and Xiang-Yang Li. 2023. Accuth: Anti-Spoofing Voice Authentication via Accelerometer. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems* (Boston, Massachusetts) *(SenSys '22)*. Association for Computing Machinery, New York, NY, USA, 637–650. https://doi.org/10.1145/3560905.3568522

[31] Monson H Hayes. 1996. *Statistical digital signal processing and modeling*. John Wiley & Sons.

[32] Xuanli He, Qiongkai Xu, Lingjuan Lyu, Fangzhao Wu, and Chenguang Wang. 2022. Protecting intellectual property of language generation apis with lexical watermark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10758–10766.

[33] HSBC. 2016. HSBC rolls out voice and touch ID security for bank customers. https://www.theguardian.com/business/2016/feb/19/hsbc-rolls-out-voice-touch-id-security-bank-customers

[34] Changshuo Hu, Xiao Ma, Dong Ma, and Ting Dang. 2023. Lightweight and Non-Invasive User Authentication on Earables. In *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications* (Newport Beach, California) *(HotMobile '23)*. Association for Computing Machinery, New York, NY, USA, 36–41. https://doi.org/10.1145/3572864.3580332

[35] Jiankun Hu. 2008. Mobile fingerprint template protection: Progress and open issues. In *2008 3rd IEEE Conference on Industrial Electronics and Applications*. IEEE, 2133–2138.

[36] Ewa Jacewicz, Robert A Fox, Caitlin O'Neill, and Joseph Salmons. 2009. Articulation rate across dialect, age, and gender. *Language variation and change* 21, 2 (2009), 233–256.

[37] Shuo Jiang, Ling Li, Haipeng Xu, Junkai Xu, Guoying Gu, and Peter B. Shull. 2020. Stretchable e-Skin Patch for Gesture Recognition on the Back of the Hand. *IEEE Transactions on Industrial Electronics* 67, 1 (2020), 647–657. https://doi.org/10.1109/TIE.2019.2914621

[38] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology* 7, 4 (2015), 396.

[39] Prerna Khanna, Tanmay Srivastava, Shijia Pan, Shubham Jain, and Phuc Nguyen. 2021. Jawsense: recognizing unvoiced sound using a low-cost ear-worn system. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 44–49.

[40] Teuvo Kohonen. 1990. The self-organizing map. *Proc. IEEE* 78, 9 (1990), 1464–1480.

[41] Bernhard Kulzer, Guido Freckmann, Lutz Heinemann, Oliver Schnell, Rolf Hinzmann, and Ralph Ziegler. 2022. Patch Pumps: What are the advantages for people with diabetes? *Diabetes Research and Clinical Practice* (2022), 109858.

[42] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D. Abowd, Nicholas D. Lane, and Thomas Ploetz. 2020. IMUTube: Automatic Extraction of Virtual on-body Accelerometry from Video for Human Activity Recognition. arXiv:2006.05675 [cs.CV]

[43] Mbient Lab. 2020. Mbient IMU. https://mbientlab.com/metamotionr/

[44] BW Lawton. 2013. Exposure limits for airborne sound of very high frequency and ultrasonic frequency. (2013).

[45] TG Leighton. 2016. Are some people suffering as a result of increasing mass exposure of the public to ultrasound in air? *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 472, 2185 (2016), 20150624.

[46] Willem JM Levelt. 1993. *Speaking: From intention to articulation*. MIT press.

[47] James R Lewis. 2018. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction* 34, 7 (2018), 577–590.

[48] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, and Wenyao Xu. 2020. VocalPrint: Exploring a Resilient and Secure Voice Authentication via MmWave Biometric Interrogation. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems* (Virtual Event, Japan) *(SenSys '20)*. Association for Computing Machinery, New York, NY, USA, 312–325. https://doi.org/10.1145/3384419.3430779

[49] Jianwei Liu, Wenfan Song, Leming Shen, Jinsong Han, Xian Xu, and Kui Ren. 2021. MandiPass: Secure and Usable User Authentication via Earphone IMU. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*. 674–684. https://doi.org/10.1109/ICDCS51616.2021.00070

[50] Rui Liu, Cory Cornelius, Reza Rawassizadeh, Ronald Peterson, and David Kotz. 2018. Vocal resonance: Using internal body voice for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–23.

[51] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. When video meets inertial sensors: zero-shot domain adaptation for finger motion analytics with inertial sensors. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 182–194.

[52] Shivangi Mahto, Takayuki Arakawa, and Takafumi Koshinak. 2018. Ear Acoustic Biometrics Using Inaudible Signals and Its Application to Continuous User Authentication. In *2018 26th European Signal Processing Conference (EUSIPCO)*. 1407–1411. https://doi.org/10.23919/EUSIPCO.2018.8553015

[53] Tania Marur, Yakup Tuna, and Selman Demirci. 2014. Facial anatomy. *Clinics in dermatology* 32, 1 (2014), 14–23.

[54] RW McAlister, EM Harkness, and JJ Nicoll. 1998. An ultrasound investigation of the lip levator musculature. *The European Journal of Orthodontics* 20, 6 (1998), 713–720.

[55] Maranda McBride, Phuong Tran, and Tomasz Letowski. 2008. Head mapping: Search for an optimum bone microphone placement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52. SAGE Publications Sage CA: Los Angeles, CA, 503–507.

[56] Assaf Hurwitz Michaely, Xuedong Zhang, Gabor Simko, Carolina Parada, and Petar Aleksic. 2017. Keyword spotting for Google assistant using contextual speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 272–278. https://doi.org/10.1109/ASRU.2017.8268946

[57] NicoDerm. 2023. NicoDerm CQ Step 2. https://www.walmart.com/ip/NicoDerm-CQ-Step-2-Extended-Release-Nicotine-Patches-to-Quit-Smoking-14-Mg-14-Count/1115576?wl13=2906&selectedSellerId=0

[58] The University of Reading. 2021. The production of speech sounds. http://www.personal.rdg.ac.uk/~llsroach/phon2/artic-basics.htm

[59] R. Parent, S. King, and O. Fujimura. 2002. Issues with lip sync animation: can you read my lips?. In *Proceedings of Computer Animation 2002 (CA 2002)*. 3–10. https://doi.org/10.1109/CA.2002.1017500

[60] Chang-Mok Park, Devinder Thapa, and Gi-Nam Wang. 2007. Speech authentication system using digital watermarking and pattern recovery. *Pattern Recognition Letters* 28, 8 (2007), 931–938.

[61] Joel E Pessa, Vikram P Zadoo, Earle K Adrian Jr, Cheng H Yuan, Jason Aydelotte, and Jaime R Garza. 1998. Variability of the midfacial muscles: analysis of 50 hemifacial cadaver dissections. *Plastic and reconstructive surgery* 102, 6 (1998), 1888–1893.

[62] David Poeppel and M Florencia Assaneo. 2020. Speech rhythms and their neural foundations. *Nature reviews neuroscience* 21, 6 (2020), 322–334.

[63] Grand View Research. 2022. GVR Report coverEarphones & Headphones Market Size, Share & Trends Report Earphones & Headphones Market Size, Share & Trends Analysis Report By Product (Earphones, Headphones), By Price, By Technology, By Application, By Region, And Segment Forecasts, 2020 - 2027.

[64] Douglas A Reynolds. 1995. Speaker identification and verification using Gaussian mixture speaker models. *Speech communication* 17, 1-2 (1995), 91–108.

[65] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible Voice Commands: The {Long-Range} Attack and Defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*. 547–560.

[66] T Sabhanayagam, V Prasanna Venkatesan, and K Senthamaraikannan. 2018. A comprehensive survey on various biometric systems. *International Journal of Applied Engineering Research* 13, 5 (2018), 2276–2297.

[67] Pouria Saidi, George Atia, and Azadeh Vosoughi. 2017. On robust detection of brain stimuli with Ramanujan Periodicity Transforms. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*. 729–733. https://doi.org/10.1109/ACSSC.2017.8335440

[68] Stefan Schneegass, Youssef Oualil, and Andreas Bulling. 2016. SkullConduct: Biometric user identification on eyewear computers using bone conduction through the skull. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1379–1384.

[69] Mojtaba Sepasian, Cristinel Mares, and Wamadeva Balachandran. 2009. Liveness and spoofing in fingerprint identification: Issues and challenges. In *Proc. 4th WSEAS Int. Conf. Comput. Eng. Appl.(CEA)*. 150–158.

[70] William A Sethares and Diego Bañuelos. 2007. *Rhythm and transforms*. Vol. 1. Springer.

[71] William A Sethares and Thomas W Staley. 1999. Periodicity transforms. *IEEE transactions on Signal Processing* 47, 11 (1999), 2953–2964.

[72] William A Sethares and Thomas W Staley. 2001. Meter and periodicity in musical performance. *Journal of New Music Research* 30, 2 (2001), 149–158.

[73] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.

[74] Cong Shi, Xiangyu Xu, Tianfang Zhang, Payton Walker, Yi Wu, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. 2021. Face-Mic: Inferring Live Speech and Speaker Identity via Subtle Facial Dynamics Captured by AR/VR Motion Sensors. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking* (New Orleans, Louisiana) *(MobiCom '21)*. Association for Computing Machinery, New York, NY, USA, 478–490. https://doi.org/10.1145/3447993.3483272

[75] Shokz. 2023. Shokz - OpenRun Bone Conduction Open-Ear Endurance Headphones - Black. https://www.bestbuy.com/site/shokz-openrun-bone-conduction-open-ear-endurance-headphones-black/6492431.p?skuId=6492431&extStoreId=498&ref=212&loc=1&gclid=Cj0KCQjw4s-kBhDqARIsAN-ipH1_M-i6_ujf5ADFF-EA4M30q1n_KNNLEcrdWYHCCdUCxcH5JmQ9zlAaAqhKEALw_wcB&gclsrc=aw.ds

[76] Prakash Shrestha and Nitesh Saxena. 2018. Listening watch: Wearable two-factor authentication using speech signals resilient to near-far attacks. In *Proceedings of the 11th ACM conference on security & privacy in wireless and mobile networks*. 99–110.

[77] Nilu Singh, Alka Agrawal, and RA Khan. 2018. Voice biometric: A technology for voice based authentication. *Advanced Science, Engineering and Medicine* 10, 7-8 (2018), 754–759.

[78] Sony. 2023. Sony Xperia Ear Duo True Wireless headset – Black. https://www.amazon.com/Sony-Xperia-True-Wireless-headset/dp/B079WDK6S3

[79] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 140 (sep 2022), 26 pages. https://doi.org/10.1145/3550281

[80] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 140 (sep 2022), 26 pages. https://doi.org/10.1145/3550281

[81] RG Stockwell. 2007. Why use the S-transform. *Pseudo-differential operators: partial differential equations and time-frequency analysis* 52 (2007), 279–309.

[82] R.G. Stockwell, L. Mansinha, and R.P. Lowe. 1996. Localization of the complex spectrum: the S transform. *IEEE Transactions on Signal Processing* 44, 4 (1996), 998–1001. https://doi.org/10.1109/78.492555

[83] SugarBEAT. 2022. SugarBEAT: "The World's First Non-Invasive Glucose Monitor". https://www.healthline.com/diabetesmine/non-invasive-sugarbeat-cgm-diabetes#Winning-at-glucose-measurement-without-needles?

[84] Saurabh Upadhyay and Sanjay Kumar Singh. 2012. Video authentication: Issues and challenges. *International Journal of Computer Science Issues (IJCSI)* 9, 1 (2012), 409.

[85] Upright. 2023. The Simplest Way to Transform Your Posture. https://www.uprightpose.com/

[86] Eric Vatikiotis-Bateson and David J Ostry. 1995. An analysis of the dimensionality of jaw motion in speech. *Journal of Phonetics* 23, 1-2 (1995), 101–117.

[87] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I. Hong, Carmel Majidi, and Swarun Kumar. 2020. RFID Tattoo: A Wireless Platform for Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 155 (sep 2020), 24 pages. https://doi.org/10.1145/3369812

[88] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. ToothSonic: Earable Authentication via Acoustic Toothprint. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 78 (jul 2022), 24 pages. https://doi.org/10.1145/3534606

[89] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 39 (mar 2021), 27 pages. https://doi.org/10.1145/3448098

[90] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.

[91] Wencheng Yang, Song Wang, Jiankun Hu, Guanglou Zheng, and Craig Valli. 2019. Security and accuracy of fingerprint-based biometrics: A review. *Symmetry* 11, 2 (2019), 141.

[92] Shibo Zhang and Nabil Alshurafa. 2020. Deep generative cross-modal on-body accelerometer data synthesis from videos. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 223–227.

[93] Yongpan Zou, Haibo Lei, and Kaishun Wu. 2021. Beyond Legitimacy, also with Identity: Your Smart Earphones Know Who You Are Quietly. *IEEE Transactions on Mobile Computing* (2021), 1–1. https://doi.org/10.1109/TMC.2021.3134654