MMGatorAuth: A Novel Multimodal Dataset for Authentication Interactions in Gesture and Voice

Sarah Morrison-Smith
smorriso@barnard.edu
Barnard College
New York, NY

Aishat Aloba*
aoaloba@ufl.edu
University of Florida
Gainesville, FL, USA

Hangwei Lu†
qslvhw@ufl.edu
University of Florida
Gainesville, FL, USA

Brett Benda*
Shaghayegh Esmaeili*
Gianne Flores*
University of Florida
Gainesville, FL, USA

Jesse Smith*
Nikita Soni*
Isaac Wang*
University of Florida
Gainesville, FL, USA

Rejin Joy†
University of Florida
Gainesville, FL, USA

Damon L. Woodard†
dwoodard@ece.ufl.edu
University of Florida
Gainesville, FL, USA

Jaime Ruiz*
jaime.ruiz@ufl.edu
University of Florida
Gainesville, FL, USA

Lisa Anthony*
lanthony@cise.ufl.edu
University of Florida
Gainesville, FL, USA

ABSTRACT

The future of smart environments is likely to involve both passive and active interactions on the part of users. Depending on what sensors are available in the space, users may make use of multimodal interaction modalities such as hand gestures or voice commands. There is a shortage of robust yet controlled multimodal interaction datasets for smart environment applications. One application domain of interest based on current state-of-the-art is authentication for sensitive or private tasks, such as banking and email. We present a novel, large multimodal dataset for authentication interactions in both gesture and voice, collected from 106 volunteers who each performed 10 examples of each of a set of hand gesture and spoken voice commands chosen from prior literature (10,600 gesture samples and 13,780 voice samples). We present the data collection method, raw data and common features extracted, and a case study illustrating how this dataset could be useful to researchers. Our goal is to provide a benchmark dataset for testing future multimodal authentication solutions, enabling comparison across approaches.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Security and privacy** → *Usability in security and privacy*.

*Dept of Computer & Information Science & Engineering (CISE)
†Dept of Electrical & Computer Engineering (ECE)

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418881

KEYWORDS

Datasets; multimodal; gesture; voice; authentication; biometrics.

ACM Reference Format:
Sarah Morrison-Smith, Aishat Aloba, Hangwei Lu, Brett Benda, Shaghayegh Esmaeili, Gianne Flores, Jesse Smith, Nikita Soni, Isaac Wang, Rejin Joy, Damon L. Woodard, Jaime Ruiz, and Lisa Anthony. 2020. MMGatorAuth: A Novel Multimodal Dataset for Authentication Interactions in Gesture and Voice. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3382507.3418881

1 INTRODUCTION

Mark Weiser's famous vision for *ubiquitous computing* was that computer technologies would be interconnected and integrate seamlessly with the environment, "fading" into the background for most users [58]. Researchers have taken this vision to imply that so-called *natural* communication modalities such as touch, voice, vision, and motion will be useful for multimodal interaction in these ubiquitous "smart environments" [2, 42]. Many of these futuristic visions imply little to no effort on the part of the user to consciously "interact" with the system, but instead that the system will be able to infer user intent from natural behaviors seamlessly in any context. In reality, based on current trends such as smart home assistants, it seems likely that the future of smart environments is likely to involve both passive and active interactions on the part of users in the space. Which modalities are used in a space by users will depend both on what sensors are available in the space, and how users feel about the comfort, usability, trustworthiness, and/or robustness of various multimodal interaction modalities.

Research into designing and developing this vision of future smart environments requires access to interaction datasets that reflect the range of potential users and variability in their interaction behaviors. Small datasets with few users can result in an algorithm that performs well on the training data, but does not generalize

to the larger population well due to bias (underfitting) and variance (overfitting) [6]. However, there is a shortage of robust yet controlled multimodal interaction datasets for smart environment applications that can support these goals.

Context of the interaction is also a critical factor in which user behavior is likely to change. One application domain of interest based on current state-of-the-art is authentication for sensitive or private tasks, such as banking and email. Smart home assistants were slow to support any type of user authentication [13], which could lead to security exploitations, accidental or purposeful. News headlines have documented the problems with this model: news anchors speaking commands on TV [7] or advertisements for current technology [35] have been recognized by the devices, resulting in unwanted purchases. Even these recently added authentication features are limited in scope, requiring users to log in only once, and emphasize passive authentication [13].

To support future research on natural interaction for smart environments, specifically for the context of user authentication, we present a novel multimodal dataset in both gesture and voice, which we call MMGatorAuth. We collected this dataset in a controlled lab setting from 106 volunteers, who each performed 10 examples of each of a set of hand gesture and spoken voice commands. We chose the set of commands based on a review of prior literature and a goal to span the range of input behaviors that may be encountered in a real-world authentication interaction. In total we have 10,600 gesture samples and 13,780 voice samples. We used the Kinect V2 sensor and a high-quality Blue Yeti microphone to record our dataset, to create a high quality dataset using sensors likely to be available in smart-home devices in the near future.

In this paper, we present the data collection method, raw data and common features extracted, and a case study illustrating the types of research questions this dataset can make possible to answer. Our goal is to provide a benchmark dataset for testing future multimodal authentication solutions, enabling comparison across approaches. Our large, robust dataset, which provides two modalities that have not been studied as frequently in the literature, can also be useful for general purpose gesture and voice recognition algorithm evaluations.

2 SELECTION OF DATASET COMMANDS

The gesture and voice commands chosen for this study were based on a review of prior research in this space [14, 24, 44, 49, 51, 52, 60, 63]. A set of 10 gesture commands were chosen, and a set of 10 voice commands plus 3 spoken English-language pangrams. We chose to ask users to perform specific commands rather than proposing their own natural command preferences in order to support the systematic testing of gesture and speech recognition algorithms, which requires a balanced sample of instances in each class to succeed. The process of identifying and selecting the commands in the dataset is detailed for each modality next.

2.1 Gestures

A review of past work revealed that three main types of gestures have been used for authentication purposes: hand shape [52], mid-air handwriting (aka, "air signature") [52], and hand wave [24, 49, 63]. Takeuchi et al.'s [52] work on recognizing hand shapes suggested gestures such as an index finger pointing up. Expanding on this example, we also included three other handshapes common in Western culture: thumbs up, thumbs down, and the okay sign, since it is likely that such common gestures would be easily learned or chosen independently by users. We also adopted four other gestures from prior work by Rico and Brewster [44] that explored acceptable body and device gestures. The hand distance, forearm touch, wrist rotation (aka, "double flip" [45]), and clap gestures were all selected due to usage in existing gesture-based interfaces [14, 51, 60] and their potential for inclusion in future interfaces [44].

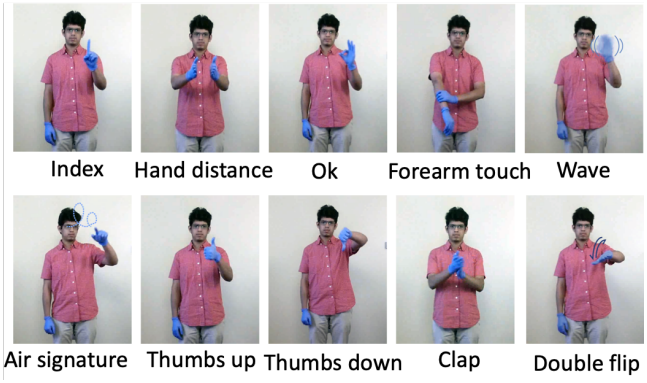

Figure 1: Gesture types represented in the dataset; participants performed 10 examples of each gesture type. Gesture types were chosen based on a review of prior related work.

Each gesture is shown in Figure 1, from top to bottom and left to right: index finger, thumbs up, thumbs down, ok sign, wave, air signature, forearm touch, hand distance, clap, and wrist rotation. Each gesture sample includes the user starting with their hands at their sides, performing the gesture, holding it for a short length of time, and then transitioning out of making the gesture by returning their hands to their sides.

2.2 Phrases

We designed the phrase set for the users to speak so that it would cover all 44 phonemes in the English language. This approach enables the dataset to be used for both voice-based authentication based on the recognition of the passphrase or utterance, and biometric authentication based on audio features of the vocal sine wave itself. We also reviewed common current use cases for smart assistants like Google Home or Amazon Alexa. We created ten phrases that spanned all 44 English phonemes, as well as a phonemic pangram in the form of a short story. A phonemic pangram is a phrase that contains all possible phonemic sounds in the target language [11]. The pangram was further subdivided into three phrases to make it easier for participants to repeat the story correctly since a mistake near the end of the story would otherwise necessitate repeating the story in its entirety. These phrases and the three-part pangram are listed in Table 1. These data can also be used in speech recognition experiments.

Table 1: List of phrases represented in the dataset; participants spoke 10 examples of each voice command and pangram. Commands were chosen based on prior work.

Abbr	Phrase
Chairs	Alexa, order a thousand more chairs.
Credit	Alexa, what is my credit score?
Message	Alexa, play my new voice messages.
Mail	Alexa, place a hold on my mail until March.
Photos	Google, delete all photos of theaters.
Doctor	Google, when is my next doctor's appointment?
Bank	Google, what is my bank balance?
Password	Google, change my email password to "Missouri".
Facebook	Google, update my Facebook status to "I love milkshakes".
Grade	Google, what's my grade for my computer vision class?
Story1	The beige hue on the waters of the large loch impressed all, including the French queen.
Story2	She sighed before she heard that pure symphony again, just as the young boy Arthur wanted.
Story3	Her eyes would wander across the open air, lost for hours.

3 METHOD

3.1 Participants

We collected data from 106 volunteers, 32 female, recruited from graduate and undergraduate computer science courses at the University of Florida[1]. Participants were aged between 18 and 32 (M = 21.6, SD = 2.63). The majority of participants were right-handed; six participants were left handed and five participants were ambidextrous. The majority (75.5%) of the participants had prior experience with motion-based interaction devices (e.g., Kinect, LeapMotion), 98.1% of participants had prior experience with spoken-language interaction devices (e.g., Amazon Alexa, Google Home, Microsoft Cortana, Apple Siri), and 66% of participants reported English as being their first language.

3.2 Procedure

Participants were compensated with 1 point of extra credit in their respective class. Sessions lasted approximately 45 to 60 minutes. To avoid fatigue effects, participants were randomly assigned to one of four groups in order to counterbalance the order of each type of task: half of the participants performed gestures first, while half repeated the phrases first. The phrase subgroups were similarly counterbalanced, with half of the participants repeating the story before the command phrases, and vice versa.

3.2.1 Gesture Tasks. Participants were asked to perform the 10 gestures 10 times in a counterbalanced order determined by using a Latin square, presented on a PowerPoint slide deck. The slides gave brief descriptions of how to perform the gesture (e.g., "Thumbs up") but did not contain images in order to capture natural variances in how participants performed the gesture. For the gesture tasks, the participant stood directly in front of a Kinect V2 placed on a desk and directly under a Kinect V2 suspended from the ceiling, facing down. To facilitate the use of RGB data to segment hands, the participant wore colored nitrile gloves and removed any watches or jewelry. Participants were instructed to stand with their arms and hands in a neutral position at their side before performing each gesture and to return to that position when they had completed the gesture. Each gesture was recorded using both Kinects simultaneously by the researcher. Participants were instructed to perform gestures consistently and were asked to repeat gestures whenever a mistake was made; recordings of erroneous gestures were removed from the database. Given the duration of the study, participants were periodically asked if they needed to take a break.

3.2.2 Phrase Tasks. Participants were asked to repeat the 10 voice commands in a counterbalanced order determined by using a Latin square, as well as to repeat the pangram story in order 10 times. As described above, half of the participants repeated the voice commands before the pangram and vice versa. As with the gesture tasks, the voice phrases were presented to the participants via a PowerPoint slide deck. Participants performed the phrase tasks while seated in front of a Blue Yeti USB Microphone. Participants were asked to repeat phrases whenever a mistake was made; recordings of erroneous phrases were removed from the database. As in the gesture tasks portion of the study, participants were periodically asked if they needed to take a break to avoid fatigue effects.

3.3 Data Collection and Extraction

Two Kinect V2 sensors were used for data collection. One sensor was placed approximately six feet in front of the participant ("bottom Kinect") and the other directly above facing downward ("top Kinect"). The sensors have frame rates of 30 fps and an image frame size of 424 × 512. Data from the top and bottom Kinects were collected in Kinect's native format, xef, which can be played back in Kinect Studio as if the Kinect was recording a live-stream. However, for our purposes, the xef data was split into the following components after it was collected: full raw video, audio, RGB video (video without audio), depth data, Kinect skeleton data, and Kinect depth skeleton data. The videos and RGB videos were recorded at a resolution of 1,920 x 1,080 pixels at 30 fps. The depth data contains the distance between the Kinect device and the objects in front of the device, in millimeters for each pixel in the frame. The Kinect skeleton data contains the x, y, z positional and rotational information for each of 25 joints in the skeleton at each timestamp, and the Kinect depth skeleton maps the Kinect skeleton to the 2D coordinates of the depth data. A total of 10,600 gestures were recorded. Due to experimenter error in operating the recording sensors, top Kinect recordings for 100 gestures (one session, P415) were discarded.

The voice data was collected in a continuous audio stream which was manually segmented at the boundaries of each utterance by the researchers after the study. A total of 13,780 utterances were recorded. However, due to experimenter error in operating the recording sensors, 130 utterances (one session, P149) were discarded. Voice data was saved in the Free Lossless Audio Codec (aka, FLAC), which is a lossless, compressed format.

[1] At our university, it is common for students to be able to earn extra credit by participating in outside-of-class human subjects research studies. While not directly related to the pedagogical goals of the class, it broadens computer science students' exposure to human-centered computing research methods and procedures.

Table 2: Characteristics of the MMGatorAuthor corpus we collected in this study. Data is available for $N = 106$ users in both the gesture and voice modality.

	Gesture	Voice
No. Files (total)	147,000	13,650
No. Files per user	1,400	130
Media Length (total) (hh:mm:ss)	12:26:03	15:02:23
Media Length (average) per user	4.22 secs	3.95 secs

4 DATASET AND FEATURE EXTRACTION

Table 2 provides details of the dataset in terms of quantity of data overall and per user. Next, we describe features extracted from each stream, gesture and voice, which are also released with the dataset.

4.1 Gesture Feature Extraction

We extract three types of gesture features established in prior work.

Trajectory Features: The trajectory feature represents the drawing path of the in-air gesture. It is obtained by tracking key-points (fingertip or palm center) of the hand region from each image frame. To obtain sufficient behavioral features, we calculate the velocity and the acceleration from the 3D coordinates of the key-point [28, 53], which leads to a 9-dimensional feature vector for each frame. The dimension of the feature vector for each sample is $N \times 9$ where N is the number of frames. The trajectory is scaled within a $1 \times 1 \times 1$ bounding box (x-, y-, and z-axis) and smoothed using a Kalman filter [23] to reduce the error caused by sensor interference and correct jagged lines caused by quickly performed in-air gestures. The difference of the raw trajectory and the smoothed trajectory is shown in Figure 2. Figure 3 presents x-axis time series of three "air signature" trajectories from two users.

Skeleton Features: The skeleton features are from the skeleton of the hand, which is represented as the center axis of the segmented hand region. The center axis is obtained by applying the Chamfer distance [27]. Figure 4 shows the center axis (center of mass) of the hand, which we segmented from the video frame, being tracked over successive frames. From this figure, we can see differences in how the same "Wave" gesture is performed between users.

Silhouette Features: We leveraged silhouette features [61] to represent hand motion. Our approach considers the whole hand object for motion feature extraction. We construct a 14-dimensional feature vector, which includes time, 3D coordinates, eight directional distances, and two time-based distances, calculated for every pixel from the segmented hand. Since the silhouette feature is considered as a "bag of features" [41], we apply the covariance metrics [57] for sparse feature selection (dimensional reduction). Similar to the approach by Wu et al. [61], we add two additional sub-tunnel silhouettes to potentially improve performance.

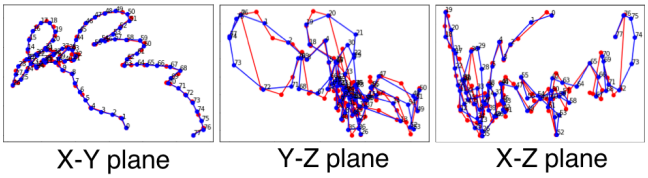

Figure 2: Example of smoothing trajectory in the x-y, y-z, and x-z planes. The red line is the raw tracked trajectory and the blue line is the smoothed trajectory.

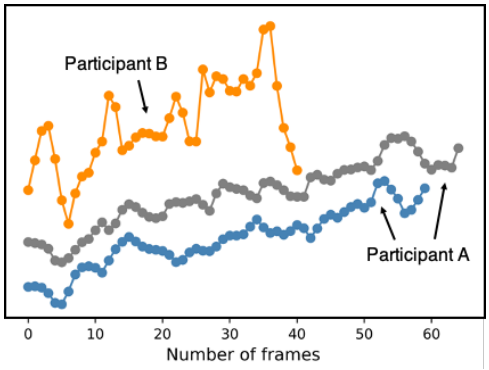

Figure 3: Examples of the trajectory feature from "air signature". The blue and gray plots that have a similar shape are trajectories from the same participant. The orange plot is a trajectory from a different participant.

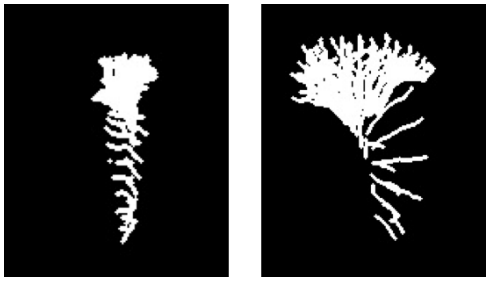

Figure 4: Gesture "Wave" from different users; the size and range of waving are different as per the skeleton features.

4.2 Voice Feature Extraction

We derive four voice-based features established in prior work.

Mel-frequency-cepstral-coefficients (MFCC): The MFCC feature Mel-scales the received signal to mimic the human hearing spectrum. A detailed explanation of MFCC was proposed by Davis et al. in the 1980s [15]. As suggested by previous research [37], we use 13 MFCC coefficients with 32 window lengths and a 16 ms window shift for speaker authentication, which leads to a 13-dimensional feature vector for each voice sample.

Linear-predictive-coding (LPC): LPC is a method to model the voice sample based on a linear combination of its past samples [16]. It is accurate in estimating voice parameters such as pitch, energy, and formant [59] using low bit rates. We use 15 LPC orders with 32 window lengths and a 16ms window shift, which lead to a 15-dimensional feature matrix for each voice sample.

Linear-predictive-cepstral-coefficients (LPCC): LPCC is derived from LPC, which also represents vocal parameters but with better performance and reliability [39] compared to LPC. We use 12

LPCC coefficients with 32 window lengths and 16 ms window shift, which lead to a 12-dimensional feature vector for each sample.

Perceptual-linear-prediction (PLP): PLP is an estimation of the auditory spectrum based on the psychophysics of the human hearing system and was first proposed by Hynek in 1989 [25]. The RASTA processing method [26] is usually applied with PLP for channel effect removal. In our dataset, we use 12th-order PLP features for feature representation.

5 CASE STUDY: USER AUTHENTICATION USING VOICE FEATURES

To demonstrate the utility of our dataset for multimodal analysis, we present baseline results of the voice-based features that we extracted for authentication. Voice-based features are widely used in authentication and have demonstrated high discriminability, allowing for use in a variety of authentication applications [36, 46]. We analyze all voice features that we extracted: Mel-frequency-cepstral-coefficients (MFCC), the linear-predictive-coding (LPC), the linear-predictive-cepstral-coefficients (LPCC), and the perceptual-linear-prediction (PLP) [15, 16, 25, 39]. The metrics we used for measuring discriminability are the intra-class similarity, inter-class similarity, equal error rate, and classification accuracy.

5.1 Experiment Methodology

We defined the interclass similarity as the similarity between features extracted from samples collected from different individuals while the intra-class similarity is defined as the similarity of features extracted from samples collected from the same individual. Ideally, for an authentication application, intra-class similarity should be high and inter-class similarity should be low, to help discriminate between different users' gestures but accept natural variance within a user's gestures.

5.1.1 Similarity Measurement. **Cross-correlation matrix:** The cross-correlation matrix is used to analyze feature similarity. We present the averaged cross-correlation score (CCS) from the cross-correlation matrix for similarity measurement. The intraclass similarity is obtained from sample-to-sample correlation by the same participant performing the same voice command, while the inter-class similarity is obtained from user-to-user correlation when performing the same voice command. The cross-correlation score is between 0 to 1. In general, two features have good correlation if the cross-correlation score is higher than 0.75 [10].

5.1.2 Matching and Evaluation. **Equal error rate (EER):** False acceptances and false rejections are two types of errors that occur in a biometric classification system. The false acceptance rate (FAR) indicates the likelihood of a system incorrectly accepting an unauthorized attempt, and the false rejection rate (FRR) indicates the likelihood of a system incorrectly rejecting an authorized attempt. The EER is defined as the value when the FAR equals the FRR. Typically, an accurate biometric system will have a low EER value. In this work, we calculate EER for each user based on dynamic time warping (DTW) score, which is a well-known algorithm for measuring the similarity between two temporal sequences (e.g., time series data with different lengths) [4].

Classification accuracy: The accuracy is obtained by a trained Gaussian mixture model (GMM). GMM is a statistical model that is commonly used in voice/speaker recognition research and applications [64], since it can efficiently and accurately represent feature distributions for a biometric system.

5.2 Experiment Results

We present the cross-correlation score (CCS) to represent the intra-class and the inter-class similarity in Table 3. This analysis explores if the features possess the discriminability required for use in authentication applications. The EER and the classification accuracy is presented to evaluate the performance of an authentication system based on voice features. The EER is calculated based on the DTW scores that are shown in Table 4. The classification accuracy is obtained by averaging the accuracy from 10,600 speaker recognition experiments for the voice features and results are shown in Table 5.

Observations from the results are as follows:

1) According to the correlation score from Table 3, the highest intra-class similarity of voice features is 0.930. The lowest inter-class similarity of the voice features is 0.085.

2) Voice features achieve an EER of 11.90% at worst using LPCC features. The best performing voice feature is for the phrase "Story3" using MFCC features.

This case study shows that our dataset can provide similar performance on speaker recognition with voice features as has been demonstrated in previous literature [22, 56], thus validating the quality of our dataset. The general utility of our dataset to other researchers beyond what has been shown in prior work is the additional modality of gesture interaction, and the ability to explore both voice and gesture modalities together, in authentication applications and beyond.

6 RELATED WORK

We motivate the need for MMGatorAuth based on prior work.

6.1 Need for Gesture and Voice Datasets

There is a long history, especially in the Multimodal Interaction research community, of collecting and releasing robust datasets that include both unimodal and multimodal user behaviors in various contexts. For example, a search of the ACM Digital Library for "multimodal dataset" published within the last five years at ICMI yielded 32 results. Many of these datasets aim to solve problems in passive affective computing by providing data streams from physiological sensors like EEG, GSR, and heart rate [5, 32, 47], RGB camera streams for facial and emotion recognition [3, 34, 40], or combinations of these [30]. Not many datasets of active user interaction episodes in modalities like gesture and voice/speech are publicly available [12]. Those that do typically focus on human-human communication episodes [8, 17, 54], even though the state-of-the-art in multimodal interaction is far from allowing users to interact with a system using the same communication strategies they would use with other people [55]. Furthermore, to the best of our knowledge, no multimodal datasets with voice and gesture exist specifically for the context of authentication. Multimodal authentication datasets have focused on wearable sensors or other biometric modalities [9, 29, 31]. The MMGatorAuth dataset we

Table 3: Averaged cross correlation score of intra-class and inter-class similarity for voice features.

	MFCC		LPC		LPCC		PLP	
Phrases	Intra-class	Inter-class	Intra-class	Inter-class	Intra-class	Inter-class	Intra-class	Inter-class
Facebook	**0.930**	0.112	**0.953**	0.115	0.844	0.112	0.871	0.115
Message	0.929	0.110	0.925	0.135	0.870	0.153	0.879	0.135
Story2	0.924	0.103	0.912	0.117	0.839	0.116	0.867	0.112
Mail	0.921	0.100	0.877	0.114	0.842	0.192	0.877	0.114
Chair	0.920	0.098	0.862	0.133	0.846	0.172	0.870	0.133
Photos	0.918	0.095	0.866	0.102	0.828	0.139	0.878	0.122
Story1	0.916	0.093	0.868	0.131	0.805	**0.109**	0.872	**0.105**
Password	0.916	0.103	0.899	0.111	0.866	0.150	0.875	0.130
Story3	0.911	0.088	0.891	0.119	0.802	0.249	0.249	0.119
Grade	0.911	0.099	0.879	0.102	0.857	0.121	0.889	0.132
Credit	0.909	0.909	0.884	0.134	**0.877**	0.127	0.883	0.109
Doctor	0.907	**0.085**	0.864	**0.094**	0.849	0.119	**0.888**	0.123

Table 4: Averaged equal error rate score for voice features.

Phrases	MFCC	LPC	LPCC	PLP
Message	1.64%±1.05%	8.14%±1.60%	8.45%±1.47%	9.24%±1.25%
Bank	1.92%±1.11%	7.76%±1.83%	7.22%±1.37%	9.23%±1.24%
Credit	2.61%±1.34%	7.55%±1.23%	7.76%±1.77%	7.72%±1.34%
Chairs	1.60%±1.06%	**5.46%±1.58%**	**6.75%±1.78%**	7.56%±1.17%
Mail	1.55%±0.73%	7.72%±0.60%	7.07%±1.72%	7.54%±1.08%
Doctor	1.89%±1.03%	4.99%±1.32%	8.04%±1.72%	**6.44%±1.42%**
Facebook	0.87%±0.57%	6.99%±1.69%	7.12%±1.54%	685%±1.25%
Photos	1.55%±1.04%	838%±1.73%	9.0%±1.21%	9.56%±1.28%
Story1	1.41%±0.70%	9.80%±1.34%	11.90%±2.23%	9.42%±1.89%
Story2	0.97%±0.73%	6.21%±1.63%	7.70%±1.52%	9.13%±1.84%
Password	0.95%±1.10%	7.17%±1.59%	7.99%±1.23%	7.77%±1.01%
Grade	0.89%±0.77%	8.01%±1.07%	9.05%±1.97%	8.00%±1.33%
Story3	**0.84%±0.69%**	8.11%±1.34%	10.40%±1.48%	8.61%±1.92%

Table 5: Averaged classification accuracy for voice features.

Phrases	MFCC	LPC	LPCC	PLP
Story1	**100%**	**100%**	**100%**	**100%**
Story3	**100%**	**100%**	**100%**	**100%**
Doctor	**100%**	99.5%	**100%**	98.8%
Message	**100%**	99.0%	98.6%	99.0%
Story2	**100%**	99.0%	99.02%	99.0%
Facebook	**100%**	98.6%	98.8%	98.4%
Mail	**100%**	97.6%	97.4%	98.0%
Photos	**100%**	96.9%	96.7%	96.9%
Password	99.8%	97.9%	98.1%	97.9%
Bank	99.5%	99.05%	99.1%	99.05%
Grade	99.2%	98.6%	98.8%	98.8%
Credit	99.1%	97.6%	97.9%	97.3%
Chair	99.0%	94.3%	94.2%	94.8%

provide fills an important gap in the literature to enable research into synchronous or asynchronous multimodal authentication with smart environments in gesture and speech.

6.2 Voice-Based Authentication

Many advancements in the area of speaker recognition have been made in the past two decades, demonstrating the effectiveness of the technology [20]. This has been continuously demonstrated in the National Institute of Standard and Technology (NIST) Speaker Recognition Evaluations (SRE) [38], which have been performed yearly since 1996. Various approaches to speaker recognition have been proposed. Low level features such has Mel-frequency cepstral coefficients (MFCC), cepstral mean subtraction, MFCC, and CMS derivatives, as well as pitch/energy averages, have been used to represent differences in an individual's speech during short utterances [19]. In addition to these low-level features, high-level features (linguistic measurements) such as word usage and pronunciation have also been used in speaker recognition but are limited to scenarios when recordings are long [50]. Researchers have recently explored the application of powerful deep learning-based methods to the problem of speaker recognition, opening new possibilities [43].

6.3 Gesture-Based Authentication

Previous research on hand gesture recognition is mainly focused on the classification of different gestures [62]. Relatively little research investigates gesture-based user authentication. Fong et al. [18] and Gupta et al. [21] used an RGB camera to perform authentication with American Sign Language; however, their methods are limited to off-line authentication within a restricted recording environment. Tian et al. [53] and Aumi and Kratz [1] both implemented methods to track fingertips to match gesture trajectories. Wu et al. [61] extracted hand silhouette features for the same purpose. However, the gestures used in Tian et al. [53] are very complex Chinese characters and Wu et al. [61] restricted gestures to those captured within a small distance from the sensor. These restricted in-air gesture commands may not be suitable for a real-time authentication system. Also, all these researchers used relatively small datasets recorded from fewer than 20 participants. Other previous work on stroke-based gesture recognition for authentication has collected more robust datasets from 50+ participants (e.g., [33, 48]), but those datasets cannot transfer to hand gestures like the ones we provide in MMGatorAuth.

7 LIMITATIONS AND FUTURE WORK

Our MMGatorAuth dataset fills an important gap in the literature due to the current shortage of robust yet controlled multimodal interaction datasets for smart environment applications. However, the dataset does have some limitations. First, the population sample in this dataset includes only computer science graduate and undergraduate students, who may not be representative of all users. Future work could obtain more samples from a broader segment of the population, including older adults, persons with disabilities, or children. We encourage other researchers to help expand this dataset by contributing samples collected from other populations. The gesture and speech commands chosen for this dataset also may be biased in favor of Western, English-speaking user groups. Future work could collect data samples of different commands more suitable for non-Western audiences. The commands in this dataset were all scripted and given to the users to perform, in order to maximize the utility of this dataset for systematic recognition testing. Future work could investigate user-defined multimodal authentication commands to understand user preferences and personalized

adaptation in the target domain. Finally, we only present a voice-feature authentication case study; future work could compare the results achieved with our dataset to others reported in prior work on authentication with Kinect-based hand gesture features, such as Tian et al. [53] or Wu et al. [61], or multimodal features.

8 CONCLUSION

This paper presents a novel, large multimodal gesture dataset collected from 106 volunteers of 10,600 gesture and 13,780 speech commands. The original intended application for the dataset is authentication in smart home environments, but the dataset might be useful to any researchers who need a benchmark to evaluate their gesture and/or voice recognition algorithms, two modalities which have been studied less frequently in the literature recently. We present the dataset collection method, dataset and feature characteristics, and a case study illustrating how the dataset could be useful to researchers. A download link for the full dataset is available here: *https://init.cise.ufl.edu/downloads/*.

ACKNOWLEDGMENTS

The authors thank Discover Financial Services, who partly funded this work. We also thank the participants who provided data and the instructors who offered extra credit in their courses.

REFERENCES

[1] Md Tanvir Islam Aumi and Sven Kratz. 2014. AirAuth: evaluating in-air hand gestures for authentication. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices & Services (MobileHCI)*. Association for Computing Machinery, New York, NY, USA, 309–318. https://doi.org/10.1145/2628363.2628388

[2] Steve Ballmer. 2010. CES 2010: A transforming trend–the natural user interface. http://www.huffingtonpost.com/steve-ballmer/ces-2010-a-transforming-t_b_416598.html

[3] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, 464–472. https://doi.org/10.1145/3136755.3136814

[4] Donald J. Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDDM)*. AAAI Press, 359–370.

[5] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, and Raffaella Lanzarotti. 2017. AMHUSE: a multimodal dataset for HUmour SEnsing. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, 438–445. https://doi.org/10.1145/3136755.3136806

[6] Damien Brain and Geoffrey I. Webb. 1999. On the effect of data set size on bias and variance in classification learning. In *Proceedings of the Australian Knowledge Acquistion Workshop (AKAW)*. 7–128.

[7] CBS Interactive Inc. 2017. 6-year-old Brooke Neitzel orders expensive dollhouse, cookies using Amazon's voice-activated Echo Dot. https://www.cbsnews.com/news/tv-news-anchors-report-accidentally-sets-off-viewers-amazons-echo-dots/

[8] Lang Che and Liqin Zha. 2019. A synergy study of metaphoric gestures on rhetorical behavior construction: based on the corpus of "AI"-themed public speeches. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, Article 3, 6 pages. https://doi.org/10.1145/3357160.3357669

[9] Girija Chetty and Michael Wagner. 2006. Audio-visual multimodal fusion for biometric person authentication and liveness verification. In *Proceedings of the 2005 NICTA-HCSNet Multimodal User Interaction Workshop - Volume 57* (Sydney, Australia) *(MMUI '05)*. Australian Computer Society, Inc., AUS, 17–24.

[10] Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6, 4 (1994), 284–290.

[11] Clagnut. 2014. List of pangrams. http://clagnut.com/blog/2380/

[12] Gradeigh D. Clark and Janne Lindqvist. 2015. Engineering gesture-based authentication systems. *IEEE Pervasive Computing* 14, 1 (2015), 18–25.

[13] Ry Crist. 2017. Amazon's Alexa can now recognize your voice. https://www.cnet.com/news/amazons-alexa-can-now-recognize-your-voice/

[14] Andrew Crossan, John Williamson, Stephen Brewster, and Rod Murray-Smith. 2008. Wrist rotation for interaction in mobile contexts. In *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*. ACM Press, New York, New York, USA, 435–438. https://doi.org/10.1145/1409240.1409307

[15] Steven B. Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 4 (1980), 357–366.

[16] Li Deng and Doug O'Shaughnessy. 2003. *Speech processing: a dynamic and optimization-oriented approach*. CRC Press.

[17] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopes, Isabelle Guyon, Vassilis Athitsos, and Hugo Escalante. 2013. Multi-modal gesture recognition challenge 2013: dataset and results. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, 445–452. https://doi.org/10.1145/2522848.2532595

[18] Simon Fong, Yan Zhuang, Iztok Fister, and Iztok Fister Jr. 2013. A biometric authentication model using hand gesture images. *Biomedical Engineering Online* 12, 111 (2013).

[19] Mark Gales and Steve Young. 2008. The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing* 1, 3 (2008), 195–304. https://doi.org/10.1561/2000000004

[20] Craig S. Greenberg, Lisa P. Mason, Seyed Omid Sadjadi, and Douglas A. Reynolds. 2020. Two decades of speaker recognition evaluation at the National Institute of Standards and Technology. *Computer Speech Language* 60 (2020), 101032. https://doi.org/10.1016/j.csl.2019.101032

[21] Anand Gupta, Ashima Arora, and Bhawna Juneja. 2013. Tag: A two-level framework for user authentication through hand gestures. In *International Conference on Contemporary Computing*. 503–509.

[22] Kartiki Gupta and Divya Gupta. 2016. An analysis on LPC, RASTA and MFCC techniques in automatic speech recognition system. In *International Conference on Cloud System and Big Data Engineering (Confluence)*. 493–497.

[23] Andrew C. Harvey. 1990. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

[24] Eiji Hayashi, Manuel Maas, and Jason I. Hong. 2014. Wave to me: user identification using body lengths and natural gestures. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM Press, New York, New York, USA, 3453–3462. https://doi.org/10.1145/2556288.2557043

[25] Hynek Hermansky. 1990. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America* 87, 4 (1990), 1738–1752.

[26] Hynek Hermansky and N. Morgan. 1994. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing* 2, 4 (1994), 578–589.

[27] Bogdan Ionescu, Didier Coquin, Patrick Lambert, and Vasile Buzuloiu. 2005. Dynamic hand gesture recognition using the skeleton of the hand. *EURASIP Journal on Advances in Signal Processing* (2005), Article No. 236190.

[28] Je-Hyoung Jeon, Beom-Seok Oh, and Kar-Ann Toh. 2012. A system for hand gesture based signature recognition. In *Proceedings of International Conference on Control Automation Robotics & Vision (ICARCV)*. 171–175.

[29] Mohamed Khamis, Mariam Hassib, Emanuel von Zezschwitz, Andreas Bulling, and Florian Alt. 2017. GazeTouchPIN: protecting sensitive data on mobile devices using secure multimodal authentication. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, 446–450. https://doi.org/10.1145/3136755.3136809

[30] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A. Neerincx, and Wessel Kraaij. 2014. The SWELL knowledge work dataset for stress and user modeling research. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, 291–298. https://doi.org/10.1145/2663204.2663257

[31] Vrishab Krishna, Yi Ding, Aiwen Xu, and Tobias Höllerer. 2019. Multimodal biometric authentication for VR/AR using EEG and eye tracking. In *Adjunct Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, Article 6, 5 pages. https://doi.org/10.1145/3351529.3360655

[32] Iulia Lefter and Siska Fitrianie. 2018. The multimodal dataset of negative affect and aggression: a validation study. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, 376–383. https://doi.org/10.1145/3242969.3243013

[33] Can Liu, Gradeigh D. Clark, and Janne Lindqvist. 2017. Guessing attacks on user-generated gesture passwords. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 1, Article 3 (March 2017), 24 pages. https://doi.org/10.1145/3053331

[34] Kaixin Ma, Xinyu Wang, Xinru Yang, Mingtong Zhang, Jeffrey M. Girard, and Louis-Philippe Morency. 2019. ElderReact: a multimodal dataset for recognizing emotional response in aging adults. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, 349–357. https://doi.org/10.1145/3340555.3353747

[35] Sapna Maheshwari. 2017. Burger King "O.K. Google" ad doesn't seem O.K. with Google. https://www.nytimes.com/2017/04/12/business/burger-king-tv-ad-google-home.htm

[36] Mitchell McLaren, Yun Lei, and Luciana Ferrer. 2015. Advances in deep neural network approaches to speaker recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4814–4818.

[37] Lindasalwa Muda, Mumtaj Begam, and I. Elamvazuthi. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) Techniques. *Journal of Computing* 2, 3 (2010), 138–143.

[38] National Institute of Standards and Technology (NIST). 1996-2020. Speaker Recognition Evaluation (SRE). https://www.nist.gov/itl/iad/mig/speaker-recognition

[39] Navnath S. Nehe and Raghunath S. Holambe. 2012. DWT and LPC based feature extraction methods for isolated word recognition. *EURASIP Journal on Audio, Speech, and Music Processing* (2012), Article No.7.

[40] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E. Hughes, and Louis-Philippe Morency. 2016. EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, 137–144. https://doi.org/10.1145/2993148.2993168

[41] Eric Nowak, Frédéric Jurie, and Bill Triggs. 2006. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision, Lecture Notes in Computer Science, vol 3954*, A. Leonardis, H. Bischof, and A. Pinz (Eds.). Springer Berlin Heidelberg, Chapter European C, 490–503.

[42] Sharon Oviatt and Philip R. Cohen. 2000. Multimodal interfaces that process what comes naturally. *Commun. ACM* 43, 3 (2000), 45–53.

[43] Fred Richardson, Douglas Reynolds, and Najim Dehak. 2015. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters* 22, 10 (2015), 1671–1675.

[44] Julie Rico and Stephen Brewster. 2010. Usable gestures for mobile interfaces: evaluating social acceptability. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 887–896.

[45] Jaime Ruiz and Yang Li. 2010. DoubleFlip: a motion gesture delimiter for interaction. In *Adjunct Proceedings of the Annual ACM Symposium on User Interface Software and Technology (UIST)*. Association for Computing Machinery, New York, NY, USA, 449–450. https://doi.org/10.1145/1866218.1866265

[46] Zia Saquib, Nirmala Salam, Rekha P. Nair, Nipun Pandey, and Akanksha Joshi. 2010. A survey on automatic speaker recognition systems. In *Signal Processing and Multimedia*, T. Kim, S. K. Pal, W. I. Grosky, N. Pissinou, T. K. Shih, and D. Ślęzak (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 134–145.

[47] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. Association for Computing Machinery, New York, NY, USA, 400–408. https://doi.org/10.1145/3242969.3242985

[48] Michael Sherman, Gradeigh Clark, Yulong Yang, Shridatt Sugrim, Arttu Modig, Janne Lindqvist, Antti Oulasvirta, and Teemu Roos. 2014. User-Generated Free-Form Gestures for Authentication: Security and Memorability. In *Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services (MobileSys)*. Association for Computing Machinery, New York, NY, USA, 176–189. https://doi.org/10.1145/2594368.2594375

[49] Babins Shrestha, Nitesh Saxena, and Justin Harrison. 2013. Wave-to-access: protecting sensitive mobile device services via a hand waving gesture. In *International Conference on Cryptology and Network Security, Lecture Notes in Computer Science, vol 8257*. Springer-Verlag New York, Inc., Chapter Proceeding, 199–217. https://doi.org/10.1007/978-3-319-02937-5_11

[50] Elizabeth Shriberg and Andreas Stolcke. 2008. The case for automatic higher-level features in forensic speaker recognition. *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 1509–1512.

[51] Steven Strachan, Roderick Murray-Smith, and Sile O'Modhrain. 2007. BodySpace: inferring body pose for natural control of a music player. In *Extended Abstracts of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM Press, New York, New York, USA, 2001–2006. https://doi.org/10.1145/1240866.1240939

[52] Akiji Takeuchi, Yusuke Manabe, and Kenji Sugawara. 2013. Multimodal soft biometrie verification by hand shape and handwriting motion in the air. In *Proceedings of the International Joint Conference on Awareness Science and Technology & Ubi-Media Computing (iCAST & UMEDIA)*. IEEE, 103–109. https://doi.org/10.1109/ICAwST.2013.6765417

[53] Jing Tian, Chengzhang Qu, Wenyuan Xu, and Song Wang. 2013. KinWrite: handwriting-based authentication using Kinect. In *Network & Distributed System Security Symposium*. 18 pp.

[54] Isaac Wang, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, J. Ross Beveridge, Bruce A. Draper, and Jaime Ruiz. 2017. EGGNOG: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 414–421.

[55] Isaac Wang, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, Bruce Draper, Ross Beveridge, and Jaime Ruiz. 2017. Exploring the use of gesture in collaborative tasks. In *Extended Abstracts of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Association for Computing Machinery, New York, NY, USA, 2990–2997. https://doi.org/10.1145/3027063.3053239

[56] Jia-Ching Wang, Chien-Yao Wang, Yu-Hao Chin, Yu-Ting Liu, En-Ting Chen, and Pao-Chi Chang. 2017. Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition. *Multimedia Tools and Applications* 76, 3 (2017), 4055–4068. https://doi.org/10.1007/s11042-016-3335-0

[57] Larry Wasserman. 2013. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

[58] Mark Weiser. 1991. The computer for the 21st century. *Scientific American* 265, 3 (1991), 94–104.

[59] Wikipedia Contributors. 2004. Linear predictive coding — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Linear_predictive_coding

[60] John Williamson, Roderick Murray-Smith, and Stephen Hughes. 2007. Shoogle: excitatory multimodal interaction on mobile devices. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM Press, New York, New York, USA, 121–124. https://doi.org/10.1145/1240624.1240642

[61] Jonathan Wu, James Christianson, Janusz Konrad, and Prakash Ishwar. 2015. Leveraging shape and depth in user authentication from in-air hand gestures. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*. 3195–3199.

[62] Ying Wu and Thomas S. Huang. 1999. Vision-based gesture recognition: A review. In *Gesture-Based Communication in Human-Computer Interaction, Lecture Notes in Computer Science, vol 1739*, A. Braffort, R. Gherbi, S. Gibet, D. Teil, and J. Richardson (Eds.). Springer Berlin Heidelberg, 103–115.

[63] Lei Yang, Yi Guo, Xuan Ding, Jinsong Han, Yunhao Liu, Cheng Wang, and Changwei Hu. 2015. Unlocking smart phone through handwaving biometrics. *IEEE Transactions on Mobile Computing* 14, 5 (May 2015), 1044–1055. https://doi.org/10.1109/TMC.2014.2341633

[64] Chang Huai You, Kong Aik Lee, and Haizhou Li. 2008. An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition. *IEEE Signal Processing Letters* 16, 1 (2008), 49–52.