IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 26, 2021, accepted October 27, 2021, date of publication November 4, 2021,
date of current version November 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3125485

Multilingual Audio-Visual Smartphone Dataset and Evaluation

HAREESH MANDALAPU[1], P. N. ARAVINDA REDDY[2], RAGHAVENDRA RAMACHANDRA[1], (Senior Member, IEEE), KROTHAPALLI SREENIVASA RAO[3], (Member, IEEE), PABITRA MITRA[3], (Member, IEEE), S. R. MAHADEVA PRASANNA[4], (Member, IEEE), AND CHRISTOPH BUSCH[1], (Senior Member, IEEE)

[1]Department of Information Security and Communication Technology, Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway
[2]Advanced Technology Development Centre, IIT Kharagpur, Kharagpur, West Bengal 721302, India
[3]Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, West Bengal 721302, India
[4]Department of Electrical Engineering, IIT Dharwad, Dharwad, Karnataka 580011, India

Corresponding author: Hareesh Mandalapu (hareesh.mandalapu@ntnu.no)

This work was supported by Department of Information Security and Communication Technology, NTNU, Gjøvik and research Council of Norway under Grant IKTPLUSS 248030/O70 and advanced Technology Development Centre, Indian Institute of Technology, Kharagpur, India.

ABSTRACT Smartphones have been employed with biometric-based verification systems to provide security in highly sensitive applications. Audio-visual biometrics are getting popular due to their usability, and also it will be challenging to spoof because of their multimodal nature. In this work, we present an audio-visual smartphone dataset captured in five different recent smartphones. This new dataset contains 103 subjects captured in three different sessions considering the different real-world scenarios. Three different languages are acquired in this dataset to include the problem of language dependency of the speaker recognition systems. These unique characteristics of this dataset will pave the way to implement novel state-of-the-art unimodal or audio-visual speaker recognition systems. We also report the performance of the bench-marked biometric verification systems on our dataset. The robustness of biometric algorithms is evaluated towards multiple dependencies like signal noise, device, language and presentation attacks like replay and synthesized signals with extensive experiments. The obtained results raised many concerns about the generalization properties of state-of-the-art biometrics methods in smartphones.

INDEX TERMS Smartphone biometrics, audio-visual speaker recognition, presentation attack detection, multilingual.

I. INTRODUCTION

With the advances in biometrics, the usage of passwords and smart cards to gain access into several control applications have been slowly depreciated. Henceforth for reliable and secure access control, biometrics have been deployed in various applications, including smartphone unlocking, banking transactions, financial services, border control, etc. The biometrics in access control applications improve trustworthiness and enhance user proficiency by verifying who they are. A biometric system aims to recognize the person based on their physiological or behavioural characteristics based on ISO/IEC 2382-37. The physiological characteristics include the face, iris, fingerprint etc., and behavioural characteristics include speech, keystroke, gait etc.

Smartphone biometrics has grown expeditiously over the years. The number of smartphone users crossed 3 billion in 2020 and is expected to increase in millions in the coming years. According to the Mercator Advisory Group report, 66% of smartphone users are expected to use biometrics for authentication by the end of 2024. In 2020, 41% of smartphone users used biometrics which was 27% in 2019. Among different biometric modalities, fingerprint-based authentication is at the top. However, the amount of users for face and biometrics has been increasing. Voice-based recognition increased to 20% in 2020, from 11% in 2019 and face recognition jumped to 30% in 2020, from 20% in 2019. The application of smartphone biometrics has been widely used in mobile banking, e-commerce, remote identification etc.

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

Different types of smartphones like Android, iPhone and blackberry provide uni-modal applications based on either fingerprint, iris or face recognition, and recently speech has been added as a biometric cue for authentication purposes. The built-in biometrics are not fixed for all smartphones. For example, some smartphones come with fingerprint, and some include face recognition. The captured uni-modal biometrics like face or iris comes with several problems like low quality, variations in pose, problem with illuminations, background noise, low spatial and temporal resolutions of video [18]. Therefore, this problem is addressed in multimodal biometrics by taking advantage of default sensors like cameras and microphones. Multimodal systems like audio-visual biometrics utilize the complementary information of face and speech and exploit the user-friendly capture of face and voice in a single recording. Audio-visual biometric data capture is cost-effective and can be carried out without additional sensors (e.g., fingerprint reader or iris camera).

The applications based on biometrics in smartphones has several advantages but also exist several challenges. The key challenges are the robustness and generalizability of a biometric system caused by algorithm dependencies and evolving presentation attacks. The aforementioned challenges are the main problems that circumscribe reliable and secure smartphone-based applications. The first challenge is the algorithm dependencies which limits the interoperability of a biometric algorithm across multiple types of smartphones. Interoperability is defined as the ability of a biometric system to handle variations introduced in the biometric data due to different capture devices. Due to different kinds of smartphone sensors, capturing conditions and human behaviour. The dependency of the biometric algorithm on particular data properties limits the robustness of optimal recognition. Therefore, it is very challenging to develop a conventional biometric method for a wide variety of smartphones.

The second challenge is from the presentation attacks or also called spoofing attacks and indirect attacks, which are comprehensively explained in [29] for face and in [18] for audio-visual. Presentation attacks are defined as the presentation to a biometric capture subsystem with the goal of interfering with the operation of the biometric system [12]. Presentation attacks have become easy to create and use as a concealer or impostor towards the target subject. Growing presentation attacks and limitations in smartphone sensors cause major problems questioning the performance of smartphone biometrics.

The factors above motivated research on the study of smartphone biometrics towards the key challenges. In this direction, to examine the challenges, we need a smartphone biometrics database with different attributes. There are few biometric databases have been created using smartphones in both uni-modal [31] and multimodal biometrics [19], [30]. However, the existing databases are limited with several devices, languages and sessions. Therefore, we have created a multilingual audio-visual smartphone (MAVS) dataset considering smartphone devices, sessions, speech languages and presentation attacks. The novel dataset contains audio-visual biometric data of 103 subjects (70 male, 33 female) captured in three sessions with variable noise and illumination. Each subject utters six sentences, each in three different languages and recorded in five different smartphones. We have also created two types of presentation attacks in both audio, video and audio-visual scenarios. The first type of attack is a physical access attack which is created by replaying an audio-visual sample on a display-speaker setup and recorded using a smartphone. The second attack is a synthesized attack where audio and video are created separately via speech synthesis and face-swapping.

Further, we have benchmarked the dataset by performing extensive experiments in two directions. The first direction is to observe the biometric algorithm dependencies concerning device, illumination, background noise and language. The second direction is to examine the vulnerability towards presentation attacks. The baseline presentation attack detection methods in both audio and visual domains are included in this work. The biometric recognition algorithms are chosen from the state-of-the-art methods from the literature. The experimental results are presented in ISO/IEC biometric standards [11] with pictorial representations and detailed discussion.

The rest of the paper is organized as follows. Section II presents the related work in audio-visual datasets with sample images and discussion of results. The detailed description of the multilingual audio-visual smartphone (MAVS) dataset created in this research is presented in Section III. Section IV describes the performance evaluation protocols used in bench-marking the MAVS dataset. Section V presents the experiments performed and results obtained and Section VI concludes this paper with discussion on the future work.

II. RELATED WORK

The sensitivity of data in smartphone utilization has made the usage of biometrics a critical feature. Therefore, the research in smartphone biometrics has obtained much attention in recent years. The built-in biometric sensors provide the necessary authentication for many smartphones. However, the inconsistency of performance in these devices encouraged a new direction of biometric recognition using the default sensors like camera and microphone. In this direction, few audio-visual smartphone biometric datasets have been developed by capturing talking subjects' videos. Multimodal biometric databases captured modalities like a finger photo, face, iris photo, and speech data. However, considering the standard sensors in all smartphones, we studied only audio-visual databases, including face and voice. In this section, we present a comprehensive study on audio-visual biometric databases. A detailed study on all audio-visual biometric databases is performed in [18] by Mandalapu *et al.* along with a comparison of best-performing algorithms. In this section, we present some audio-visual databases in detail.

Early audio-visual biometric datasets are created by the advanced multimedia processing (AMP) lab of Carnegie Melon University (CMU).[1] With ten subjects, each speaking 78 isolated words, the recording is taken by a digital camcorder with a tie-clip microphone [42]. The dataset is made publicly available with sound files and lip parameters. Although the number of subjects is low, this dataset assisted in developing a visual shape-based feature vector for audio-visual speaker recognition in [1]. Biometrics Access Control for Networked and E-Commerce Applications (BANCA)[2] [2] is developed for E-Commerce applications. Important features in this database are multiple European languages captured using both high and low-quality devices under three different scenarios: controlled, degraded, and adverse. Also, the total number of subjects was 208, with an equal number of men and women. Figure 1 shows the sample images of this database from three different scenarios.

FIGURE 1. Example BANCA database images Up: Controlled, Middle: Degraded and Down: Adverse scenarios [2].

The goal of multimodal biometrics is to improve the robustness of the recognition/verification process. The VALID database was created in a realistic audio-visual noisy office room under uncontrolled lighting and acoustic noise. The VALID database is publicly available to research purposes.[3] The MultiModal Verification for Teleservices and Security (M2VTS) applications database has been developed for granting access to secure regions using audio-visual person verification [27]. An extension to the M2VTS database is XM2VTS (extended M2VTS) with focus on high-quality biometric samples [20]. It contains high-quality face images, 32 kHz 16-bit audio files, video sequences, and a 3D Model. The database is publicly available at cost price.[4]

Video recordings of people reading sentences from Texas Instruments and Massachusetts Institute of Technology (TIMIT) corpus (VidTIMID)[5] is a publicly available dataset presented in [36]. A distinctive part of VidTIMIT dataset is that it also contains head rotation sequence for each person in each session [35]. BioSecure[6] is a popular multimodal database that also comprises of audio-visual dataset [25]. The database consists of data from 600 subjects recorded in three different scenarios. The sample images from the database are shown in Figure 3.

FIGURE 2. Front profile shots of a subject from four sessions of XM2VTS database [20].

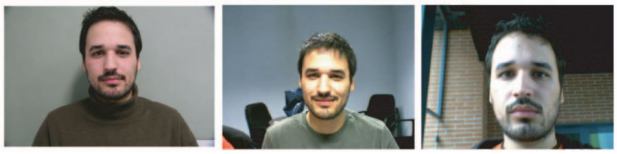

FIGURE 3. Face samples acquired in BioSecure database in three different scenarios. Left: indoor digital camera (from DS2), Middle: Webcam (from DS2), and Right: outdoor Webcam (from DS3) [25].

FIGURE 4. Talking face samples from SWAN database one frame from each session [30].

The aforementioned audio-visual datasets are captured with different types of sensors. In some cases, the audio and video capturing sensors are two different devices, and the data is presented separately. However, in smartphones, the built-in camera and microphone can be used to create audio-visual data. The MOBIO database[7] [19] is an audio-visual data created using a mobile phone (NOKIA N93i) and a laptop computer (2008 MacBook). MOBIO dataset helped in the study of person identification in a mobile phone environment [22]. In a similar fashion, the MobBIO database is developed by Sequeira *et al.* in [38]. The sensors used in this work are the rear camera of the Asus Transformer Pad TF 300T.

[1]The AMP/CMU dataset: http://amp.ece.cmu.edu/
[2]The BANCA database: http://www.ee.surrey.ac.uk/CVSSP/banca/
[3]The VALID database: http://ee.ucd.ie/validdb/
[4]The XM2VTS database: http://www. ee.surrey.ac.uk/CVSSP/xm2vtsdb/
[5]The VidTIMTI dataset: http://conradsanderson.id.au/vidtimit/
[6]BioSecure: https://biosecure.wp.tem-tsp.eu/biosecure-database/
[7]The MOBIO database: https://www.idiap.ch/dataset/mobio

TABLE 1. Details of audio-visual biometric verification databases.

Dataset	Year	Devices	No. of subjects	Biometric	Availability
AMP/CMU [42]	2001	Digital Camcorder, tie-clip microphone	10 (7 M, 3 F)	Face, voice	Free
BANCA [2]	2003	Webcam and Digital Camera	208 (104 M, 104 F)	Face, voice	Free
VALID [8]	2005	Canon 3CCD XM1 PAL	106 (77 M, 29 F)	Face, voice	Free
M2VTS [27]	2005	Hi8 camera, D1 digital recorder	37	Face, voice	Free
XM2VTS [20]	2005	Sony VX1000E, DHR1000UX	295	Face, voice	Free
VidTIMIT [36]	2009	Digital video camera	43 (24 M, 19 F)	Face, voice	Free
BioSecure [25]	2010	Samsung Q1, Philips SP900NC HP iPAQ hx2790 Webcam, PDA	DS1: 971 DS2: 667 DS3: 713	Face, Fingerprint Voice, Signature	Paid
MOBIO [19]	2012	Nokia N93i Mac-book	152	Voice, Face periocular	Free
MobBIO [38]	2014	Asus Transformer Pad TF 300T	105	-	-
Hu *et al.* [9]	2015	-	11	Audio-Visual	Free
SWAN database [30]	2019	iPhone 6 iPad Pro	88	Face, Periocular, Multilingual Voice Presentation Attack dataset	Free
MAVS dataset	2021	iPhone 6, iPhone 10, iPhone 11 Samsung S7 and Samsung S8	103 (70 M, 33 F)	Face, Multilingual Voice Presentation Attack dataset	Free

The Smartphone Multimodal Biometric database was collected for the application of mobile banking [30]. The real-world scenarios are attributed in this database with multiple sessions and languages using iPhone 6s and iPad Pro. Along with audio-visual data, the SWAN database also contains face, eye region, finger photo and voice data. Presentation attacks are also provided as a part of this database. Figure 4 shows the sample images of subjects from six sessions.

The existing databases on audio-visual biometrics provide limited variance in addressing the problem of robustness—most databases on session variance but not on device variance and language dependency. Alongside, presentation attacks are growing widely and displaying a huge impact on the optimal performance of biometric algorithms. We have formulated advanced protocols to create a multilingual audio-visual smartphone (MAVS) database considering all these problems. In this direction, the significant contributions of this paper are mentioned as follows.

1) A novel multilingual audio-visual smartphone dataset will be made available for research purposes. The uniqueness of this dataset is described below.
   - Biometric data from 70 male and 33 female subjects from various backgrounds.
   - Three language speeches and three sessions (variable illumination and background noise) for all the subjects.
   - Data recorded on multiple smartphone devices: iPhone 6s, iPhone 10, iPhone 11, Samsung S7 and Samsung S8.
   - Three unique and three common sentences for each subject, each device, each language and each session.
   - Two types of presentation attacks are created, each in physical access and logical access scenarios.
2) Benchmarking the dataset with state-of-the-art face recognition, speaker recognition algorithms and score-level fusion biometric methods.
3) Evaluating the vulnerability of presentation attacks on state-of-the-art biometric verification and testing baseline presentation attack detection methods.

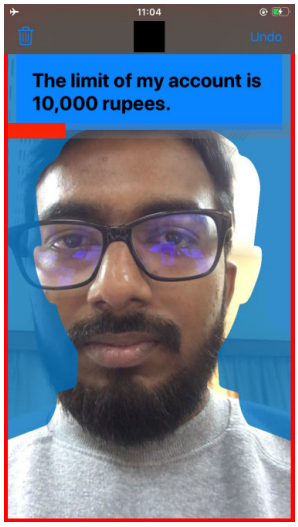

FIGURE 5. Mobile application (iOS) interface for data capturing.

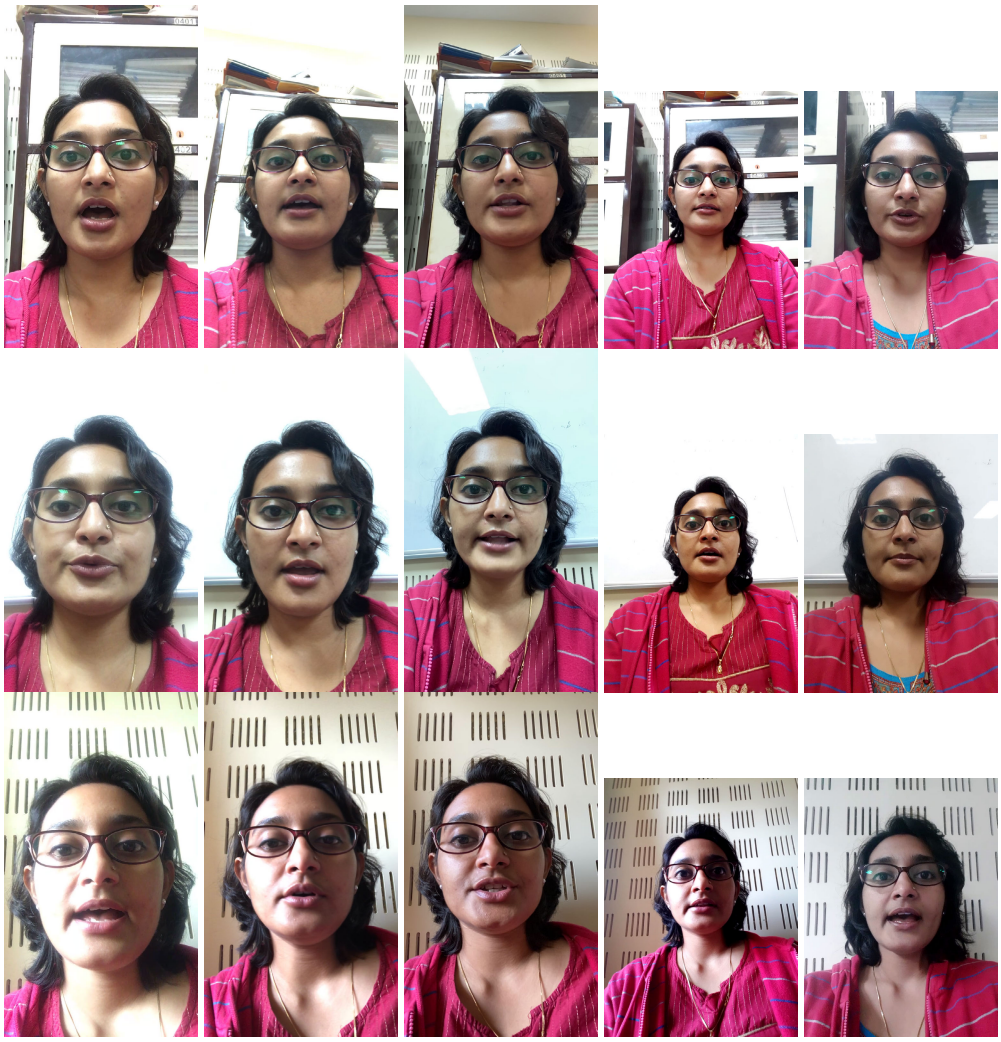

FIGURE 6. Audio-visual data samples (1 frame of a talking face). Left to Right: iPhone 6s, iPhone 10, iPhone 11, Samsung S7 and Samsung S8. Top row: Session 1, middle: Session2, bottom: Session3.

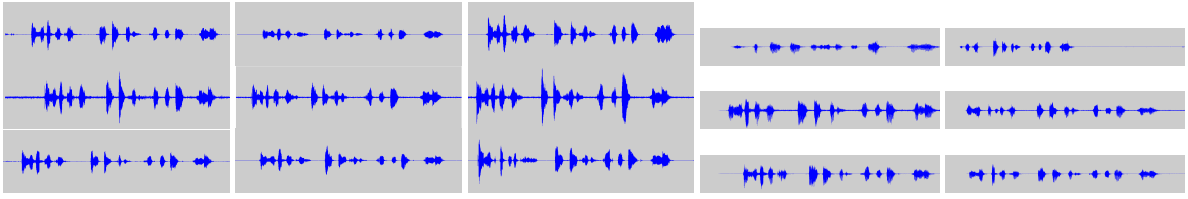

FIGURE 7. Audio data sample for speaker recognition. Left to right: iPhone 6s, iPhone 10, iPhone 11, Samsung S7 and Samsung S8. Top row: Session 1, middle: Session 2, bottom: Session 3.

III. MULTILINGUAL AUDIO-VISUAL SMARTPHONE (MAVS) DATASET

A. ACQUISITION

In data acquisition, we have used five smartphone devices, namely iPhone 11, iPhone10, iPhone 6s, Samsung S7 and Samsung S8. The data capturing is a self-assisted process where the speaker handles the mobile device and records the biometric data. For the process of data capturing, a mobile application has been used in both iOS and Android devices. The application provides a simple interface that assists the speaker to provide audio-visual data, as shown in Figure 5. A pre-defined text appears on the screen for a limited time for each sample. The speaker reads the text while the data is being recorded.

B. PARTICIPANT DETAILS

We have obtained 70 male and 33 female participants for the data collection. The average age of the participants

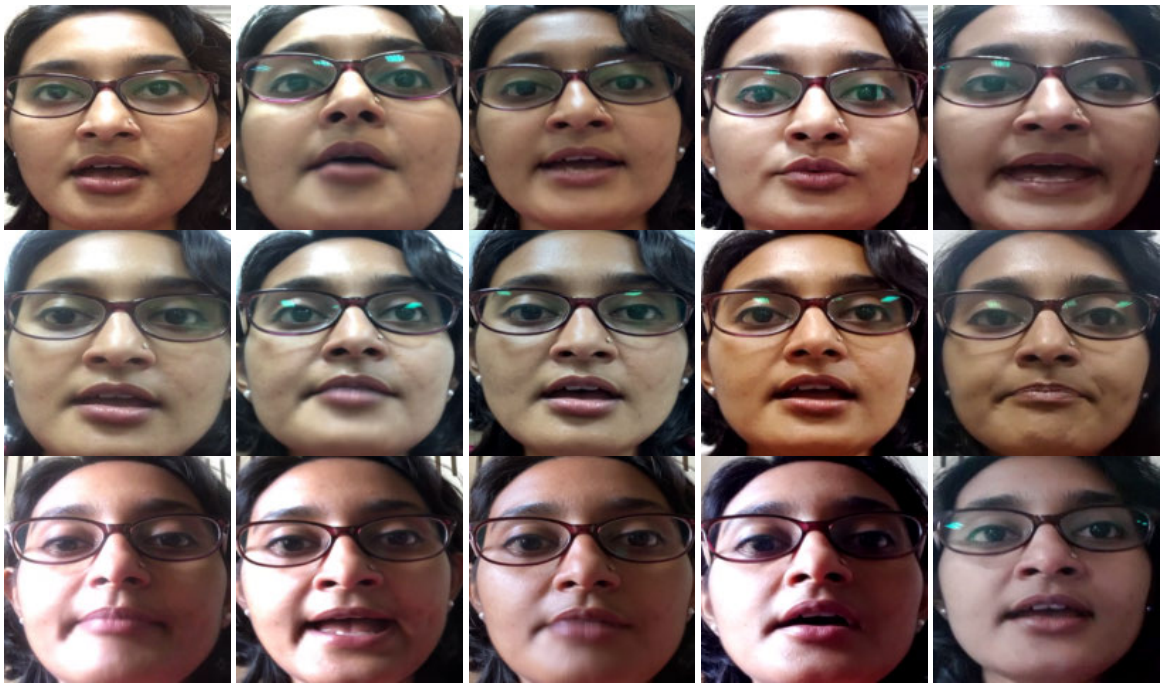

FIGURE 8. Detected face using MTCNN for face recognition. Left to right: iPhone 6s, iPhone 10, iPhone 11, Samsung S7 and Samsung S8. Top row: Session 1, middle: Session2, bottom: Session3.

is 27 years. All participants are of Indian origin with medium to expert range fluency in speaking the three languages (English, Hindi and Bengali). All participants are informed about the data acquisition protocol and are instructed to use the mobile application by self-assisting the data capture. Each session, the participant is given five mobile devices, one after the other, and audio-visual data of 6 sentences in three languages is recorded.

C. DATA DETAILS

Each participant records six sentences in each language. Three of the sentences are the same for all subjects, and the other three sentences have a unique part for each subject. The six sentences in the English language are mentioned below, and the blank spaces are filled with unique fake text for each subject. Similarly, translated sentences for the other two languages are presented in their corresponding script.

1) My full name is `fake name`.
2) I live at the address `fake address`.
3) I am working at IIT Kharagpur.
4) My bank account number is `fake number`.
5) The limit of my account is 10,000 rupees.
6) The code for my bank is 9876543210.

Data is captured in three sessions with three different lighting and noise environments. In session1, there is no noise, and uniform lighting is used. This data can be used as clean data for enrollment purposes. Session2 has continuous controlled noise from a portable fan intentionally put near the data capturing process and different lighting than session1 but with uniform illuminance. Session3 has uncontrolled noise from natural background and nonuniform lighting where certain parts of the participant's face are dark. The order of sentences, languages, and mobile devices used during data capture is kept the same for all the sessions. The sample video data can be seen in Figure 6 (one frame per session, the device is presented for convenience). The waveform of audio samples is presented in Figure 7. In Figure 8, the segmented face images (using MTCNN, see Section IV-B1) of each session and device are presented.

D. PRESENTATION ATTACKS

We have created two types of presentation attacks: replay attacks and synthesized attacks.

1) REPLAY ATTACKS

The replay attacks are created by synchronized capture of audio-visual playback using Dell office monitor and Logitech speakers recorded on Samsung S8 phone. Figure 9 show the replay attacks samples created in this work. The spectrograms of audio replay attacks are presented in figure 10.

2) SYNTHESIZED ATTACKS

Deep learning has been successfully applied to solve complex problems ranging from big data analysis to computer vision tasks and human level control. Advanced deep learning concepts have also been used to create threats to privacy, democracy and national security. One such deep-learning based application that loomed recently is ''deepfake'' (derived from 'deep learning' and 'fake'). For creating synthesized attacks, we have used deepfake approaches in this work.

FIGURE 9. Replay attack data sample. Left: Bona fide, right: Replay attack.

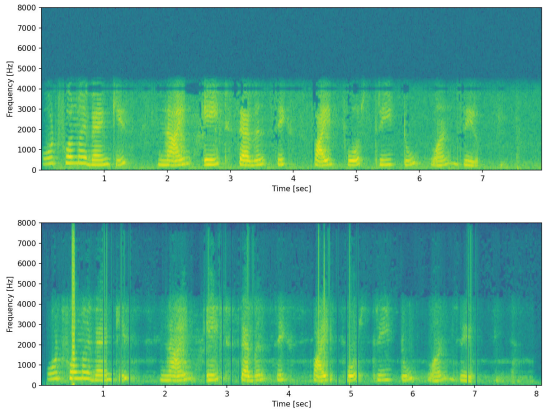

FIGURE 10. Spectrograms of bona fide and corresponding replay attack audio. Top: Bona fide, bottom: Replay attack.

One of the approaches for creating face deepfakes is a technique where the face image of a source person is superimposed onto a target person to create a video/image of the target person. In this direction, the face-swapping model is proposed by Nirkin *et al.* [23] where swapping of face images are done in three stages. Reenactment and face segmentation is carried out in the first stage, followed by in-painting and blending. Reenactment, face transfer, or puppeteering uses facial expressions and assists in transforming the face in one video to guide the motions and deformations of the face appearing in another video or image. Face segmentation is performed using U-Net [32] and reenactment is performed using generative model named pix2pixHD [43]. In the second step, the occluded regions of the source face are mitigated using the same in-painting generator [43]. In the last step, a Gaussian Poisson Generative Adversarial Network (GP-GAN) [44] is used for high-resolution image blending for combining the gradient and colour information.

In our work, we have utilized FSGAN for swapping similar faces.[8] The face-swapping approach preserves the context of the target video by digitally overlaying the source's face landmarks. Therefore, the target video contains the key biometric characteristics of the source subject, which can efficiently be used as a presentation attack for the source's identity. Multiple deepfake datasets in the literature [14], [33], [45], and dolhansky2019deepfake used a manual selection of faces for swapping. However, we have employed an automatic way to find a pair of similar faces in this work. We used cosine similarity of ArcFace embeddings to find a similar face for each of the male and female subjects (more on ArcFace in section IV-B4). We have generated 97 face swapped videos for sentence 6 of bona fide data from session1 data of the Samsung S8 device.

FIGURE 11. Face swap using FSGAN. Left: Source face, middle: Target face, right: Swapped face.

WaveNet vocoder is used to generate high-quality raw speech samples conditioned on acoustic features [24]. The WavNet-based vocoder is popularly used in ASVSpoof 2019 challenge to create logical access presentation attacks [41]. In our work, we have used MFCC features as acoustic features in synthesizing 16-bit raw audio. We have adapted the implementation of WaveNet vocoder form the github[9] and pre-trained models from LJSpeech [13]. The figures 11 and 12 show the images samples and spectrograms of synthesized attacks respectively.

IV. PERFORMANCE EVALUATION PROTOCOLS

The dataset is benchmarked with various face recognition, speaker verification and presentation attack detection methods. In this section, we explain briefly the baseline biometric systems employed along with evaluation metrics.

[8]FSGAN: https://github.com/YuvalNirkin/fsgan
[9]WaveNet Vocoder: https://github.com/r9y9/wavenet_vocoder

A. AUTOMATIC SPEAKER VERIFICATION

1) I-VECTOR BASED SPEAKER VERIFICATION

The I-vector based ASV method is a Joint Factor Analysis (JFA) approach proposed in [5]. It models the channel effects and also speaker voice characteristics. The speech sample is represented as a low-dimensional super vector called i-vector. The i-vector represents the total factor in a speech utterance, including channel compensation which is carried out in a low-dimensional total variability space.

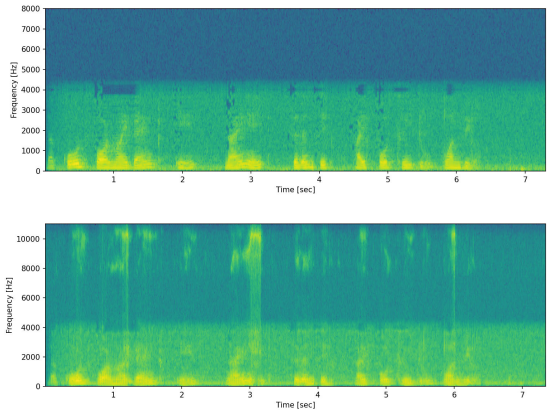

FIGURE 12. Spectrograms of bonafide and corresponding wavenet-vocoder synthesized audio. Top: Bona fide, bottom: Synthesized audio.

2) X-VECTOR BASED SPEAKER VERIFICATION

The deep neural networks (DNN) and end-to-end speaker verification approaches are state-of-the-art research methods that overcome handcrafted methods' drawbacks. The x-vector based speaker verification is a recent approach showing promising results in automatic speaker verification [39]. This method uses deep neural network (DNN) embeddings as features. The variable-length speech utterances are mapped to a fixed low-dimensional embedding (called x-vectors), and a deep network is trained to differentiate speakers. The training process requires a large amount of training data. Therefore, data augmentation is used along with added noise and reverberation to increase training data size. The implementations in Kaldi are employed in our work, and the pre-trained Universal Background Models, i-vector extractor and x-vector extractor are adapted to our experiments.[10] Probabilistic linear discriminant analysis (PLDA) [28] is used as a classifier for the i-vectors and x-vectors of enrollment and test samples. The log-likelihood score is computed between the enrolled and test speech sample pair.

3) DILATED RESIDUAL NETWORK (DltResNet)

Extended ResNet implementation from [15] named dilated residual network (DltResNet) is used as the third speaker verification methods. The implementation is publicly available.[11] The DltResNet model is one of the state-of-the-art systems on the Voxceleb1 database evaluations achieving 4.8% EER on the dataset. The Euclidean distance between the DltResNet features is used for obtaining scores between enrolled and test samples.

B. FACE RECOGNITION

1) FACE DETECTION

Face detection is performed as a prepossessing step on the video frames to detect and crop the face image. We have employed multitask cascaded convolutional networks (MTCNN) approach from Zhang *et al.* [46] for efficient face detection. The face recognition and face PAD methods used in this work used segmented face images.

2) LOCAL BINARY PATTERNS (LBP)

Local Binary Patterns (LBP) are a textual operator that labels the pixels in a face image according to neighbouring pixels' values and assigns a binary number. LBP for an image is calculated by assigning 0 or 1 to the pixel depending on the neighbour's pixel having high or low value. The resultant binary test is stored in an 8-bit array and later converted to decimal. This thresholding process, accumulating binary strings, and storing the decimal value is repeated for every pixel in the input image. Further, the LBP histogram is computed over the LBP output array. For a block, one of the $2^8 = 256$ possible patterns is possible. The advantage of LBP features is high discriminative power, computational simplicity, and invariance to grey-scale changes. LBPs have shown a prominent advantage in face recognition approaches. We used LBP histograms as features for face images and cosine distance to compute the score between the enrolled and test samples.

3) FaceNet FACE EMBEDDINGS

The deep learning approaches have evolved into image processing and pattern recognition applications. In face recognition methods, FaceNet embeddings displayed an excellent image representation for facial features [37]. This is a deep face recognition approach that adapted the ideas from [26]. In this work, we have used the pretrained model on the VGGFace2 dataset using Inception ResNet v1. This model displayed an accuracy of 99.65% on the Labeled Faces in the Wild (LFW) dataset [10]. We have obtained FaceNet embeddings[12] for face detected images in our dataset and used cosine distance between the samples to obtain the verification scores.

4) ArcFace FACE DESCRIPTOR

ArcFace face features are proposed in [6] for the large scale face recognition with enhanced discriminative power. ArcFace features emphasize the loss function in deep

[10] Kaldi GitHub: https://github.com/kaldi-asr/kaldi

[11] DltResNet: https://www.idiap.ch/software/bob/docs/bob/bob.learn.pytorch/v0.0.4/guide_audio_extractor.html

[12] FaceNet: https://github.com/davidsandberg/facenet

convolutional neural networks (DCNN) for clear geometric interpretation of face images. The proposed descriptor is evaluated over ten face recognition benchmarks, and results show consistent performance improvement. We have employed the ArcFace implementation provided in Github.[13] The training data contains cleaned MS1M, VGG2 and CASIA-Web face datasets. ArcFace face descriptors are computed over detected face images, and similar to other face recognition methods, we have used cosine distance as a classifier.

In addition to the face recognition, we have used ArcFace face embeddings to obtain similarity scores between subjects in creating attacks in FSGAN face swapped videos (see section III-D2).

C. PRESENTATION ATTACK DETECTION (PAD)

1) VOICE PAD

The PAD methods used to evaluate the attacks created using speech are chosen from the baseline methods in the ASVSpoof 2019 challenge [41]. The two baseline methods are available in ASVSpoof 2019 evaluation protocols. Features used in these two methods are based on cepstral coefficients in the front-end and Gaussian Mixture Models (GMM) in the back-end. Linear Frequency Cepstral Coefficients (LFCC) and Constant Q Cepstral Coefficients (CQCC) are two features used to represent speech samples.

The LFCC features are similar to the Mel-frequency cepstral coefficients (MFCCs), with filters placed linearly in the exact sizes. The initial approach of LFCCs is used for the detection of synthetic speech in [34]. In this work, we used LFCC features are extracted with a frame length of 25ms and a 20-channel linear filter bank. An LFCC feature comprises 19 cepstral coefficients, a zeroth coefficient, static, delta, and delta-delta coefficients. The CQCC features are extracted with the toolkit provided in ASVSpoof 2019. The maximum frequency is set to fs/2, where fs is the sampling frequency, and the minimum frequency is fixed at $fs/2/2^9$ $15Hz$ (where 9 is the number of octaves) [40]. The number of bins per octave is set to 96, and re-sampling is applied with a period of 16. The dimension of features is 29 coefficients along with zeroth, static, delta, and delta-delta coefficients.

The front-end provides the cepstral coefficients, which are used to train 2-class GMMs in the back-end. The training process is carried out on the bonafide and attack speech samples with 512-component GMM models. An expectation-maximization (EM) algorithm is employed in training with random initialization. For testing, the scores of samples are calculated from the log-likelihood ratio with the help of trained bona fide and the attack speech models.

2) FACE PAD

The face recognition PAD methods are chosen from the baseline methods used in smartphone dataset evaluation in [30]. The two best-performing methods from five baseline methods are taken for evaluation in this work. These methods utilize local binary patterns (LBP) [4] and color texture features [3]. The support vector machines (SVM) are trained for different attacks and test for attack detection.

The LBP features are experiments for PAD in [4] for face attacks in a full biometrics verification system. In [29], the LBP features displayed a consistent performance of detecting attacks in different protocols of smartphone biometric data. Similarly, the experiments using colour texture features [3] resulted in the best-performing face PAD on smartphone face images. Therefore, we have included these methods in our evaluation of detection attacks.

[13] ArcFace: https://github.com/deepinsight/insightface

D. PERFORMANCE METRICS

The performance evaluation metrics from ISO/IEC [11] are utilized in our experiments to present and compare the results of different methods.

1) VERIFICATION METRICS

- False Match Rate (FMR) is the proportion of the completed biometric non-mated comparison trials that result in a false match.
- False Non-Match Rate (FNMR) is the proportion of the completed biometric mated comparison trials that result in a false non-match.

In addition to ISO/IEC metrics mentioned above, we have also presented an equal error rate (EER) to represent FMR and FNMR metrics in a single value. EER is the error rate at the point where FMR and FNMR are equal.

2) PRESENTATION ATTACK DETECTION METRICS

- Impostor-Attack Presentation Match Rate (IAPMR) is the proportion of impostor attack samples (replay attacks) that are matched with bona fide samples. To compare ASV methods' performance, we have fixed FMR at 0.1% and presented FNMR and IAPMR for zero-effort impostors and attacks, respectively.
- Attack Presentation Classification Error Rate (APCER) is the proportion of attack presentations that are incorrectly classified as bona fide presentations, and Bonafide Presentation Classification Error Rate (BPCER) is the ratio of bona fide presentations incorrectly classified as attacks. This work presents the BPCER_5 and BPCER_10 of PAD methods: the BPCER values at APCER are 5% and 10%, respectively.

Also, we used Detection Equal Error Rate (D-EER) to present PAD methods' performance, a single value representation of APCER and BPCER. The score distributions of bona fide, zero-effort impostors and attacks are plotted along with the threshold of FMR = 0.1% to observe the impact of presentation attacks. Detection error trade-off (DET) curves plot the relationship between false match rate (FMR) and false non-match rate (FNMR) for bona fide samples or impostor attack presentation match rate (IAPMR) for attack samples, respectively.

TABLE 2. Inter-session speaker recognition evaluation (EER%).

Inter-session	i-Vector			X-Vector			DltResNet		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
S1	**5.31**	11.52	10.35	**5.31**	11.18	10.84	**4.85**	10.69	9.56
S2	11.70	**4.13**	10.51	11.20	**3.51**	9.96	10.63	**4.32**	9.50
S3	10.48	10.65	**5.16**	10.70	9.96	**5.23**	9.51	9.59	**4.53**

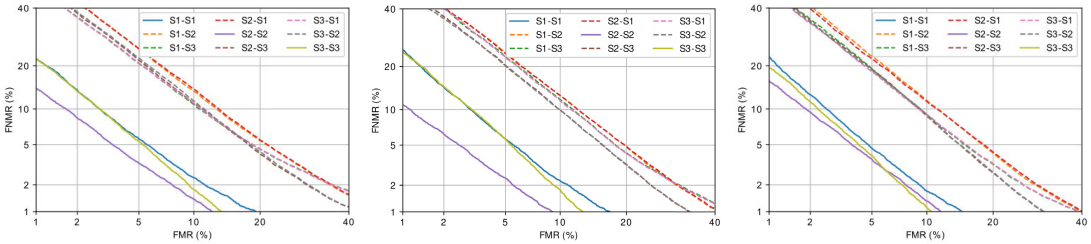

FIGURE 13. DET curves of inter-session speaker recognition experiments. Left: i-vector, middle: X-vector and right: DltResNet.

V. EXPERIMENTAL RESULTS

The main focus of this dataset is to provide scope for developing generalized biometric algorithms in face and speech-based recognition. The generalizability of a biometric algorithm can be achieved by considering multiple dependencies like session variance, device dependency and language. Therefore, in our work, we have performed experiments to demonstrate how these dependencies affect the state-of-the-art face and speaker recognition algorithms mentioned in IV. The benchmarking of the dataset is carried out by performing different experiments and presenting the results.

A. AUTOMATIC SPEAKER VERIFICATION

Automatic Speaker Verification methods display variable performance depending on the channel used to acquire and the noise present in the audio samples. In the following experiments, we have evaluated the performance of the ASV methods in correspondence to the session, device and language.

1) INTER-SESSION SPEAKER RECOGNITION

The MAVS dataset contains data from three different sessions as explained in section III. We have examined the session dependency by performing the inter-session speaker recognition. In this process, we have used the samples from one session to enroll and each of the other sessions to test. Table 2 presents the EER values displaying the comparison of three ASV methods on inter-session experiments.

- Session 2 data contains an added noise in all data samples. Therefore, it is seen that higher EER values are observed in all the results where session 2 data is used to enroll.
- However, when the same noise is present in test data, the ASV methods tend to perform better than the session with clean data (session 1). This concludes that ASV methods characterize the noise in the data and use it for recognition.
- Similarly, session 3 contains natural noise, which is not consistent in all samples, but it helps recognise the speaker better than the data with no noise.
- Alongside, DltResNet based ASV method displayed better performance compared to other methods.

2) INTER-DEVICE SPEAKER RECOGNITION

The properties of the data capturing device are key attributes for speaker recognition [5]. Although state-of-the-art ASV methods accommodate the channel characteristics, the change in devices from enrollment to test can still affect the speaker recognition performance. Our dataset used five different smartphones in data collection to examine the dependency of the device on ASV methods. Tables 3, 4, 5 show the EERs of all device combinations of enrollment and testing from the three ASV methods.

The results from inter-device experiments output some key points. These observations conclude the impact of channel dependency on state-of-the-art speaker recognition methods.

- The DltResNet method gave out the highest EER in most of the combinations even though it worked better with noisy data as shown in Section V-A2.
- The DNN based X-vector methods performed better than other methods.
- It is observed that the combinations of smartphones from the same manufacturer (Apple or Samsung) correlate with speaker recognition. When the enrollment and testing data are from the same manufacturer, the speaker recognition performs better than the cross-manufacturer combination.

3) INTER-LANGUAGE SPEAKER RECOGNITION

The language difference in the audio sample for ASV has been a hot topic in recent years. Although there are datasets with utterances of the same person in different languages, the

TABLE 3. Inter-device speaker recognition evaluation (EER%) on i-vector method.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**1.86**	5.76	6.67	15.46	14.37
iPhone 10	5.88	**1.62**	4.74	15.02	13.97
iPhone 11	6.73	4.67	**1.47**	15.90	14.76
Samsung S7	15.51	14.90	15.70	**10.01**	13.26
Samsung S8	14.51	13.98	14.78	13.34	**8.77**

TABLE 4. Inter-device speaker recognition evaluation (EER%) on x-vector method.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**1.45**	5.82	6.55	15.33	14.09
iPhone 10	5.85	**1.81**	4.37	13.56	12.37
iPhone 11	6.54	4.30	**1.81**	14.27	13.10
Samsung S7	15.50	13.69	14.13	**8.55**	12.97
Samsung S8	14.04	12.25	12.93	13.30	**7.37**

TABLE 5. Inter-device speaker recognition evaluation (EER%) on DltResNet method.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**2.08**	6.52	7.07	16.56	16.38
iPhone 10	6.62	**2.03**	4.09	15.00	15.66
iPhone 11	7.06	4.03	**2.02**	15.92	16.14
Samsung S7	16.68	15.07	15.83	**7.04**	10.44
Samsung S8	16.51	15.52	16.11	10.63	**7.73**

TABLE 6. Inter-language speaker recognition evaluation (EER%).

Inter-language	i-vector			x-vector			DltResNet		
	English	Hindi	Bengali	English	Hindi	Bengali	English	Hindi	Bengali
English	**5.47**	5.50	6.72	**4.98**	5.55	6.93	**4.88**	5.26	6.27
Hindi	5.58	**4.16**	5.33	5.45	**4.0**	5.60	5.32	**3.95**	5.15
Bengali	6.78	5.92	**5.08**	6.93	5.67	**5.21**	6.34	5.19	**4.87**

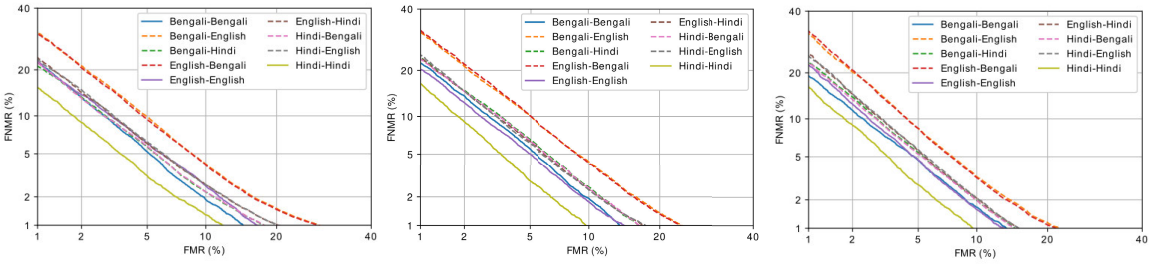

FIGURE 14. DET curves of inter-language speaker recognition experiments. Left: i-vector, middle: X-vector and right: DltResNet.

problem of language dependency is not benchmarked [30]. The degradation of biometric recognition due to language mismatch is presented in some previous works [16], [17], [21]. Our dataset comprises of the same subjects speaking three different languages, therefore, providing scope for inter-language speaker recognition evaluation. Table 6 shows the inter-language speaker recognition evaluations.

- The problem of language mismatch from enrollment to testing is observed in all three ASV methods.
- However, the drop in EER is not high, but it is consistent across all the methods.
- It is important to notice that the training dataset contains multiple languages, and we assume that the extracted features contain language factors.
- Therefore, in the scenario of a small subset of languages in training data, the language mismatch problem would be considerable.

B. FACE RECOGNITION

The robustness of face recognition algorithms in smartphones is evaluated in this section. Similar to speaker recognition, we have performed two dependency experiments, namely inter-session and inter-device. The three face recognition systems are examined in these experiments by taking 20 equally distributed frames in each video.

1) INTER-SESSION

The session variability in face recognition is observed in this experiment.

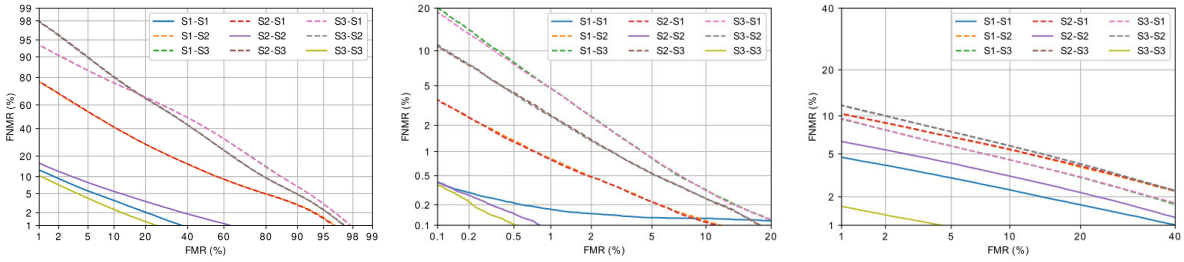

FIGURE 15. DET curves of inter-session face recognition experiments. Left: LBP, middle: FaceNet and right: ArcFace.

TABLE 7. Inter session face recognition evaluation EER (%).

Inter-session	LBP			FaceNet			Arcface		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
S1	**5.39**	24.28	44.73	**0.26**	0.89	2.22	**3.42**	6.68	5.60
S2	24.28	**6.81**	41.55	0.87	**0.24**	1.65	6.42	**4.34**	6.81
S3	44.67	41.43	**4.43**	2.21	1.63	**0.21**	5.59	6.81	**1.43**

TABLE 8. LBP face recognition performance EER(%) in inter-device scenario.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**6.96**	19.50	19.60	22.94	31.21
iPhone 10	19.55	**5.32**	18.72	31.69	37.95
iPhone 11	19.70	18.76	**5.09**	25.67	32.60
Samsung S7	22.96	31.69	25.70	**5.05**	21.04
Samsung S8	31.13	37.87	32.65	21.10	**5.04**

TABLE 9. FaceNet face recognition performance EER (%) in inter-device scenario.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**0.20**	0.44	0.64	0.66	0.48
iPhone 10	0.45	**0.28**	0.51	0.69	0.53
iPhone 11	0.64	0.51	**0.3**	0.92	0.71
Samsung S7	0.67	0.68	0.90	**0.25**	0.34
Samsung S8	0.49	0.54	0.71	0.33	**0.16**

- Session 2 and session 3 data has non-uniform lighting on the face region. Therefore, the cross-session face recognition displayed a clear drop in the performance.
- FaceNet performed better in attributing the problem of session variability among the three face recognition methods while displaying near-zero error rates in the same session.
- Table 7 present the EER values for inter-session face recognition experiments.

2) INTER-DEVICE

The results from inter-device experiments on face recognition are shown in Tables 8, 9, 10.

- The LBP features based face recognition displayed a high dependency on devices. When the device is the same in enrollment and testing, LBP features performed better face recognition. However, the recognition error has increased by three times when there is a miss-match in devices.

TABLE 10. Arcface face recognition performance EER (%) in inter-device scenario.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**3.30**	4.14	4.03	4.79	4.36
iPhone 10	4.10	**3.10**	3.76	4.76	4.31
iPhone 11	4.04	3.79	**3.01**	4.60	4.03
Samsung S7	4.80	4.76	4.55	**2.98**	3.78
Samsung S8	4.39	4.30	4.03	3.78	**2.72**

- Another observation is that the change in device manufacturer has also impacted face recognition similar to speaker recognition.
- FaceNet has displayed better face recognition considering the problem of device dependency. The drop in performance is observed, but it is not as consistent as other methods.
- ArcFace performed similarly to FaceNet in an inter-device face recognition scenario.
- Although the EER is higher in ArcFace than FaceNet; the device mismatch has not impacted the performance very much.

C. AUDIO-VISUAL SPEAKER RECOGNITION

The audio-visual speaker recognition is performed by score-level fusion of best-performing face recognition and speaker recognition methods, FaceNet and X-vector methods, respectively. The score fusion approach used in this work is a simple averaging of scores obtained in individual verification methods.

1) INTER-SESSION

- The combination of audio and visual data displayed similar results as that of individual biometric algorithms. This is because of the simple score-level fusion method employed in our work.
- We assume that an adaptive fusion approach would improve the performance.
- However, it introduces a new dependency on biometric algorithms in the form of a fusion approach.
- Table 11 show the results of inter-session audio-visual fusion experiments. Figure 16 present the corresponding DET curves.

2) INTER-DEVICE

The inter-device experiments on audio-visual biometric recognition are carried out similar to the inter-session approach. The obtained results display the same observations

TABLE 11. Inter session audio-visual speaker recognition evaluation EER (%).

Inter-session	S1	S2	S3
S1	**4.99**	10.73	10.46
S2	10.74	**3.21**	9.56
S3	10.34	9.55	**4.90**

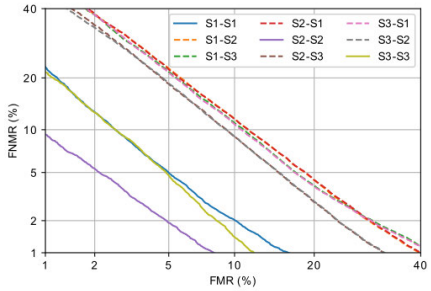

FIGURE 16. DET curves of inter-session experiments on audio-visual fusion of FaceNet and X-vector methods.

as that of audio-visual inter-session biometric recognition. It is clear from these experiments that an efficient fusion approach is required to take advantage of bi-modal biometrics. Table 12 display the EER values of inter-device experiments using audio-visual fusion.

D. VULNERABILITY FROM PRESENTATION ATTACKS

The vulnerability of biometric recognition towards presentation attacks is examined in this section. The two types of presentation attacks created in this work are explained in Section III-D. The biometric recognition performance before and after the attacks is compared to check the robustness. When a presentation attack is not carried out, the performance is expressed in false non-match rate (FNMR) caused by zero-effort impostors. In presentation attacks, the vulnerability is presented as impostor attack presentation match rate (IAPMR).

1) REPLAY ATTACKS

The replay attacks are created by replaying an audio-visual biometric sample on a display and loudspeaker combination. The playback sample is recorded on one of the smartphones, namely the Samsung S8. The audio and face channels of replay attacks are examined for vulnerability individually on the two best performed biometric methods from the previous sections. For face recognition, FaceNet features are used, and for speaker recognition, X-vector features are employed.

- The impact of replay attack is presented in Table 13 in FNMR and IAPMR rates for zero-effort impostors and replay attacks, respectively.
- In face recognition, the vulnerability is observed as 96.87% IAPMR, representing the number of attacks being matched with bonafide samples.

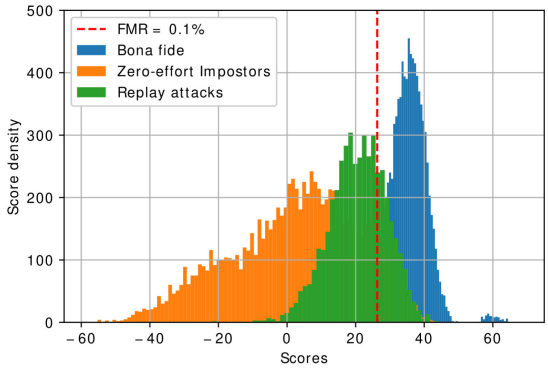

FIGURE 17. Audio replay attacks score distribution tested on X-vector method.

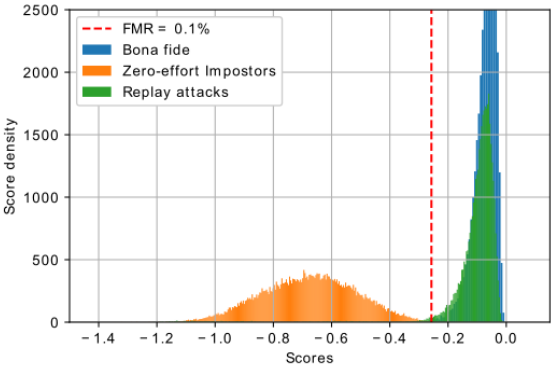

FIGURE 18. Video replay attacks score distribution tested on FaceNet method.

- The speaker recognition method displayed 25.93% IAPMR when compared to 6.4% FNMR.
- The score distributions of bona fide, zero-effort impostors and replay presentation attacks are presented in Figures 17 and 18.

2) SYNTHESIZED ATTACKS

Synthesized attacks are logical access attacks where the attack sample is presented digitally to the biometric system. Table 14 shows the vulnerability of synthesized attacks on face and voice modalities.

- The vulnerability evaluation on FaceNet based face recognition shows a 38.77% IAPMR, and the score distributions are presented in Figure 19.
- The speech synthesis is carried out using wavenet-vocoder, and the attacks displayed 99.68% IAPMR.
- The score distributions are presented in Figure 20.

3) AUDIO-VISUAL PRESENTATION ATTACKS

The vulnerability of audio-visual presentation attacks is examined with the help of fusion of presentation attacks on AV recognition methods explained in Section V-C. The replay attacks and synthesized attacks are performed in individual

TABLE 12. Inter-device performance (EER%) of score-level fusion of FaceNet and X-vector methods.

Inter-device	iPhone 6s	iPhone 10	iPhone 11	Samsung S7	Samsung S8
iPhone 6s	**1.31**	5.53	6.24	14.55	13.30
iPhone 10	5.57	**1.65**	4.18	12.82	11.74
iPhone 11	6.25	4.15	**1.70**	13.53	12.41
Samsung S7	14.75	12.93	13.34	**7.92**	12.30
Samsung S8	13.28	11.54	12.30	12.59	**6.81**

TABLE 13. Replay attack vulnerability on face and voice at FMR = 0.1%.

Biometric Algorithm	Zero-Effort impostors	Replay Attacks
	FNMR	IAPMR
FaceNet	0.09%	96.87%
X-vector	6.4%	25.93%

TABLE 14. Synthesized attack vulnerability on face and voice at FMR = 0.1%.

Biometric Algorithm	Zero-Effort impostors	Synthesized Attacks
	FNMR	IAPMR
FaceNet	0.21%	38.77%
X-vector	5.59%	99.68%

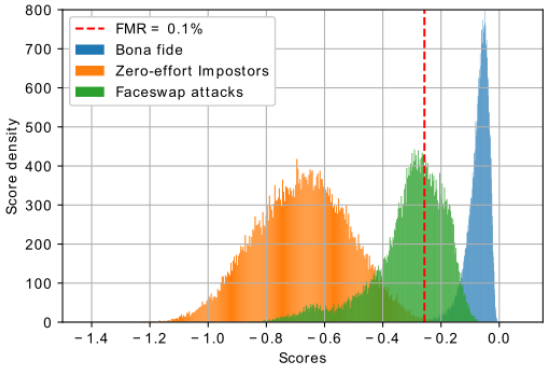

FIGURE 19. Score distribution of face swap attacks.

TABLE 15. Audio-visual replay attacks vulnerability on AV fusion method at FMR = 0.1%.

Attack Type	Zero-Effort impostors	Presentation Attacks
	FNMR	IAPMR
Replay Attacks	5.29%	28.46%
Synthesized Attacks	4.64%	99.83%

biometric modalities, and the attack scores are fused to calculate the final scores. The impact of the audio-visual attacks is presented in Table 15 on two different attacks. Unlike unimodal biometric matching, the results of audio-visual biometrics are presented in False Rejection Rate (FRR) because it represents the system-level performance. Similarly, the score distributions are shown in Figures 21, 22.

- The results indicate that audio-visual fusion is vulnerable to presentation attacks.
- The problem of replay attacks is less compared to the synthesized attacks.

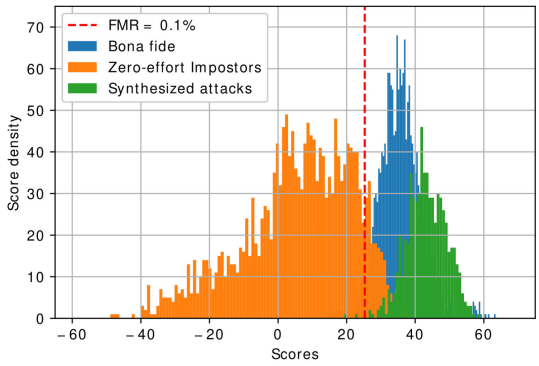

FIGURE 20. Score distributions of wavenet speech synthesized attacks.

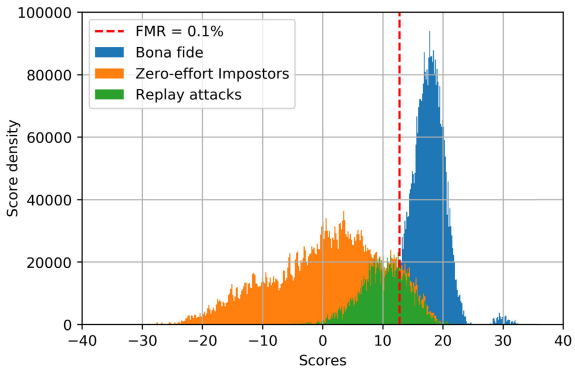

FIGURE 21. Audio-visual replay attacks score distribution.

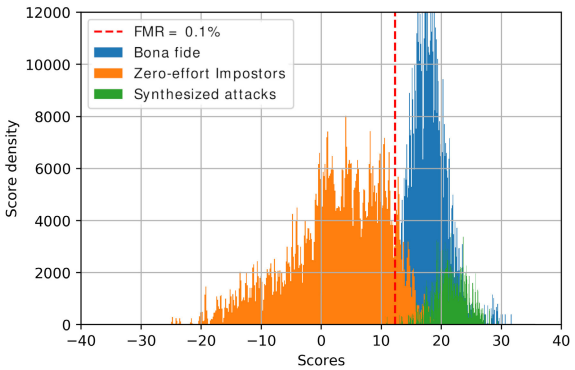

FIGURE 22. Audio-visual synthesized attacks score distribution.

- Although the replay attacks on face recognition displayed the highest vulnerability; the AV fusion approach appears to have the ability to overcome this problem.

TABLE 16. Results of speaker recognition presentation attack detection.

Attack type	LFCC-GMM			CQCC-GMM		
	D-EER	BPCER_5	BPCER_10	D-EER	BPCER_5	BPCER_10
Replay Attacks	44.14%	100%	93.15%	20.49%	45.63%	36.89%
Speech Synthesis	14.00%	39.82%	20.38%	14.08%	40.77%	22.33%

TABLE 17. Results of face recognition presentation attack detection.

Attack type	LBP-SVM			Color texture-SVM		
	D-EER	BPCER_5	BPCER_10	D-EER	BPCER_5	BPCER_10
Replay Attacks	4.96%	5.07%	1.28%	2.15%	1.35%	0.32%
FaceSwap	2.99%	1.74%	1.15%	2.54%	0.83%	0.26%

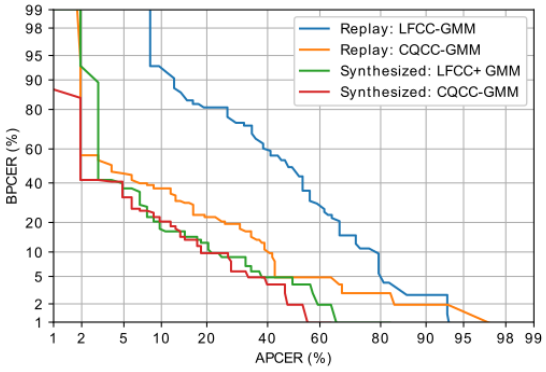

FIGURE 23. DET curves of voice PAD evaluation using baseline methods.

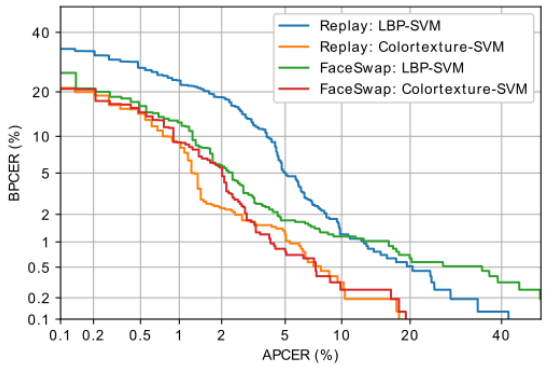

FIGURE 24. DET curves of face PAD evaluation using baseline methods.

However, a similar observation is not seen in synthesized attacks.

- Thus, the AV fusion recognition approach has the vulnerability due to combined AV presentation attacks.

E. PRESENTATION ATTACK DETECTION

The presentation attack detection experiments are performed using baseline PAD methods. The attack data is partitioned into three sets: training, developing and testing, with 35%, 35% and 30% of bona fide and attack samples, respectively. Each partition includes data from a unique set of subjects. We have chosen the baseline approaches used in Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVSpoof) for speaker recognition PAD in 2019. See Section IV-C. For face recognition, we opted the two best-performing methods from the face PAD methods used in [30]. Tables 16 and 17 show the results of the PAD methods in terms of D-EER, BPCER at APCER = 5% and BPCER at APCER = 10%. The DET curves in figures 23 and 24 present the performance of PAD methods.

- The voice PAD results indicate that the baseline methods are not able to detect the attacks.
- Alongside, replay attacks are difficult to detect when compared to synthesized attacks. In contrast, both face PAD methods performed well in detecting the attacks.
- The voice PAD methods are tested on the whole speech sample, where the face PAD methods are performed on detected face images in individual frames.

TABLE 18. Results of audio-visual PAD methods.

Attack type	Fusion PAD		
	D-EER	BPCER_5	BPCER_10
Replay Attacks	16.99%	38.83%	30.10%
Synthesized	11.87%	32.04%	15.54%

- Therefore, it is reasonable to assume that this could be the reason for the difference in performance.

1) MULTIMODAL PAD

The presentation attacks on both modalities are possible with sophisticated equipment. The PAD methods should be able to detect the attacks before the verification process. In this experiment, we have fused the PAD scores from the CQCC-GMM method and the Color texture-SVM method to compute multimodal PAD scores. We have used a sum rule based fusion to combine two PAD methods. The table 18 shows the results of multimodal PAD approach and Figure 25 shows the PAD performance on two different types of attacks.

- The replay attacks are observed to be difficult to detect compared to synthesized attacks. The performance of multimodal PAD is similar to individual PAD in regards to the types of attacks.
- The multimodal PAD does not improve the attack detection performance. The reason for this could be the usage of simple sum rule based fusion.
- The co-related and complementary information between audio and visual domains is not taken into account in this

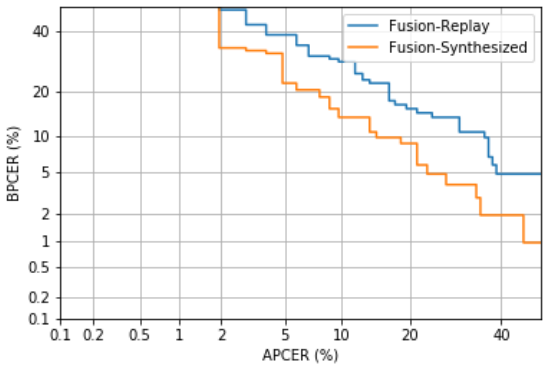

FIGURE 25. DET curves of audio-visual PAD of CQCC and Color texture methods.

fusion approach. Therefore, multimodal PAD does not show any promising improvement over individual PAD approaches.

VI. CONCLUSION

Smartphone biometrics have emerged into advanced security applications like banking transactions and identity verification. The built-in biometric systems by smartphone manufacturers can be utilized for this purpose. However, it is difficult to entirely rely on the built-in systems due to the variance in sensors and unknown algorithms embedded into smartphones. In this direction, it is possible to use the default sensors in smartphones like cameras and microphones. Therefore, we have developed a multidimensional smartphone audio-visual dataset that includes different languages, devices, sessions, and texts in this work. We have presented in this paper some of the previous works on building an audio-visual dataset and discussed our multi-lingual smartphone audio-visual (MAVS) dataset.

Further, we have performed experiments on examining the robustness of state-of-the-art biometric algorithms in two directions. The first direction concerns the problem of algorithm dependencies that include signal noise, capturing device and speech language. We have prepared inter-session, inter-device and inter-language experiments and presented the results. In the second direction, presentation attacks are evaluated for the vulnerability of biometric algorithms and the performance of baseline PAD algorithms. The results show the requirement of robust audio-visual biometrics algorithms to deal with the problems of multiple dependencies and presentation attacks. The proposed dataset would help the research community in developing advanced biometric algorithms and presentation attack detection approaches.

A. FUTURE WORK

The MAVS dataset is made publicly available for research purposes.[14] The proposed dataset can be used in multiple directions in smartphone audio-visual research. The future work in this research direction using the dataset is as follows.

1) Novel biometric algorithms are modelled by identifying various problems that question the robustness of smartphone authentication.
2) The authentication technology through biometrics can be improved via Audio-visual person recognition through the efficient usage of complementary information between audio and visual modalities.
3) The dataset contains subjects of different ages ranging from 18 to 48 years and gender labels (70 male and 33 female). Therefore, the dataset can be used for studying gender classification and fairness. Further, the audio data from three different languages can be used for language detection.
4) The correlated information between biometric cues are used to propose advanced presentation attack detection algorithms towards unknown and unseen attacks. E.g. lip-sync, correlated biometric data.
5) Generalizable biometric algorithms are developed in smartphone environments for real-world applications across different devices and capturing conditions.

[14]MAVS dataset request form: https://docs.google.com/forms/d/e/1FAIpQLSfTMqnQj8KNoUi1Ms1tx8Ewgil2l4wAAJVaKUJs6VkWfjAo4w/viewform?usp=sf_link

ACKNOWLEDGMENT

The authors acknowledge the Idiap Research Institute and Prof. Sébastien Marcel for the data capture mobile application developed as a part of the Secured access over Wide Area Network (SWAN) Project.

REFERENCES

[1] P. S. Aleksic and A. K. Katsaggelos, *An Audio-Visual Person Identification and Verification System Using FAPs as Visual Features*. Santa Barbara, CA, USA: Works. Multimedia User Authentication, 2003.
[2] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, "The BANCA database and evaluation protocol," in *Proc. 4th Int. Conf. Audio Video-Based Biometric Person Authentication (AVBPA)*. Berlin, Germany: Springer-Verlag, 2003, pp. 625–638.
[3] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2636–2640.
[4] I. Chingovska, A. R. D. Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2264–2276, Dec. 2014.
[5] N. Dehak, P. J. Kenny, R. Dehak, D. Pierre, and O. Pierre, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.
[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[7] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.
[8] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "Valid: A new practical audio-visual database, and comparative results," in *Audio- and Video-Based Biometric Person Authentication*, T. Kanade, A. Jain, and N. K. Ratha, Eds. Berlin, Germany: Springer, 2005, pp. 777–786.
[9] Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, "Deep multimodal speaker naming," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1107–1110.
[10] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces Real-Life'Images, Detection, Alignment, Recognit.*, Oct. 2008, pp. 1–11.

[11] *Information Technology Biometric Performance Testing and Reporting—Part 4: Testing Methodologies for Technology and Scenario Evaluation*, Standard ISO/IEC JTC1 SC37 Biometrics. ISO/IEC 19795-4:2008, International Organization for Standardization and International Electrotechnical Committee, 2008.
[12] *Information Technology—Biometric Presentation Attack Detection—Part 1: Framework*, Standard ISO/IEC JTC1 SC37 Biometrics. ISO/IEC 30107-1, International Organization for Standardization, 2016.
[13] K. Ito and L. Johnson. (2017). *The LJ Speech Dataset*. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/
[14] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*.
[15] N. Le and J.-M. Odobez, "Robust and discriminative speaker embedding via intra-class distance variance regularization," in *Proc. Interspeech*, Sep. 2018, pp. 2257–2261.
[16] L. Li, D. Wang, A. Rozi, and T. F. Zheng, "Cross-lingual speaker verification with deep feature learning," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 1040–1044.
[17] H. Mandalapu, T. M. Elbo, R. Ramachandra, and C. Busch, "Cross-lingual speaker verification: Evaluation on X-vector method," in *Intelligent Technologies and Applications*, S. Y. Yayilgan, I. S. Bajwa, and F. Sanfilippo, Eds. Cham, Switzerland: Springer, 2021, pp. 215–226.
[18] H. Mandalapu, A. Reddy P N, R. Ramachandra, K. S. Rao, P. Mitra, S. R. M. Prasanna, and C. Busch, "Audio-visual biometric recognition and presentation attack detection: A comprehensive survey," *IEEE Access*, vol. 9, pp. 37431–37455, 2021.
[19] C. McCool and S. Marcel, "Mobio database for the ICPR 2010 face and speech competition," Idiap, Tech. Rep., 2009.
[20] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Pers. Authentication*, vol. 964, 1999, pp. 965–966.
[21] A. Misra and J. H. L. Hansen, "Spoken language mismatch in speaker verification: An investigation with NIST-SRE and CRSS bi-ling corpora," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2014, pp. 372–377.
[22] P. Motlicek, L. E. Shafey, R. Wallace, C. McCool, and S. Marcel, "Bi-modal authentication in mobile environments using session variability modelling," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 1100–1103.
[23] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7184–7193.
[24] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
[25] J. Ortega-Garcia, J. Fierrez, F. Alonso-Fernandez, J. Galbally, M. R. Freire, J. Gonzalez-Rodriguez, C. Garcia-Mateo, J. L. Alba-Castro, E. Gonzalez-Agulla, E. Otero-Muras, and S. Garcia-Salicetti, "The multiscenario multienvironment biosecure multimodal database (BMDB)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1097–1111, Jun. 2010.
[26] O. Parkhi, A. Vedaldi, and A. Zisserman, *Deep Face Recognition*, vol. 1. British Machine Vision Association, Jan. 2015, pp. 41.1–41.12.
[27] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database (release 1.00)," in *Audio- and Video-based Biometric Person Authentication*, J. Bigün, G. Chollet, and G. Borgefors, Eds. Berlin, Germany: Springer, 1997, pp. 403–409.
[28] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. D. Prince, "Probabilistic models for inference about identity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 144–157, Jan. 2012.
[29] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Comput. Surv.*, vol. 50, no. 1, pp. 1–37, Apr. 2017.
[30] R. Ramachandra, M. Stokkenes, A. Mohammadi, S. Venkatesh, K. Raja, P. Wasnik, E. Poiret, S. Marcel, and C. Busch, "Smartphone multimodal biometric authentication: Database and evaluation," 2019, *arXiv:1912.02487*.
[31] A. Rattani and R. Derakhshani, "A survey of mobile face biometrics," *Comput. Elect. Eng.*, vol. 72, pp. 39–52, Nov. 2018.
[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, N. Nassir, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.
[33] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.
[34] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech*, Sep. 2015, pp. 1–5.
[35] C. Sanderson, "The vidtimit database," IDIAP, Tech. Rep., 2002.
[36] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *Advances in Biometrics*, M. Tistarelli and M. S. Nixon, Eds. Berlin, Germany: Springer, 2009, pp. 199–208.
[37] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[38] A. F. Sequeira, J. C. Monteiro, A. Rebelo, and H. P. Oliveira, "Mobbio: A multimodal database captured with a portable handheld device," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, vol. 3, Jan. 2014, pp. 133–139.
[39] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, Apr. 2018, pp. 5329–5333.
[40] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," *Odyssey*, vol. 2016, pp. 283–290, Jun. 2016.
[41] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. Aik Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," 2019, *arXiv:1904.05441*.
[42] T. Chen, "Audiovisual speech processing," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 9–21, Jan. 2001.
[43] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.
[44] H. Wu, S. Zheng, J. Zhang, and K. Huang, "GP-GAN: Towards realistic high-resolution image blending," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2487–2495.
[45] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8261–8265.
[46] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

HAREESH MANDALAPU received the M.Tech. degree in computer science from the University of Hyderabad, in 2015, and the M.S. degree in Erasmus Masters CIMET from Université Jean Monnet, France, in 2017. He is currently pursuing the Ph.D. degree in information security and communication technology from the Norwegian University of Science and Technology, Gjøvik, Norway. His research interests include audio-visual biometrics, presentation attack detection, and multilingual speaker recognition.

P. N. ARAVINDA REDDY received the M.Tech. degree in signal processing from Visvesvaraya Technological University, Belgaum, in 2014. He is currently pursuing the Ph.D. degree with the Advanced Technology Development Centre, IIT Kharagpur, Kharagpur, West Bengal, India. His research interests include automatic speech recognition, audio-visual biometrics, and presentation attack detection.

RAGHAVENDRA RAMACHANDRA (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from the University of Mysore, Mysore India and Institute Telecom, and Telecom Sudparis, Evry, France (carried out as a collaborative work), in 2010. He was a Researcher with the Istituto Italiano di Tecnologia, Genoa, Italy, where he worked with video surveillance and social signal processing. He is currently appointed as a Full Professor with the Institute of Information Security and Communication Technology (IIK), Norwegian University of Science and Technology (NTNU), Gjøvik, Norway. He has authored several articles and is a reviewer for several international conferences and journals. He also holds several patents in biometric presentation attack detection. His main research interests include deep learning, statistical pattern recognition, data fusion schemes, and random optimization, with applications to biometrics, multimodal biometric fusion, human behavior analysis, and crowd behavior analysis. He has received several best paper awards. He was/is also involved in various conference organizing and program committees and serving as an associate editor for various journals. He was/is participating (as a PI/Co-PI/contributor) in several EU projects, IARPA USA, and other national projects. He has served as an editor for ISO/IEC 24722 standards on multimodal biometrics and an active contributor for ISO/IEC SC 37 standards on biometrics.

KROTHAPALLI SREENIVASA RAO (Member, IEEE) received the B.Tech. degree in electronics and communication from the RVR College of Engineering, in 1990, the Ph.D. degree from the Department of Computer Science and Engineering, IIT Madras, Chennai, India, in 2004, and the M.E. degree in communication systems from the PSG College of Technology, Coimbatore, India, in 2006. Currently, he is working as a Professor with the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, West Bengal, India. He has supervised seven Ph.D. students and 14 M.S. students (by research) in different issues related to speech processing.

PABITRA MITRA (Member, IEEE) received the B.Tech. degree in electrical engineering from IIT Kharagpur, Kharagpur, West Bengal, India, in 1996, and the Ph.D. degree from the Department of Computer Science and Engineering, Indian Statistical Institute, Kolkata, India, in 2005. Currently, he is working as a Professor with the Department of Computer Science and Engineering, IIT Kharagpur. He has supervised eight Ph.D. students and 12 M.S. students (by research) in different issues related to AI and machine learning.

S. R. MAHADEVA PRASANNA (Member, IEEE) received the B.Tech. degree in electronics and communication from SSIT Tumakuru, Tumakuru, Karnataka, India, in 1994, the M.Tech. degree in industrial electronics from NIT Surathkal, Surathkal, Karnataka, in 1997, and the Ph.D. degree from the Department of Computer Science and Engineering, IIT Madras, Chennai, India, in 2004. Currently, he is working as a Professor with the Department of Electrical Engineering, IIT Dharwad, Dharwad, Karnataka. He has supervised 13 Ph.D. students in different issues related to speech processing.

CHRISTOPH BUSCH (Senior Member, IEEE) is currently a member of the Department of Information Security and Communication Technology (IIK), Norwegian University of Science and Technology (NTNU), Norway. He also holds a joint appointment with the Faculty of Computer Science, Hochschule Darmstadt (HDA), Germany. Furthermore, he has been a Lecturer of biometric systems with the Technical University of Denmark (DTU), since 2007. He has coauthored more than 400 technical articles and has been a speaker at international conferences. He is also a Convenor of WG3 in ISO/IEC JTC1 SC37 on biometrics and an Active Member of CEN TC 224 WG18. Furthermore, on behalf of Fraunhofer, he chairs the Biometrics Working Group of the TeleTrusT Association and the German Standardization Body on Biometrics (DIN-NIA37). He served for various program committees, such as NIST IBPC, ICB, ICHB, BSI-Congress, GI-Congress, DACH, WEDELMUSIC, and EUROGRAPHICS, and served for several conferences, journals, and magazines as a Reviewer, such as ACM-SIGGRAPH, ACM-TISSEC, the IEEE Computer Graphics and Applications, the IEEE Transactions on Signal Processing, the IEEE Transactions on Information Forensics and Security, the IEEE Transactions on Pattern Analysis and Machine Intelligence, and the *Computers & Security* journal (Elsevier). He is also an Appointed Member of the Editorial Board of the *IET Biometrics* journal and the IEEE Transactions on Information Forensics and Security journal.