Fraud Detection in Voice-based Identity Authentication Applications and Services

Saeid Safavi, Hock Gan, Iosif Mporas and Reza Sotudeh
Information Engineering and Processing Architectures Group
Electronics Communications and Electric Division
School of Engineering and Technology, University of Hertfordshire
Hatfield, AL10 9AB, United Kingdom
s.safavi@herts.ac.uk

***Abstract*—Keeping track of the multiple passwords, PINs, memorable dates and other authentication details needed to gain remote access to accounts is one of modern life's less appealing challenges. The employment of a voice-based verification as a biometric technology for both children and adults could be a good replacement to the old fashioned memory dependent procedure. Using voice for authentication could be beneficial in several application areas, including, security, protection, education, call-based and web-based services. Voice-based biometric applications are subject to different types of spoofing attacks. The most accessible and affordable type of spoofing for a voice-based biometrics system is a replay attack. Replay, which is to playback a pre-recorded speech sample, presents a genuine risk to automatic speaker verification technology. This work presents two architectures for detecting frauds caused by replay attacks in a voice-based biometrics authentication systems. Experimental results confirmed that obtained performances from both methods could further improve by applying a machine learning algorithm for performing fusion at the score level. The performance of both methods further improved by fusion using independent sources of scores in different architectures.**

Keywords—voice biometrics; speaker verification; spoofing; counter-measure; machine learning.

I. INTRODUCTION

Over the last decade the use of services and applications over the web have become increasingly popular. Examples are social networking sites, e-banking, online shopping, e-meeting and online management systems. The social networking sites are most popular with teenagers and young adults.Almost half of children aged from 8 to 17 who use the internet set up their own profile on a social networking site [1,2]. Due to the extensive use of these applications and services as well as the need for security and convenience of the users, biometrics are used by the users to log on and use the provided services. The most widely used biometrics in such online applications and services are face detection, iris scan and voice [3].

Voice is considered as one of the most convenient biometric interface for the user, especially when accessing data through telephone networks, the web, including social media applications on the move. Instead of typing passwords with memorable dates and other authentication details the user's voice print itself is used to verify the identity of the claimed user, and thus conveniently provide secure access to data, information and social network services. Since the number of applications and services is growing, voice-authentication is a must for the convenience of the user.

The technology in voice based biometrics has now reached a mature level. Measurement of the speaker verification performance has been standardized [4] and the recent technological advancements in the areas of signal processing, machine learning and linear algebra has resulted to high performing voice biometric systems not only in research level [5,6,7], but also in commercial applications [8]. Modern speaker verification engines use statistical speaker-dependent models, such as mixtures of Gaussian distributions [9,10] and speech signal representations using super-vectors or more compact representations like i-vectors [7,11] combined with powerful machine learning algorithms for classification [12]. Despite the tremendous progress in the speaker verification task and the achieved high verification scores there still is space for improvement when the system or application is under a spoofing attack, i.e. robust fraud detection methodologies are essential.

In voice-based identity authentication, fraud is considered as the attempt of an impostor speakerto get access to the application or service. Fraud in voice-based identity authentication interfaces is mainly performed with replay attacks, synthetic speech, voice conversion technologies, or their combinations. In replay attacks, recordings of the target speaker are used to cheat the identity authentication interface by replaying the audio recording from the target speaker. In speech synthesis attacks, text-to-speech (TTS) engines are used to create synthetic speech of the target speaker, while in voice conversion (VC) attacks a speech sample of the impostor speaker is processed by a voice conversion engine in order to produce synthetic speech similar to the acoustic characteristics of the target speaker.

In real-life voice-based identity authentication interfaces, the speech input provided by the user may vary. In detail, the interface can operate in fixed pass-phrase, text-dependent or text-independent mode of operation. In the fixed pass-phrase mode the user is

This work was supported in part by the OCTAVE H2020 Project entitled "Objective Control for TAlker VErification" funded by the EC with Grand Agreement number 647850.

asked to provide a pass-phrase, which has been setup and stored in his/her user profile. This mode of operation achieves high speaker verification performance (the target speaker models are trained with the specific pass-phrases) but is also highly vulnerable to replay attacks, since an impostor could record the voice of the target speaker voicing the fixed pass-phrase. In the text-dependent mode of operation, the user is prompted with a text message (a short utterance) from a list of pre-selected utterances to read. The text-dependent mode of operation achieves less speaker verification performance but is more robust to spoofing attacks comparing to the fixed-passphrase mode. In the text-independent mode, the user is prompted a text message produced by a random text generator. In this case the speaker models are not trained with the same test utterances, and thus achieve lower performance. However, when using the text-independent prompts mode along with an utterance verification engine, replay attacks are practically impossible, because synthetic speech (either from TTS or from VC engines) cannot beprecisely fitted with the acoustic characteristics of the target speaker, and thus this mode is less vulnerable to fraud (mainly replay attacks).

In real applications with voice-based identity authentication interfaces the detection of fraud is a must. In this paper we present two methodologies for detecting frauds in voice-based user authentication services. The methodologies are based on statistical modeling and classification of genuine and impostor speech as well as on modeling of the differences between genuine and impostor speakers in the score distributions from different modes of operation for speaker verification.

The remainder of this paper is organized as follows. In Section II we present the fraud detection methodologies. In Section IIIcontains the experimental results. Finally, Section IV summaries finding of this work.

II. FRAUD DETECTION METHODOLOGIES

In this Section we present two methodologies for fraud detection. The first methodology is based on the classification of genuine and replayed speech using a statistical modeling approach. The second methodology relies on the differences in terms of performance and robustness of different modes of operation for speaker verification.

A. Fraud detection using direct binary modeling

The first proposedfraud detection methodology consists of two stages; (i) a general verification stage that estimates the likelihood of a speech frame being a genuine speaker or a replay attack, and (ii) a fusion stage which does a low-level fusion of the results from two independent statistical engines (Gaussian Mixture Model – GMM classification and Hidden Markov Model – HMM classification). The GMM system is identical to [15], and 128 Gaussian mixtures are used to fit to the 57 dimensional (19 MFCC+19Δ+19ΔΔ) feature vectors.

In the HMM-UBM [16] system the feature input was the same as the GMM-UBM system. The 24 state HMM(including the initial state and the emitting observations, 8mixtures per state, and diagonal covariance matrices)is trained without any speech transcriptions,where a dummy word (e.g. ‘HELLO’) is assigned (forced) as a label for all training data. TheHMM-UBM is initialized with a flat start and then its parameters are re-estimated with fewiterations of the Baum-Welch algorithm. During the training phase, target speaker models arederived from the HMM-UBM with three iterations of MAP using their respective training data. Inthe test phase, test utterances are forced aligned against the claimant and HMM-UBM for loglikelihood ratio.

Fig.1 shows the setup of the models required for the general verification stage. Two models have to be set up; the genuine speech model is a compact representation of a genuine speech frame from a speaker, and the replay speech model is a representation of replay attacks. Fraud detection is described here as a binary classification problem where the system indicates if the claimant is a real speaker or a replay of the speaker’s voice.

As Fig.1 shows, the voice input is captured and segmented into frames and in turn a feature extraction process converts the speech frames into compact speech signatures represented by feature vectors. These feature vectors in turn are compacted even further by machine learning algorithms that fit the vectors to a closest match of a parametric statistical model. The models, simply put, are generated by an iterative process of matching features with the model and adjustment of the statistical parameters. Samples from different speakers or replay speech files are used to refine the two models respectively.

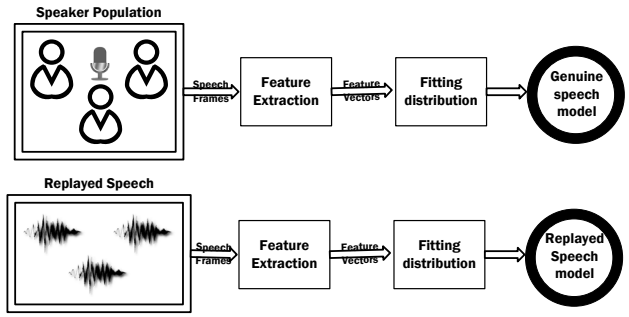

Fig.1. The architecture for the training phase for the models to be used in a general fraud (replay attacks) detection process for voice-based biometrics engines (Proposal 1).

Having generated the two models, the general verification stage can be constructed. Figure 2 shows the general fraud detection stage being used to train a fusion model. The input of the general detection stage consists of speech frames which can come from either a genuine speaker or a replay speech file. The output of the fraud detection stage is produced by a likelihood estimator which produces a likelihood ratio that compares the likelihood of the received speech frame being associated with a genuine speaker or a replay-speech file.

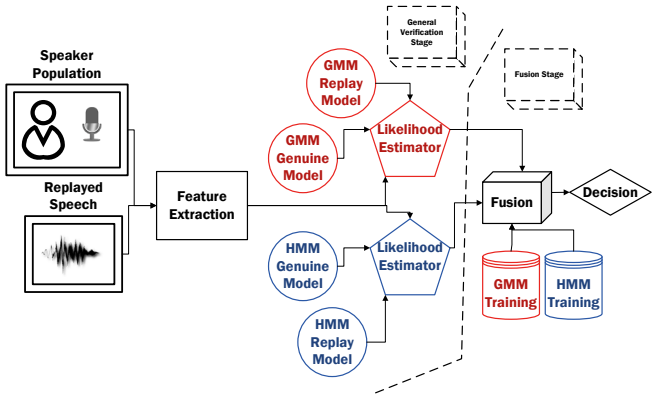

Fig.2. The architecture for training and evaluating a fusion model (Proposal 1).

The core of the fusion stage is a fusion model, shown in Figure 2. The fusion model is a supervised learning system and in order to train the model, databases are initially set up that contain the training data for the model. The number of databases depend on the number of independent statistical engines to be investigated or used for fusion. In this case, there are two databases; one for the outputs (scores) of a GMM engine and another for an HMM engine. A variety of replay speech frames and target speaker frames are used to generate likelihood scores that are stored in their respective databases. Each database holds a record of the scores and their associated origin (which is either a genuine speaker or a replay speech file). The fusion model is trained to map a two-value set to a single decision score; one value from the GMM database and the other from the HMM database. The score values are outputs generated by the same input source (be it a genuine speaker or a replay speech file). The fusion model distinguishes the genuine speakers from the spoof files based on the acceptance score it produces.

B. Fraud detection using fusion of scores from different modes of speaker verification operation

Fig.13 shows a setup for a typical speaker verification system which is used as a component of the hybrid speaker verification system, i.e. a system that combines different modes of speaker verification operation. A Universal Background Model (UBM) is trained using all speech files from the TIMIT corpora [13]. The background model is used as a template for generating individual speaker models, which is a compact representation of specific speaker identities, by maximum a posteriori (MAP) adaptation using speaker specific enrolment data. Speaker verification is a two class problem where the system indicates if the claimant is a specific speaker or "somebody else" and the background model also represents a universal speaker which is the "somebody else". A

number of distinct speaker models are constructed out of the AVSpoof database and they form sets of speaker models that are distinguished by gender and the way the speech phrase is formulated in different operational modes. The protocols and trials sets are designed for three operational modes; namely the fixed-phrase (mode A), text dependent (mode B) and text independent (mode C) modes.

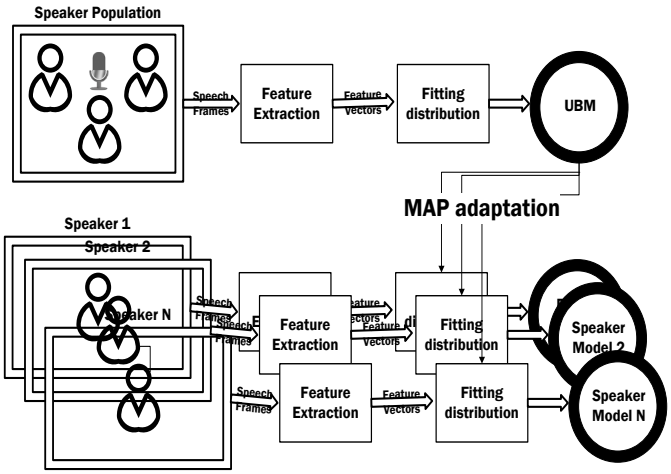

Fig.3. The architecture for the training phase of a bio-metric system is shown (Proposal 2). UBM and MAP stand for Universal Background Model and Maximum A Posteriori, respectively.

Figure 3illustratesspeech signal frame blocksand a feature extraction process which converts the speech frames into compact speech signatures represented by feature vectors. These feature vectors in turn are used by machine learning algorithms that fit the vectors to a closest match of a parametric statistical model. The background model is trained by a representative sample of the population distinct from the individual speaker models which come from a different population. Different patterns of speech phrases used in training (operational modes) result in separate models and the figure therefore depicts the training of any operational mode in general and the difference would be the content of the training materials.

The fusion model has to be trained with likelihood ratios from a variety of replayed speech frames and genuine speech frames against that from a speaker model trained on genuine speaker's speech. The likelihood ratios indicate the likelihood of a match of the spoofed frame with the claimed speaker model. The scores from all three operational modes are used to train the fusion model and form the fusion process. The trained fusion model maps a set of three inputs (one from each mode) to a single fused decision score which is used to determine if the inputs represent a genuine or replayed input voice signal.

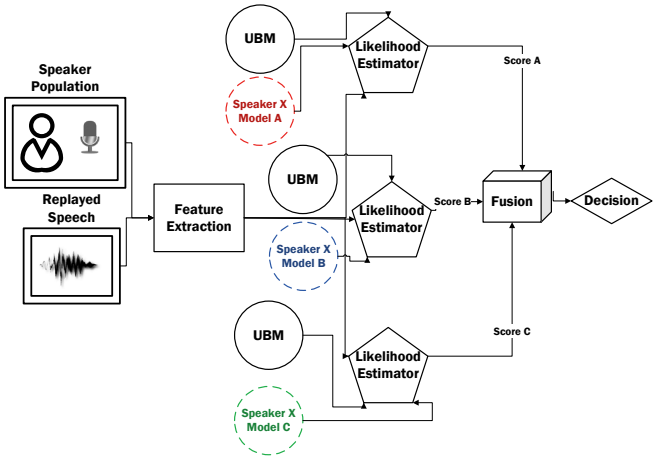

Fig.4. The architecture of the hybrid speaker verification methodology for fraud detection (Proposal 2).

Figure 4illustrates the architecture of a hybrid biometric speech verification system. The speech frames from a claimant has to be converted in order to carry out a matching process with the model of speaker that the claimant claims to be. The claimant is prompted to provide inputs for three operational modes (mode A, B and C) so that more data from diverse sources can be used to

detectreplayed inputs. The fusion counter-measure exploits the strengths of both security and accuracy provided by each individual operational mode with the ability of the system to detect significant differences between spoofed and genuine inputs. The fusion system assumes that the attacker would not have the resources to provide strong attacks on all three operational modes at the same time and relies on a mixture of weak and strong attacks in different modes that will play into the strengths of the system. In reality the probability that the fake user can replay an appropriate input for all three modes is very low and it is almost impossible. The use of the strongest and most accurate types of attackswas used to investigate the worst-case scenario.

Figure 5illustrates the usage of scores from the development and evaluation sets for the training and testing of the hybrid countermeasure architecture for the second proposal. In order to follow the state of the art procedure for fusion we have used two separate sets, development and evaluation, to make sure that none of the speakers from the development set appeared in the evaluation sets.

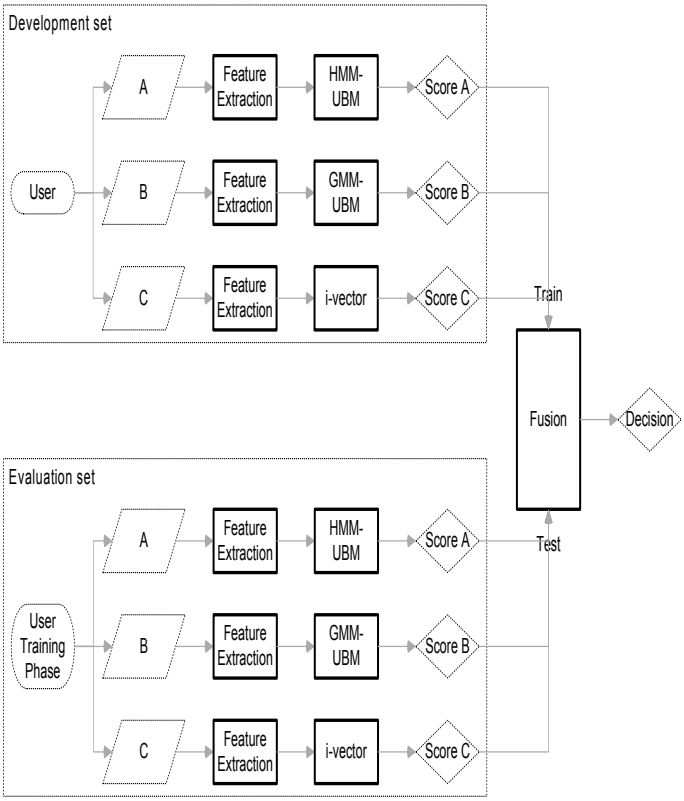

Fig. 5. Train and test input to fusion model (Proposal 2).

Fig. 6 illustrates the flowcharts of the stages of training the fusion model. The fusion model is composed of a machine learning system that maps the likelihoods of a claimant belonging to the speaker model in three operational modes to a score that determines if the claim is valid. The fusion model is a supervised learning system and a database is initially set up that contains the training data for the model. The left flowchart shows the database set up. This is a record of the results of a general speaker verification process for a variety of inputs which is labelled with the intended result. The labels are only of interest to the fusion model which has to learn which results belong to genuine inputs and which donot. Once the training data of all claimants have been processed, the database is ready to be used by the next stage which is described by the flowchart on the right. One database is established for each operation mode and once the three databases have been set up, the next stage can begin. The fusion model is trained with a specific machine learning algorithm and the threshold which separate the output scores is established. This system is then ready for evaluation.

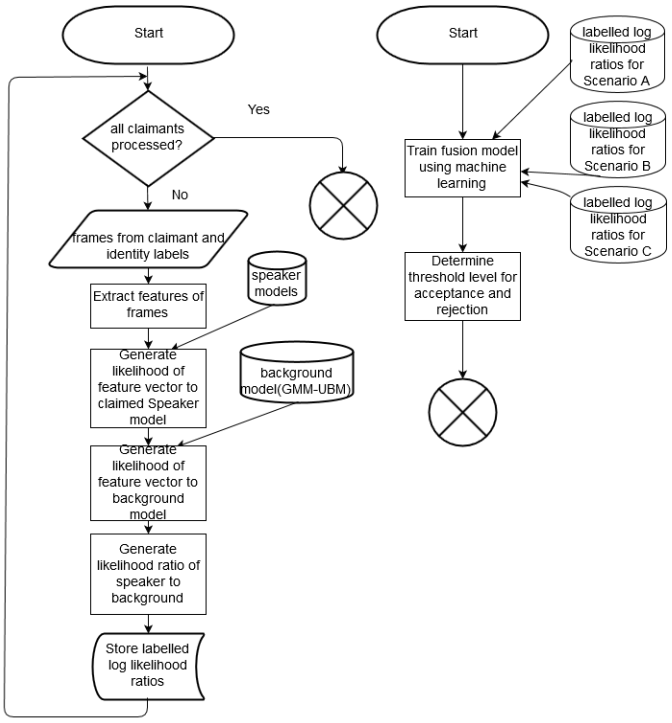

Fig.6. Flowcharts for training the fusion model are shown. Two processes are shown; the one on the left describes the flow of setting up the fusion database for a specific operational mode or scenario, the one on the right describes how all the databases are used to train a fusion model using machine learning. GMM-UBM is used as an example of a Background model.

III. Experimental Setup and Results

This section contains information about the speech corpus which were used for experiments, definition of metrics for performance evaluations, and results.

A. Sppech corpus

The AVspoof [14] database is based upon a standard database consisting of both genuine and spoofed speech. It contains speech data from 31 male and 13 female speakers. In order to use the data, we need to divide the data of each gender into three subsets, but as the number of female speakers is small, we only considered male speakers for the experiments in this report.

All the data from the 31 male speakers were divided into three sections, called Counter Measure (CM), Development and Evaluation sets, and each of these sets contain training and testing subsets. The data for the first 10 speakers (m001 to m013, i.e., first 10 when sorted alphabetically) werenot used in the experiment but were used for training the countermeasure models in proposal 1. The Development and Evaluation set had data for10 (m014 to m030) and 11 (m033 to m044) speakers, respectively and they have been used for proposal 2.

The CM set (which were used to train two models for genuine recorded and replayed speech in Proposal 1)hadonly training and testing protocols. None of the speakers in training appeared in the test set of CM set. On the other hand each of evaluation and development set (Proposal 2)hadthree sub protocols attached to them, one for each of the three operational modes of speaker verification. Themodes were, Fixed-pass phrase (A), Text-prompted text-dependent (B), Text-prompted text-independent (C). The CM test trial contained 2204 instances. Statistics for the CM set are tabulated in Table1.

In the following subsections the division of evaluation and development sets for three different modes of speaker verification operation is explained.

1) Fixed-pass phrase (A)

Fixed-pass phrase protocols weredesigned such that each target speaker hadfive different phrase-specific models created separately for five pass phrases (Sentence01-5). For training only, one speech segment was recorded to form a pass phrase using a laptop and wasused for each enrolment. During the test phase each test phrase wasscored against the trained segment model that containedthe same text from the same speaker (referred toas target trials) and different speaker (non-target trials) and the tests were carried outusing a laptop, phone1 and phone2.

Two types of non-target trials are indicated in Table I as, zero-effort spoofed (denoted as 'non-target genuine') and replay spoofed (denoted as 'non-target replay'). These relate to inputs from genuine speakers that claim to be someone else and inputs from replay files respectively.

TABLE I. TRIAL STATISTICS FOR COUNTER MEASURE (CM) SET AND THREE MODES OF OPERATION OF SPEAKER VERIFICATION, MODE A, B AND C.

Statistics	**Data sets**				
	CM	***Mode A***		***Mode B & C***	
		Eval.	***Dev.***	***Eval.***	***Dev.***
Target models	2	55	50	11	10
Target trials (genuine)	945	494	450	494	450
Non-target trials(replay spoofed)	1260	600	600	600	600
Non-target trials (zero-effort)	-	4940	4050	4940	4050

2) Text-prompted text-dependent and text-independent (B), (C)

For mode (B) each target model wasenrolled with data from five pass phrases of session 1, while in Mode (C) each target model wasenrolled with first five sentences of free speech (free01.wav to free05.wav), recorded using a laptop. But for both modes, B and C, are sharing a unique test trial list. The trial statistics for mode B and C is tabulated in Table I.

B. Metrics

The metric used for the evaluation of the performance of a variety of spoofing conditions and background models in this work is the 'threshold-free' equal error rate (EER). Let $P_{fra}(\theta)$ and $P_{frr}(\theta)$ denote the false acceptance rate and false rejection rate at threshold θ, respectively:

$$P_{far}(\theta) = \frac{\#\{spoof\ trials\ with\ score > \theta\}}{\#\{total\ spoof\ trials\}}$$

EQUATION 1. FALSE ACCEPTANCE RATE (FAR)

$$P_{frr}(\theta) = \frac{\#\{genuine\ trials\ with\ score \leq \theta\}}{\#\{total\ genuine\ trials\}}$$

EQUATION 2 – FALSE REJECTION RATE (FRR)

$P_{far}(\theta)$(which is the probability that a spoof is incorrectly accepted) and $P_{frr}(\theta)$ (which is the probability that a genuine speaker is incorrectly rejected) are monotonically decreasing and increasing functions of θ respectively. The EER is computed as the point θ_{EER} where $P_{far}(\theta)$ and $P_{frr}(\theta)$ are equal: $EER = P_{far}(\theta_{EER}) = P_{frr}(\theta_{EER})$

C. Results using direct binary modeling method

Table II contains the obtained performance from different CM setup in terms of EER. Their corresponding DET-Plotis plotted on Figure 7.

The result confirms that the HMM-UBM system outperforms GMM-UBM and the usage of fusion algorithm does not always improve the performance.Only in the case of LR was there a small gain of 2.27% and 0.38% relative and absolute improvement with respect to the performance of HMM-UBM system and 20.53% and 4.23% relative and absolute improvement compare to the performance obtained using GMM-UBM system.

TABLE II. EER IN (%) ON SPOOF DETECTION TASK USING SINGLE AND FUSEED ARCHITECTURES.

Mode of Operation	Approach	EER
Single mode scores	GMM-UBM	20.6
	HMM-UBM	16.75
Score level fusion	Linear regression	16.37
	Multi-layer perceptron	17.15
	Support vector regression	16.95

The poor performance of SVR and MLP seems to be related to the small number of available test instances in the CM set trial list. There were only 2204 test trial instances and in order to be able to compare the performance with the single system scores the 3-fold cross validation setup has been used. In this setup the verification step was repeated 3 times; at each instancetwo folds wereused for training and one for testing. So in this setup for each set of verification only 2/3 of the 2204 instances were used for training the fusion parameters, and this could be the reason thatLR performed better than MLP and SVR.

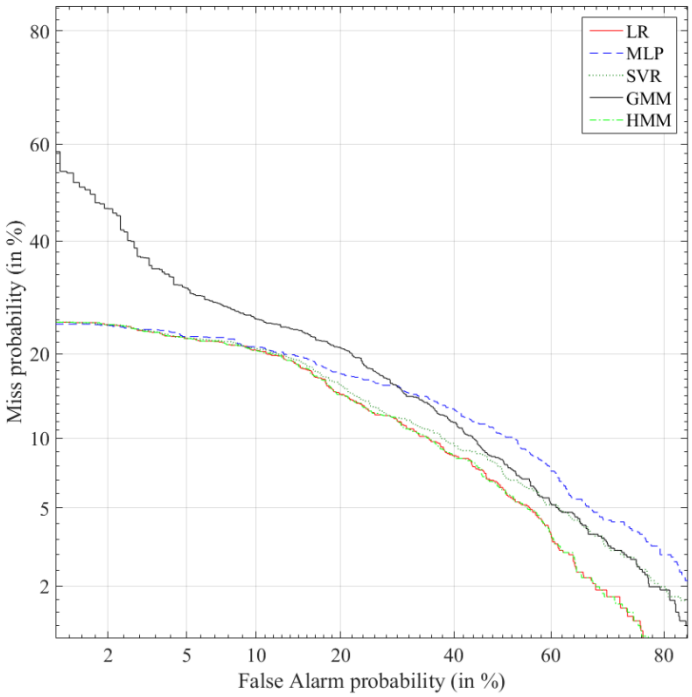

Fig. 7. DET-plot from CM set of AVspoof corpus for different CM architectures.

D. Results using fusion of scores from different modes

Table III shows the equal error rates obtained during replay attacks within a single operation mode or Scenario. The replay attack uses recorded phrases of genuine speakers that correspond to the mode of the evaluation. For example, A' is a replay of the claimed speaker uttering the fixed-phrase used for mode A. For this architecture we relied on the GMM-UBM modelling approach. Looking down the columns of the table, it is noted that the EER increases for Scenarios A, B and C respectively. The EER for Scenarios A and B are close and a jump is noted for Scenario C. There is a higher complexity in the data to be learnt by the models as the Scenarios progress from A to C. The EER is expected to rise for normal verification in those cases. If the replay attack is regarded as "noisier" versions of the genuine speaker, then the increase in EER is also expected for replay attacks.

If we examine the actual EER values presented for GMM, they come close to 50% (41%, 45%) with Scenario C exceeding 50% (53%). This suggests that GMM is ineffective on its own against replayattacks for each single mode of speaker verification.

Table IV shows the performance of different fusion algorithm for fusion of the scores from different modes of speaker verification operation. All three approaches obtained performance improvements comparable to all single modes of speaker verification engine. The best performance was obtained using the multi-layer perceptron approach, which was 35.3%. Compared to the performance obtained by mode A, fusion results in a relative and absolute improvement of 12.84% and 5.20% in terms of EER, respectively.

TABLE III. THE TABLE SHOWS THE EQUAL-ERROR RATE (%) PERFORMANCE OF REPLAY ATTACK DETECTION IN DIFFERENT OPERATIONAL MODES. THE USE OF GMM-UBM WERE INVESTIGATED.

Mode of Operation	Replay Attack	GMM-UBM
Fixes pass-phrase (A)	A’	40.5
Text-dependent prompt (B)	B’	45.1
Text-independent prompt (C)	C’	53.2

Fig. 8 shows the DET curves that corresponds to the performance of fraud detection in each single mode of speaker verification operation and the score level fusion of operational modes.

TABLE IV. THE TABLE SHOWS THE PERFORMANCE OF A HYBRID-BIOMETRIC ARCHITECTURE IN TERMS OF THE EQUAL-ERROR RATE (%).

Fusion algorithm	Replay Attack/mode	GMM-UBM
Linear regression	A’,B’,C’/A,B,C	37.4
Multi-layer perceptron	A’,B’,C’/A,B,C	35.3
Support vector regression	A’,B’,C’/A,B,C	37.1

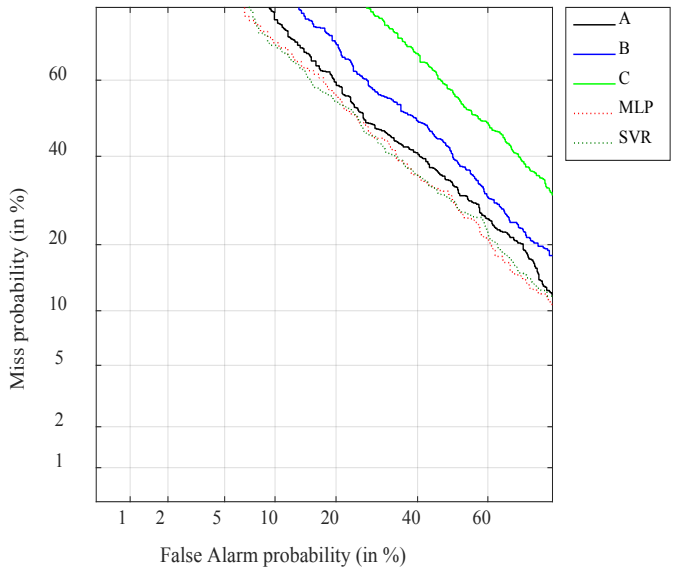

Fig. 8. Spoofing detection performance using different speaker verification operational modes and their fusion.

IV. Conclusion

In this study, we evaluate the vulnerability of different modes ofoperation of speaker verification systems under replay attacks using a standard benchmarking database, and also propose an anti-spoofing technique to safeguard the voice-based authentication services.The work introduces the use of two biometric engines as a means of providing inherent robustness. Two proposed methods are first, a method based on the direct modelling of genuine and spoofed speech, and second one using scores from different modes of operation of speaker verification task to detect fraud in the voice-based biometric authentication systems.The first solution uses the combined scores fromtwo independent sources (GMM and HMM) to provide information that highly reduces the effectiveness of spoofing in general. The notion is based on the assumption that the strengths of each model will make up for the individual weaknesses of each. The analogy in this case would be the expectation that somebody with the mind of a genius and the body of a super athlete would be able to win the Nobel Prize and get an Olympics medal at the same time whereas if they were stuck with their original attributes, they would only be able to attain a single notable achievement but not both.

The second solution uses the combined scores of different modes of speaker verification operation to provide information that highly reduces the effectiveness of spoofing in general. The notion is based on the assumption that an attacker will not have the resources to develop a spoof of multiple systems. The analogy to bio-modal systems is that it may be possible to develop measures to fool a finger-printing system but if the system is capable of uniquely associating the fingerprint, voiceprint, retinal scan and facial recognition of a person, the chances of spoofing the identity of the person is much reduced.

Both methods improved the performance of the fraud detection task in the voice-based biometrics authentication systems. The experiments conducted on the AVSpoof database and performance of each architecture presented in terms of the equal error rate (EER) and illustrated on DET-plots. The first proposed fraud detection architecture could be used in wide range of applications, while the second proposed method is specifically designed for sensitive applications, for example online banking which needs high level of security.

Acknowledgment

The authors would like to thank Professor Aladdin Ariyaeeinia,,Dr Tomi Kinnunen,Dr Md Sahidullah, and Achintya Sarkar for their support in this work.

References

[1] Anonymous "Engaging with social networking sites," vol. 2011.

[2] S. Safavi, M. Najafian, A. Hanani, M. Russell, P. Jan˘covi˘c, andM. Carey, “Speaker recognition for children’s speech,”Inter-speech, 2012.

[3] Jain A, Flynn P, Ross AA, editors. Handbook of biometrics. Springer Science & Business Media; 2007 Oct 23.

[4] Doddington GR, Przybocki MA, Martin AF, Reynolds DA. The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective. Speech Communication. 2000 Jun 30;31(2):225-54.

[5] Kenny, P., Boulianne, G., Ouellet, P. and Dumouchel, P., 2007. Joint factor analysis versus eigenchannels in speaker recognition. Audio, Speech, and Language Processing, IEEE Transactions on, 15(4), pp.1435-1447.

[6] Campbell, W.M., Sturim, D.E. and Reynolds, D.A., 2006. Support vector machines using GMM supervectors for speaker verification. Signal Processing Letters, IEEE, 13(5), pp.308-311.

[7] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P., 2011. Front-end factor analysis for speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on, 19(4), pp.788-798.

[8] Connaughton R, Sgroi A, Bowyer KW, Flynn P. A cross-sensor evaluation of three commercial iris cameras for iris biometrics. InCVPR 2011 WORKSHOPS 2011 Jun 20 (pp. 90-97). IEEE.

[9] S. Safavi, A. Hanani, M. Russell, P. Jancovic and M. J. Carey, "Contrasting the Effects of Different Frequency Bands on Speaker and Accent Identification," in IEEE Signal Processing Letters, vol. 19, no. 12, pp. 829-832, Dec. 2012.

[10] Safavi, Saeid, Maryam Najafian, Abualsoud Hanani, Martin J. Russell, Peter Jancovic, and Michael J. Carey. "Speaker Recognition for Children's Speech." In INTERSPEECH, pp. 1836-1839. 2012.

[11] H. Sun, K. A. Lee and B. Ma, "A new study of GMM-SVM system for text-dependent speaker recognition," Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, South Brisbane, QLD, 2015, pp. 4195-4199.

[12] Byun H, Lee SW. Applications of support vector machines for pattern recognition: A survey. InPattern recognition with support vector machines 2002 (pp. 213-236). Springer Berlin Heidelberg.

[13] Zue V, Seneff S, Glass J. Speech database development at MIT: TIMIT and beyond. Speech Communication. 1990 Aug 31;9(4):351-6.

[14] Ergünay SK, Khoury E, Lazaridis A, Marcel S. On the vulnerability of speaker verification to realistic voice spoofing. InBiometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on 2015 Sep 8 (pp. 1-6). IEEE.

[15] D. Reynolds, T. Quatieri and R. Dunn, “Speaker verification using adapted gaussian mixture models,” Digital Signal Process, 2000.

[16] L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,,” Proc. of the IEEE acoust., Speech, Signal Processing Mag., vol. 77, no. 2, p. 257–285, 1989.