

ORIGINAL PAPER

An analysis of speaker dependent models in replay detection

GAJAN SUTHOKUMAR,^{1,2}  KAAVYA SRISKANDARAJA,¹  VIDHYASAHARAN SETHU,¹ 
ELIATHAMBY AMBIKAI RAJAH^{1,2}  AND HAIZHOU LI³ 

Most research on replay detection has focused on developing a stand-alone countermeasure that runs independently of a speaker verification system by training a single spoofed model and a single genuine model for all speakers. In this paper, we explore the potential benefits of adapting the back-end of a spoofing detection system towards the claimed target speaker. Specifically, we characterize and quantify speaker variability by comparing speaker-dependent and speaker-independent (SI) models of feature distributions for both genuine and spoofed speech. Following this, we develop an approach for implementing speaker-dependent spoofing detection using a Gaussian mixture model (GMM) back-end, where both the genuine and spoofed models are adapted to the claimed speaker. Finally, we also develop and evaluate a speaker-specific neural network-based spoofing detection system in addition to the GMM based back-end. Evaluations of the proposed approaches on replay corpora BTAS2016 and ASVspoof2017 v2.0 reveal that the proposed speaker-dependent spoofing detection outperforms equivalent SI replay detection baselines on both datasets. Our experimental results show that the use of speaker-specific genuine models leads to a significant improvement (around 4% in terms of equal error rate (EER)) as previously shown and the addition of speaker-specific spoofed models adds a small improvement on top (less than 1% in terms of EER).

Keywords: Speaker Dependent Models, Replay Attack, Spoofing Detection, Speaker Verification, Speaker Adapted Neural Networks

Received 29 August 2019; Revised 27 February 2020

I. INTRODUCTION

Automatic speaker verification (ASV) is the process of verifying a speaker's identity based on their voice [1]. The remote nature of ASV systems makes them highly vulnerable to spoofing attacks, which aim to mimic the valid claimants. Spoofing attacks are a serious threat, where an attack could lead to a severe loss to credibility and significant financial costs.

Spoofing attacks, which can target a system before or after the microphone sensor, and are referred to as physical attacks and logical attacks, respectively [1]. Logical attacks need system-level access and are somewhat less of a threat and are not the main focus of this paper.

Spoofing attacks can be broadly divided into one of four different categories: impersonation, replay, speech synthesis, and voice conversion. Among them, replay attacks are known to be the simplest attack type and can be easily

initiated compared to the other three types [1]. Moreover, state-of-the-art ASV systems have been shown to be highly vulnerable to replay attacks [2].

Anti-spoofing countermeasures can either be stand-alone, in which case they work independently of an ASV system, or integrated, in which case they try to make the ASV system itself more robust to the spoofing attack. The integrated approach allows the use of shared sub-systems (i.e., front end, modeling techniques, etc.), which could be computationally efficient, while a standalone approach can operate independently without modifying the ASV system and also allows the use of different front-ends and modeling techniques. More importantly, an integrated approach allows for more information to be brought to bear on the spoofing detection problem (and possibly the ASV problem as well). In particular, an integrated approach can make use of prior knowledge about the claimed speaker's characteristics, which would be available to the ASV system, to detect potential spoofing attacks [3].

Few studies have investigated the integrated approach for speech synthesis, voice conversion, and replay spoofed speech [4–7]. In [4], this is achieved by introducing an additional common spoofed model into an ASV system and in [5, 6], i-vectors and a probabilistic linear discrimination analysis (PLDA) back-end are used to formulate a joint spoofing detection, and ASV system. The study on

¹School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, N.S.W. 2052, Australia

²Data61, CSIRO, Eveleigh, N.S.W. 2015, Australia

³Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583

Corresponding author:

Gajan Suthokumar

Email: g.suthokumar@unsw.edu.au

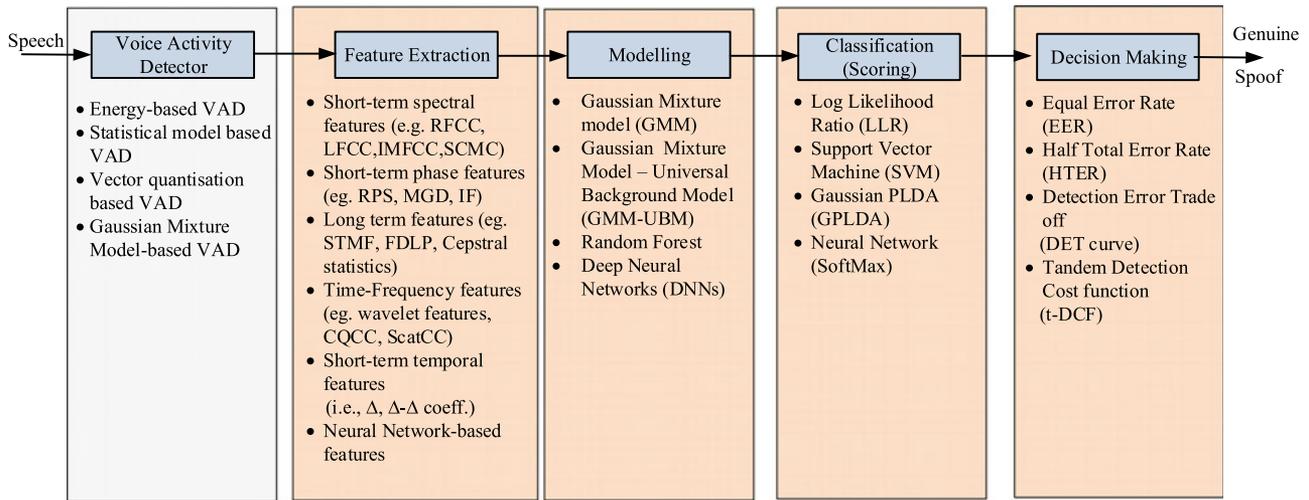


Fig. 1. Summary of principal components of a spoofing detection system. Here VAD block is often optional.

integrated spoofing detection with the use of shared single front-end for ASV and replay detection with score fusion has been reported [7].

A standalone replay detection system typically consists of a front-end and a back-end classifier to identify whether a given utterance is genuine or spoofed and the key components of a spoofing detection system are shown in Fig. 1. The most commonly employed front-ends in spoofing detection systems are typically based on spectral features, such as spectral centroid magnitude coefficient [8], constant-Q cepstral coefficient (CQCC) [9], single frequency filter cepstral coefficient [10], linear prediction cepstral coefficients, inverse-Mel cepstral coefficients [11], rectangular filter cepstral coefficients [8], scattering coefficients [11], spectral slope features [12], cochlear features [13, 14], and voice source features [15, 16]. Among them, uniform linear frequency filters have shown to be superior in capturing the replay artefacts than the human auditory inspired Mel scale and constant-Q scale filters [8].

In addition, a number of phase-based front ends, such as instantaneous frequency features [15, 17], frequency modulation [18], modified group delay (MGD) [19], and relative phase shift have also been investigated [19]. Finally, it can be observed from the literature that long-term features, such as long-term spectral statistics [20], frequency domain linear prediction features [21], and state-of-the-art spectro-temporal modulation features (STMF) [22] have also been successfully used in replay detection systems.

Among the classifiers that have been investigated, such as Gaussian mixture models (GMMs), support vector machines, PLDA, and random forest [23], GMM classifier remains the dominant back-end in replay detection [24]. Additionally, several variants of neural network architectures have also been investigated for use as the front-end [25, 26], the back-end [27, 28], and in an end-to-end [29] manner. Attention mechanisms [29, 30] and residual networks [31] have shown the most promise for replay detection because they help to emphasize salient regions of the

input and help the network to generalize well for the small amount of training data.

All the feature sets employed in spoofing detection also exhibit variability due to a number of other factors, such as acoustic variability (including channel effects), speaker variability, phonetic variability, etc. These sources of unwanted variability can subsequently lead to less effective models and reduce the accuracy of spoofing detection systems. To mitigate this, the unwanted variability can either be incorporated into the models or can be normalized. This is supported by recent work whereby the use of cepstral mean variance normalization (CMVN) improved the reliability of spoofing detection across the diverse variations in replay attacks [8, 24]. However, in our previous work, we have shown that the cepstral mean and variance have replay related information [20]. A less explored approach is the incorporation of this variability into the back-end models. Authors have studied the consequences of the phonetic variability in the spoofing detection system and proposed a framework to incorporate the phonetic variability into the system, which is showed highly beneficial [32] to the replay detection task.

The authors have previously demonstrated that incorporating speaker variability into the back-end in the form of speaker-specific genuine speech models with spectral feature front-ends, is highly effective and can significantly improve spoofing detection performance [3].

In this paper, we address the natural follow on question of whether further improvements in spoofing detection can be obtained by making both genuine and spoofed models speaker-specific. Specifically, this study is motivated by the observation that when the genuine model (of the feature distribution) is specific to the target speaker, it has less variability. Subsequently, *if the variability in the spoofed model is also reduced, by making it speaker specific, can further improvements in spoofing detection be obtained?*

In order to answer this question, here we: (a) attempt to quantify speaker variability; (b) investigate if speaker variability affects genuine and spoofed models differently;

(c) compare the discriminability of speaker-specific distributions of genuine and spoofed speech features and the corresponding speaker-independent (SI) distributions; and (d) quantify the improvement in spoofing detection performance for spectral and non-spectral feature front-ends, when incorporating speaker specific-spoofed models, one relying on a generative and another one on discriminative paradigm.

Finally, we also attempt to address the limitation that speaker specific spoofed data may not be available during system development by investigating the use of simulated spoofed data for training speaker specific models.

The rest of the paper is organized as follows: In section II, an analysis of speaker variability is provided. Then the two proposed speaker-dependent spoofing detection systems using the GMM and neural network back-ends are explained in sections III and IV, respectively. Section V describes the database preparation and section VI provides the details of front-end features. The key experiment settings and evaluation techniques are explained in section VII and the results and discussions are given in section VIII.

II. ANALYSIS OF SPEAKER VARIABILITY

The analyses of speaker variability presented in this section aimed to address the following: (a) Is speaker variability present in both genuine and spoof class models?; (b) Does speaker variability affect the genuine and spoofed models differently?; and (c) Are speaker-specific distributions of genuine and spoofed speech easier to distinguish than their corresponding SI distributions?

A) Visualizing the speaker variability in the feature space

Initially, we visualize the distribution of features from both genuine and spoofed speech corresponding to multiple speakers by projecting the feature space onto a 2-D plane (refer Fig. 2) via t-SNE. Based on this, we can observe distinct clustering of the features corresponding to different speakers in both genuine and spoofed speech indicating that speaker variability is a significant factor in both. Moreover, it is interesting to observe that the speaker clusters are somewhat more distinct in genuine speech features than in spoofed speech features suggesting that genuine speech models may be more affected by speaker variability.

The above visualization just shows a very abstract idea about speaker variability. To quantify the speaker variability, we should analyze the differences in probabilistic distributions over the feature space, which captures the differences in the acoustic characteristics of different speakers in genuine and spoofed classes.

B) Quantifying the speaker variability

While the above visualization (Fig. 2) provides a clear indication that speaker variability affects both genuine and spoofed speech, it makes no attempt to quantify the degree of variability. In order to do so, we now model the underlying feature distributions for each speaker and estimate the differences between these distributions. Specifically, given the joint distribution $P(x, y|spk)$ of features x and class y for each speaker spk , it can be expected that the differences in the speech characteristics between two speakers, $spk = i$ and $spk = j$ will correspond to the differences between the

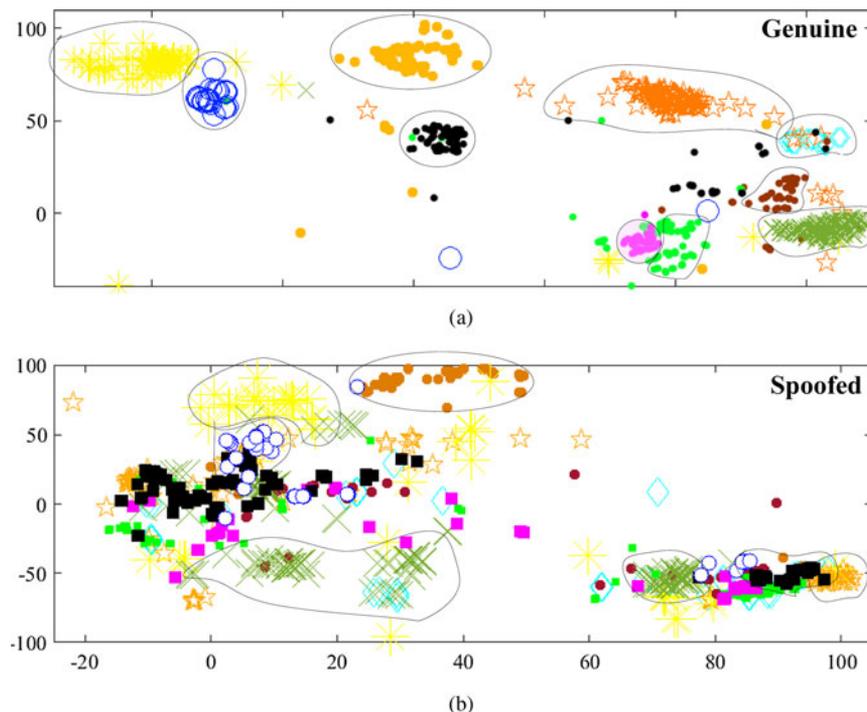


Fig. 2. Feature space (STMF) projected onto 2-D via t-SNE indicates the presence of clusters corresponding to speakers in both genuine and spoofed speech from the ASVspoof2017 v2.0 corpus. The same markers are used for corresponding speakers in genuine and spoofed classes.

distributions $\mathcal{P}_i = P(x|spk = i)$ and $\mathcal{P}_j = P(x|spk = j)$. For example, if speaker variability is a significant confounding factor for a model, we would expect that the differences between the distributions corresponding to any two target speakers to be greater than the differences between the distribution of a target speaker, $P(x|spk)$ and the SI distribution, $P(x)$.

In this work, the differences between distributions are estimated as the Kullback–Leibler (*KL*) divergence between the corresponding GMMs. *KL* divergence is generally used to measure the distance between two probabilistic models, $(\mathcal{P}_1, \mathcal{P}_2)$. Given a D -dimensional feature vector, $X \in \mathbb{R}^D$, the *KL* divergence is of \mathcal{P}_2 from \mathcal{P}_1 is defined as [33]:

$$KL(\mathcal{P}_1, \mathcal{P}_2) = \int_X \mathcal{P}_1(X) \ln \left(\frac{\mathcal{P}_1(X)}{\mathcal{P}_2(X)} \right) dX \quad (1)$$

As $KL(\mathcal{P}_1, \mathcal{P}_2)$ is an asymmetric divergence measure, i.e., $KL(\mathcal{P}_1, \mathcal{P}_2) \neq KL(\mathcal{P}_2, \mathcal{P}_1)$, a symmetric *KL* divergence, S_{KL} , is defined as [33]:

$$S_{KL}(\mathcal{P}_1, \mathcal{P}_2) = \frac{1}{2} (KL(\mathcal{P}_1, \mathcal{P}_2) + KL(\mathcal{P}_2, \mathcal{P}_1)) \quad (2)$$

A Monte Carlo approximation-based symmetric *KL* divergence [34] is used to measure the distance between two probability distributions. Average inter-speaker *KL* divergence (between a target speaker’s distribution and all other target speakers’ distributions) of a speaker is estimated to measure the separation between speaker distributions, $P(x|spk)$. The average inter-speaker *KL* divergence, I_{KL} , for each speaker, i , is given as follows:

$$I_{KL}(i) = \frac{1}{N_{spk} - 1} \sum_{j=1}^{N_{spk}} S_{KL}(\mathcal{P}_i, \mathcal{P}_j); \quad i \neq j \quad (3)$$

where N_{spk} is the total number of claimed speakers. $S_{KL}(\mathcal{P}_i, \mathcal{P}_j)$ is the *KL* divergence between i^{th} and j^{th} speaker distributions as shown in equation 2. In addition to that, the *KL* divergence between a target speaker’s feature distribution and the SI distribution, U_{KL} , is then given by:

$$U_{KL}(i) = S_{KL}(P(x|spk = i), P(x)) \quad (4)$$

where i is a speaker and $P(x)$ is the SI distribution and $P(x|spk = i)$ is i^{th} speaker distribution.

All the feature distributions are modeled as four mixture GMMs of the STMF space for this analysis. Target speaker’s feature distributions are then estimated via Maximum a Posteriori (MAP) adaptation from SI distributions. Genuine and spoofed speaker distributions and SI distributions are trained separately with genuine and spoofed data, respectively.

The average inter-speaker *KL* divergences (I_{KL} as in equation (3)) and the *KL* divergence between a target speaker’s feature distribution and the SI distribution (U_{KL} as in equation (4)), for multiple speakers, are compared in Fig. 3 for both genuine and spoofed classes, separately.

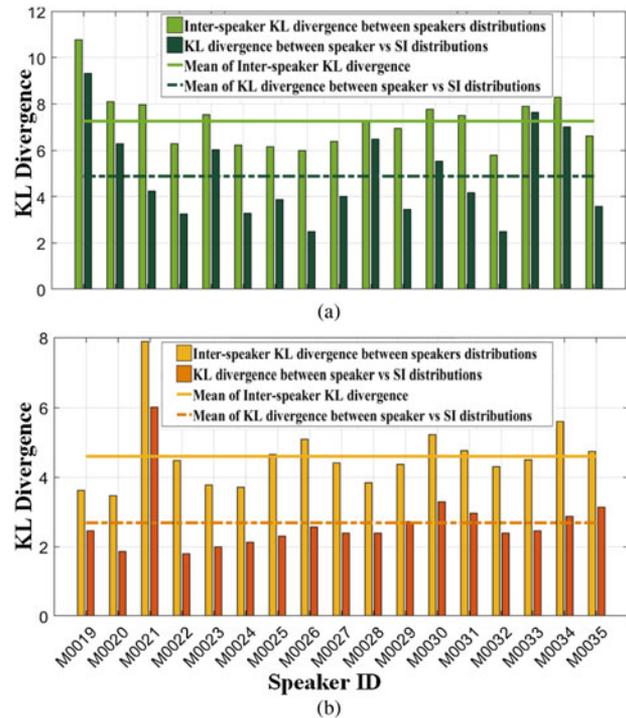


Fig. 3. Symmetric *KL* divergence for speaker-dependent distributions and SI distributions of probability distributions for all speakers in (a) genuine class and (b) spoofed class. This is used to quantify the difference between different speakers as well as the difference between speaker-specific and SI models (speaker variability). The STMF are used as the front-end in this analysis carried out on the ASVspoof2017 v2.0 corpus.

Additionally, the mean values of both measures across all considered speakers also drawn in Fig. 3. This comparison was carried out on the ASVspoof2017 v2.0 speaker-specific enrolment set (refer to section IV for details about the database and the enrolment set). From Fig. 3, it can be seen that the average inter-speaker *KL* divergence between one speaker and all other speakers (I_{KL}) is greater than the one between that speaker and SI distributions (U_{KL}), for both genuine and spoofed classes, which indicates that there is clear separation between probability distributions corresponding to different target speakers.

Also, genuine and spoofed classes showed different degrees of speaker variability. Specifically, (1) *KL* between genuine speaker distribution *versus* genuine SI distribution is higher than spoofed speaker distribution *versus* spoofed SI distribution for each corresponding speaker (refer Figs. 3(a) and 3(b), U_{KL} ’s and mean of U_{KL}); (2) inter-speaker *KL* divergences of genuine classes are greater than inter-speaker *KL* divergences of spoofed classes (refer Figs. 3(a) and (b), I_{KL} ’s and mean of I_{KL}). Thus, the implication of this observation can be interpreted as spoofed class containing less speaker variability than genuine which is in line with our feature space visualization.

C) Analysis of speaker-specific information in the features

In addition to quantifying the degree of speaker variability based on models of the feature distributions of target

Table 1. Speaker identification accuracies on genuine speech and replayed speech evaluated on ASVspoof2017 v2.0 using a GMM-UBM speaker identification system.

| Features | Genuine (%) | Replay (%) |
|----------|-------------|------------|
| STMF | 94.67 | 37.33 |
| CQCC | 99.67 | 58.67 |

speakers, we also quantify it in terms of how well speaker identification can be carried out. Specifically, we build simple GMM-based speaker identification systems on both genuine speech as well as replayed speech and use the speaker identification accuracy as an indicator of the level of speaker variability.

The speaker identification systems, for both genuine and replayed speech, were set up as 512 and 4 mixture GMM-universal background model (UBM) systems using both CQCC and STMF front-ends, respectively. The accuracies of these systems are reported in terms of identification rates (ratio of number correctly identified to total number of test utterances) and estimated on held out test sets comprising 20 utterances per speaker each, for genuine speech and spoofed (replayed) speech. There were a total of 15 speakers in the test set, which corresponds to a chance accuracy of $\sim 6.6\%$. From the results reported in Table 1, it can be seen that for both genuine speech and replayed speech, the speaker identification rate is significantly higher than chance indicating that a high level of speaker variability is present in the features. It is also interesting to note that the speaker identification rate on genuine speech is much higher than replayed speech.

D) Analysis of model separation: genuine versus spoofed classes

Even though the analyses in the previous three sections strongly suggest that both genuine and spoofed speech is greatly affected by speaker variability, it does not directly imply that the use of target speakers' genuine and spoof distributions will lead to better spoofing detection. In order to discern this, we now quantify the ability to discriminate between genuine and spoofed speech with both target speaker's distributions and SI distributions.

It should be noted that a set of GMMs that perform well as a classifier will have a large degree of mutual dissimilarity and consequently a large KL value when compared with a set of GMMs that are more similar to each other. To analyze the discriminability three KL divergence measures are estimated. (1) KL divergence between genuine SI distribution (genuine SI) and spoofed SI distribution (spoofed SI), DU_{KL} , is given by:

$$DU_{KL} = S_{KL}(P(x|y = G), P(x|y = S)) \quad (5)$$

where features x and class y with two classes genuine, G and spoofed, S ; (2) the spoofed SI and each of the genuine target speaker's distributions (genuine speaker models), $D1_{KL}$,

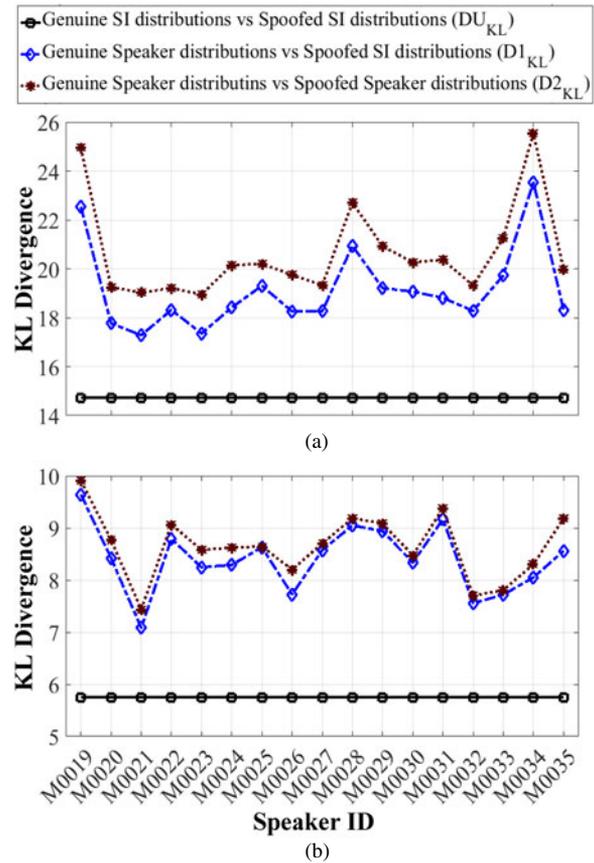


Fig. 4. Comparison of KL divergence between genuine and spoofed models of: (a) STMF and (b) CQCC features for both SI distributions and speaker-dependent distributions on the ASVspoof2017 v2.0 corpus. This is used to quantify the discriminability between genuine and spoofed speech (for each speaker) and compare between speaker-specific and speaker-independent models.

as follows:

$$D1_{KL}(i) = S_{KL}(P(x|y = G, spk = i), P(x|y = S)) \quad (6)$$

where features x and class y of genuine, G for i^{th} speaker, spk and spoofed, S ; (3) each of the genuine speaker models and each of the spoofed speaker models, $D2_{KL}$, is given by:

$$D2_{KL}(i) = S_{KL}(P(x|y = G, spk = i), P(x|y = S, spk = i)) \quad (7)$$

where features x and class y of genuine, G and spoofed, S for i^{th} speaker, spk . These three KL measures, such as DU_{KL} , $D1_{KL}$, and $D2_{KL}$ are compared in Fig. 4 for two different features, such as CQCC and STMF.

As observed in Fig. 4, both $D1_{KL}$ and $D2_{KL}$ constantly show higher KL than DU_{KL} for both CQCC and STMF features. Also the difference between $D2_{KL}$ and $D1_{KL}$ is not considerably higher than that between DU_{KL} and $D1_{KL}$. This might be due to the scarcity of speaker-related information in the spoofed class over the genuine class which is in line with the observation from Fig. 3. Taken together, the results shown in Table 1 and Fig. 4, we can notice that short-term CQCC have rich speaker-related information than long-term STMF.

The analysis of this section concurs with the observations made in the case of feature distributions and further

support that the speaker variability is a significant confounding factor and make use of these speaker-dependent distributions for spoofing detection will be beneficial. Thus, we proposed a framework based on the fact that the use of claimed speaker genuine and/or spoofed models extinguishes speaker variability and leads to better discrimination between genuine and spoofed classes.

III. PROPOSED SPEAKER DEPENDENT SPOOFING DETECTION - GMM BACKEND

Current replay detection systems typically employ a “genuine” speech model and a “spoofed” speech model that is common across all test utterances and independent of the claimed target speaker (refer Fig. 5(a)). Such SI models of genuine and spoofed speech would always be affected by speaker variability, since they would be trained on data from multiple speakers. In this section, we introduce a GMM back-end for spoofing detection that employs speaker-dependent genuine and spoofing models that are specific to each claimed target speaker.

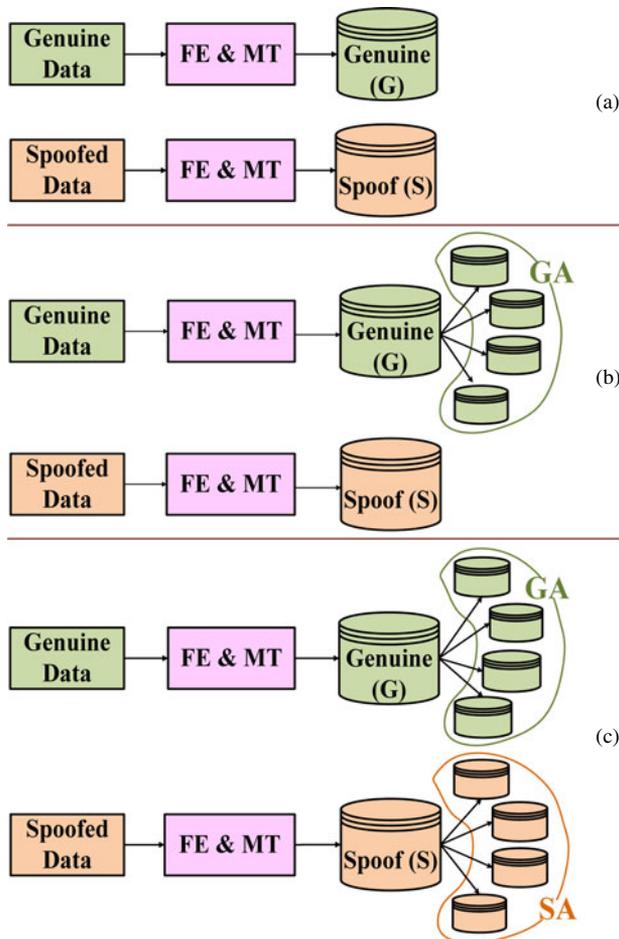


Fig. 5. An overview of SI and speaker-dependent GMM-based spoofing detection systems: (a) SI genuine and spoofed models; (b) speaker-dependent genuine models and SI spoofed models; and (c) both genuine and spoofed models are speaker-dependent. Here FE & MT refer the process of feature extraction and model training.

In the context of replay detection for ASV, the test utterance is always accompanied by a claimed speaker identity and a genuine speech model specific to the claimed speaker will not be affected by any speaker variability. Moreover, enrolment data used to generate speaker models for the ASV system can also be utilized to train speaker dependent models of genuine speech for spoofing detection. In our preliminary study [3], we demonstrated the advantage of incorporating speaker specific information by adopting this approach (refer Fig. 5(b)), whereby instead of the common genuine model (G), we employ claimed speaker dependent genuine speaker models (GA). In this approach, initially, a UBM for genuine speech is trained on genuine speech from multiple speakers, using the EM algorithm. Subsequently, claimed speaker dependent genuine models are adapted from genuine UBM using the enrolment data via MAP adaptation.

Here we propose the use of speaker dependent spoofed speech models in addition to speaker dependent genuine models (refer Fig. 5(c)). Specifically, two background GMMs are initially trained on genuine speech and spoofed speech from multiple speakers, using the EM algorithm, and are referred to as the genuine universal background model (G) and spoof universal background model (S), respectively. Following this, claimed target speaker-dependent models are adapted from the corresponding UBMs using the relevant genuine and spoofed enrolment data via MAP adaptation. During the test phase, the log-likelihood ratios between the claimed speaker models for genuine and spoof speech are used for classification.

IV. PROPOSED SPEAKER DEPENDENT DEEP NEURAL NETWORK BACKEND

GMM based back-ends for spoofing detection have thus far demonstrated good performance and provided a principled method for speaker adaptation. However, in other speech processing tasks, deep learning-based systems have also been shown to be highly effective back-ends that can be adapted to target speakers [35]. In this section, we explore the efficacy of adapting the final layers of a DNN back-end for speaker-dependent spoofing detection.

Similar to the training of the GMM back-end, we train the deep neural network (DNN) back-end in two stages. Initially, we train a SI DNN back-end to distinguish between genuine and spoofed speech trained on data from multiple speakers. This SI back-end is then “adapted” to each target speaker by retraining the final two layers of the DNN using only genuine and spoofed data corresponding to that target speaker (refer Fig. 6).

V. DATABASES AND DATA PREPARATION

We evaluate our system on two datasets that are widely employed in studies of spoofing detection systems, namely,

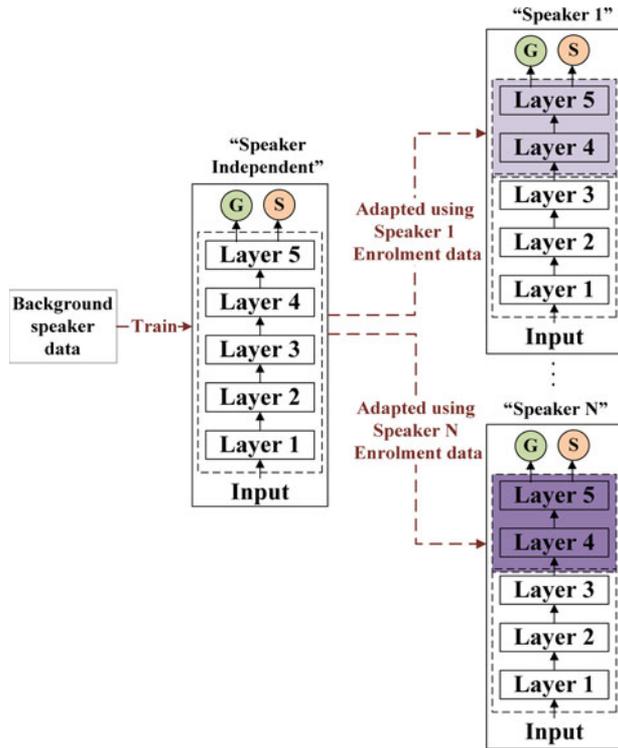


Fig. 6. Speaker-dependent DNN back-ends are obtained by retraining the final layers of a SI DNN back-end.

the ASVspoof2017 v2.0 dataset [24] and the BTAS2016 dataset [36]. However, in both datasets the training, development, and evaluation partitions are non-overlapping in terms of the speakers. This does not reflect conditions under which real ASV systems would operate where data from all target speaker would be present in the training set. Specifically, ASV systems will be developed using a certain amount of enrolment data from each target speaker. Consequently, we repartition both datasets as shown in Fig. 7 and use a small portion of the evaluation partition as enrolment data

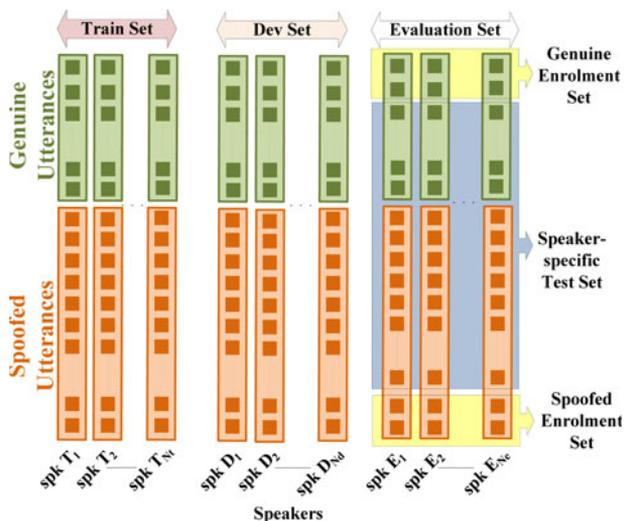


Fig. 7. Schematic diagram showing the repartitioning of the evaluation set into enrolment and test sets for both the ASVspoof2017 v2.0 and BTAS2016 corpora. Training and development partitions are not modified.

and keep the rest of the evaluation partition as a held-out “test set.” We have strived to keep the this held out test set as similar as possible to the original evaluation set in terms of replay configurations (RCs) and conditions.

A) ASVspoof2017 (V2.0) corpus

The evaluation partition of the ASVspoof2017 v2.0 corpora comprises genuine speech corresponding to multiple utterances of 10 passphrases from 24 speakers (same 10 passphrases across all speakers). In addition, replayed versions of these utterances under 57 different replay conditions (combination of recording device, environmental acoustic channel, and playback device) from 17 of the 24 speakers are also part of the evaluation partition. In addition, these 57 RCs are marginalized as low, medium, and high threats in terms of recording, environment, and playback conditions [24].

As noted in Fig. 8, replayed speech was not available for seven speakers. Consequently, we only use the remaining 17 speakers in our evaluation. When repartitioning this evaluation set (of 17 speakers) into enrolment and test sets, we select one utterance corresponding to each passphrase as the “genuine enrolment set.” The remaining genuine utterances are incorporated into the held out “test set.”

Of the 57 RCs present in the evaluation partition, only seven overlap with the train and dev sets. Consequently, we include the spoofed data corresponding to these seven replay conditions into the “spoofed enrolment set” and retain the data corresponding to the remaining 50 unseen RCs (unseen in terms of training the spoofing detection models) in the held-out, “test set.” The number of utterances in the enrolment and test sets is highlighted in Table 2.

B) BTAS 2016 corpus

BTAS 2016, a text-independent database, contains genuine and different kinds of spoofing attacks where genuine, speech synthesis and voice conversion speech samples were recorded and played back using high-quality devices [36]. We have only used “replay subset” of BTAS for our experiments which comprises replayed versions of the genuine speech utterances as in [37]. The number of utterances for each of the train, development and evaluation partitions is presented in Table 3 and information about speaker data given in Fig. 9. Unknown RCs are present in the evaluation partition to make the spoofing detection task more challenging.

Similar to ASVspoof2017 v2.0, we partition the evaluation set of the BTAS corpus into an enrolment set and a test set by taking 30 genuine and 30 spoofed utterances from every speaker to form the enrolment set (rest constitute the “test set”). The data statistics of both data sets are summarized in Table 3. Once again, the enrolment set only contains known RCs (all unknown RCs are in the test set) and held out “test set” kept as the same characteristics (in terms of replay conditions) as the original BTAS 2016 evaluation set.

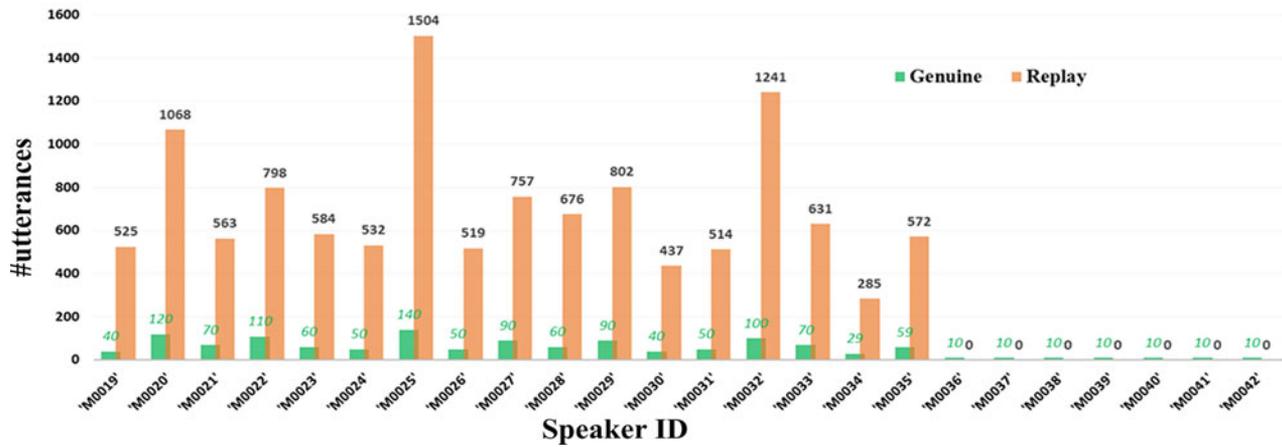


Fig. 8. Number of utterances for each speaker in the ASVspoof2017 v2.0 evaluation set. The speaker IDs provided in the dataset are indicated along the x-axis.

Table 2. ASVspoof2017 Version 2.0 [24]

| Subset | # Speakers | # Utterances | | #RC conditions |
|------------|----------------|--------------|-------|----------------|
| | | Genuine | Spoof | |
| Train | 10 | 1507 | 1507 | 3 |
| Dev | 8 | 760 | 950 | 10 |
| Evaluation | Enrolment set* | 17 | 1169 | 7 |
| | Test set* | 17 | 1058 | 50 |

*The details of “Enrolment set” and the “Speaker-specific Test set” can be found in <http://www2.ee.unsw.edu.au/ASVspoof/>.

Table 3. “Replay Subset” of the BTAS2016 corpus.

| Subset | # Speakers | # Utterances | |
|------------|----------------|--------------|-------|
| | | Genuine | Spoof |
| Train | 14 | 4973 | 2800 |
| Dev | 14 | 4995 | 2800 |
| Evaluation | Enrolment set* | 16 | 480 |
| | Test set* | 16 | 5096 |

*The details of “Enrolment set” and the “Speaker-specific Test set” can be found in <http://www2.ee.unsw.edu.au/ASVspoof/>.

VI. FRONT-END FEATURES

A key difference between ASV systems and spoofing detection systems pertains to the fact the ASV systems try

to capture speaker cues while normalizing other factors (channel factors, phonetic factors, etc), whereas spoofing detection aims to model the channel differences between genuine and spoofed speech. Consequently, a significant portion of recent research on replay detection has focused on developing suitable front-ends that encode information most relevant to identifying replay channels. In our experiments reported in this paper we adopt two front-ends that have previously shown to be effective for replay detection, namely, CQCC [2] and spectro-temporal modulation feature (STMF) [22].

Finally, in addition to CQCCs and STMFs, we also employ a third front-end in the form of log compressed modified group delay (LMGD) in our investigations. While CQCCs and STMFs are both spectral features, LMGDs are extracted from frame-based group delay and including them in our experiments allow us to study the effect of speaker variability in non-spectral features as well. It is also worth noting that all three are extracted over differing time periods (~ 8 ms for CQCC, 20 ms for LMGD, and entire utterance for STMF).

MGD function was initially developed as an alternative to typical spectral features [38]. However, recent results suggest that their high dynamics range and spiky nature may make them ineffective for replay attack detection [19]. To offset these disadvantages, we propose the use of log



Fig. 9. Number of utterances from each speaker in the BTAS 2016 evaluation set. The speaker IDs provided in the dataset are indicated along the x-axis.

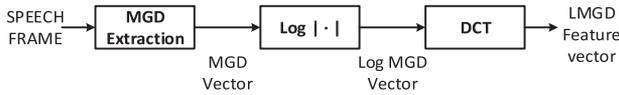


Fig. 10. LMGD feature extraction.

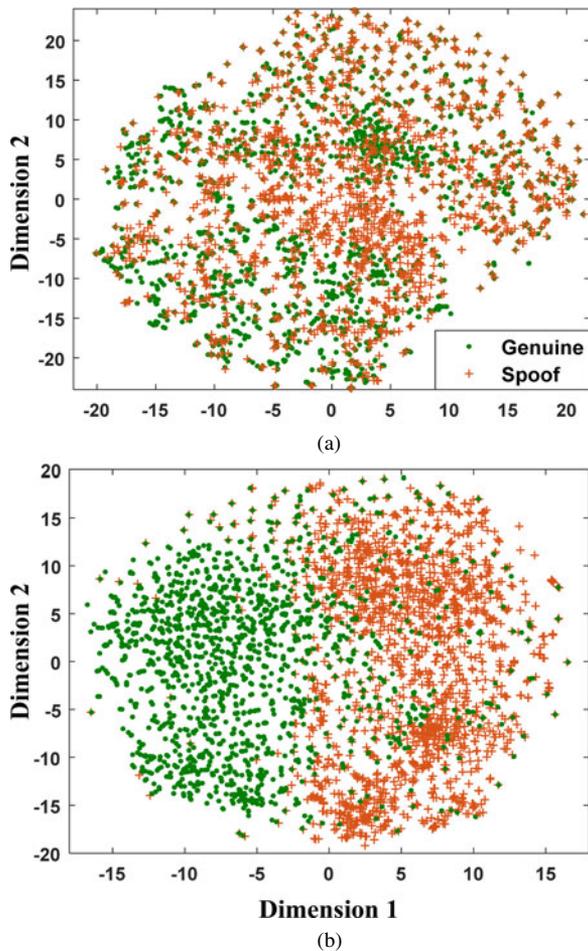


Fig. 11. t-SNE plot depicting the distribution of (a) MGD and (b) LMGD features for a subset of ASVspoof2017 v2.0 train.

compressed magnitude of MGD with DCT as an alternative feature representation which we refer to as LMGD (refer Fig. 10). To ascertain if this proposed advantage is realized, we projected both MGD and LMGD features (both 120 dimensional feature space) obtained from a subset of the ASVspoof v2.0 training set onto 2-dimensions using t-SNE and plot them in Fig. 11. It is clear that spoofed and genuine speech classes are better separated with LMGD features than with MGD features. The log compressed MGD vector is defined as,

$$\log \text{MGD} = \log \left| \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S(\omega)^{2\gamma}} \right|^\alpha \quad (8)$$

where the subscripts R and I denote the real and imaginary parts of $X(\omega)$ and $Y(\omega)$, which in turn correspond to the Fourier transforms of $x(n)$ and $nx(n)$ respectively. The parameters α and γ vary from 0 to 1.

VII. EXPERIMENTAL SETTING

A number of experiments were carried out to evaluate the proposed approaches for speaker-dependent spoofing detection against the SI baselines. The metric employed to quantify performance in these comparisons are: (a) “Overall EER”, which is the equal error rate for spoofing detection evaluated over all utterances in the test set; (b) “Speaker-wise EER”, which is the equal error rate computed separately speaker-by-speaker; and finally (c) “Average EER” is average speaker-wise equal error rate (EER) (averaged over all speakers).

A) Front-end feature configurations

Prior to all feature extraction, the speech was pre-emphasized with 0.97 factor. It should be noted that no normalization is applied to any of the features in this work.

STMF parameters: Pre-emphasized speech is framed with a 50% overlap between them using a Hamming window. The STMF features are extracted using the same parameters and MCF-CC and MSE-CC features are chosen with 15 and 30 dimensions respectively and feature-level concatenation is performed to obtain 45 dimensions for each utterance as in [22].

LMGD parameters: LMGD features are extracted from frames of 50% overlap followed by Hamming window. α, γ were empirically chosen to be 0.4 and 0.9, respectively, based on the development set. LMGD feature is 120 dimensions which consists of static, velocity, and acceleration coefficients.

CQCC parameters: For the derivation of CQCC features, we have used the same configuration as used in the ASVspoof2017 challenge baseline [2]. A Constant-Q Transform is applied with a maximum frequency of $f_{\max} = 8$ kHz, which is the Nyquist frequency and the minimum frequency, $f_{\min} = f_{\max}/2^9 \approx 15$ Hz, where 9 is the number of octaves. The number of bins per octave is set to 96. Resampling is applied with a sampling period of 16 bins in the first octave. The CQCC feature dimension is set to 90 coefficients.

B) GMM backend configuration

In the proposed GMM system, the spoofed and genuine UBMs were both implemented as 512 mixture GMMs for CQCC and LMGD features and as 4 mixture GMMs for STMF features. All UBMs were trained using the EM algorithm with random initialization. In the experiments carried out on the ASVspoof2017 v2.0 corpus, the Train and Dev sets were used to model the genuine and spoofed UBMs (no overlap with test speaker data). For the system evaluated on BTAS 2016, the train set was used to model the UBMs. The speaker-dependent genuine and spoofed models were then estimated from the UBMs via MAP adaptation (mean and weights only) using the Enrolment set (refer section V for details on the enrolment set).

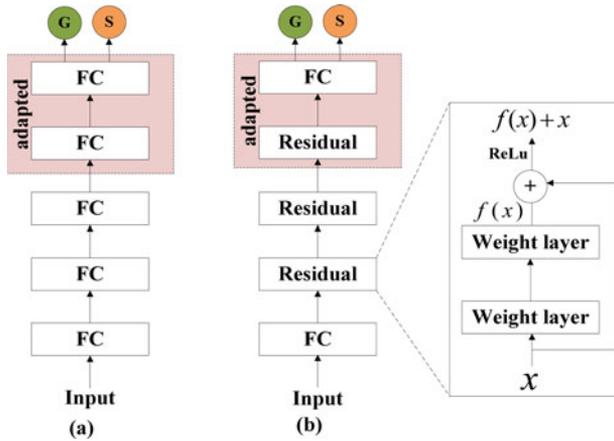


Fig. 12. Two variants of the DNN that can be employed as the SI and speaker-dependent DNNs in Fig. 6: (a) five FC layers; and (b) three residual layers in between two FC layers.

C) DNN backend configuration

The SI baseline systems, using either the CQCC or the LMGD front-ends, were trained using Train and Dev sets for the ASVspoof2017 v2.0 system and the Train set for the BTAS 2016 system. Two variations of this DNN system were developed and evaluated, one with five fully-connected (FC) layers and the other with three residual network layers sandwiched between two FC layers as shown in Fig. 12. In both systems, the final FC layer has 64 neurons and the first FC layer has as many neurons as the input feature dimension. The middle three FC layers of the first system (Fig. 12(a)) have 1024, 512, and 256 neurons, respectively (going from the input to output). The three residual layer blocks of the second system (Fig. 12(b)) all have an identical architecture with each block comprising two layers of weights comprising 256 and 128 neurons each. All hidden layer neurons employ a ReLU activation function and the final layer uses softmax activation. Dropouts and batch normalization were employed during model training, using a cross-entropy loss function, and the dropout rate was set as 0.4. Learning rates were decayed uniformly across the network per epoch from 10^{-2} to 10^{-5} . Early stopping based on validation loss on the dev set was used to avoid overfitting. All network weights are initialized from a normal distribution with zero-mean and a standard deviation of 10^{-2} . Layerwise L2 regularization was also performed.

Finally, to obtain speaker-dependent DNN back-ends for our proposed system, the trained (SI models were used as initial models and the final two layers were retrained for three additional epochs using the enrolment data from the target speaker. When using CQCC and LMGD features, the average posterior predicted by the back-end over all the frames in an utterance is used as the score to calculate EERs.

The DNN back-end Architecture chosen for the STMF front-end were constrained to be a smaller version of that used by the CQCC and LMGD systems to avoid over parameterization. Since STMF is an utterance-based feature, a single set of posteriors is predicted by the back-end and used as the score to estimate EERs.

VIII. RESULTS AND DISCUSSION

In this section, we report experimental results obtained when comparing the proposed speaker-dependent back-ends to corresponding SI ones. In addition to overall error rates, we also report comparisons per speaker and under different RCs.

A) Speaker dependent versus speaker independent spoofing detection

1) GMM BACKEND

The proposed speaker-dependent GMM back-end for spoofing detection outlined in Fig. 5(c), which employs speaker-dependent genuine and spoofed models (herein referred to as GA + SA), are compared to the earlier system [3] that employed only speaker-dependent genuine models (shown in Fig. 5(b) and herein referred to as GA + S) as well as a SI back-end (Fig. 5(a), herein referred to as G + S). Table 4 reports the comparison carried out on the ASVspoof2017 v2.0 dataset and Table 5 shows the results obtained on the BTAS2016 dataset. From these results, it can be seen that the speaker-dependent approach outperforms a SI one. These improvements are in line with observations noted in section II that indicate short term CQCC features encode more speaker information compared to longer term STMFs. However, these improvements obtained by making the spoofed models speaker specific are all relatively small, especially when compared to the significant improvements obtained when the genuine model was made speaker specific.

In addition to the overall EER, we also compare speaker-wise EERs (as well their averages) and show this comparison in Figs 13 and 14 for ASVspoof2017 and BTAS2016 respectively. The speaker-wise results indicate that the speaker-dependent approach is superior to the SI one for almost all speakers in the test set. Furthermore, the average of speaker wise EER of GA + SA is consistently lower than the average speaker wise EER of GA + S across all three features on both ASVspoof2017 v2.0 and BTAS 2016.

Table 4. Comparison of SI and speaker-dependent GMM back-ends evaluated on the ASVspoof2017 v2.0 “test set” in terms of the overall EER %.

| Feature set | G + S (baseline) | GA + S [3] | GA + SA (proposed) |
|-------------|------------------|------------|--------------------|
| CQCC | 25.1 | 12.33 | 11.14 |
| LMGD | 29.79 | 17.16 | 14.84 |
| STMF | 8.12 | 3.98 | 3.63 |

Table 5. Comparison of SI and speaker-dependent GMM back-ends evaluated on the BTAS 2016 “test set” in terms of the overall EER %.

| Features | G + S (baseline) | GA + S [3] | GA + SA (proposed) |
|----------|------------------|------------|--------------------|
| CQCC | 8.36 | 2.41 | 1.77 |
| LMGD | 0.92 | 0.57 | 0.31 |
| STMF | 1.12 | 0.42 | 0.37 |

Here it should be noted that the number of trials (per speaker) used to estimate the speaker-wise EERs is significantly lower than the number of trials in the test set used to evaluate the overall EER and consequently these speaker-wise EERs should be considered as indicative only.

In addition, it is also worth noting that when evaluating the SI baseline systems (without CMVN) on the original ASVspoof2017 v2.0 evaluation set we obtained EERs of 24.5%, 29%, and 7.9% for the CQCC, LMGD, and STMF front-ends respectively. These are comparable to all previously published results (since the original evaluation set

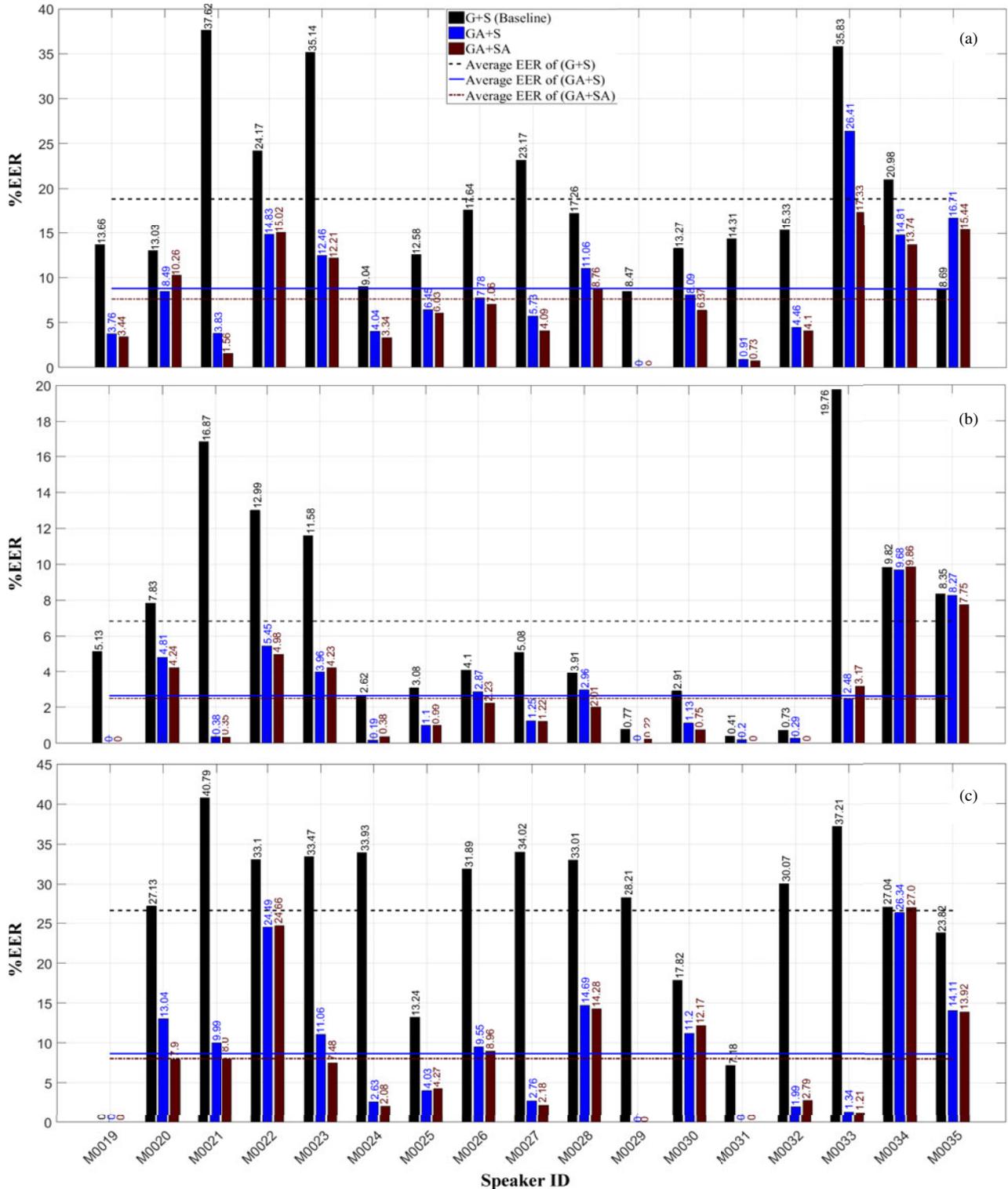


Fig. 13. Comparison of SI and speaker-dependent GMM back-ends evaluated on the ASVspoof2017 v2.0 "test set" in terms of speaker-wise EERs for the three different front-ends: (a) CQCC; (b) STMF; and (c) LMGD. In addition the graphs also show the average EERs (obtained by averaging across speakers).

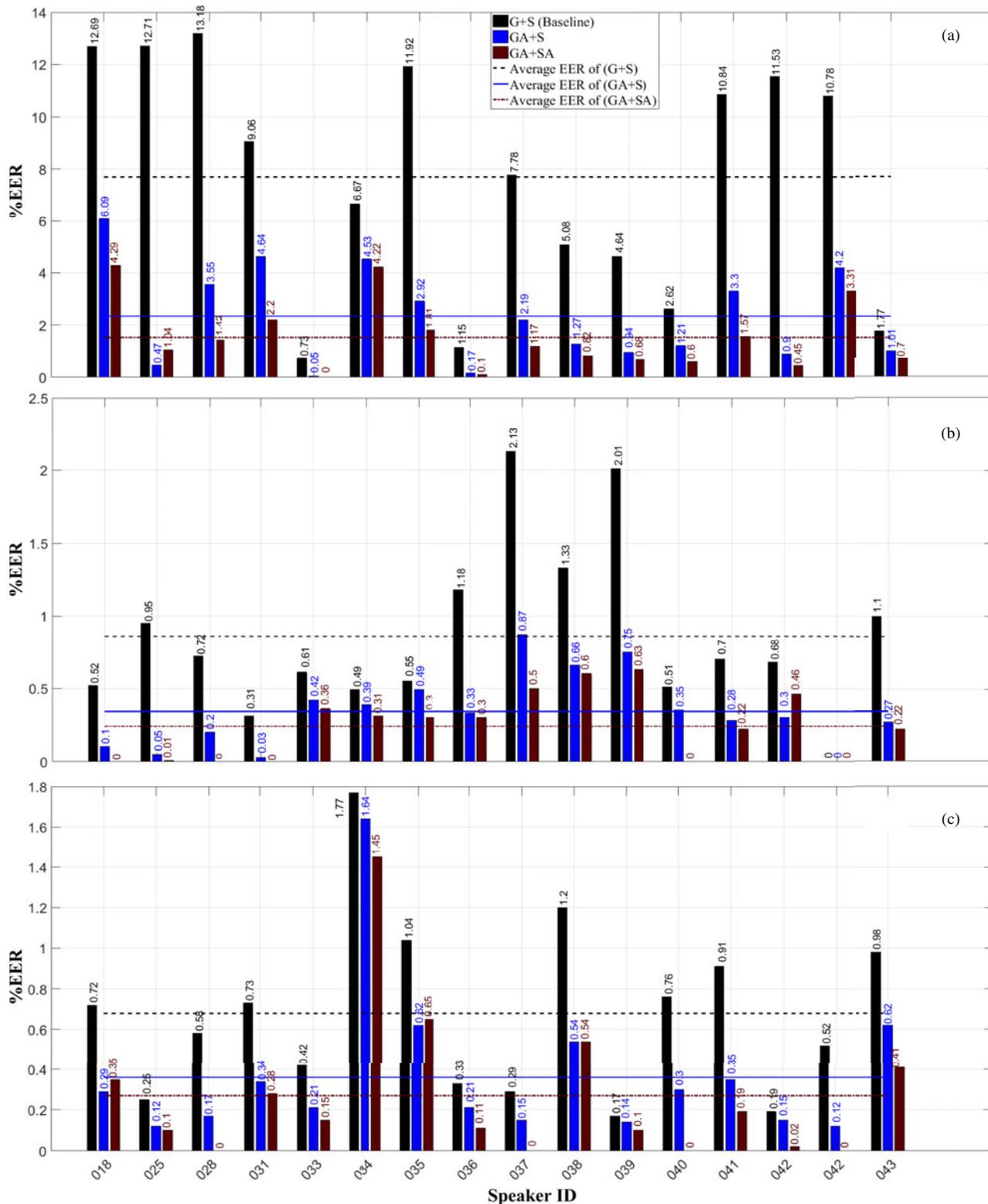


Fig. 14. Comparison of SI and speaker-dependent GMM back-ends evaluated on the BTAS 2016 “test set” in terms of speaker-wise EERs for the three different front-ends: (a) CQCC; (b) STMF; and (c) LMGD. In additions the graphs also show the average EERs (obtained by averaging across speakers).

is employed) and the 7.9% for STMF systems is the lowest EER currently reported on v2.0 [22]. Also, the EERs on the original evaluation set and the modified evaluation set employed in this paper are very similar which demonstrates that the results obtained using the modified evaluation

set are also equally valid. Finally, in order to determine if the reported results are biased in any way due to the larger number of spoofed enrolment data compared to genuine enrolment data (specifically in the ASVspoof2017 v2.0 enrolment set), we repeated the experiment with an equal

Table 6. Comparison of speaker-dependent and SI DNN back-ends on the ASVspoof2017 v2.0 “Test set” in terms of overall EER (%).

| Features | | Speaker-independent (baseline) | Speaker-dependent (proposed) |
|----------|------|--------------------------------|------------------------------|
| FC | STMF | 13.69 | 9.91 |
| | CQCC | 32.96 | 24.57 |
| | LMGD | 32.31 | 22.17 |
| Residual | STMF | 13.2 | 11.12 |
| | CQCC | 29.1 | 22.32 |
| | LMGD | 27.2 | 19.80 |

Table 7. Comparison of speaker-dependent and SI DNN back-ends on the BTAS2016 “Test set” in terms of overall EER (%).

| Features | | Speaker-independent (baseline) | Speaker-dependent (proposed) |
|----------|------|--------------------------------|------------------------------|
| FC | STMF | 2.92 | 1.97 |
| | CQCC | 4.23 | 2.72 |
| | LMGD | 3.35 | 2.34 |

amount of genuine and spoofed enrolment data. We set up this condition by reducing the spoofed enrolment data set to match the size of the genuine enrolment set. The results obtained matched those from the other experiments suggesting the results are not biased in any way due to the spoofed enrolment set being larger than the genuine enrolment set.

2) DNN BACKEND

Similar to the GMM back-end comparisons, we compare the speaker-dependent DNN back-end to the SI DNN back-end (details in section VII-C) on both the ASVspoof2017 v2.0 dataset and the BTAS2016 dataset and report the results in Tables 6 and 7, respectively. Since two variants of the DNN back-end were considered (as shown in Fig. 12), we report both results. These results also indicate that a speaker-dependent approach outperforms a SI one.

B) Investigating the system performance in terms of different replay configurations for ASVspoof2017 v2.0

To understand the behavior of the proposed speaker-dependent systems under different qualities of environments, play-back, and recording devices, we compare the speaker-dependent STMF GMM system to a SI STMF GMM system in different RCs. The ASVspoof2017 v2.0 database contains recordings collected with diverse RCs, each comprising one recording device, one playback device, and one acoustic environment. In order to aid analysis, the distinct RCs were previously grouping together overlapping configurations [24]. We adopt the same groupings in our analyses. The overall EER evaluated under each condition (group) is reported in Fig. 15. These results also suggest that a speaker-dependent back-end would be preferable to a SI one.

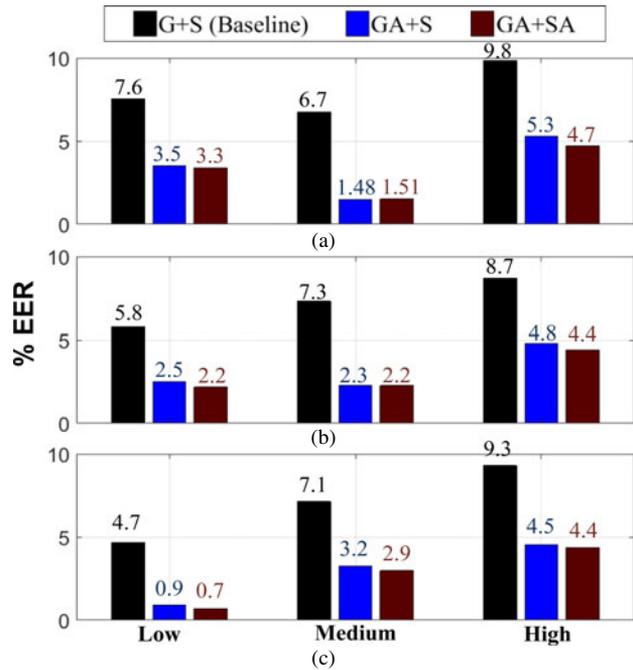


Fig. 15. Comparison of overall % EER of STMF grouped by the (a) acoustic environment, (b) playback, and (c) recording device into low, medium, and high threat attacks for speaker independent (baseline) and speaker dependent (proposed) systems. For each group, the quality of the other two parameters is a mix of all three qualities (for e.g., low quality playback test utterances come from a mix of low, medium, and high quality acoustic environment and recording devices).

C) Simulated spoofed data

In practice, speaker specific spoofed data is not likely to be available during the system training phase. However, it may still be possible to create speaker specific spoofed models using simulated spoofed data. Specifically, one or more spoofing environment models can be employed to generate simulated spoofed speech from the available genuine speech. To investigate this hypothesis, we carried out spoofing detection experiments where speaker specific spoofed models were trained using only simulated spoofed data (created as outlined above).

The simulated replayed speech has been created using the pyroomacoustic simulator [39]. The pyroomacoustic tool is a recent open software package which facilitates the simulation of room acoustic conditions as well as loudspeakers and microphones. The model parameters employed in our simulation are shown in Table 8. These simulations were made in line with the assumptions inherent in the ASVspoof2017 replay data [2] and we do not include any model of room or channel acoustics that would have been present when the original recording of the speech might have been made surreptitiously.

All 170 genuine speech utterances (refer Table 2) in the genuine enrolment set was put through this model to simulate a corresponding “replayed spoofed speech” utterance. This simulated data was then used to adapt the speaker specific spoofed models. The ASVspoof2017 v2.0 evaluation set was retained as the test set and the results have been tabulated in Table 9. Comparing Table 4 with Table 9, we see

Table 8. Configuration of the Pyroomacoustic [39] replay simulation.

| | |
|--------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| Room corners | [(0,0), [0,3], [5,3], [5,1], [3,1], [3,0]], |
| Source | Unidirectional ideal loudspeaker at position (1,1) |
| Room | Uniform absorption of 0.01 |
| Mic | Circular 6 channel microphone array at (2,2) with a radius of 0.1 and main directivity of $\Phi = 0$. Channel 1 is taken as the replay |

Table 9. Comparison of SI and speaker-dependent GMM back-ends evaluated on the ASVspoof2017 v2.0 “test set” in terms of overall EER (%). Here the speaker specific spoofed models (SA) are all obtained using simulated replayed speech data.

| Features | G + S (baseline) | GA + S [3] | GA + SA (simulated) |
|----------|------------------|------------|---------------------|
| CQCC | 25.1 | 12.33 | 11.22 |
| LMGD | 29.79 | 17.16 | 15.25 |
| STMF | 8.12 | 3.98 | 3.64 |

that almost identical results are obtained even when we use simulated spoofed data to train the speaker specific spoofed models.

IX. CONCLUSION

Even though most of the recent research efforts in developing spoofing detection systems has focused on their development independent of the ASV system, in practice, they will always be used in conjunction with speaker verification. Consequently, some speech data (enrolment data) from each target claimed speaker will always be available since they are required to develop any speaker verification system. In this paper, we tackle the question of whether this data can be used to improve the spoofing detection system.

Specifically, we have carried out multiple analyses to quantify speaker variability in the feature space, which revealed that adoption of speaker-dependent models for spoofing detection can be expected to lead to better accuracy. Following this, we develop two straight-forward approaches, one for a GMM back-end and one for a neural network back-end, to adapt spoofing detection back-ends to each claimed target speaker using the available enrolment speech as well as spoofed versions of these obtained by replaying them via a few different conditions (available in the training sets of both speech corpora used in our experiments). Validation on the evaluations sets of both ASVspoof2017 v2.0 and BTAS2016 reveals that this improves spoofing detection accuracy over the corresponding SI back-ends under all conditions, even when the spoofed utterances in the test came via unseen replay channels. Our experiments were carried out over three different front-ends: CQCCs, STMF, and LMGDs, as well as two different back-ends: GMM and DNN, and consistently the speaker-dependent versions outperformed the SI versions.

Our previous work showed that significant gains can be obtained by incorporating this information to train speaker specific genuine speech models. Experiments reported in this paper demonstrate that small additional gains may be

obtained by adopting speaker-specific spoofed models as well. Furthermore, simulated spoofed speech can be used to train these speaker-specific spoofed models.

These experimental results suggest that reducing variability in the spoofed model (by eliminating speaker variability) only leads to small improvements in spoofing detection. In future work, we plan to study the interaction between speaker variability and spoofing channel effects to extend this investigation.

REFERENCES

- [1] Wu Z.; Evans N.; Kinnunen T.; Yamagishi J.; Alegre F.; Li H.: Spoofing and countermeasures for speaker verification: A survey. *Speech Commun.*, **66** (2015), 130–153.
- [2] Kinnunen T. *et al.*: The ASVspoof2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection, in *INTERSPEECH, Stockholm*, 2017, 2–6.
- [3] Suthokumar G.; Sriskandaraja K.; Sethu V.; Wijenayake C.; Ambikairajah E.; Li H.: Use of Claimed Speaker Models for Replay Detection, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, Hawaii*, 2018, 1038–1046.
- [4] Sarkar A.K.; Tan Z.; Kinnunen T.: Improving Speaker Verification Performance in Presence of Spoofing Attacks Using Out-of-Domain Spoofed Data, in *INTERSPEECH, Stockholm*, 2017, 2611–2615.
- [5] Khoury E.; Kinnunen T.; Sizov A.; Wu Z.: “Introducing I-Vectors for Joint Anti-spoofing and Speaker Verification,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014, 61–65.
- [6] Sizov A.; Khoury E.; Kinnunen T.; Wu Z.; Marcel S.: Joint speaker verification and anti-spoofing in the i-vector space. *IEEE Trans. Inf. Forensics Secur.*, **10** (4) (2015), 821–832.
- [7] Todisco M. *et al.*: Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion, in *INTERSPEECH, Hyderabad, India*, 2018.
- [8] Font R.; Espin J.M.; Cano M.J.: Experimental Analysis of Features for Replay Attack Detection — Results on the ASVspoof2017 Challenge, in *INTERSPEECH, Stockholm*, 2017, 7–11.
- [9] Todisco M.; Delgado H.; Evans N.: Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Comput. Speech Lang.*, **45** (2017), 516–535.
- [10] Alluri K.N.R.K.R.; Achanta S.; Kadiri S.R.: Detection of Replay Attacks using Single Frequency Filtering Cepstral Coefficients, in *INTERSPEECH, Stockholm*, 2017, 2596–2600.
- [11] Sriskandaraja K.; Suthokumar G.; Sethu V.; Ambikairajah E.: Investigating the use of scattering coefficients for replay attack detection, in *APSIPA ASC, Kuala Lumpur*, 2017, 1195–1198.
- [12] Saranya M.; Murthy H.A.; and H S.M.S.; Murthy A.: Decision-level feature switching as a paradigm for replay attack detection, in *INTERSPEECH, Hyderabad, India*, 2018, 686–690.
- [13] Gunendradasan T.; Irtza S.; Ambikairajah E.; Epps J.: Transmission Line Cochlear Model Based AM-FM Features for Replay Attack Detection, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton*, 2019, 6136–6140.
- [14] Patel T.B.; Patil H.A.: Cochlear filter and instantaneous frequency based features for spoofed speech detection. *IEEE J. Sel. Top. Signal Process.*, **11** (4) (2017), 618–631.

- [15] Jelil S.; Das R.K.; Prasanna S.R.M.; Sinha R.: Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features, in *INTER-SPEECH, Stockholm, 2017*, 22–26.
- [16] Das R.K.; Li H.: Instantaneous Phase and Excitation Source Features for Detection of Replay Attacks, in *APSIPA ASC, Honolulu, 2018*.
- [17] Patil H.A.; Kamble M.R.; Patel T.B.; Soni M.: Novel Variable Length Teager Energy Separation Based Instantaneous Frequency Features for Replay Detection, in *INTER-SPEECH, Stockholm, 2017*, 12–16.
- [18] Sailor H.B.; Kamble M.R.; Patil H.A.: Auditory Filterbank Learning for Temporal Modulation Features in Replay Spoof Speech Detection, in *INTER-SPEECH, Hyderabad, India, 2018*, 666–670.
- [19] Srinivas K.; Patil H.A.: Relative Phase Shift Features for Replay Spoof Detection System, in *6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages, 2018*, 98–102.
- [20] Suthokumar G.; Sriskandaraja K.; Sethu V.; Wijenayake C.; Ambikairajah E.: Independent Modelling of Long and Short Term Speech Information for Replay Detection, in *Speech Science and Technology Conference (SST), Sydney, Australia, 2018*, 49–53.
- [21] Wickramasinghe B.; Irtza S.; Ambikairajah E.; Epps J.: Frequency Domain Linear Prediction Features for Replay Spoofing Attack Detection, in *INTER-SPEECH, Hyderabad, India, 2018*, 661–665.
- [22] Suthokumar G.; Sethu V.; Wijenayake C.; Ambikairajah E.: Modulation Dynamic Features for the Detection of Replay Attacks, in *INTER-SPEECH, Hyderabad, India, 2018*, 691–695.
- [23] Ji Z.; Li Z.; Li P.; An M.; Gao S.; Data B.: Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017, in *INTER-SPEECH, Stockholm, 2017*, 2–6.
- [24] Todisco M.; Evans N.; Kinnunen T.; Lee K.A.; Yamagishi J.: ASVspoof2017 Version 2.0: meta-data analysis and baseline enhancements, in *Odyssey, 2018*, 296–303.
- [25] Lavrentyeva G.; Novoselov S.; Malykh E.; Kozlov A.; Kudashev O.; Shchemelinin V.: Audio replay attack detection with deep learning frameworks, in *INTER-SPEECH, Stockholm, 2017*, 82–86.
- [26] Sriskandaraja K.; Sethu V.; Ambikairajah E.: Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric, in *INTER-SPEECH, Hyderabad, India, 2018*, 671–675.
- [27] Nagarsheth P.; Houry E.; Patil K.; Garland M.: Replay Attack Detection using DNN for Channel Discrimination, in *INTER-SPEECH, Stockholm, 2017*, 97–101.
- [28] Huang L.; Pun C.-M.: Audio Replay Spoof Attack Detection Using Segment-based Hybrid Feature and DenseNet-LSTM Network, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, 2019*, 2567–2571.
- [29] Tom F.; Jain M.; Dey P.; Kharagpur I.: End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention, in *INTER-SPEECH, Hyderabad, India, 2018*, 681–685.
- [30] Lai C.-I.; Abad A.; Richmond K.; Yamagishi J.; Dehak N.; King S.: Attentive filtering networks for audio replay attack detection, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019*, 6316–6320.
- [31] Cai W.; Cai D.; Liu W.; Li G.; Li M.: Countermeasures for Automatic Speaker Verification Replay Spoofing Attack: On Data Augmentation, Feature Representation, Classification and Fusion, in *INTER-SPEECH, Stockholm, 2017*, 17–21.
- [32] Suthokumar G.; Sriskandaraja K.; Sethu V.; Wijenayake C.; Ambikairajah E.: Phoneme specific modelling and scoring techniques for anti spoofing system, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, 2019*, 6106–6110.
- [33] Kullback S.: *Information Theory and Statistics*, 2nd ed., Dover Publications, Mineola, NY, 1968.
- [34] Sethu V.; Epps J.; Ambikairajah E.: Speaker variability in speech based emotion models – Analysis and normalisation, *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc.*, 2013, 7522–7525.
- [35] Miao Y.; Zhang H.; Metze F.: Speaker adaptive training of deep neural network acoustic models using I-vectors. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, **23** (11) (2015), 1938–1949.
- [36] Korshunov P. *et al.*: Overview of BTAS 2016 speaker anti-spoofing competition, in *IEEE 8th International Conference on Biometrics Theory, Applications and Systems, BTAS, 2016*, 1–6.
- [37] Xie Z.; Zhang W.; Chen Z.; Xu X.: A Comparison of Features for Replay Attack Detection, in *Journal of Physics: Conference Series (JPCS)*, 2019, vol. **1229**, 1–8.
- [38] Murthy H.A.; Ramana V.; Gadde R.: The Modified Group Delay Function and Its Application to Phoneme Recognition, in *ICASSP, 2003*, 68–71.
- [39] Scheibler R.; Bezzam E.; Dokmanić I.: Pyroomacoustics: A Python package for audio room simulations and array processing algorithms, in *ICASSP, 2018*, 351–355.

Gajan Suthokumar received the B.Sc. (Hons.) degree in engineering from the University of Moratuwa, Sri Lanka, in 2015. He is currently working towards the Ph.D. degree with the speech processing research group in the School of Electrical Engineering and Telecommunications, University of New South Wales (UNSW), Sydney, Australia. He has previously worked as a research intern at National University of Singapore (NUS). He is the recipient of the Best Student Paper Award in APSIPA 2018. His research interests include the application of signal processing techniques and machine learning in voice biometrics, anti-spoofing and ubiquitous computing. He is a member of IEEE, ISCA and APSIPA.

Kaavya Sriskandaraja received the B.Sc.(Hons.) degree in engineering from the University of Peradeniya, Peradeniya, Sri Lanka, in 2012, and Ph.D. degree from the University of New South Wales(UNSW), Sydney, Australia, in 2018. She is currently a Post-doctoral Research Fellow at the School of Electrical Engineering and Telecommunications, UNSW, Sydney, Australia. Her research interests are on applying machine learning and signal processing techniques in voice biometrics, voice anti-spoofing, healthcare and therapeutic applications. She is a Reviewer of IEEE conferences and journals. Dr Kaavya is a Member of IEEE, ISCA and APSIPA.

Vidhyasaharan Sethu received the B.E. degree with Distinction from Anna University, Chennai, India, in 2005, and the M.Eng.Sc. degree with High Distinction in signal processing and the Ph.D. degree from the University of New South Wales (UNSW), Sydney, Australia, in 2006 and 2010, respectively. Following this, at UNSW he received a postdoctoral fellowship from 2010 to 2013. He is currently a Senior Lecturer at the School of Electrical Engineering and Telecommunications, UNSW. In addition to his role as one of four principle investigators in the Speech Processing Research Group at UNSW, he is also a co-director of the Signal, Information and Machine Intelligence Lab [<https://www.simi.unsw.edu.au>]. His research interests include the application of machine learning to speech processing and affective computing, voice biometrics, and computational paralinguistics. Vidhyasaharan is a member of ISCA, IEEE, AAC, ASSTA and APSIPA and serves

on the editorial board for Computer Speech and Language (CSL).

Professor Eliathamby Ambikairajah received his BSc (Eng) (Hons) degree from the University of Sri Lanka and received his PhD degree in Signal Processing from Keele University, UK. He was appointed as Head of Electronic Engineering and later Dean of Engineering at the Athlone Institute of Technology in the Republic of Ireland from 1982 to 1999. His key publications led to his repeated appointment as a short-term Invited Research Fellow with the British Telecom Laboratories, U.K., for ten years from 1989 to 1999. Professor Ambikairajah is currently serving as the Acting Deputy Vice-Chancellor Enterprise, after previously serving as the Head of School of Electrical Engineering and Telecommunications, University of New South Wales (UNSW), Australia from 2009 to 2019. As a leader he has firmly established the School as the top Electrical Engineering school in Australia and among the top 50 in the world. His research interests include speaker and language recognition, emotion detection and biomedical signal processing. He has authored and co-authored approximately 300 journal and conference papers and is the recipient of many competitive research grants. For his contributions to speaker recognition research, he was a Faculty Associate with the Institute of Infocomm Research (A*STAR), Singapore in 2009–2018, and is currently an Advisory Board member of the AI Speech Lab at AI Singapore. Professor Ambikairajah was an Associate Editor for the IEEE Transactions on Education from 2012–2019. He received the UNSW Vice-Chancellor's Award for Teaching Excellence in 2004 for his innovative use of educational technology and innovation in electrical engineering teaching programs, and again in 2014 he received the UNSW Excellence in Senior Leadership Award and in 2019 he was the recipient of the People's Choice Award as part of the UNSW President's Awards. Professor Ambikairajah was an APSIPA Distinguished Lecturer for the 2013–14 term. He is a Fellow and

a Chartered Engineer of the IET UK and Engineers Australia (EA) and is a Senior Member of the IEEE and a Life Member of APSIPA

Haizhou Li received the B.Sc., M.Sc., and Ph.D degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. Dr Li is currently a Professor at the Department of Electrical and Computer Engineering, National University of Singapore (NUS). His research interests include human language technology, and neuromorphic computing. Prior to joining NUS, he taught in the University of Hong Kong (1988–1990) and South China University of Technology (1990–1994). He was a Visiting Professor at CRIN in France (1994–1995), Research Manager at the Apple-ISS Research Centre (1996–1998), Research Director in Lernout & Hauspie Asia Pacific (1999–2001), Vice President in InfoTalk Corp. Ltd. (2001–2003), and the Principal Scientist and Department Head of Human Language Technology in the Institute for Infocomm Research, Singapore (2003–2016). Dr Li has served as the Editor-in-Chief of IEEE/ACM Transactions on Audio, Speech and Language Processing (2015–2018), a Member of the Editorial Board of Computer Speech and Language since 2012, a Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the President of the International Speech Communication Association (2015–2017), the President of Asia Pacific Signal and Information Processing Association (2015–2016), and the President of Asian Federation of Natural Language Processing (2017–2018). He was the General Chair of ACL 2012, INTERSPEECH 2014, ASRU 2019. Dr Li is a Fellow of the IEEE, and a Fellow of the ISCA. He was a recipient of the National Infocomm Award 2002, and the President's Technology Award 2013 in Singapore. He was named Nokia Visiting Professors in 2009 by the Nokia Foundation, and Bremen Excellence Chair Professor in 2019 by the University of Bremen, Germany.