Use of Claimed Speaker Models for Replay Detection

Gajan Suthokumar[*†], Kaavya Sriskandaraja[*], Vidhyasaharan Sethu[*], Chamith Wijenayake[*], Eliathamby Ambikairajah[*†], Haizhou Li[‡]
[*] The University of New South Wales, Sydney, Australia
[‡] The National University of Singapore, Singapore
[†] DATA61, CSIRO, Sydney, Australia
E-mail: g.suthokumar@unsw.edu.au, k.sriskandaraja@unsw.edu.au, v.sethu@unsw.edu.au, c.wijenayake@unsw.edu.au, e.ambikairajah@unsw.edu.au, haizhou.li@nus.edu.sg

***Abstract*— Replay attacks are the simplest form of spoofing attacks on automatic speaker verification (ASV) systems and consequently the detection of these attacks is a critical research problem. Currently, most research on replay detection focuses on developing a stand-alone countermeasure that runs independently of a speaker verification system by training a single common spoofed model as well as a single common genuine model. This paper investigates the potential advantages of sharing speaker data between the speaker verification system and the replay detection system. Specifically, it explores the benefits of using the claimed speaker's model in place of the common genuine model. The proposed approach is validated on a modified evaluation set of the ASVspoof 2017 version 2.0 corpus and show that the use of adapted speaker models is far superior to the use of a single common genuine model.**

I. INTRODUCTION

The development of effective anti-spoofing countermeasures for use with automatic speaker verification (ASV) is an area of rapidly growing research interest. There are four broad approaches to spoofing, namely, impersonation [1], replay [2], speech synthesis [3] and voice conversion [4]. Here, impersonation refers to one person mimicking another; replay refers to recording the speech of a person and playing it back; speech synthesis refers to text-to-speech waveform generation; and voice conversion refers to an automatic system that transforms the speech of one person to sound like that of another. Among these, replay attacks are the simplest and the most easily accessible forms of attack compared to other three types. Studies on the vulnerabilities of state-of-the-art automatic ASV systems to replay attacks [3, 8] show that replay attacks are highly effective leading to significant increases in both equal error rate (EER) and false acceptance rate (FAR). Consequently, the development of techniques for the detection and prevention of replay spoofing attacks becomes a critical area of research and is the focus of this paper.

There are two broad approaches to incorporating anti-spoofing countermeasures [5]. One approach is to have a 'standalone countermeasure' that operates independently of the ASV system. The alternative approach is to make the ASV system itself more robust to a spoofing attack; this is called the 'integrated approach'. Both approaches have their merits: the integrated system allows for a shared front-end [6], which could be computationally more efficient; while the standalone countermeasure can operate independently without modifying the ASV system and also allows the use of different front-ends and modelling techniques [5].

There are very few studies that have investigated the integrated approach for speech synthesis and voice conversion spoofed speech [6]–[8], while no studies on integrated spoofing detection have been reported for replay attacks. In [7] authors focussed on GMM-UBM (Gaussian mixture model – universal background model) framework. It uses an additional UBM trained on spoofed speech. Their experiments on the ASVspoof 2015 [9], using out-of-domain data (IDIAP AVspoof [10]) for training, showed that the proposed method was able to considerably improve the ASV performance for spoofing impostors compared to the baseline with or without a spoofing detector, without compromising the performance under zero-effort spoofing. On the other hand, Khoury et al. [6] and Sizov et al. [8] adopted standard i-vectors and a PLDA (probabilistic linear discrimination analysis) back-end for joint analysis to create spoofing detection and ASV system.

Since ASVspoof 2017 challenge [11], the public availability of the dataset has led to an increased focus on standalone countermeasures for replay detection. Front ends based on variants of spectral features, long-term spectral statistics [12], time-domain features, voice source [13] and different variants of deep neural network based systems [14]–[16] have been investigated, extensively. Such features include spectral centroid magnitude coefficient (SCMC) [17], constant-Q cepstral coefficient (CQCC) [18], single frequency filter cepstral coefficient [19], inverse-Mel cepstral coefficients (IMFCC) [20], rectangular filter cesptral coefficients (RFCC) [17], scattering coefficients [21], and variable length teager energy separation based instantaneous frequency cesptral coefficients (VESA-IFCC) [22].

Current research on spoofing detection system has primarily focused on either improving the back-end modelling, or the development of novel features to detect spoofed speech. In practice, all features employed in spoofing detection also exhibit variability due to a number of factors such as acoustic variability (including channel effects), speaker variability, phonetic variability, etc. These sources of variability can subsequently lead to less precise models and reduce the accuracy of spoofing detection systems. To mitigate this, the variability should either be incorporated into the models or it should be normalized. This is supported by recent work where by the use of cepstral mean variance normalization (CMVN)

improved the reliability of spoofing detection across the diverse variations in replay attacks [17], [18]. In this paper, we propose an approach where speaker variability if explicitly modelled rather than normalized.

Given that replay detection will always run in conjunction with an ASV system, it is reasonable to expect that information about the claimed speaker (available to the ASV system in the form of enrolment data) can also be used by the replay detection system. Specifically, it is proposed that the target speaker enrolment data can be used to estimate claimed speaker specific models (herein referred to as speaker dependent models) of genuine speech that are unaffected by speaker variability and in turn improve the performance of replay detection systems. Current systems do not adopt this approach and instead focus on an 'in-wild' type of spoofing detection which is evaluated without taking into consideration the speaker verification aspect. This is reinforced by databases such as ASVspoof 2015 and ASVspoof 2017 where the speakers in enrolment and evaluation set are non-overlapped.

II. Proposed Use of Claimed Speaker Models

Current replay detection systems employ a 'genuine' speech model and a 'spoofed' speech model that is common across all test utterances as shown in Fig. 1. In this paper, we propose an alternative approach where instead of the common genuine model, we employ test utterance specific genuine models based on the claimed speaker identity (refer Fig. 2).

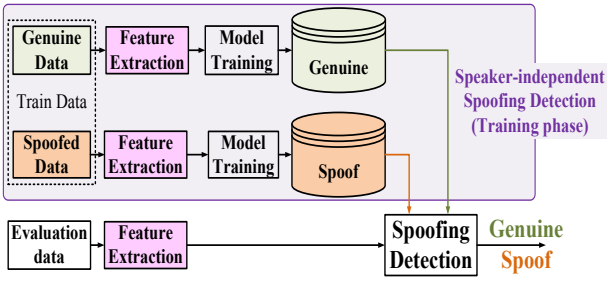

Fig. 1: Schematic diagram of a typical stand-alone spoofing detection system

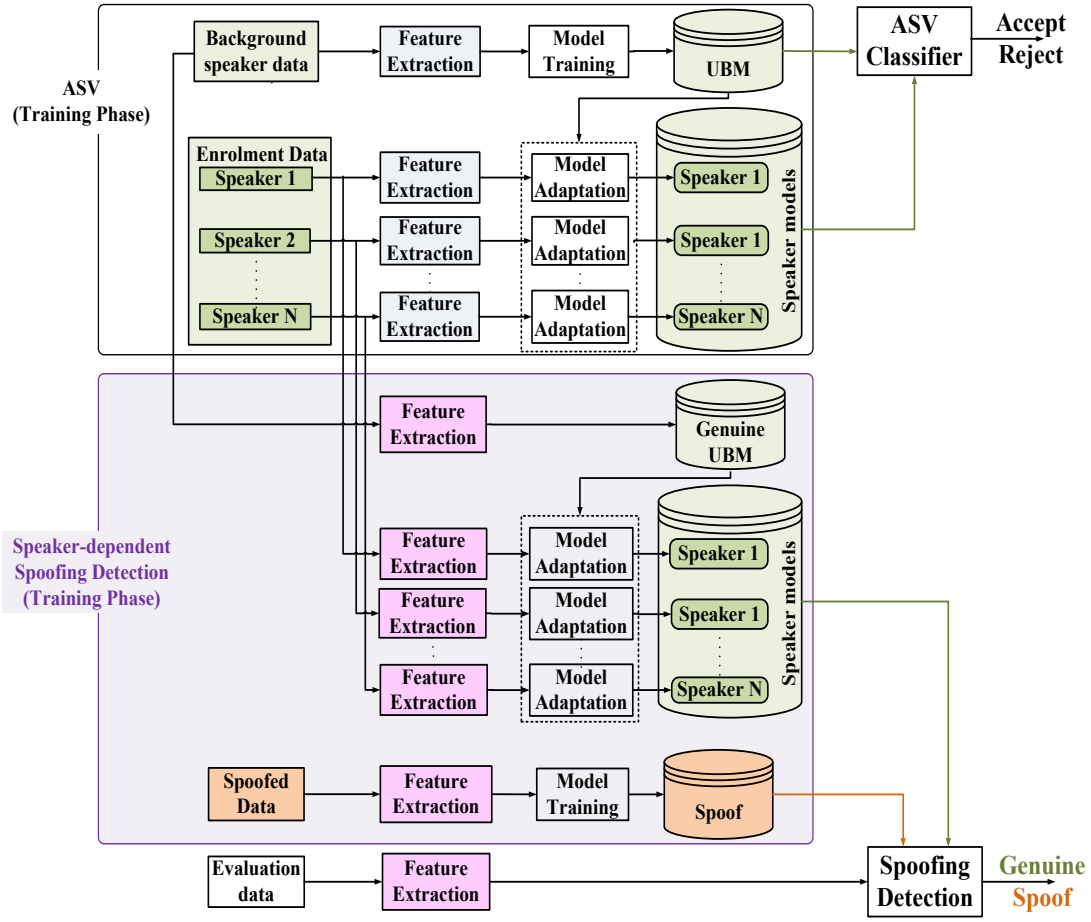

Fig. 2: Schematic of proposed modelling framework to incorporate claimed speaker models

The proposed approach is motivated by the fact that a common genuine model, typically trained on speech from multiple speakers, will be affected by speaker variability. However, 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 be utilised to train genuine speech models specific to every possible claimed speaker.

The approached proposed in this paper employs GMM based models of spoofed and genuine speech. Specifically, a background model for genuine speech is initially trained on genuine speech from multiple speakers, using the EM algorithm, and is referred to as the genuine universal background model (UBM). Following this, claimed speaker specific genuine models are adapted from genuine UBM using the enrolment data via MAP adaptation (refer Fig. 2). Given these models and a test utterance, the final decision about whether it is genuine speech or replayed speech can be based on the log-likelihood ratio between the claimed speaker model and the spoof model given by:

$$LLR(X) = \log P\left(X|\theta_{speaker}\right) - \log P\left(X|\theta_{spoof}\right) \quad (1)$$

where X denotes the set of feature vectors from a test utterance, $\theta_{speaker}$ denotes the GMM corresponding to claimed speaker, θ_{spoof} denotes the model of spoofed speech that is common to all test utterances.

III. Databases and Data Preparation

A. *ASVspoof 2017 (Version 2.0)*

The ASVspoof 2017 corpus [11] makes use of the RedDots corpus [23], as well as replayed versions of the same data [24]. The main technical aim of the ASVspoof 2017 challenge was to assess spoofing attack detection accuracy 'in-wild' conditions, thereby advancing research towards generalised spoofing countermeasures to detect replay attacks in particular. This database is partitioned into training, development and evaluation sets as shown in Table 1 (more details can be found in [11], [18]). ASVspoof 2017 Version 2.0, is released in 2018 by the challenge organisers, is an updated version of ASVspoof 2017 version 1.0, correcting several data anomalies found in the original. All reported experiments and analysis in this paper are conducted on ASVspoof 2017 version 2.0 corpus (herein referred to as ASVspoof 2017). Fig. 3 shows the number of utterances per each speaker in the evaluation set of ASVspoof 2017 corpus.

Table 1: ASVSpoof 2017 Version 2.0

Subset	# Speakers	# Utterances	
		Genuine	Spoof
Train	10	1507	1507
Dev	8	760	950
Evaluation	24	1298	12008

The replayed speech in these partitions was created using different playback and recording devices in various environments. Metadata of this corpus includes the ground-truth labels, which indicate genuine/replayed speech, along with speaker ID, phrase ID, and replay configuration details (details about replay and recording devices and acoustic environments). All three subsets are non-overlapping in terms of data collection locations. However the same 10 phrases appear in all the three sets.

B. *Repartitioning of ASVspoof 2017 Evaluation set*

All three partitions of ASVspoof 2017 corpus are also non-overlapping in terms of speakers. However, this in turn means that these partitions cannot be used to study the proposed approach since 'Enrolment data' corresponding to 'claimed speakers' in the test utterances would not be available during model training. Consequently the corpus was repartitioned to allow for the proposed approach to be studied.

As shown in Fig. 4, one utterance of each passphrase of genuine data for every speaker in ASVspoof 2017 evaluation set are separated as an 'Enrolment set' that can be used to learn the speaker models (i.e. totally 10 utterances per speaker). Since there are only 10 genuine utterances (see Fig. 3) for speakers from 'M0036' to 'M0042', these speakers are

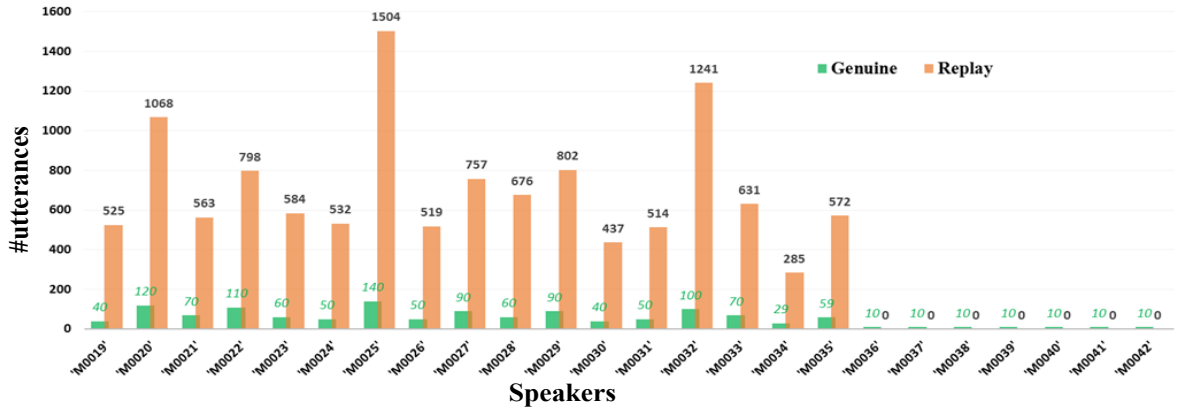

Fig. 3: Number of utterances for each speaker in ASVspoof 2017 evaluation set

excluded from 'Enrolment set'. Remaining utterances of speakers (from 'M0019' to 'M0035'), which are not included in the 'Enrolment set' is partitioned off as 'Speaker-specific Test set'. Table 2 shows the number of utterances and number of speakers in each data partitions which are used in this paper.

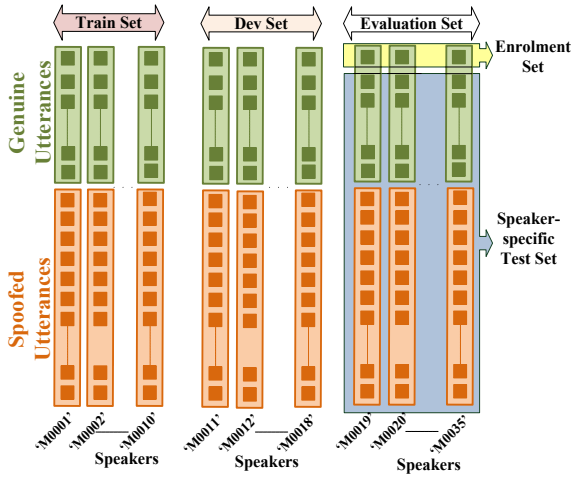

Fig. 4: Schematic diagram of repartitioned 'Enrolment set' and 'Speaker-specific Test set' from ASVspoof 2017 evaluation set. ASVspoof 2017 Train and Dev set remains same. Total number of utterances for each speaker is not represented in this figure, which is not equal to every speaker

Table 2: Statistics of the ASVspoof 2017 partition which used to create 'Enrolment set' and 'Speaker-specific Test set'

Subsets	#Speakers	#Utterances	
		Genuine	Replay
ASVspoof 2017 evaluation set	24	1298	12008
Enrolment set[1]	17	170	0
Speaker-specific Test set[1]	17	1058	12008

IV. Front-End Features

Three features are considered in this study: CQCCs [11], RFCCs [17] and joint acoustic temporal modulation spectrum based features. These three different features are selected for two main reasons: (i) they provide state-of-the-art performance in replay attacks; (ii) they are all extracted from windows of different durations - CQCCs use a very short feature window (~8ms), RFCCs use a standard window (20ms) and joint acoustic temporal modulation spectrum based features are extracted from an entire utterance.

[1]The details of 'Enrolment set' and the 'Speaker-specific Test set' can be found in http://www2.ee.unsw.edu.au/ASVspoof/

A. *Joint acoustic (spectro-temporal) modulation spectrum based features*

Long-term temporal modulation static and dynamics features are derived from the 'Joint acoustic modulation spectrum' (Fig. 5): modulation centroid frequency cosine coefficients (MCF-CC) and modulation static energy cepstral coefficients (MSE-CC) [25]. MCF-CC features capture the variation of the modulation peak energy within acoustic frequency bins due to spoofed speech channels. As illustrated in Fig. 5, the 0th modulation bin energies ($m = 0$) of the normalized modulation spectrum along the acoustic frequencies are retained as a feature vector, which is referred to as the modulation static energy (MSE). The concatenation of these two features is referred as the spectro-temporal modulation features (STMFs) for rest of this paper. This used as one of the front-end for the proposed framework. For more details readers are referred to [25].

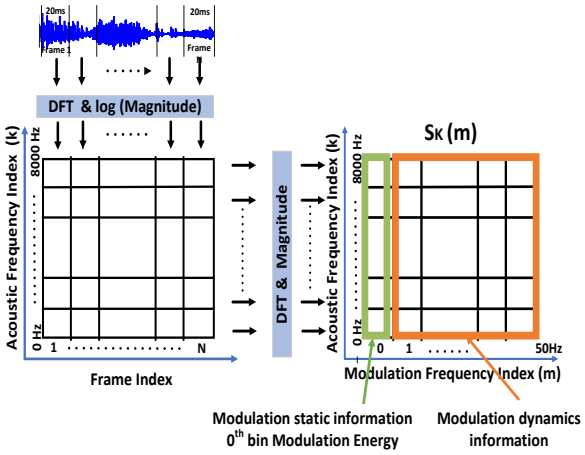

Fig. 5: Computation of Short term log Spectrogram (left) and Joint acoustic modulation spectrum (Right); Regions of Modulation Static (Green) and Dynamics information (Orange)

B. *Constant-Q Cepstral Coefficients (CQCC)*

CQCC features [26] are derived using a Constant-Q transform [27]. These features were first introduced to speech synthesis and voice conversion spoofing detection and have since becomes a standard feature set in spoofing detection system.

C. *Rectangular Filter Cepstral Coefficients (RFCC)*

RFCC features are similar to conventional MFCC but are extracted using a filter bank of equally spaced rectangular filters. This feature set has been shown to be effective in spoofing detection for not only replay attacks but also for speech synthesis and voice conversion attacks [17], [28].

V. Experimental Settings

A number of experiments were carried out to evaluate the proposed approach. In these experiments three different

metrics are employed to quantify performance. Namely, (a) 'Speaker-wise EER' which is calculated by pooling only the genuine and spoof trials of the target speaker; (b) 'Overall EER' which is derived as considering all the trials in the corresponding evaluation set; and (c) 'Average EER' which is calculated as the average of the 17 speaker-wise EERs.

A. Front-end configurations

The STMF features are extracted using the same parameters as in [25]. 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.

RFCC features are derived along 20ms window with 50% overlap from speech utterances pre-emphasised by factor of 0.97 using 30 linearly spaced rectangular filters. This feature vector is 90 dimensional with static, delta and delta-delta coefficients.

For the derivation of CQCC features we have used the same configuration as used in the ASVspoof 2017 challenge baseline [11]. A Constant-Q Transform is applied with a maximum frequency of $f_{max} = 8kHz$, which is the Nyquist frequency and the minimum frequency

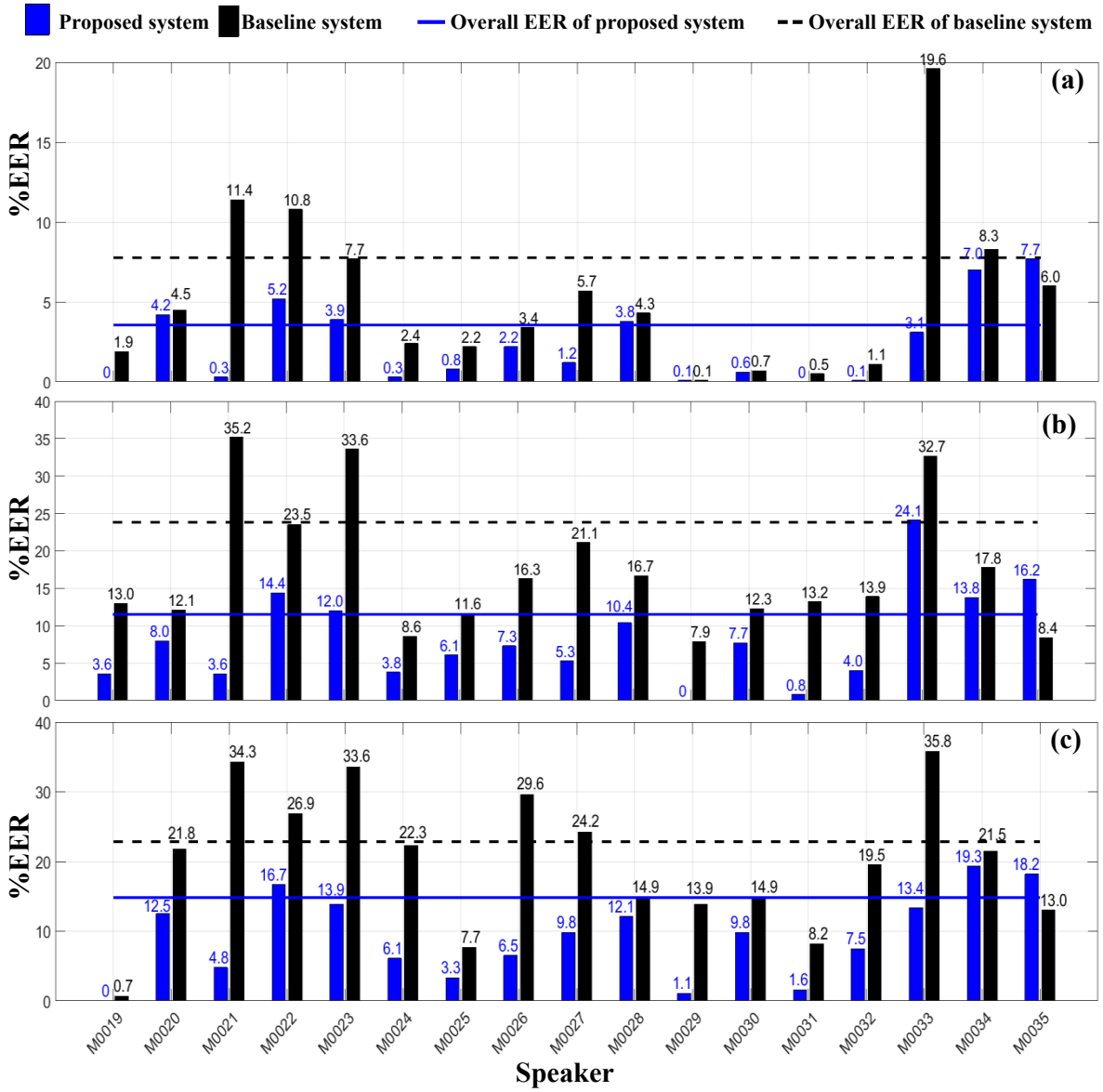

Fig. 6: Comparison of (a) STMF (b) CQCC and (c) RFCC features as front-end for proposed and baseline approaches with respect to speaker-wise EER and Overall EER in 'Speaker-specific Test set'

$f_{min} = f_{max}/2^9 \approx 15\text{Hz}$, 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 (static including c_0, delta and delta-delta coefficients).

Normalisation techniques are not applied to any of the features in this work. The Cepstral mean and variance normalization (CMVN) is commonly employed as a feature normalisation technique [17]. However, CMVN generally aims to reduce channel effects, which is counter-intuitive for replay detection which is essentially based on channel discrimination.

B. Back-end configurations

In the proposed system, the spoofed speech model and the genuine UBM are implemented as 512 mixture Gaussian mixture models (GMMs) for RFCC and CQCC features and 4 mixture GMMs for STMF features trained using the EM algorithm with random initialization. ASVspoof 2017 train and dev set used to model the genuine UBM and spoofed speech models (no overlap with test speaker data). The genuine speech models specific to each claimed speaker are then estimated from the genuine UBM via MAP adaptation (mean, variance and weights with a relevance factor of 1 were chosen based on dev set results) based on the available Enrolment data (refer section III).

The baseline systems (using a single common genuine model) also use 512 mixtures (identical to the ASVspoof 2017 baseline) for modelling RFCC and CQCC features. The baseline system employing STMF features uses GMMs with 4 mixture components.

VI. Results and Analysis

The primary metric for evaluation is the equal error rate (%EER). Overall EER (i.e pooled EER) is derived by using the entire 'Speaker-specific Test Set' (refer section V-B), unless otherwise indicated.

A. Comparison of experimental results with state-of-the-art features

The performance of the proposed and baseline approaches are compared in Table 3 for all three front-ends. It can be seen from these results that the proposed approach outperforms the baseline by a significant margin in all three cases. It is also interesting to note the long term STMF features are superior to short term RFCC and CQCC features, both when using the baseline approach and when using the proposed approach.

Fig. 6 illustrates individual performances of the proposed and baseline systems using STMF, CQCC and RFCC features. These plots reveal that the proposed approach of using claimed speaker models of genuine speech is consistently superior to the baseline. Here it should be noted that the number of trials (per speaker) used to estimate the speaker specific EERs is significantly lower than the number of trials in the standard ASVspoof test set and consequently these speaker specific EERs should be considered as indicative results only. However, the overall EERs shown in Fig 6 are estimated from more or less the same number of trials as the standard ASVspoof test set and should have similar confidence intervals.

Table 3: Comparison of the STMF, CQCC and RFCC features with proposed and baseline approaches in terms of %EER in ASV spoof 2017 'Speaker-specific Test set'

Features	Speaker-independent (Baseline) system	Speaker-dependent (Proposed) system
	Overall EER	Overall EER
STMF	7.75	3.54
CQCC	23.81	11.51
RFCC	22.84	14.82

B. Analysis of Model Separation

The proposed approach is based on the hypothesis that the use of genuine models specific to the claimed speaker eliminates speaker variability from the genuine model. Results reported in the previous section support this hypothesis with the proposed approach consistently outperforming the baseline system. Additionally, in an attempt to discern if the proposed approach leads to better discriminability, we directly estimate the separation between genuine and spoofed models as the Kullback-Leibler (KL) divergence between the corresponding GMMs. Specifically, we compare the KL divergence between genuine models corresponding to the claimed speakers and the spoofed model to the KL divergence between the single genuine model and the spoofed model of the baseline system.

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 $X \epsilon \mathbb{R}^D$, the KL divergence is of $\mathcal{P}_2$ from $\mathcal{P}_1$ is defined as [29]:

$$KL(\mathcal{P}_1, \mathcal{P}_2) = \int_X \mathcal{P}_1(X) \ln\left(\frac{\mathcal{P}_1(X)}{\mathcal{P}_2(X)}\right) \quad (2)$$

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 is defined as [29]:

$$S\_KL(\mathcal{P}_1, \mathcal{P}_2) = \frac{1}{2}(|KL(\mathcal{P}_1, \mathcal{P}_2) + KL(\mathcal{P}_2, \mathcal{P}_1)|) \quad (3)$$

A Monte Carlo approximation based symmetric KL divergence [30] is used to measure the distance between spoofed and genuine speech models. 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. The KL divergence between the spoofed model and each of the claimed speaker genuine models in the proposed approach are shown in Fig. 7. This can be compared to the KL divergence between the spoofed and genuine models of the baseline system shown as a red line in Fig. 7. In

this analysis both the proposed and the baseline systems used the STMF front-end since that had the best performance among the three front-ends.

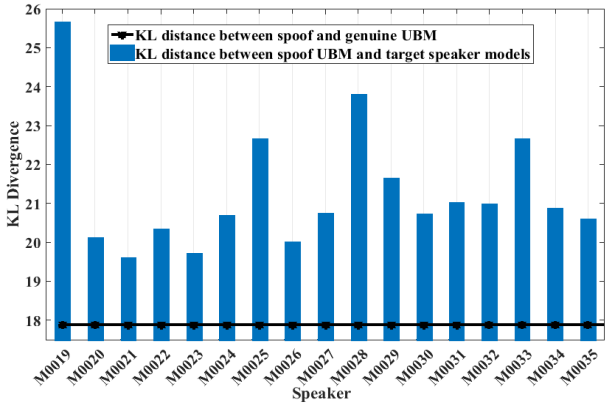

Fig. 7: Comparison of KL divergence of STMF features from the spoof UBM to the genuine UBM and from the spoof UBM to speaker models

C. *Investigating the system performance in terms of different replay configuration*

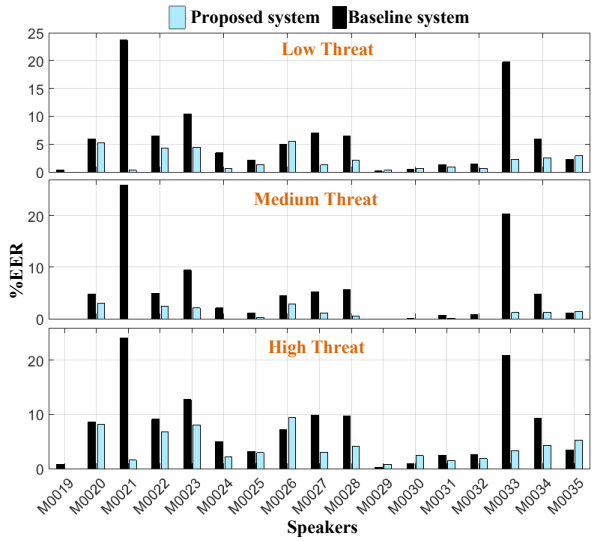

Fig. 8: Comparison of results of STMF features grouped by the acoustic environment into low, medium and high threat attacks for speaker independent (baseline) and speaker-dependent (proposed) systems.

The ASVspoof 2017 database contains recordings collected with diverse replay configurations (RCs), each comprising of one recording device, one playback device, and one acoustic environment. In order to aid analysis, the distinct RCs were reduced by grouping together overlapping configurations [18]. To analyse how well the proposed speaker-dependent system performs compared to speaker independent system, replay detection performance in terms of %EER for different qualities of environments, play-back and recording devices are evaluated.

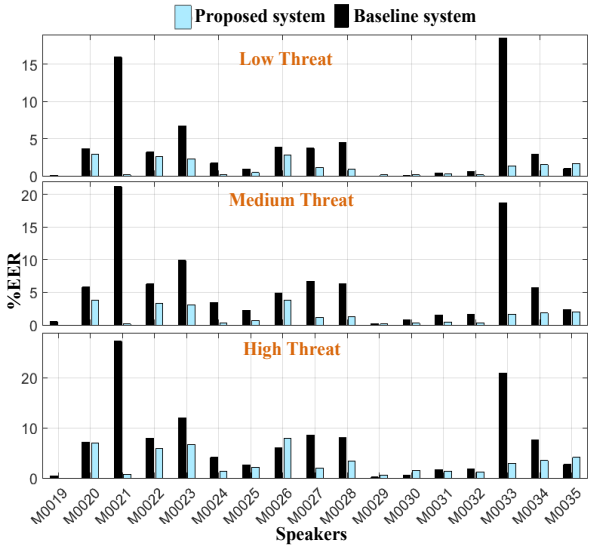

Fig. 9: Comparison of the results of STMF features grouped by the playback device into low, medium and high threat attacks for speaker independent (baseline) and speaker-dependent (proposed) systems.

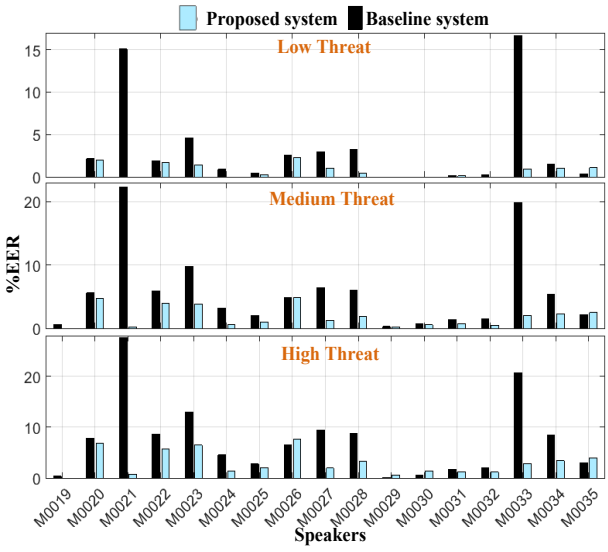

Fig. 10: Comparison of the results of STMF features grouped by the recording device into low, medium and high threat attacks for speaker independent (baseline) and speaker-specific (proposed) systems.

Fig. 8-10 show results of STMF features grouped by acoustic environment, playback device, recording device into low, medium and high threat attacks for speaker independent (baseline) and speaker-dependent (proposed) systems respectively. It is evident that the proposed speaker-dependent system not only superior in terms of overall EER and speaker-dependent EER (refer Fig. 6), also it outperforms the baseline

in most of the replay configurations for all of the speakers. However, number of test trails in all the replay configurations is not same for each speaker, so that the direct comparison can be biased to the group which has more trials. In addition to that, high threat replay configurations show higher EER compared to low and medium replay configurations, similar to baseline systems. Hence, a system that could detect the artefacts for high threat replay configurations will be interesting future area to work with.

D. Investigating the effect of amount of adaptation data

All of the previous experiments reported in this paper are based on the 'Speaker-specific Test set' from ASVspoof 2017 where each speaker model is adapted using 10 utterances (one per passphrase). In this section we analyse the effect of using a larger enrolment set for adaptation. To analyse this effect, experiments are carried out with different set of enrolment data. As further utterances are added to the enrolment set partition, they should be excluded from the 'speaker-specific Test set' to ensure both sets are non-overlapped. To compare the performance with the constant test set, smallest one correspond to the largest adaptation data (50%), is used as 'speaker-specific test set' for only the experiments reported in section.

Table 4: Comparison of STMF feature performance for the different amount of data used for the speaker adaptation (in terms of %EER). Constant test set, smallest one correspond to the 50% adaptation data, is used as 'speaker-specific test set' for evaluation.

#utterance use for adaptation for each speaker	Speaker-independent (Baseline) system		Speaker-dependent (Proposed) system	
	Avg EER	Overall EER	Avg EER	Overall EER
No adaptation [25]	5.25	7.32	N/A	N/A
10 utterance			2.52	3.12
20% of the total utterance			2.38	2.48
20 utterance			1.87	2.18
50% of the total utterance			1.45	1.66

Since STMF features performed better than CQCC and RFCC in the previous experiments, only the performance of STMF features is tabulated in Table 4. However, the same trends as STMF features, which is larger the adaptation data, better the performance, were observed for CQCC and RFCC features when increasing the number of utterances used for speaker model adaptation. Also, it is observed that the performance of the proposed system for all nine replay configurations improved for all the speakers with the proportional to the amount of adaptation data. Hence, high threat conditions of replay attack can be detected with high accuracy, if larger amount of adaptation data used.

VII. CONCLUSIONS

This work investigates the effect of incorporating speaker specific information into a replay spoofing detection system by using claimed speakers' models which can be estimated from enrolment data that would be available to any speaker verification system. A single model of genuine speech, as would be employed in most current replay detection systems, would always be affected by speaker variability, since it would be trained on data from multiple speakers. However, in the proposed approach claimed speaker specific models of genuine speech are employed (trained on enrolment data corresponding to that speaker), thus significantly reducing speaker variability. Experimental results based on the ASVspoof 2017 corpus show that the proposed approach reduces equal error rates by a factor of two when compared to the use of a single common genuine model. Further, this study proved that STMF features show superior performance in replay detection not only for speaker independent models but also for speaker-specific models. Outcomes from our work motivate the study of spoofing detection systems that include speaker-specific information.

REFERENCES

[1] Y. W. L. Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," *Proc.of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, no. 5. pp. 145–148, 2004.

[2] J. Lindberg and M. Blomberg, "Vulnerability In Speaker Verification - A Study Of Technical Impostor Techniques," *Proc. Eurospeech*, vol. 3, no. MARCH 2001. pp. 1211–1214, 2001.

[3] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 8, pp. 2280–2290, 2012.

[4] T. Kinnunen, Z. Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. pp. 4401–4404, 2012.

[5] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.

[6] E. Khoury, T. Kinnunen, A. Sizov, Z. Wu, and S. Marcel, "Introducing I-vectors for joint anti-spoofing and speaker verification," *INTERSPEECH*, no. September, pp. 61–65, 2014.

[7] A. K. Sarkar, M. Sahidullah, Z.-H. Tan, and T.

Kinnunen, "Improving Speaker Verification Performance in Presence of Spoofing Attacks Using Out-of-Domain Spoofed Data," *Interspeech*, pp. 2611–2615, Aug. 2017.

[8] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint Speaker Verification and Antispoofing in the i-Vector Space," *IEEE Trans. Inf. Forensics Secur.*, 2015.

[9] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, pp. 1–5, 2015.

[10] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the Vulnerability of Speaker Verification to Realistic Voice Spoofing," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2015, pp. 1–8.

[11] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Interspeech*, 2017, pp. 2–6.

[12] H. Muckenhirn, P. Korshunov, M. Magimai-Doss, and S. Marcel, "Long-Term Spectral Statistics for Voice Presentation Attack Detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 11, pp. 2098–2111, 2017.

[13] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features," 2017.

[14] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Interspeech*, 2017, pp. 82–86.

[15] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, vol. 2017–Augus, pp. 102–106.

[16] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," in *Interspeech*, 2017, pp. 97–101.

[17] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-Results on the ASVspoof 2017 Challenge," in *Interspeech*, 2017, pp. 7–11.

[18] M. Todisco, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," in *Odyssey*, 2018, pp. 296–303.

[19] K. N. R. K. Raju Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "Detection of replay attacks using single frequency filtering cepstral coefficients," in *Interspeech*, 2017, pp. 2596–2600.

[20] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, pp. 27–31.

[21] K. Sriskandaraja, G. Suthokumar, V. Sethu, and E. Ambikairajah, "Investigating the use of scattering coefficients for replay attack detection," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1195–1198.

[22] H. A. Patil, M. R. Kamble, T. B. Patel, and M. H. Soni, "Novel variable length teager energy separation based instantaneous frequency features for replay detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, vol. 2017–Augus, no. August, pp. 12–16.

[23] K. A. Lee, A. Larcher, G. Wang, P. Kenny, D. Van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, H. Li, T. Stafylakis, M. J. Alam, A. Swart, and J. Perez, "The RedDots Data Collection for Speaker Recognition," *Interspeech*, pp. 2996–3000, 2015.

[24] T. Kinnunen, M. Falcone, L. Costantini, R. Gonz, D. Thomsen, A. Sarkar, Z. Tan, M. Todisco, N. Evans, V. Hautam, K. A. Lee, and F. U. Bordoni, "Reddots Replayed : a New Replay Spoofing Attack Corpus for Text-Dependent Speaker Verification Research," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2017, pp. 5395–5399.

[25] G. Suthokumar, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Modulation Dynamic Features for the Detection of Replay Attacks," in *Interspeech*, 2018, pp. 691–695.

[26] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Comput. Speech Lang.*, vol. 45, pp. 516–535, 2017.

[27] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Am.*, vol. 89, no. 887, pp. 2698–662, 1991.

[28] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," *INTERSPEECH*, vol. 2015–Janua, pp. 2087–2091, 2015.

[29] S. Kullback, *Information theory and statistics*, 2nd ed. Dover Publications, 1968.

[30] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in speech based emotion models - Analysis and normalisation," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 7522–7525, 2013.