Independent Modelling of Long and Short Term Speech Information for Replay Detection

Gajan Suthokumar
UNSW Sydney

Kaavya Sriskandaraja
UNSW Sydney

Vidhyasaharan Sethu
UNSW Sydney

Independent Modelling of Long and Short Term Speech Information for Replay Detection

Gajan Suthokumar[1,2], *Kaavya Sriskandaraja*[1], *Vidhyasaharan Sethu*[1], *Chamith Wijenayake*[1], *Eliathamby Ambikairajah*[1,2]

[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2]DATA61, CSIRO, Sydney, Australia
g.suthokumar@unsw.edu.au

Abstract

Cepstral normalisation is widely employed in replay detection systems. However, incorporating some information that is lost during normalisation may be useful. Additionally, anti-spoofing systems may further benefit from not treating all speech frames identically. In this paper, we separate speech information based on two different criteria and model them independently. Three novel approaches are proposed based on (1) long and short term information; (2) high and low energy frames; and (3) a combination of the two. Experiments were conducted on the ASVSpoof2017 (V2.0) corpus and the best results correspond to an EER of 8.67% with a relative improvement of 29%.

Index Terms: speaker verification, spoofing detection, replay

1. Introduction

Significant improvements have been made in automatic speaker verification (ASV) in recent decades; however, concerns about security vulnerabilities continue to form a barrier to their widespread use. Speaker verification uses voice biometrics to verify the claimed speaker from a given speech utterance [1]. ASV systems are vulnerable to a diverse range of spoofing attacks, including speech synthesis (SS), voice conversion (VC), impersonation and replay [1], all of which have been shown to heavily degrade the robustness of ASV. Replay attacks, the playback of pre-recorded genuine speech, are arguably the most common ASV spoofing technique since they do not require attackers to have any special speech technology knowledge and can be mounted with relative ease using common consumer devices.

Developing anti-spoofing techniques to effectively mitigate the replay attacks generally aims to exploit one of the several factors: the fact that replayed speech would be a copy of a previous speech utterance [2]; differences in the transmission channel; or differences in the spectral properties of replayed speech. In literature, transmission channel differences are identified in the form of pop-noise [3], acoustic channel artefacts [4], and the detection of far-field recording [5]. Most of other techniques are based on the short term spectral cue differences. These include rectangular filter cepstral coefficients (RFCCs) [6] , spectral centroid magnitude coefficients (SCMCs) [6] , constant-Q cepstral coefficients (CQCCs) [7], scattering cepstral coefficients (SCCs) [8] and inverse Mel frequency cepstral coefficients (IMFCCs) [9]. In addition to the spectral based features, phase [10] and voice source features [10] have also been investigated. Spectral cues captured by the short term features derived from a linear frequency scale have dominated over warped frequency scales [6]. Different variants of neural network (NN) based systems [11] have also been investigated. Gaussian mixture models (GMMs) remain superior to other classifiers [6], [11]. Apart from these individual features and classifiers, the literature also indicates that the score fusion of different types of features and systems can perform better in replay detection.

Cepstral normalisation shown to be highly beneficial in replay detection in the form of cepstral mean normalisation (CMN) [6] and cepstral mean and variance normalisation (CMVN) [12], which normalise the temporal cepstral statistics (mean and variance) of short term (ST) features across each dimension. Even though, the cepstral normalisation of features in replay detection may at first seem counter-intuitive, the study of [12] argued that this may help to align both genuine and replayed speech distributions on to a similar scale. However, cepstral normalisation might also remove the information that could be useful for replay detection. Because, normalisation techniques are used in many other speech applications to normalise nuisance channels [13][14], which arise due to heterogeneous conditions, e.g. recording device and environment. Also, long term (LT) spectral statistics have been shown to be effective in spoofing detection [15]. In addition to that, temporal features based on amplitude modulation have demonstrated the significance of the long term temporal dynamics in our previous work [16].

Apart from that, standard replay detection systems model all the frames identically, however voiced and unvoiced regions could mask channel information differently and a speaker’s voice might mask the channel information in voiced regions, so if unvoiced portions are focused on, the channel information may become more pronounced [17]. This further suggests that voiced and unvoiced frames might not contain identically emphasized discriminative information for replay detection.

In this paper, we propose systems based on the following two hypotheses: (a) Cepstral statistics of short term features that are removed during normalisation may contain discriminative information for replay detection; and (b) voiced and unvoiced region artefacts might not be emphasized in the same manner in the presence of replay channels. Specifically, we propose separating speech into regions with non-overlapping and complimentary information and modelling the differences between replayed and genuine speech in these regions independently and that are then fused at the score level. It is noted that the state-of-the-art replay detection systems model the cepstral normalised (short term) features only.

2. Proposed System Architectures

We proposed three architectures to individually model the distribution of (1) cepstral normalised features (short term) and cepstral statistics (long term) as shown in Figure 2; (2) high energy (HE) and low energy (LE) frames as shown in Figure 4, to form two spoofing detection systems that are then fused at the score level. The third proposed hybrid architecture is a combination of the independent parallel paths of high energy, low energy and cepstral statistics of the short term feature information.

2.1. Short and Long Term Based Separation

In order to determine the effectiveness of the cepstral mean and variance in replay detection, a t-SNE representation of cepstral mean and cepstral variance of RFCCs from genuine (green) and spoofed speech (orange) is shown in Figure 1, where the feature spaces within genuine and spoof classes of the entire training set is compared. It can be seen that the feature space of the cepstral mean of RFCC feature depicts good discriminability, while the cepstral variance of RFCC feature space shows less. It is arguable that the incorporation of either cepstral mean or cepstral variance (long term) of the RFCC feature independently with the cepstral normalised RFCC features (short term) as depicted in Figure 2 could be helpful for improved replay detection.

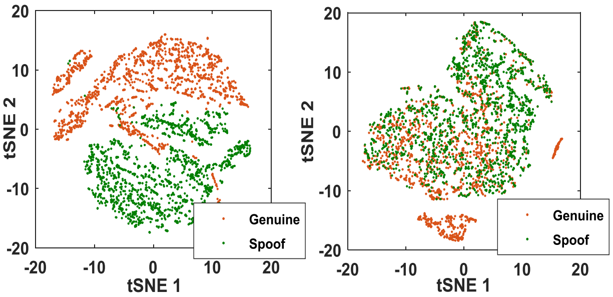

Figure 1: *t-SNE plot for genuine (orange) and spoof (green) RFCCs for Cepstral Mean (left) and Cepstral Variance (right) on the entire training set.*

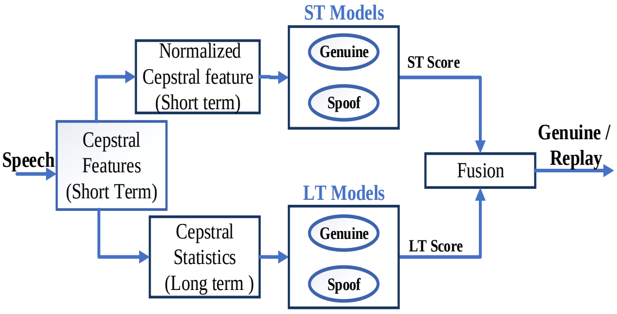
ST Models
Normalized Cepstral feature (Short term)
Genuine
Spoof
ST Score
Speech
Cepstral Features (Short Term)
Fusion
Genuine / Replay
LT Models
Cepstral Statistics (Long term)
Genuine
Spoof
LT Score

Figure 2: *System architecture based on short-term (ST) and long-term (LT) separation*

2.2. High Energy (HE) and Low Energy (LE) Based Separation

Speech frames are initially categorised as either high energy (HE) or low energy (LE) frames using a voice activity detector (VAD). Here we expect the LE frames to contain unvoiced speech. In order to determine effectiveness of separating high and low energy frames, we show the t-SNE plots of the RFCCs from genuine and spoofed speech in terms of HE and LE frames in Figure 3 and compare the feature spaces of genuine and spoofed speech. It can be seen that the features for genuine HE and LE frames occupy different spaces, and a similar pattern is observed for spoofed speech as well. Thus, it is also arguable that separate modelling of HE and LE frames could be helpful for improved replay detection, with an architecture as depicted in Figure 4. Similar technique has proved to be effective in synthetic speech detection in our previous work [18] since the voiced speech and unvoiced speech are not synthesized in same manner in speech synthesis and voice conversion algorithms.

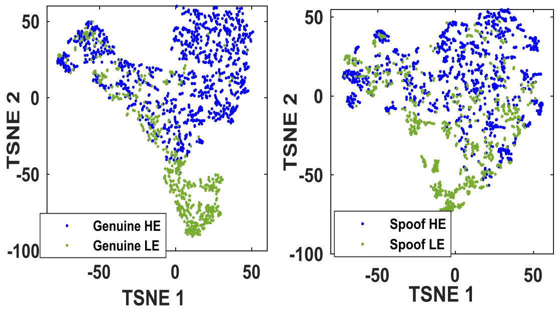

Figure 3: *t-SNE plots of the RFCC static features from a subset of the training set for genuine (left) and spoofed speech (right), in terms of high energy (HE) (blue) and low energy (LE) (green) frames.*

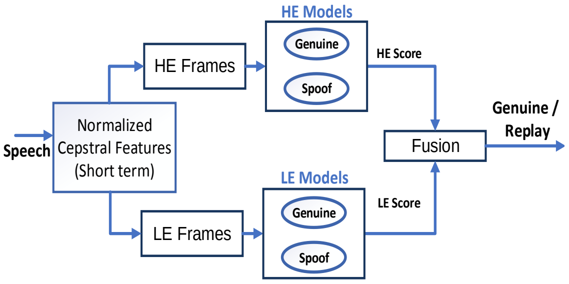

Figure 4: *System architecture based on high energy (HE) and low energy (LE) separation.*

2.3. Hybrid Architecture

A depiction of the proposed hybrid architecture is shown in Figure 5, to combine the individual gains of both of the proposed architectures.

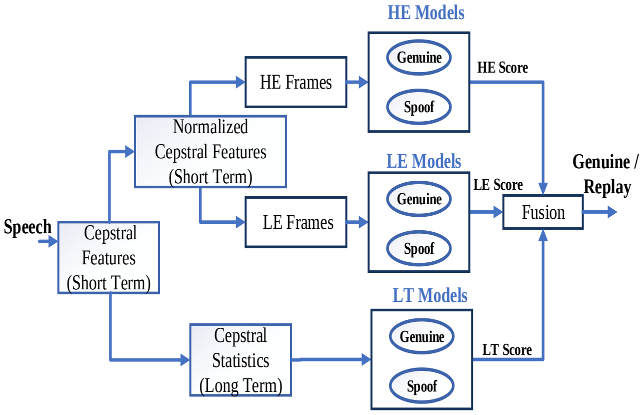

Figure 5: *Hybrid system.*

3. Experimental Setup

3.1. Database

The ASVspoof 2017 database consists of genuine recordings and their replayed versions as spoofed speech. Genuine utterances are sourced from the RedDots corpus. Spoofed utterances are created through recording and replaying genuine utterances using a variety of recording devices, playback devices and acoustic environments. This database consists of three non-overlapping subsets for training, development and evaluation. The ASVspoof 2017 Version 2.0 (V2.0) database [12] and the CQCC baseline are released earlier this year by the challenge organisers, is an updated version of the original ASVSpoof 2017 Challenge database Version 1.0 (V1.0) [7], correcting several data anomalies found in the original. As such, previously reported results using the V1.0 database are not directly comparable with the V2.0 results. In addition to this, the meta data of recording and playback devices, and the environmental conditions of the evaluation set is also released in V2.0. All reported experiments in this paper are conducted on the V2.0 database.

3.2. Voice Activity Detection

Vector quantization based VAD (VQVAD) [19] was employed in preference to other VAD's because of its unsupervised nature as well as it does not depend on a pre-determined threshold. Also, it shows better performance on short duration utterances which is suitable for ASVspoof 2017 V2.0 corpus [19]. The VQVAD is tuned for the ASV microphone environment with default parameters. HE frames are expected to include voiced frames, while LE frames are expected to include unvoiced frames and silence. The ASVSpoof 2017 V2.0 corpus consists of 54% HE frames and 46% LE Frames.

3.3. Front-End

RFCCs [6] and SCMCs [6] were used as the front-end for our experiments, as they are the state-of-the-art short term features for replay attack detection. They are extracted with a frame size of 20ms with 50% overlap. 40 dimensional static and dynamic features (i.e velocity and acceleration) are utilized. CMN is carried out for all our short term features unless otherwise specified.

3.4. Classifier and Score level Fusion

A 2-class GMM back-end was employed for genuine and spoofed speech detection. The GMMs were trained using the expectation maximization (EM) criterion, for genuine and spoofed speech with random initialization. The proportion of the amount of HE and LE frames in development set is considered in the selection of number of GMM mixtures. We investigated a range of numbers of GMM mixtures, eventually employed 512 mixtures for the baseline (i.e. no separation) and chose 256 each for the HE and LE models. A small number of mixtures to model the cepstral statistics (LT) is sufficient as it is an utterance level feature and 4 mixtures are found to be optimal. A linear regression based score level fusion from the Bosaris toolkit [20] is employed in order to combine the independent classifier scores, since the features associated with each models are highly complementary.

4. Results and Discussion

4.1. Long term Cepstral Statistics Features

Table 1 shows the development set performance for the long term cepstral statistics features (i.e. mean and variance) of static (S) and combined static and dynamic (i.e velocity and acceleration) (SD) for RFCCs and SCMCs. Firstly, the systems using cepstral mean features (i.e. LT$_M$ (S) & LT$_M$ (SD)) consistently performed better than those using cepstral variance features (i.e. LT$_V$ (S) & LT$_V$ (SD)). Secondly the cepstral means systems, LT$_M$ (S) performs better than LT$_M$ (SD), in contrast to the variance systems, while LT$_V$ (SD) performs better than LT$_V$ (S). The success of the cepstral statistics features is reasonable as the temporal dynamics of the replayed speech tend to be affected, presumably due to the heterogenous replay channels and environments.

Table 1. *Development results in terms of % EER for individual long term cepstral mean (LT$_M$) & variance (LT$_V$) systems, for RFCC and SCMC static (S) features, and combined static & dynamic (SD) features.*

System	RFCC	SCMC
LT$_M$ (S)	**13.65**	**13.64**
LT$_V$ (S)	22.53	23.04
LT$_M$ (SD)	**18.34**	**16.84**
LT$_V$ (SD)	20.68	21.20

4.2. Comparative Performance

Table 2 shows the development set performance for different combinations of the proposed systems and the baseline systems for RFCCs and SCMCs. It is noted that, our baseline front-ends are better than the improved CQCC system [12] released with ASVSpoof 2017 V2.0. The proposed HE+LE, ST+LT$_{M,}$ and hybrid systems outperform the baseline, which models only the normalised short term features. Again, the incorporation of the static LT$_M$ (S) features seems superior to the LT$_M$ (SD).

Table 3 shows the evaluation set performance for different combinations of the proposed best systems identified on the development set, as well as the baseline systems for RFCC and SCMC features. All of the proposed systems are superior to the baseline system, which does not independently model neither the separated speech information nor the cepstral statistics.

Table 2. *Development set results in terms of %EER for RFCC and SCMC features for the proposed systems.*

Architecture	System	RFCC	SCMC
Baseline	ST [6]	7.76	8.66
Proposed 1	ST+ LT$_M$(S)	6.50	7.35
	ST+ LT$_M$(SD)	6.68	7.68
Proposed 2	HE+LE	7.15	8.41
Proposed 3	HE+LE+ LT$_M$(S)	**6.12**	**6.99**
	HE+LE+LT$_M$(SD)	6.38	7.27

Table 4 compares the evaluation set performances of our best proposed system (HE+LE+LT$_M$(S)) with the previously reported best results [12], for different threat conditions as defined in [12]. The meta data of ASVSpoof 2017 V2.0 defines the different replay threat conditions for recording device, playback device and acoustic environments. It can be noticed that the proposed system is superior under all threat conditions.

Table 3. *Evaluation set results in terms of %EER for RFCC and SCMC features for the proposed systems.*

Architecture	System	RFCC	SCMC
Baseline	ST [6]	11.22	12.23
Long Term	LT_M(S)	16.96	17.05
	LT_M(SD)	16.94	15.46
Proposed 1	ST+ LT_M(S)	**10.28**	**10.09**
	ST+ LT_M(SD)	11.20	11.65
Proposed 2	HE+LE	10.42	11.01
Proposed 3	HE+LE+ LT_M(S)	**9.03**	**8.67**
	HE+LE+LT_M(SD)	9.88	9.82

Table 4. *Proposed best systems results in terms of % EER of RFCC/SCMC features for different threat conditions (low, medium and high) as defined in* [12] *in terms of % EER. (The best results previously reported on version 2.0* [12] *are given within parentheses).*

Conditions	Low	Medium	High
Environment	**8.44/8.28** (16.68)	**8.13/7.29** (18.73)	**14.52/13.13** (21.86)
Playback Device	**8.24/8.58** (16.64)	**7.04/6.6** (16.44)	**10.65/10.28** (18.37)
Recording Device	**7.77/7.27** (10.80)	**8.24/8.01** (15.69)	**9.99/9.73** (17.77)

5. Conclusions

A novel framework is proposed in this paper to independently model speech information separated based on two different criteria, prior to score fusion. This approach places a greater emphasis on relevant speech information compared to the standard approach. This is beneficial since this information encompasses complementary discriminative ability for replay detection, which is not well emphasized in the standard approach whereby all genuine and spoofed speech information is described by one GMM each. The incorporation of the long term cepstral statistics of the short term features that are discarded during cepstral normalization is proved to be beneficial for replay detection. The approach of independently modelling the high energy and low energy frames and cepstral statistics of the short term feature was found to be superior to the standard approach. The proposed framework has been validated on the ASVSpoof 2017 V2.0 corpus and the results consistently showed that the proposed approach is superior to the standard approach with a 29% relative improvement.

6. References

[1] 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.

[2] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," *2014 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA 2014*, pp. 92–96, 2014.

[3] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice Liveness Detection for Speaker Verification Based on a Tandem Single / Double-Channel Pop Noise Detector," pp. 259–263, 2016.

[4] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *2011 International Conference on Machine Learning and Cybernetics*, 2011, pp. 1708–1713.

[5] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *2011 Carnahan Conference on Security Technology*, 2011, pp. 1–8.

[6] 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.

[7] T. Kinnunen *et al.*, "ASVspoof 2017: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," vol. 0, no. 1, pp. 1–5, 2016.

[8] 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.

[9] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio Replay Attack Detection Using High-Frequency Features," in *Interspeech*, 2017, pp. 27–31.

[10] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Spoof Detection Using Source , Instantaneous Frequency and Cepstral Features," in *Interspeech*, 2017, pp. 22–26.

[11] 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.

[12] 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.

[13] O. Strand and A. Egeberg, "Cepstral mean and variance normalization in the model domain," *COST278 ISCA Tutor. Res. Work. Robustness Issues Conversational Interact.*, pp. 2–5, 2004.

[14] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, pp. 133–147, 1998.

[15] 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.

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

[17] Z. H. Lim, X. Tian, W. Rao, and E. S. Chng, "An investigation of spectral feature partitioning for replay attacks detection," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1570–1573.

[18] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Independent Modelling of High and Low Energy Speech Frames for Spoofing Detection," in *Interspeech*, 2017, pp. 2606–2610.

[19] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7229–7233.

[20] N. Brümmer and E. de Villiers, "The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF," Apr. 2013.