Interspeech 2018
2-6 September 2018, Hyderabad

Detection of Replay-Spoofing Attacks using Frequency Modulation Features

Tharshini Gunendradasan[1], *Buddhi Wickramasinghe*[1,2], *Phu Ngoc Le*[1,2], *Eliathamby Ambikairajah*[1,2], *Julien Epps*[1,2]

[1]School of Electrical Engineering and Telecommunications, UNSW, Australia
[2]ATP Research Laboratory, DATA61, CSIRO, Australia
tharshini.gunendradasan@student.unsw.edu.au,b.wickramasinghe@student.unsw.edu.au
phule@unsw.edu.au, e.ambikairajah@unsw.edu.au, j.epps@unsw.edu.au

Abstract

Prevention of malicious spoofing attacks is currently acknowledged as a priority area of investigation for the deployment of automatic speaker verification systems. Various features of speech signals have been used to fight counterfeit attacks. Among the different spoofing attack variants, replay attacks pose a significant threat as they do not require any expert knowledge and are difficult to detect. This paper proposes the use of a spectral centroid based frequency modulation (FM) features that we term spectral centroid deviation (SCD) for replay attack detection. Spectral centroid frequency (SCF) and spectral centroid magnitude coefficient (SCMC) features extracted from the same front-end as SCD are also investigated as complementary features. Evaluations on the ASVspoof 2017 dataset indicate that the proposed SCD features with a Gaussian Mixture Model (GMM) back-end is highly capable of discriminating genuine from replay spoofed speech, providing an equal error rate improvement greater than 60% relative to the CQCC baseline system from the ASVspoof 2017 challenge. Interestingly, experiments also reveal that the proposed SCD features exhibit an increased variance for replay spoofed speech relative to genuine speech, particularly for the lowest and highest frequency subbands.

Index Terms: Frequency modulation, speaker verification, spoofing attack, spectral centroid magnitude, spectral centroid frequency, ASVspoof 2017

1. Introduction

Voice-based authentication, known as Automatic Speaker Verification (ASV), has received significant attention from researchers in the last two decades due to its convenience, low cost, and remote operability with simple devices like mobile phones. Hence, ASV is widely used as a security measure in telephone banking, building access systems, call centers, and many other areas. In some applications, such as restricting access to confidential information and financial transactions, ensuring that ASV is highly reliable against spoofing attacks is compelling, thus increasing the need for research in this area.

Spoofing attacks are mainly categorized into four different types, speech synthesis (SS), voice conversion (VC), replay, and impersonation. SS and VC are used to generate artificial speech to fool an ASV system. Replay attacks are the easiest form of attacks to create, whereby the pre-recorded speech of the target speaker is played back to ASV. This does not require any knowledge of speech processing, and the abundance of high-quality recording devices such as smartphones increases the chances of this attack. The ASVspoof 2017 Challenge [1] was organized with a focus on the limitations of existing preventive measures against replay attacks. The moderate results reported in this challenge indicate the restrictions of current techniques to detect unknown and diverse replay conditions.

Systems submitted to the ASVspoof 2017 challenge investigated different front-end features and the usage of different classifiers to detect replay attack under diverse conditions. Among them, the best-performing system [2], with an Equal Error Rate (EER) of 6.73%, used a light convolutional neural network (LCNN) to extract high-level features from the log power spectrum, together with a Gaussian Mixture Model (GMM) for a classifier. Variable length Teager energy operator-energy separation algorithm-instantaneous frequency cosine coefficients (VESA-IFCCs) [3] were proposed as a single system with the motivation of capturing the spectral changes due to the transmission and channel characteristic of replay devices. To capture the channel information embedded in the low signal to noise ratio region, a single frequency filtering feature with high spectro-temporal resolution was proposed [4].

Most of the other ASVspoof 2017 challenge systems used a combination of different features and classifiers to improve the performance of replay detection, such as Constant-Q Cepstral Coefficients (CQCCs), Mel-Frequency Cepstral Coefficients (MFCCs), Linear Frequency Cepstral Coefficients (LFCCs), Rectangular Filter Cepstral Coefficients (RFCCs), Perceptual Linear Predictive (PLP) and deep features as front-ends [6, 7]. GMM, support vector machine (SVM) and i-vector Gaussian Probabilistic Linear Discriminant Analysis were employed as back-end classifiers [5].

Magnitude-based features are widely used in replay attack detection, discarding phase-related information in the speech signal. However, phase-based features are effective in related applications: group delay features have been utilized in emotion recognition, language recognition and speaker recognition [6, 7], and frequency modulation (FM) features have been used in speech recognition and speaker recognition [8, 9].

Phase-based features have been successfully used to detect VC and SS spoofing attacks: they are effective in detecting anomalies and may be complementary to magnitude-based subsystems [10, 11]. However, little work has been done on replay attack detection using phase-based features or other features motivated by phase and frequency information.

In this paper, we explore the discriminating ability of FM features, extracted using two quite different methods that have

10.21437/Interspeech.2018-1473

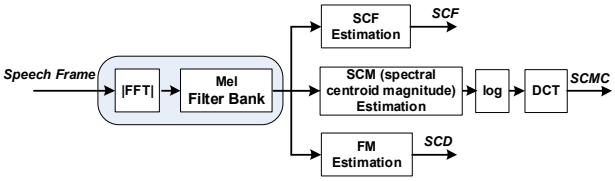


Figure 1: *Block diagram of SCF, SCMC and SCD feature extraction using spectral centroid method.*

both shown promise in the speaker recognition context [12], in detecting replay attacks. This work also explores the use of detailed spectral features, namely spectral centroid frequency (SCF [13]), spectral centroid magnitude coefficient (SCMC [13]) and spectral centroid deviation (SCD), which can all be extracted from a single front-end, to improve the performance of replay detection systems.

2. FM Feature Extraction

In this section, we introduce a novel method to extract FM features as an alternative to Linear Predictive Coefficients (LPC) based FM extraction [14]. Further, SCF and SCMC features, together with the proposed FM based features SCD all shown in Figure 1, are fused to provide a detailed characterization of the speech spectrum to detect replay spoofing attacks.

2.1. FM feature extraction using LPC model

An AM-FM model for the speech signal was proposed in [12] incorporating frequency modulation components, where the time series of the speech signal $s[n]$ is modeled as a summation of AM-FM signals notionally corresponding to the vocal tract resonances. The total speech is considered as the sum of all terms of resonant frequencies, given in [12], is

$$s[n] = \sum_{k=1}^{K} a_k[n] \cos\left(\frac{2\pi f_{ck} n}{f_s} + \frac{2\pi}{f_s}\sum_{j=1}^{n} q_k(j)\right) \quad (1)$$

where k is the resonance index, K is the total number of resonances, $a_k[n]$ is the time varying AM component, f_{ck} is the resonant frequency, f_s is the sampling frequency, $q_k(j)$ is the time-varying frequency modulation component and n is the speech sample index.

The resonant frequency of each bandpass-filtered signal was extracted as described in [15]. The output of each bandpass filter can be modelled using LPC analysis as a second-order all-pole resonator. The pole frequency of the resonator was then calculated from the linear prediction coefficients. The FM component at the output of the bandpass filter was then taken as the deviation of the pole frequency f_p from the resonant frequency of the bandpass filter. If the speech signal is filtered using P bandpass filters, it will result in a P-dimensional FM feature.

2.2. FM feature extraction using the proposed spectral centroid method

For each speech frame, the process for the proposed FM feature extraction is shown in Figure 1. The FM component in the k^{th} frequency band is obtained as the deviation between the spectral centroid frequency (SCF_k) and the center frequency f_{ck} of that band, as illustrated in Figure 2. The SCF is an estimate of the 'center of gravity' of the spectrum providing the

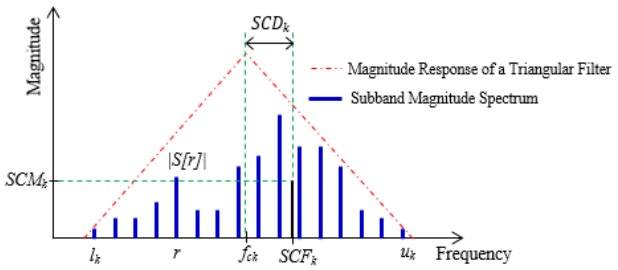


Figure 2: *Illustration of SCF, SCM and SCD features in the k^{th} sub-band.*

formant frequency in each frequency band [16]. For the k^{th} subband, the equation to compute SCF_k is given by

$$SCF_k = \frac{\sum_{r=l_k}^{u_k} r|S[r]|}{\sum_{r=l_k}^{u_k} |S[r]|} \quad (2)$$

where r is a frequency, l_k and u_k are the lowest and highest frequencies in the band, and $S[r]$ is the spectral magnitude at the frequency r. The FM component of the k^{th} frequency band leads to the spectral centroid deviation, SCD_k, by taking its magnitude and is given as

$$SCD_k = |SCF_k - f_{ck}| \quad (3)$$

where f_{ck} is the centre frequency of the k^{th} subbband. The SCD feature vector for a frame is then formulated by concatenating all individual subband SCD components into a single vector. It should be noted that a similar vector can also be extracted using the LPC based FM outlined in Section 2.1 and is referred to as Frequency Modulation Deviation (FMD).

2.3. SCD feature characteristics for genuine and spoofed speech

The FM based features (SCD and FMD) extracted from a speech frame using fifty Mel scale subband filters are plotted in Figure 3, showing that while both feature extraction methods result in relatively similar values, the different approaches provide different estimates in the lowest and highest frequency bands.

The statistics of SCD features, extracted using fifty Mel scale subband filters, for all genuine and replayed speech utterances from the ASVspoof 2017 version 1 training dataset are illustrated using boxplots in Figure 4. This suggests the discriminative ability of SCD features to separate genuine speech from replayed speech, especially in the high and low

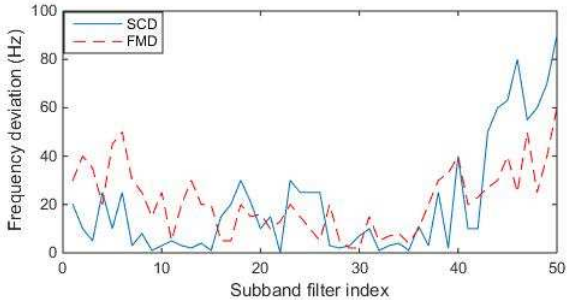
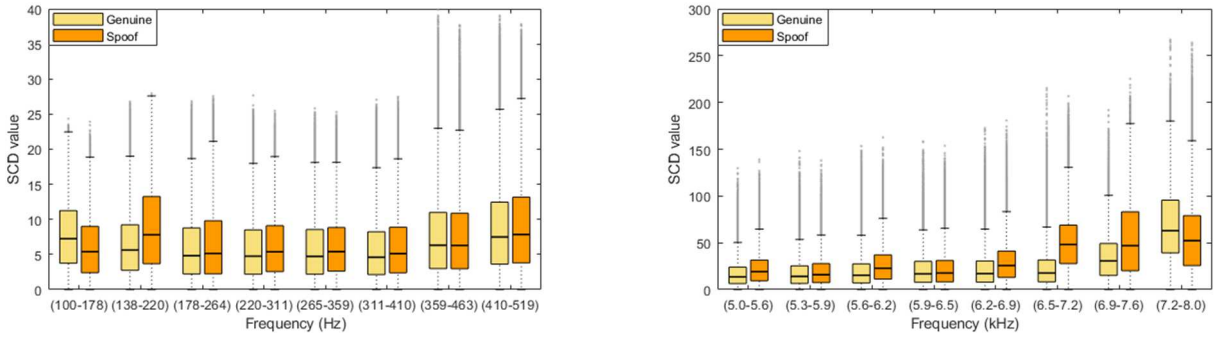


Figure 3: *FMD (dashed red) and SCD (blue) features extracted using LPC and Spectral Centroid methods respectively.*



(a) (b)

Figure 4: *Boxplot of SCD features of all the genuine and replayed utterances of ASVspoof 2017 version 1 training data, extracted from the (a) first eight subband filters, and (b) last eight subband filters. The middle strip of each box is the median, the edges of the box represents the 25th and 75th percentiles, the whiskers extend to the most extreme data points that are not considered to be outliers, and the outliers are plotted individually.*

frequency subbands.

The increased discriminating capacity of the high-frequency bands can be explained by some of the following properties of replayed speech. As the replay attack involves multiple recordings of the speech, the effect of the anti-aliasing filters of the recording devices [17] is to attenuate the high-frequency components, which tends to push the centroid frequencies lower. Hence, the SCD features will be larger, which can be seen particularly in the 2nd and 3rd highest filter boxplots for replay attack speech relative to genuine speech in Figure 4(b).

Figure 5 shows an example of frequency variation in conjunction with SCD features for genuine and replayed speech. It illustrates the highest four frequency bands out of the fifty Mel scale subbands, where SCD components have been offset by their corresponding center frequency in each subband. A similar pattern to that discussed in the previous paragraph, of the SCD values for genuine and replayed speech, can be observed here. In general, SCD values from replayed speech are higher than the corresponding values estimated from genuine speech.

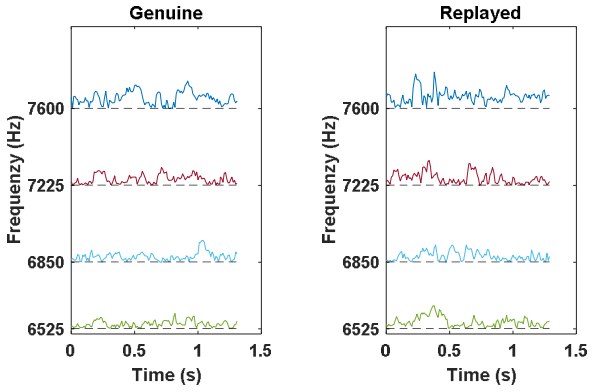


Figure 5: *An example of feature values of genuine and replayed speech, extracted using SCD for the high frequency subbands, of the same person speaking the same sentence. Each subband has been offset by its bandpass filter's center frequency, shown with a dashed line. In general, the SCD values of spoofed speech are higher than genuine speech.*

3. Experimental Results

3.1. Database

All experiments in this paper were conducted on the ASVspoof 2017 version 1 database, which was released in 2017 as part of a Challenge [18] to detect replay spoofing attacks in speaker verification systems. All speech utterances in this database, sampled at 16kHz, were derived from the Red Dot Corpus [19]. Genuine speech was obtained directly from the original corpus and replay attack speech utterances are replayed versions of the Red Dot speech in different recording environments, using different recording and replay devices. Table 1 presents the details of the training, development and evaluation subsets. Replayed utterances in the development and evaluation data have large differences from the training data to emphasize the importance of implementing generalized spoofing countermeasures.

According to the ASVspoof 2017 Challenge protocol, the Equal Error Rate (EER) was used as the primary metric to evaluate the system. The Challenge also provided a baseline system that used CQCCs as the front-end features and GMM as a classifier [18].

3.2. Experimental Setup

A stand-alone spoofing detection system was implemented using the SCD features and GMM as a classifier. Two Gaussian models were created and used as statistical models for genuine and spoofed speech. For the test utterances, scores were calculated using log-likelihood ratios. In our experiment, the Vlfeat toolkit [20] was used to model a GMM with 512-mixture components.

During the experiments, the training data was used to build statistical models of genuine and spoofed speech and the development data was used to tune their parameters. To incorporate diverse spoofing attacks into the system modeling, models are reconstructed using both training and development data with the same tuned parameters. All the results in this section are reported on evaluation data.

Initially the front-end feature extraction parameters were tuned for the proposed SCD features and then the same parameters were used to extract the complementary SCMC and SCF features. These features were extracted from the

Table 1: *ASVspoof 2017 version 1 database details.*

Subset	# speakers	# utterances		
		Genuine	Spoofed	Total
Training	10	1508	1508	3016
Development	8	760	950	1710
Evaluation	24	1298	12922	14220

short-term analysis of speech signals using 512-point fast Fourier transform with 40 ms Hamming window and 10 ms shift. Fifty sub-band filters were used to extract the features from each frame. As a preprocessing step, zero mean and unit variance normalization was performed for the features extracted from every utterance. We used linear score level fusion to fuse all three sub-systems.

3.3. Results and Discussion

This section compares our proposed method with the recent replay attack countermeasures that use single and fused systems with GMMs as back-end classifiers. We choose systems that use GMM backend in order to have a fair comparison of the discriminating ability of the front-end feature itself. The front-end features of single systems considered were CQCCs [18] as the baseline system, MFCCs [21], VESA-IFCCs) [3], and a single frequency filtering (SFF) based feature [4]. The fused system in [22] used voice source, instantaneous frequency and cepstral features, and [21] uses LFCCs and RFCCs.

Experimental results from the evaluation data are shown in Table 2. Compared with other single systems from the ASVspoof 2017 Challenge considered in this paper, the SCD front end exhibits significant improvement, suggesting the potential for FM based features to discriminate between genuine and replayed speech. This is likely due to the detailed spectral information contained in the SCD features, which carry information about the distribution of spectral energy within bands.

Table 2: *Results for the evaluation data of the ASVspoof 2017 version 1 database*

		Systems	Evaluation EER (%)
ASVspoof 2017 challenge	**Single systems**	Baseline (CQCC with GMM) [18]	24.60
		CQCC (6-8 kHz) with GMM [17]	17.31
		MFCC with GMM [21]	27.12
		VESA-IFCC with GMM [3]	15.50
		SFFCC-D with GMM [4]	20.2
	Fused systems	Voice source + instantaneous frequency + cepstral features with GMM [22]	13.95
		RFCC + LFCC with GMM [21]	10.52
Single systems		FMD with GMM	13.30
		A: SCMC with GMM	15.68
		B: SCF with GMM	12.34
		C: SCD with GMM	**11.45**
Fused system		**A + B + C (score fusion)**	**9.20**

As can be noted from Table 2, fusing SCD with SCMC and SCF further improves the performance of replay attack detection. This suggests that SCD, which mainly carries frequency modulation-based information, and SCF, which gives information on dominant frequency in each subband and the magnitude information SCMC, together are good complementary features for replay detection.

The EER of the best-performing system from the ASVspoof 2017 Challenge was 6.73% [2], based on a convolutional neural network to extract deep features from the spectrogram. In this paper, we did not consider any deep neural network architectures for bottleneck feature extraction or as a classifier. Instead, the proposed features were used directly with a simple GMM back-end, in order to gain an understanding of which speech features are mainly affected by replay attacks. This type of analysis contributes to further investigation of features to have clear discrimination between genuine and replayed speech and to reduce the EER.

4. Conclusions

This paper presents an effective set of spectral centroid based methods for extracting FM features from speech for blocking counterfeit access attempts to ASV systems, with a focus on replay attacks. These features are used to discriminate replay spoofing attacks from the genuine speech for ASV. Compared with other single-model and fused selected spoofing detection methods from the ASVspoof 2017 Challenge, which use voice source, magnitude and phase-based features, the proposed SCD exhibits better performance. Further, when it is fused with complementary SCMC and SCF features, the performance of the system improved again.

Our focus has been on model-based detection of replay-spoofing using conventional methods. Our future work will also focus on employing deep neural network architectures to explore the discriminating capacity of the proposed front-end features to further improve the performance of replay detection systems.

5. References

[1] T. Kinnunen *et al.*, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," *Proc. Interspeech,* pp.2-6, 2017.

[2] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," *Proc. Interspeech,* 2017.

[3] H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, "Novel Variable Length Teager Energy Separation Based Instantaneous Frequency Features for Replay Detection," *Proc. Interspeech,* pp. 12-16, 2017.

[4] K. Raju Alluri and A. K. V. Gangashetty, "SFF Anti-Spoofer: IIIT-H Submission for Automatic Speaker Verification Spoofing and Countermeasures Challenge 2017," *Proc. Interspeech,* pp. 107-111, 2017.

[5] Z. Ji *et al.*, "Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017," *Proc. Interspeech,* pp. 87-91, 2017.

[6] V. Sethu, E. Ambikairajah, and J. Epps, "Group delay features for emotion detection," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[7] F. Allen, E. Ambikairajah, and J. Epps, "Warped magnitude and phase-based features for language identification," *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 201-204, 2006.

[8] T. Thiruvaran, "Automatic speaker recognition using phase based features," Doctor of Philosophy, School of Electrical

Engineering and Telecommunications, The University of New South Wales, 2009.
[9] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letters,* vol. 12, no. 9, pp. 621-624, 2005.
[10] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, "Spoofing Speech Detection Using Modified Relative Phase Information," *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, no. 4, pp. 660-670, 2017.
[11] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
[12] T. Thiruvaran, E. Ambikairajah, and J. Epps, "FM features for automatic forensic speaker recognition." *Proc. Interspeech*, *1497-1500,* 2008.
[13] P. N. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. H. Choi, "Investigation of spectral centroid features for cognitive load classification," *Speech Communication,* vol. 53, no. 4, pp. 540-551, 2011.
[14] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition," in *Odyssey*, p. 7, 2010.
[15] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE transactions on signal processing,* vol. 41, no. 10, pp. 3024-3051, 1993.
[16] K. K. Paliwal, "Spectral subband centroid features for speech recognition," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* , vol. 2, pp. 617-620, 1998.
[17] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio Replay Attack Detection Using High-Frequency Features," *Proc. Interspeech,* pp. 27-31, 2017.
[18] T. Kinnunen *et al.*, "ASVspoof 2017: automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training,* vol. 10, no. 1508, p. 1508, 2017.
[19] K. A. Lee *et al.*, "The RedDots data collection for speaker recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
[20] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1469-1472: ACM, 2010.
[21] R. Font, J. M. Espın, and M. J. Cano, "Experimental analysis of features for replay attack detection–Results on the ASVspoof 2017 Challenge," *Proc. Interspeech,* pp. 7-11, 2017.
[22] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, "Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features," *Proc. Interspeech,* pp. 22-26, 2017.