Investigating the use of scattering coefficients for replay attack detection

Conference Paper · December 2017

DOI: 10.1109/APSIPA.2017.8282211

4 authors, including:

Kaavya Sriskandaraja
UNSW Sydney
12 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Gajan Suthokumar
UNSW Sydney
12 PUBLICATIONS **49** CITATIONS
SEE PROFILE

Vidhyasaharan Sethu
UNSW Sydney
79 PUBLICATIONS **570** CITATIONS
SEE PROFILE

Investigating the use of Scattering Coefficients for Replay Attack Detection

Kaavya Sriskandaraja*†, Gajan Suthokumar*†, Vidhyasaharan Sethu*, Eliathamby Ambikairajah*†
*School of Electrical Engineering and Telecommunicaitons, UNSW, Australia
†DATA61, CSIRO, Sydney, Australia
E-mail: k.sriskandaraja@unsw.edu.au, g.suthokumar@unsw.edu.au, v.sethu@unsw.edu.au, e.ambikairajah@unsw.edu.au

***Abstract*— Widespread adoption of speaker verification for security relies on the existence of effective anti-spoofing countermeasures. This paper presents a countermeasure based on spectral features to detect replay spoofing attacks on automatic speaker verification systems. In particular, the use of hierarchical scattering decomposition coefficients and inverse-mel frequency cepstral coefficients are explored. Our best system achieved a relative improvement of around 70% in terms of equal error rate on the development set and 20% on the evaluation set, when compared to the baseline on the ASVspoof 2017 database. In addition, we show that features with a shorter window can be beneficial to detecting replayed speech, in contrast to speech synthesis and voice conversion attack.**

I. INTRODUCTION

Biometric authentication systems, such as speaker verification systems, are expected to have countermeasures against spoofing attacks. However, while speaker verification is attractive, it has been shown to be vulnerable to spoofing attacks. Spoofing attacks can be broadly categorized as one of four types [1]; impersonation [2], replay [3], voice conversion [4] and speech synthesis [5]. Among these, replay attacks are the most accessible and can be highly effective. To address this threat, the ASVspoof 2017 challenge [6] was organised as a benchmark to quantify progress in the field of spoofing detection involving replayed speech. This paper describes the UNSW team's submission to this challenge.

There have been a few studies that have assessed the vulnerability of speaker verification systems to replay attacks [3, 8]. All of these have shown that replay attacks are highly effective, by observing significant increases in both equal Error Rate (EER) and False Acceptance Rates (FAR). For example, authors investigated the replay attack method in [8] with the considerably large dataset, RSR2015 [9], and showed that EER and FAR increased from both 2.92% to 25.56% and 78.36% respectively as a result of replay attack. Recent work compared the effect of replay attack on six different Automatic Speaker Verification (ASV) systems in three different recording and playback environments [10]. The ASV included GMM supervector, linear kernel with factor analysis, and state-of-the-art i-vector probabilistic linear discriminative analysis and the authors concluded that the EER of even the most resistant systems increased significantly. Hence there is an urgent need for effective countermeasures against replay attacks to deploy speaker verification in commercial applications.

Countermeasures against replay attacks generally aim to exploit the exact reproduction due to replay, differences in the speech transmission channel, or differences in spectral properties of replayed speech. The first approach involves the storing of previous access attempts and their comparison to new attempts [11]. Every new access attempt is verified by checking for similarities to previous attempts; too great a similarity means the access attempt is rejected [8, 11]. The second approach involves distinguishing between the transmission channels of genuine and replayed speech, such as the detection of pop-noise [12], of far-field recordings based on reverberation levels [13], channel differences [14] or local binary patterns [10]. Most other techniques developed for the detection of replayed speech take the final approach of using features indicative of spectral cues, which is the approach followed in this paper. Specifically, this paper presents a spoofing detection system for the ASVspoof 2017 challenge [6], which makes use of scattering spectral coefficients and Inverse Mel Filter Cepstral Coefficients (IMFCC) in the front-end.

II. FEATURES

Fig 1 compares the spectrogram of a genuine speech utterance to a replayed speech utterance from the ASVspoof 2017 challenge corpus [6]. The comparison suggests there may be some differences at the low and high ends of the frequency spectrum. In order to investigate this, we develop systems that make use of one of two front-ends. The first makes use of scattering spectral decomposition features, which provide a high resolution description of the low frequency end and have previously been shown to be effective in detecting synthetic speech attacks [19]. The second utilises inverse mel cepstral coefficients, which provide a higher resolution representation at the high frequency end of the spectrum and have been used in synthetic speech detection [25].

A. Scattering Spectral Decomposition

The recently proposed scattering spectrum [15], can be viewed as a unified framework for constant-Q cepstral coefficients [16] and cochlear filters [17]. The scattering transform [15] is a hierarchical spectral decomposition of a

signal based on wavelet filter-banks (constant-Q filter-banks), $\{\psi_{k_j}[n];\ j = 1, 2, \ldots, N_k\}$, where N_k is the number of filters at each level, k, of decomposition, followed by the modulus operator (absolute value), $|\cdot|$, and finally averaging with a low pass filter, $\phi[n]$ as shown in Fig 2.

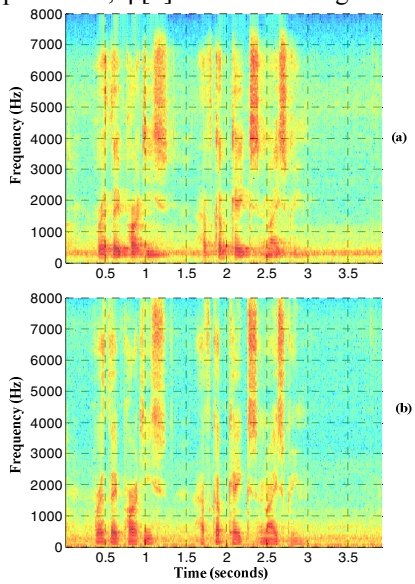

Fig 1: Spectrogram of (a) genuine speech (b) replayed speech.

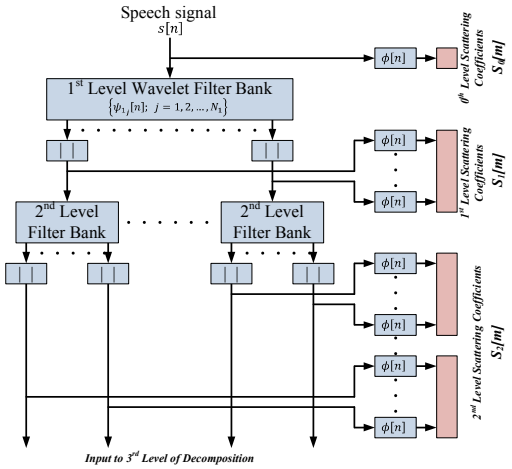

Fig 2: Two-level scattering decomposition

A simplified view of scattering decomposition is described in Fig 2, where the k^{th} level of decomposition runs the input signals (outputs from the previous levels) through a wavelet filter-bank and takes the absolute value of the filter outputs leading to a scalogram. The scattering coefficients, $S_k[\text{m}]$, at that level are estimated by windowing the scalogram signals and computing the average value within these windows. Finally, the scattering coefficients, $\boldsymbol{V}$, are computed by concatenating the logarithms of the scattering coefficients from all levels as given in (1). More details related to scattering decomposition can be found in [18, 19].

Fig 3 shows the log scattering coefficients obtained from a speech signal, s[n], at the first level of decomposition, along with the corresponding log-scalograms, $\{\log|\psi_{1_j} * s|[n]; \forall j\}$. It should be noted that in Fig 3 (c) and 3(f), the larger the filter number, the lower the center frequency and smaller the bandwidth. A window size of 2ms was found to be suitable for the detection of replayed speech. The log-scalograms and scattering coefficients shown in Fig 3, suggest that scattering decomposition features may be able to differentiate between the spectral characteristics of the genuine and replayed speech.

$$\boldsymbol{V} = \{[\boldsymbol{S_0}[\boldsymbol{m}], \boldsymbol{S_1}[\boldsymbol{m}], \ldots \boldsymbol{S_k}[\boldsymbol{m}], \ldots., \boldsymbol{S_{\breve{K}}}[\boldsymbol{m}]]\} \quad (1)$$

where $\breve{K}$ is the desired level of decomposition.

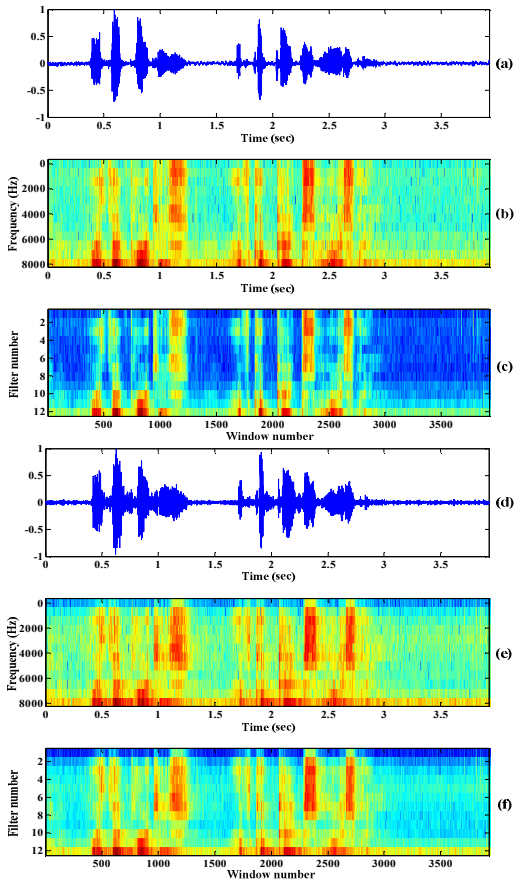

Fig 3: (a) Genuine speech waveform, (b) log-scalogram of genuine speech, (c) first level log scattering coefficients of the genuine speech signal (2ms window), (d) replayed speech waveform, (e) log-scalogram of replayed speech, (f) first level log scattering coefficients of the replayed speech signal (2ms window)

B. Inverse-Mel Cepstral Coefficient (IMFCC)

As illustrated in Fig 1, discriminative information between genuine and replayed speech may also be present in the high frequency region. Consequently, we investigate IMFCCs in addition to the scattering decomposition features. These IMFCCs are similar to standard MFCCs, but utilise an inverted filter bank comprised of a large number of narrow band filters at the high frequency regions and a small number of wide band filters at the low frequency regions [20].

III. BACK-END

Finally, in order to compare the system directly to a baseline, a 2-class GMM back-end was used to obtain the log-likelihood ratio between genuine and replayed speech. The GMM back-end was implemented as maximum likelihood estimates using 512 mixture components for the scattering features and 256 mixture components for IMFCCs. GMMs are trained using the HTK [21] and VLfeat toolkits [22].

IV. SYSTEM DESCRIPTION AND CONFIGURATION

A. Spoofing detection with scattering coefficients

The Morlet wavelet was used to obtain the scattering decomposition features using 8 filters per octave at first level and 1 filter per octave at the second level. This configuration was chosen to match the frequency resolution of mel filters at the first level. These scattering coefficients are computed using 50% overlapping windows using a publicly available toolbox [23]. Preliminary experiments were carried out on the development set of the ASVspoof 2017 challenge corpus to choose a window size for the scattering coefficients. Based on this, a window size of 2ms was chosen for all subsequent experiments (Table 1). Finally, a 75-dimensional scattering decomposition feature vector was computed by merging the scattering coefficients with their deltas and delta-deltas. It should be noted that deltas and delta-deltas are calculated over a 42ms window, i.e. 21 consecutive frames, 10 either side of the current window.

TABLE 1
SPOOFING DETECTION PERFORMANCE OF SCATTERING COEFFICIENTS USING WINDOW SIZE EVALUATED ON THE DEVELOPMENT SET OF THE ASVSPOOF 2017 CORPUS

Window Size	%EER
2ms (2^5 samples)	3.16
8ms (2^7 samples)	8.81
64ms (2^{10} samples)	16.19
256ms (2^{12} samples)	25.98

B. Spoofing detection using IMFCC

IMFCC features were extracted using a bank of 20 Gaussian filters in the 4-8kHz region in order to emphasise information from the high frequency end of the spectrum [7]. The IMFCC front-end employed 20ms windows with 50% overlap. The first 10 DCT coefficients and deltas and delta-deltas were taken to form the 30-dimensional feature vector. The delta and delta-delta coefficients were computed over 11 consecutive frames.

V. EXPERIMENTAL RESULTS

A. ASVspoof 2017 Challenge

The ASVspoof 2017 corpus [6] makes use of the RedDots corpus [24], as well as a replayed version of the same data [6]. This data is partitioned into training, development and evaluation sets. The replayed speech in these partitions was created using different playback and recording devices in various environments. The primary metric of the challenge was defined as Equal Error Rate (EER). Finally, a baseline system was provided by the organizers and made use of a Constant-Q Cepstral Coefficients (CQCC) front-end with a GMM back-end. More details on challenge baseline can be found in [6].

B. Results and Discussion

Table 2 shows the %EER obtained on the development set. These systems were trained using only the challenge training set.

TABLE 2
COMPARISON OF PROPOSED SPOOFING DETECTION SYSTEMS ON THE 'DEVELOPMENT' SET OF THE ASVSPOOF 2017 CORPUS

Systems	%EER
Baseline	10.8
Scattering Coefficients	3.16
IMFCC	2.48
Fusion (Scattering Coefficients + IMFCC)	2.11

The results on the evaluation set were obtained using systems identical to those reported in Table 2, but trained on both the Training and Development sets of the challenge database. These results are tabulated in Table 3 and show that both the scattering coefficients and IMFCC features outperformed the challenge baseline. The significant differences in performance between the development (Table 2) and evaluation (Table 3) set results suggest that all the systems investigated in this work appear to suffer from the generalization of the previously unseen recording and playback conditions in the evaluation set.

TABLE 3
COMPARISON OF PROPOSED SPOOFING DETECTION SYSTEM ON THE 'EVALUATION' SET OF THE ASVSPOOF 2017 CORPUS

Systems	%EER
Baseline	24.65
Scattering Coefficients	19.79
IMFCC	20.05
Fusion (Scattering Coefficients + IMFCC)	17.88

TABLE 4
COMPARISON OF THE PERFORMANCE OF DIFFERENT LEVELS OF SCATTERING COEFFICIENTS ON THE ASVSPOOF 2017 CORPUS EVALUATION SET

Levels	%EER
First level Scattering Coefficients	27.3
First and Second level Scattering Coefficients	19.79

In order to quantify the benefit of incorporating information contained in the second level of the scattering decomposition, the system based on scattering coefficients obtained up to second level scattering decomposition was compared to a system based only on the first level scattering decomposition on the ASVspoof 2017 challenge evaluation set (Table 4). The scattering coefficients based only on the first level decomposition are obtained by concatenating only the zeroth and first level scattering coefficients, their deltas and delta-deltas. It can be seen that the addition of information from second level scattering coefficients substantially improves performance, which demonstrates the importance of scattering coefficients from subsequent levels.

Finally, it was observed that the optimal window size for the scattering decomposition features was around 2ms, which is in sharp contrast to the 256ms optimal window size

previously used for the detection of voice conversion and speech synthesis based spoofing attacks in [19]. This is perhaps because a shorter window emphasises channel cues more than speaker cues. Unlike voice conversion and speaker synthesis based spoofing attacks, replay attacks reproduce speaker cues optimal window size, so detection is based predominantly on channel cues.

VI. CONCLUSIONS

In this work we have compared two front-ends for replay spoofing detection, one that focuses on the low-frequency region of the spectrum, scattering coefficients, and one that focuses on the high-frequency region, inverse-mel frequency cepstral coefficients. The experimental results suggest that both front-ends can capture information relevant for replay detection and outperform the challenge baseline. In addition, we showed that features with a shorter window can be beneficial to detecting replayed speech, in contrast to speech synthesis and voice conversion attack, which can be lead to the investigation of the front-ends with shorter windows.

REFERENCES

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication,* vol. 66, pp. 130-153, 2015.

[2] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*, 2004, pp. 145-148.

[3] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management*, ed: Springer, 2011, pp. 274-285.

[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," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4401-4404.

[5] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "The AHOLAB RPS SSD Spoofing Challenge 2015 Submission," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco*, et al.*, "Asvspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training,* vol. 10, p. 1508, 2016.

[7] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "Investigation of Sub-Band Discriminative Information between Spoofed and Genuine Speech," in *Interspeech*, San Francisco, USA, 2016.

[8] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, 2014, pp. 1-5.

[9] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication,* vol. 60, pp. 56-77, 2014.

[10] A. Janicki, F. Alegre, and N. Evans, "An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks," *Security and Communication Networks,* vol. 9, pp. 3030-3044, 2016.

[11] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 1678-1681.

[12] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in *INTERSPEECH*, 2015, pp. 239-243.

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

[14] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 28, pp. 2037-2041, 2006.

[15] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics,* vol. 65, pp. 1331-1398, 2012.

[16] M. Todisco, H. Delgado, and N. Evans. (2016). *CQCC Features For Spoofed Speech Detection*. Available: http://audio.eurecom.fr/content/software

[17] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. Interspeech*, 2015, pp. 2062-2066.

[18] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing,* vol. 62, pp. 4114-4128, 2014.

[19] K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li, "Front-End for Anti-Spoofing Countermeasures in Speaker Verification: Scattering Spectral Decomposition," *IEEE Journal of Selected Topics in Signal Processing,* 2016.

[20] S. Chakroborty, A. Roy, and G. Saha, "Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks," *International Journal of Signal Processing,* vol. 4, pp. 114-122, 2007.

[21] S. J. Young and S. Young, *The HTK hidden Markov model toolkit: Design and philosophy*: University of Cambridge, Department of Engineering, 1993.

[22] 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*, 2010, pp. 1469-1472.

[23] (2016). *Scattering*. Available: http://www.di.ens.fr/data/scattering/

[24] K.-A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. A. van Leeuwen*, et al.*, "The reddots data collection for speaker recognition," in *Interspeech*, 2015, pp. 2996-3000.

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