Interspeech 2018
2-6 September 2018, Hyderabad

Frequency Domain Linear Prediction Features for Replay Spoofing Attack Detection

Buddhi Wickramasinghe[1,2], *Saad Irtza*[1], *Eliathamby Ambikairajah*[1,2], *Julien Epps*[1,2]

[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2] Data61, CSIRO, Sydney, Australia
b.wickramasinghe@student.unsw.edu.au

Abstract

Automatic speaker verification (ASV) systems are vulnerable to various types of spoofing attacks such as speech synthesis, voice conversion and replay attacks. Recent research has highlighted the need for more effective countermeasures for replay attacks, which can be very challenging to detect, however replayed speech has previously shown frequency band-specific differences when compared with genuine speech. In this paper, we propose the use of long-term temporal envelopes of subband signals using a frequency domain linear prediction (FDLP) framework. This flexible framework makes use of temporal envelope information, which has not previously been investigated for replay spoofing detection. Evaluations of the proposed system and its fusion with other subsystems were carried out on the ASVspoof 2017 database. Interestingly, smoother temporal envelopes, based on very long windows of up to 1 second, seem to be most successful and show good prospects for performance improvements via fusion.

Index Terms: ASVspoof 2017, frequency domain linear prediction, convolutional neural network, replay attack

1. Introduction

Automatic Speaker Verification (ASV) systems have become a prominent form of biometric authentication for many reasons. With the development of ASV technology, however, malicious attacks that try to deceive an ASV system are also getting smarter. These attacks, commonly known as spoofing attacks, are of four main types: speech synthesis, voice conversion, replay and impersonation [1]. The amount of research that has been done on speech synthesis and voice conversion attacks is significantly higher than on replay and impersonation attacks. While impersonation has practical limitations, replay attacks are capable of posing a greater threat than speech synthesis and voice conversion [2, 3]. A replay attack simply means an adversary recording speech from a target speaker and playing it back to deceive an ASV system. Due to the availability of high quality consumer devices, a replay attack can be conducted with relative ease by a person with no technical expertise.

Some earlier studies on replay attack detection countermeasures have involved calculating a similarity score between an incoming utterance and a model based previous input utterances, and rejecting the input if the score exceeds a threshold [4, 5]. Other studies have focused on discriminating replayed speech based on the added channel noise patterns [6] and increased reverberation [7]. Although these studies have shown good results, detection of replay attacks has practical limitations such as dealing with a large number of unknown acoustic conditions [8].

The ASVspoof 2017 Challenge was organized as a means to assess these limitations [8]. The database provided in the challenge consists of replayed utterances with a variety of replay configurations [9]. Even though many successful systems were proposed in the challenge, the overall results from ASVspoof 2017 version 1 show that the generalizability of countermeasures for diverse replay attacks is still an open problem.

Most of the anti-spoofing systems currently proposed for replay attack detection utilize conventional short-term spectral variations of speech signals to extract discriminating features. These include Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients [10] and novel features such as Instantaneous Frequency Cepstral Coefficients [11] and Constant Q Cepstral Coefficients [9]. Although short-term spectral features have proven to be effective, it is also important to explore the effect of long-term temporal domain features, an alternative way of analyzing a speech signal, in the case of replay attack detection.

Evidence of the use of long-term temporal features in speech systems can be found in speech synthesis and voice conversion attack countermeasures. Modulation features extracted from magnitude/phase spectrum of a speech signal can capture long-term temporal information [12]. Tian *et al.* have utilized delta and acceleration coefficients of MFCC features to gather temporal information up to 0.1s [13].

The idea of linear prediction in the frequency domain was first proposed in [14] in the context of audio coding. Kumaresan *et al.* [15, 16] have also explored this concept, treating it as linear prediction in the spectral domain. In their approach, the envelope of the signal was obtained without computing the Hilbert transform of the signal, using linear prediction in the discrete Fourier transform (DFT) domain. Athineos *et al.* [17] have investigated the same problem by considering finite length discrete time signals. One important contribution is the use of the (real-valued) discrete cosine transform (DCT) with a long window in place of the DFT on the signal. The residual component of this process captures the frequency modulation component.

This paper investigates the use of temporal envelopes extracted from long frames. These envelopes can be obtained from frequency subbands of the signal using frequency domain linear prediction (FDLP) [18]. Even though FDLP based features have been investigated in the context of speech synthesis attacks [19], they have not been explored in a replay spoofing attack context.

10.21437/Interspeech.2018-1574

2. FDLP Feature Extraction

2.1. FDLP Temporal Envelope

The duality property of time-domain and frequency-domain properties of signals allows the application of linear prediction concepts in the frequency domain [17]. In the proposed approach, linear prediction is applied to the frequency domain representation of the signal. The resultant signal is an approximation of the temporal envelope of the signal.

The residual signal from time domain linear prediction applied to a speech signal contains the excitation source information of the signal. Thus, the spectral envelope and residual signal are expected to contain complementary information. Hence, it is safe to assume that the residual signal obtained from FDLP could contain complementary characteristics to the temporal envelope obtained using the FDLP process.

Similarly to the process in [17], this paper uses a DCT to transform the time domain signal to the frequency domain. The DCT is applied to each long speech frame, and the DCT coefficients are grouped per subband, following which a linear prediction analysis is conducted in the frequency (DCT coefficient) domain on a per-subband basis. The all-pole magnitude response of the linear predictive filter thus obtained is taken as the temporal envelope $w[n]$. In Figure 1, the FDLP temporal envelopes for the first subband of a genuine utterance, and the same utterance recorded and replayed using two different playback devices are shown. It can be seen that the temporal envelopes are markedly different between all three examples. These are typical, and in preliminary experiments less variability was observed in the replayed speech than the genuine speech.

2.2. FDLP Temporal Envelope Features

There are different ways in which features can be extracted from the FDLP temporal envelope and residual. In this paper, we propose two features that are extracted from the FDLP envelope and residual, and an overview of the feature extraction process is given in Figure 2(a).

Temporal Centroid Amplitude (TC)

This section introduces the weighted time average amplitude of the temporal envelope as a derived feature, named as temporal centroid amplitude (TC), to represent the FDLP envelope. If the temporal envelope is $w[n]$, TC_k for each subband k is given by,

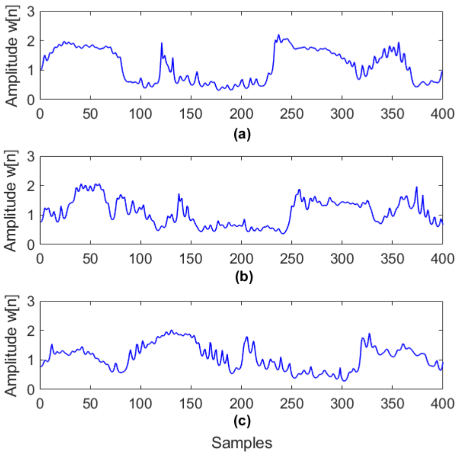

Figure 1*: Example subband FDLP temporal envelopes w[n] of the first subband for (a) genuine speech; (b) and (c) two different examples of replay attack speech from the same speaker.*

$$TC_k = \frac{\sum_{n=n_l}^{n_u} n \cdot w[n]}{\sum_{n=n_l}^{n_u} n}, \tag{1}$$

where n is a (subsampled) time index and n_l and n_u are lower and upper time limits respectively. The positioning of TC depending on the form of the envelope is illustrated in Figure 3. Here, a 400-point temporal envelope is divided into five equal non-overlapping segments and TC is calculated for each segment. It is evident that variations of the temporal envelope can be captured by the TC feature.

Residual Centroid Amplitude (RC)

The variations in the residual are modelled to provide a by-product of FDLP process as shown in Figure 2(a). The FDLP residual $e[k]$ is computed as the difference between subband signal and the envelope estimated from frequency-domain linear prediction in the discrete cosine transform domain. The variations in the residual are modelled by taking an amplitude spectrum (DFT) of $e[k]$ to produce $|E[n]|$, as shown in Figure 4. Finally, the residual centroid amplitude (RC) is computed as

$$RC_k = \frac{\sum_{n=n_l}^{n_u} n.|E[n]|}{\sum_{n=n_l}^{n_u} n}, \tag{2}$$

where k is the subband index, n is a (subsampled) time index

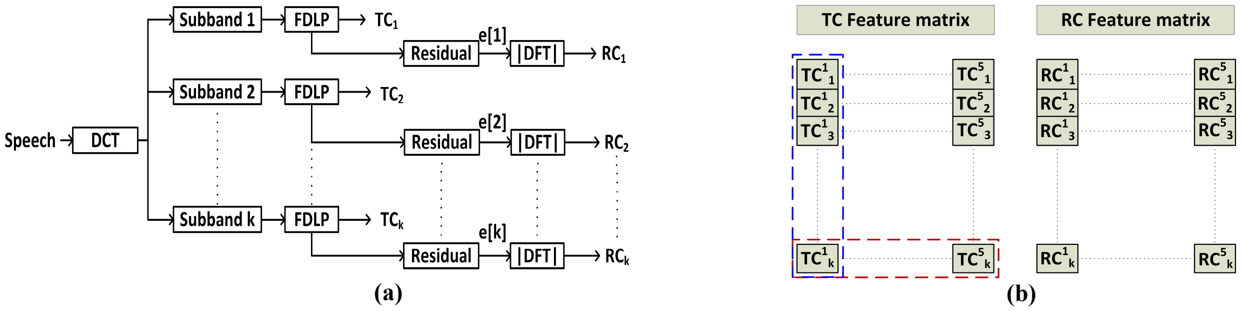

Figure 2: (*a) shows the block diagram of the FDLP temporal envelope and residual feature extraction process and (b) shows the TC and RC feature matrices of one frame*

and n_l and n_u are the lower and upper time limits of the segment.

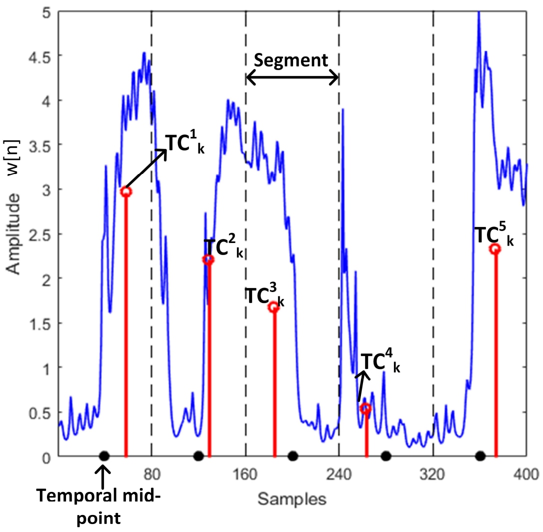

Figure 3: *Illustrative temporal envelope for the k^{th} subband, showing extraction of the TC features on a per-temporal segment basis. TC shifts right or left from the temporal mid-point of each segment of the FDLP envelope depending on whether the envelope is rising or falling respectively.*

3. Experimental Setup

3.1. Database

The experiments reported in this paper were conducted on the ASVSpoof 2017 dataset [9], which consists of training data of 1.09 hours of genuine and 1.03 hours of replayed speech utterances sampled at 16 kHz. The evaluation set comprises speech utterances from several playback and recording devices, which are unseen in the training and development sets.

3.2. Feature extraction

The speech utterances were segmented into long-term speech frames of length 1s, with a frame shift of 250ms. The DCT was performed using 16000 points on each speech frame followed by subband decomposition using 50 equal band triangular-weighted filters. The choices of uniformly spaced filters and the number of filters were determined using the performance on the development set. Frequency domain linear prediction with 160 poles was performed on each subband and the FDLP envelope was determined, then sampled 400 times per frame. The 400-sample FDLP envelope was segmented into five non-overlapping segments (see Figure 3) and the temporal centroid feature was computed for each segment. These five feature values when concatenated represent each 1s frame for a subband as shown in Figure 2(b) (Red-dashed box). Here, TC_k^i represents the temporal centroid amplitude of the ith segment within the k^{th} subband of the selected speech frame. Hence, $[TC_k^1, TC_k^2, \ldots TC_k^5]$ is the temporal centroid amplitude vector of the k^{th} subband of the frame. Following this, the DCT of the logarithm of the temporal centroid amplitudes is computed along the subbands as shown in Figure 2(b) (Blue-dashed box) and these are concatenated across segments to produce the final feature matrix (TC) for that frame. Similarly, a residual centroid amplitude feature matrix (RC) is also extracted for each frame as shown in Figure 2(b).

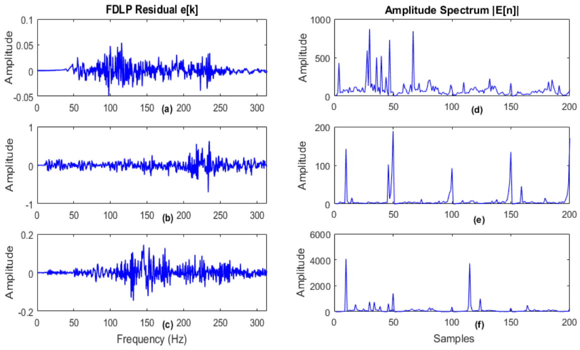

Figure 4: *Example FDLP residuals e[k] for subband 1 of (a) genuine speech; (b) and (c) two different examples of replay attacks from the same speaker; and their amplitude spectra |E[n]| for (d) genuine speech; (e) and (f) for the replayed speech*

3.3. Backend Classifiers

Two modelling approaches were employed, namely Gaussian Mixture Models (GMMs) and Convolutional Neural Networks (CNN), followed by fully connected layers. The GMM modelling approach was used with the two proposed features, TC and RC, whereas the CNN approach was used with the raw FDLP envelope, comprising of 400 points in each subband. The GMMs for genuine and spoofed data were trained with 512 mixture components each. The CNN subsystem was used to investigate the potential of the FDLP envelope to discriminate between genuine and replayed speech.

The proposed CNN architecture consisted of 4 convolutional layers with a filter size of 5x5 and a stride of 2x2, followed by two fully connected layers. Each convolutional layer was followed by a max pooling layer, a batch normalization and a dropout of 0.5. L2 kernel regularization was employed to further reduce the overfitting. Rectified Linear Units (ReLU) were used as activations. A Xavier normal kernel initializer was applied to each convolutional layer. The final convolutional layer outputs two likelihood scores as probabilities for the two classes, taking 256 outputs from the previous fully connected layer as inputs. A SoftMax activation function was used for the calculation of likelihood probabilities.

4. Experimental Results

Preliminary experiments were carried out using the development set of the database (Section 3.1). Equal error rate was used as the evaluation measure for all experiments.

Special attention was given to the importance of the type of filterbank to be used, since this provides an insight into the significance of each frequency band in the DCT domain for FDLP based replay attack detection. It was found that the type of filterbank has a major effect on the performance of the system. Experiments on TC features, using mel-scale, inverse-mel scale and uniform band scale on the development set resulted in the EER (%) values given in Table 1. From Table 1, it is evident that the equal band frequency scale is the most suitable for FDLP based features. This may be due to the importance of the high frequency spectral roll-off of replay devices and possibly the recording environment.

Table 1*: Effect of the frequency scale on EER (%) for the development set of the database*

Feature	Mel-scale	Inverse-Mel	Uniform
TC	22.68	9.21	8.03

Having chosen the system parameters, experiments were conducted on the evaluation set of the database. Both training and development datasets were pooled together for model training. EER values obtained for the two systems using TC and RC and the FDLP envelope are given in Table 2 below. A spectral centroid magnitude (SCM) [20] based system [10] was used as the baseline for comparison. This system extracts the spectral centroid magnitudes of the signal using short-term frames and classifies the speech files using two 512 component GMMs. It employs 40 SCM values with their delta and acceleration coefficients as features. The delta and acceleration coefficients represent some of the temporal structure of the signal. The proposed FDLP envelope features (with CNN) show slightly higher performance compared to the baseline. Although TC and RC based systems have performed somewhat more poorly than the baseline system, the dimensionality of the proposed features are also much lower than that of the baseline SCM features.

It is also interesting to note that the CNN based system performed better than the GMM based systems employing the proposed TC and RC features, as well as the baseline. This suggests that the temporal envelopes carry further information that can be used for replay attack detection.

Having obtained promising results with individual systems, score-level fusion of the three systems was investigated, to assess the complementary nature of each of them. The "FoCal Toolkit" [21] was used to carry out a score level fusion and the results of the fused system are also given in Table 2.

Table 2: *Results on the evaluation set*

	System	EER (%)
	Baseline – SCM (single system) [10]	11.49
	Baseline – Light CNN (fused system) [22]	6.73
S1	TC with GMM (single system)	14.89
S2	RC with GMM (single system)	15.90
S3	FDLP envelope with CNN (single system)	11.13
	Fusion: S1 + S2 + S3 (fused system)	9.70

The error rate of the fused system was less than that of the three individual systems. This suggests that the information provided by each of the three systems may be complementary. When these results are compared with the SCM based baseline system, there is a relative improvement of 15.51%. A second baseline [22], which is a score level fusion of three other systems, is also provided for comparison. However, it should be noted that this system is a very complex fused system that comprises of an i-vector front-end based support vector machine implementation, a light CNN [23] system with log-magnitude spectrum as the front-end and a stacked recurrent neural network (RNN) and CNN system.

5. Conclusion

This paper has presented an investigation of temporal envelope features, extracted using the frequency domain linear prediction framework, for the detection of replay spoofing attacks. Two components of interest in this framework are the temporal envelope, which represents the amplitude modulation component, and the residual component, which represents the frequency modulation component. Results presented here show that both components are sensitive to replay attacks, providing reasonable detection as individual low-feature-dimension systems. When applying the high-dimensional full temporal envelope representation to a convolutional neural network, the system performance outperformed a recent baseline. Finally, combining all three systems using score-level fusion brought significant reductions in error rate relative to any of the individual systems. This work demonstrates the potential of this framework, which can be investigated further to exploit the rich time-frequency information, in particular smooth temporal envelopes representing modulation components.

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 Communication,* vol. 66, pp. 130-153, 2015.

[2] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *IEEE Int. Conf. on Biometrics Special Interest Group (BIOSIG)*, 2014, pp. 1-6.

[3] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications,* vol. 75, no. 9, pp. 5311-5327, 2016.

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

[5] 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.

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

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

[8] T. Kinnunen *et al.*, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Annual Conference of the International Speech Communication Association (Interspeech) 2017,* 2017, pp 2-6.

[9] T. Kinnunen *et al.*, "ASVspoof 2017: automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training,* vol. 10, no. 1508, p.1508, 2017.

[10] R. Font, J. M. Espın, and M. J. Cano, "Experimental analysis of features for replay attack detection–Results on

the ASVspoof 2017 Challenge," in *Annual Conference of the International Speech Communication Association (Interspeech) 2017,* pp. 7-11, 2017.

[11] 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," in *Annual Conference of the International Speech Communication Association (Interspeech) 2017,* pp. 12-16, 2017.

[12] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2013, pp. 7234-7238.

[13] X. Tian, S. Du, X. Xiao, H. Xu, E. S. Chng, and H. Li, "Detecting synthetic speech using long term magnitude and phase information," in *IEEE China Summit and International Conference on*, *Signal and Information Processing (ChinaSIP),* 2015, pp. 611-615.

[14] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *Audio Engineering Society Convention 101*, 1996: Audio Engineering Society.

[15] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *The Journal of the Acoustical Society of America,* vol. 105, no. 3, pp. 1912-1924, 1999.

[16] R. Kumaresan, "An inverse signal approach to computing the envelope of a real valued signal," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999, vol. 3, pp. 1349-1352.

[17] M. Athineos and D. P. Ellis, "Autoregressive modeling of temporal envelopes," in *IEEE Transactions on Signal Processing 2007,* vol. 55, no. 11, pp. 5237-5245, 2007.

[18] M. Athineos and D. P. Ellis, "Frequency-domain linear prediction for temporal features," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2003 (ASRU'03),* 2003, pp. 261-266.

[19] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Sixteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2015, pp. 2087-2091.

[20] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition," in S*peaker and Language Recognition Workshop, IEEE Odyssey 2010,* 2010, pp. 7.

[21] N. Brümmer, "FoCal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition Scores—Tutorial and user manual—," *Software available at http://sites.google.com/site/nikobrummer/focalmulticlass,* 2007.

[22] 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, 2017.

[23] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Journal of Selected Topics in Signal Processing,* vol. 1511.02683, 2015.