International Journal of Speech Technology
https://doi.org/10.1007/s10772-019-09604-x

Combining evidences from Hilbert envelope and residual phase for detecting replay attacks

Madhusudan Singh[1] · Debadatta Pati[1]

Received: 1 October 2018 / Accepted: 26 February 2019

Abstract

In this work, the Hilbert envelope of the linear prediction (LP) residual and the residual phase have been explored for detecting replay attacks. The two source features namely, LP residual Hilbert envelope mel frequency cepstral coefficient (LPRHEMFCC) and residual phase cepstral coefficient (RPCC) are used for replay detection. From the signal perspectives, Hilbert envelope represents the amplitude information of LP residual samples. Residual phase represents to excitation information present in the sequence of LP residual samples. Hence, both can be considered as two components of the raw LP residual signal. In this direction, score level fusion of LPRHEMFCC and RPCC features is compared with a third source feature named as, residual mel frequency cepstral coefficient (RMFCC) derived from the raw LP residual using LP analysis. Comparative analysis has been performed using Gaussian mixtures model-universal background model (GMM-UBM) ASV experiments (IITG-MV replay database) and spoof detection experiments (ASVspoof 2017 database). For IITG-MV database, relative (RFAR-ZFAR) improvements of 86.10% (males), 27.45% (females) and 54.14% (whole-set) are achieved for (LPRHEMFCC + RPCC) + MFCC combination over RMFCC + MFCC combination. The RFAR and ZFAR stands for false acceptance rate under replay attacks and zero effort impostor attacks, respectively. In terms of tandem-detection cost function (t-DCF) metrics, the obtained relative improvements are 40.50%, 13.13% and 26.16%, respectively. For ASVspoof 2017 database, relative EER improvements of 11.72% and 6.74% are achieved for (LPRHEMFCC + RPCC) + MFCC and (LPRHEMFCC + RPCC) + CQCC over RMFCC + MFCC and RMFCC + CQCC, respectively. These observations justify the usefulness of exploring Hilbert envelope and residual phase components of the LP residual over direct processing of the LP residual signal for detecting replay attacks. Moreover, score level fusion of LPRHEMFCC, RPCC and CQCC provides 8.86% EER.

Keywords Speaker verification (SV) · Replay attacks detection · Hilbert envelope · Residual phase · Replay attack false acceptance rate (RFAR)

1 Introduction

Automatic speaker verification (ASV) systems accept/reject claimed identities on the basis of provided speech samples (Campbell 1997; Kinnunen and Li 2010). Modern automatic speaker verification (ASV) systems are vulnerable to spoof attacks (Wu et al. 2015a). Consequently, the development of efficient spoof detection algorithms is currently an active area of research (Wu et al. 2015b; Kinnunen et al. 2017). Spoof attack is an attempt of altering ASV system decisions by providing artificial speech samples of any target speaker (Evans et al. 2013). Generally four types of spoof attack methods are addressed in the literature, named as impersonation, replay, speech synthesis and voice conversion. *Impersonation* is a technique where an impostor himself sounds like a target speaker through voice mimicry (Hautamäki et al. 2013, 2015). *Replay attack* involves presentation of pre-recorded target speaker's speech samples through a playback device in front of the ASV system (Lindberg and Blomberg 1999; Villalba and Lleida 2010). *Speech synthesis (SS)* generates artificial target speech for a given input text (De Leon et al. 2010a, b). *Voice conversion (VC)*

✉ Madhusudan Singh
madhusudan_niit@yahoo.co.in

Debadatta Pati
debapati2003@yahoo.com

[1] Department of Electronics and Communication Engineering, National Institute of Technology Nagaland, Dimapur 797103, India

Published online: 04 March 2019

converts the impostor's voice characteristics towards the target speaker's voice (Bonastre et al. 2007; Kinnunen et al. 2012). Among all, currently low technology replay attack has acquired serious attention in the literature due to its effectiveness and accessibility (Wu et al. 2015a). It can be implemented using only a high quality recording and playback device, and hence may pose greater risk to ASV system (Wu et al. 2015a). Figure 1 shows a common ASV system with replay spoofing scenario. In genuine attempt, speech sample is directly presented to the ASV system, and is called as *actual speech*. In case of replay attempt, speech sample is passed through attacker's recording and playback devices before presented to the ASV system, and is known as *replay speech*. As consequence, replay speech gets affected by attacker's recording and playback device characteristics. Thus, detection of replay attacks is mainly rely on tracing the recording and playback device artifacts present in the replay signals.

Various approaches have been proposed for replay detection in literature. Most of them have utilized acoustic level differences between actual and replay signals caused by the playback devices, the recording devices and the environment acoustic conditions of the surroundings. In Villaba and Lieida (2011), increased spectral flatness of the replay signal due to inclusion of noise and reverberations from the surroundings was used to detect replay attacks using spectral ratio and modulation index based features. In Wang et al. (2011), the differences between the channel pattern noise computed from original and replay recordings was used for detecting replay signals. Study (Witkowski et al. 2017) explored the distortion occurs in the high frequency regions due to additional anti-aliasing filtering process during re-recording via microphone for replay detection. In Nagarsheth et al. (2017), a multi-class deep neural network (DNN) was trained to discriminate between different available channel conditions under replay attacks scenario. It was shown that multi-class DNN back end gives superior performance over two-class GMM back-end. The work in Raju Alluri and Gangashetty (2017), captures the changes in energy levels for detecting replay signals using single frequency filtering cepstral coefficients (SFFCC) features computed from the energy level corresponding to low signal-to-noise ratio (SNR) time instant in every 10 ms. In Jelil et al. (2017), source features namely, epoch feature (EF) and peak-to-side lobe ratio mean and skewness (PSRMS) were proposed to capture the variations in the source information for detecting replay attacks. Moreover, a comprehensive study on a set of different conventional and non-conventional features for the development of replay detection system was reported in Font et al. (2017). In ASVspoof 2017 challenge, the best reported replay detection system (Lavrentyeva et al. 2017) (EER = 6.73%, S01) is based on convolutional neural networks (CNN) and recurrent neural networks (RNN) learning approaches.

Moreover, few more recent works have been proposed in *Interspeech 2018* related to replay attacks detection. In Jelil et al. (2018), the authors have proposed compressed integrated linear prediction residual (CILPR) feature for discriminating between genuine and replayed signals. The CILPR feature models the changes in the temporal dynamics of the integrated linear prediction residual signal between two glottal closure instants (GCIs). In Li et al. (2018), two features namely, mel-scale relative phase (Mel-RP) and phase based source-filter vocal tract are combined at score level for replay detection. It has been shown that the phase based features outperforms conventional magnitude based features in replay detection context. In Tapkir and Patil (2018), empirical mode decomposition cepstral coefficients and linear frequency modified group delay cepstral coefficients have been proposed for detecting replay attacks. In Sailor et al. (2018), high frequency regions have been exploited using amplitude and frequency modulation (AM & FM) based temporal cepstral features for detecting replay attacks. The features are derived from convolutional restricted boltzmann machine (ConvRBM) using auditory filterbank learning approach. In Kamble et al. (2018), energy separation algorithm (ESA)-based features namely, instantaneous amplitude (IA) and instantaneous frequency (IF)

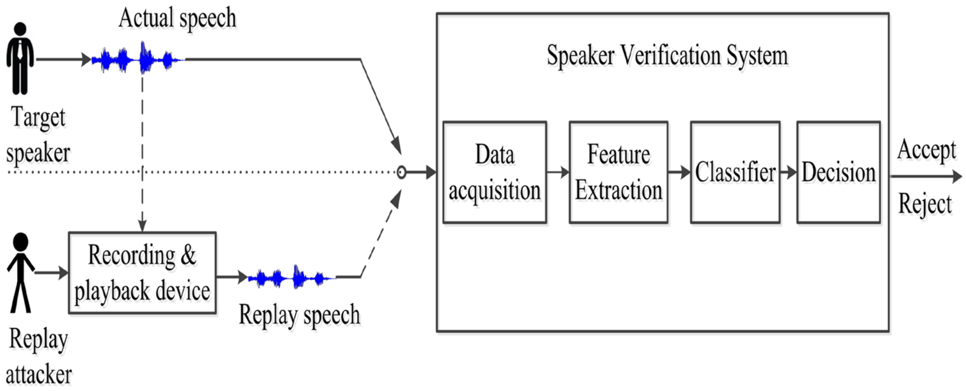


Fig. 1 Example ASV system demonstrating difference between a genuine and replay attempt

cosine coefficients have been proposed to classify between actual and replay signals. The features are extracted using linearly-spaced Gabor filtersbank to capture useful informative evidences for spoof detection from the entire spectrum. In Suthokumar et al. (2018), the authors have explored increment in the noise and reverberations from replay transmission channels for detecting replay attacks using two novel features namely, modulation centroid frequency cepstral coefficients and modulation static energy cepstral coefficients. Table 1 provides a summary of prior works related to replay attacks detection.

This work is the extension of our recent prior work (Singh and Pati 2018), published in *Interspeech 2018*. In that work, two source features namely, residual mel frequency cepstral coefficient (RMFCC) and a novel feature linear prediction residual Hilbert envelope mel frequency cepstral coefficient (LPRHEMFCC) were proposed for replay detection task. The RMFCC feature is obtained by short-term processing of the LP residual and provides compact modelling of excitation source information using mel-cepstral analysis. The LPRHEMFCC feature is obtained by short-term precessing of Hilbert envelope of the LP residual, and provides compact representation of the source information present specially around GCIs using mel-cepstral analysis. As extension, in this work another source feature called residual phase cepstral coefficient (RPCC) has been proposed along with LPRHEMFCC for detecting replay attacks. The RPCC feature is obtained by short-term processing of the residual phase, and provides compact representation of perceptually meaningful excitation source-like information using mel cepstral analysis. From the signal perspectives, Hilbert envelope of the LP residual mainly contains the amplitude information of LP residual samples. However, information about the sequence of the LP residual samples is lost in Hilbert envelop of the LP residual. It is the residual phase that represents the excitation source information present in the sequence of the LP residual samples (Murty and Yegnanarayana 2006). Thus, Hilbert envelope and residual phase can be considered as two components of the LP residual signal. Hilbert envelope is the magnitude function of the analytical signal derived using LP residual and Hilbert transform (Murty and Yegnanarayana 2006). The residual phase is the cosine of the phase function of the analytic signal. Even though analytical signal is common, LPRHEMFCC and RPCC features are derived from magnitude and phase function of the analytical signal using mel-cepstral analysis, respectively. Accordingly, the source information representations of LPRHEMFCC and RPCC features may be different. The RPCC feature explored in this work has been previously used for speaker verification task (Wang and Johnson 2012). The novelty of this work lies in exploring RPCC and LPRHEMFCC feature together for detecting replay attacks. The LPRHEMFCC and RPCC in together with CQCC provide 8.86% EER, and hence approaches to results reported in the state-of-the-art (refer Table 1). Further, intuitively, it can be predicted that the combination of individual source representations from the Hilbert envelope and residual phase would be comparable to stand-alone source representation from the raw LP residual signal. In this direction, combination of Hilbert envelope (LPRHEMFCC) and residual phase (RPCC) is compared with the raw LP residual (RMFCC) signal. With this motivation, a comparative analysis has been performed to investigate the usefulness of processing Hilbert envelope and residual phase (in together) over direct processing of raw LP residual signal. This is achieved through Gaussian mixtures model-universal background model (GMM-UBM) ASV experiments and spoof detection experiments using self-developed IITG-MV replay database and standard ASVspoof 2017 database, respectively. More on this, contrarily to current trends on development of stand-alone countermeasures, the proposed work aims to reject replay trials directly on the ASV system. The genuine trials and either zero-effort impostor trials or replay trials are classified using EER decision threshold of ASV system.

Table 1 This table represents a summary of prior works related to replay attacks detection

Study	Features	Classifier	EER (%)
Lavrentyeva et al. (2017)	FFT features	CNN, RNN	6.73 (S01)
Ji et al. (2017)	MFCC,PLP, CQCC	GMM based classifiers set	12.34 (S02)
Nagarsheth et al. (2017)	HFCC,CQCC	DNN-SVM	11.50
		GMM	18.10
Jelil et al. (2017)	EF, PSRMS, CQCC	GMM	17.61
Jelil et al. (2018)	CILPR, CQCC	GMM	9.41
Tak and Patil (2018)	LFCC, LFRCC, CQCC	GMM,CNN	9.06
Sailor et al. (2018)	AM-ConvRBM-CC FM-ConvRBM-CC	GMM	8.89

FFT Fast fourier transform, *PLP* perceptual linear prediction, *LFCC* linear frequency cepstral coefficient, *LFRCC* linear frequency residual cepstral coefficient

The significant point of interest here in that the proposed approach (in IITG-MV database context) does not require any separate countermeasure for detecting replay attacks and hence valuable contribution of this work.

The rest of the paper is organized as follows: Sect. 2 presents description of proposed LP residual features for replay detection. Comparative analysis among proposed source features using SV and spoof detection experiments has been provided in Sect. 3. The summary and future scopes of the work are reported in Sect. 4. The acknowledgment and list of refereed articles are reported at the end.

2 LP residual based features

In LP model of speech, a speech sample $s(n)$ is predicted as a linear weighted combination of previous p speech samples, and is expressed as (Makhoul 1975; Rabiner and Schafer 1978),

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

where, $\hat{s}(n)$ is the predicted speech sample, $a_k s$ are LP coefficients (LPCs) and p is the order of prediction. The difference (error) between actual and predicted speech signal is known as LP residual $r(n)$ and is given by,

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

The complex analytic signal $r_a(n)$ corresponding to $r(n)$ is given by Murty and Yegnanarayana (2006)

$$r_a(n) = r(n) + jr_h(n) \tag{3}$$

where $r_h(n)$ is the Hilbert transform of LP residual $r(n)$.

The Hilbert envelope $h(n)$ is the magnitude of analytical signal and is expressed as,

$$h(n) = \sqrt{r^2(n) + r_h^2(n)} \tag{4}$$

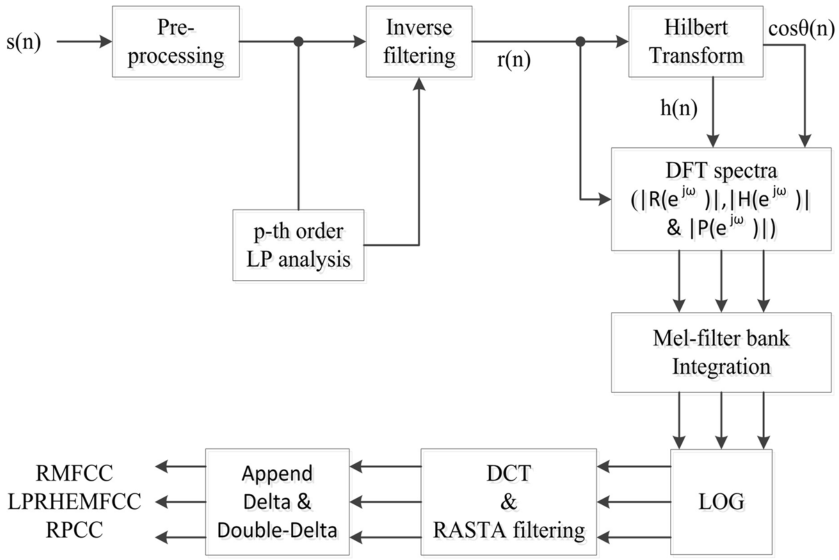
The residual phase is defined as cosine of the phase function of the analytical signal and is given by,

$$cos\theta(n) = \frac{real(r_a(n))}{h(n)} = \frac{r(n)}{h(n)} \tag{5}$$

2.1 RMFCC feature

RMFCC feature is extracted from the LP residual $r(n)$ via short-term mel cepstral analysis as shown in Fig. 2. Preprocessing stage involves the frame splitting of speech signal with 20 ms framesize and 10 ms frameshift. No voice-activity detection (VAD) has been performed as unvoiced frames may also persist useful channel information for replay detection (Font et al. 2017). Then, LP residual signal is obtained using LP analysis method. Discrete fourier transform (DFT) is performed to obtained LP residual spectrum. The magnitude of LP residual spectra is passed through non-uniform triangular band pass filter banks placed on the mel-frequency scale. At the end, discrete cosine transform (DCT) and RASTA (relative spectra) filtering is applied on the logarithm of the sub-band energies obtained from mel-filters bank to extract RMFCC features. If $R(e^{j\omega})$ is the spectrum of the LP residual $r(n)$, the magnitude of which is passed through mel-filters bank (M_{el}) for sub-band energy calculations. Then RMFCC feature ($R(k)$) is computed as,

Fig. 2 Block diagram for RMFCC, LPRHEMFCC and RPCC features extraction process. No. of mel-filters = 24, framesize=20 ms, frameshift=10 ms. (p = 10) for 8 kHz, (p = 18) for 16 kHz



$$R(k) = DCT[\log(M_{el}(|R(e^{j\omega})|)] \quad (6)$$

The source feature RMFCC involve frame based cepstral domain processing of the raw LP residual using 20 ms framesize and 10 ms overlap, and thereby models the glottal information averaged over two to three pitch periods (Das and Mahadeva Prasanna 2016).

2.2 LPRHEMFCC feature

LPRHEMFCC feature involves short-term mel cepstral processing of Hilbert envelope of the LP residual (Fig. 2). If $H(e^{j\omega})$ is the spectrum of the Hilbert envelope $h(n)$ of the LP residual, then LPRHEMFCC feature ($H(k)$) is computed in the following way,

$$H(k) = DCT[log(M_{el}(|H(e^{j\omega})|)] \quad (7)$$

Hilbert envelope of the LP residual represents amplitude information of the LP residual samples. In comparison the LP residual, the amplitudes are emphasized in Hilbert envelope at each epoch locations (instant of major excitations) in a pitch period. As a result, the high-amplitude values at the epochs in the signal dominate the computation of subband energies obtained from the mel filters banks (Raykar et al. 2005). Due to this, the obtained LPRHEMFCC feature mainly represents the source information especially at the epoch locations averaged over two to three pitch periods.

2.3 RPCC features

The source feature *RPCC* is obtained from short-term mel cepstral domain processing of the residual phase ($\cos\theta(n)$). It represents source information present in the sequence of LP residual samples, and provides perceptually meaningful excitation specific evidences for replay detection. The feature extraction steps to obtained RPCC feature are given in Fig. 2. If $P(e^{j\omega})$ is the spectrum of the residual phase, then RPCC feature ($P(k)$) is computed in the following way,

$$P(k) = DCT[log(M_{el}(|P(e^{j\omega})|)] \quad (8)$$

2.4 Usefulness of LP residual features for replay detection

Actual and corresponding replay parts of an speech signal, its LP residual, Hilbert envelope of the LP residual and residual phase are shown in Fig. 3. It seems a bit difficult to discriminate between actual and replay speech signals from the time domain representations (Fig. 3a, b). However, large fluctuations in amplitude levels occur in the LP residual of actual signal as compared to replay signal especially around GCIs (Fig. 3c, d). This corresponds to higher SNR around GCIs for actual speech as compared to replay speech. The similar patterns are observed in case of Hilbert envelopes of the LP residual (Fig. 3e, f). Similarly, a close observation of Fig. 3g and h (especially in 10–20 msec time span) can reveal that the residual phase corresponding to replay signal looks more denser as compared to actual case. This is due to the effect of intermediate devices of the replay mechanism. The replay setup contains a series of different components,

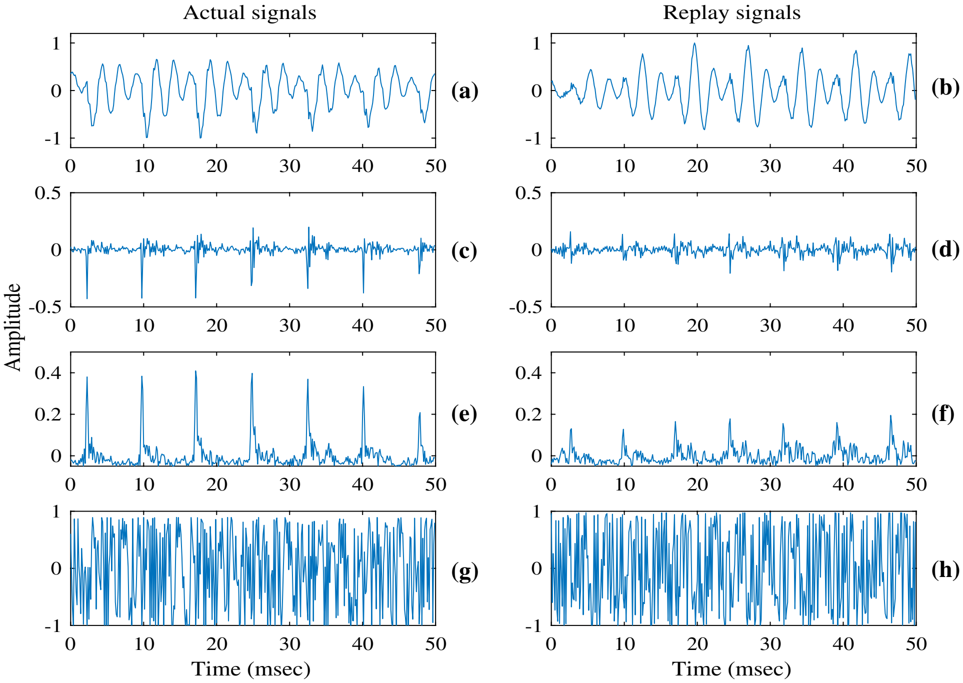


Fig. 3 **a** and **b** actual and replay speech signals, **c** and **d** corresponding LP residuals, **e** and **f** Hilbert envelopes of the LP residuals and **g** and **h** residual phase signals. The LP residuals are estimated using tenth-order LP analysis at 8 kHz sampling frequency. The speech files are taken from the ASVspoof 2017 database named as T_1000196.wav and T_1003010.wav

such as microphone, loudspeakers, amplifiers, input/output filters etc., and performs sequence of transformations on the input speech signal. These operations make some changes in phase of the input (actual) signal, resulting denser residual phase of the output (replay) signal. Thus, proposed LP residual based features in this work are able to provide compact modelling of theses discriminative characteristics from their respective residual signals using mel cepstral analysis, and hence would be useful in replay detection context.

3 Experimental study

This section demonstrates the usefulness of proposed residual based source features in detecting replay signals through ASV and spoof detection experiments.

3.1 Databases

The following two sub-sections provide detailed description of the databases used in this study.

3.1.1 IITG-MV

In this study, the replay database is manually developed by using publicly available Indian Institute of Technology Guwahati Multi-Variability (IITG-MV) speaker recognition database (Haris B. C. et al. 2012). The Phase-I (office) and Phase-II (laboratory) datasets of IITG-MV database are collected using five different microphone sensors in multiple environment conditions and in different sessions. Therefore suitable for robust speaker verification, to design database for replay attack and anti-spoofing studies like RSR database (Larcher et al. 2012). The speech samples are collected in two modes: In *read speech* mode, the participant reads pre-defined paragraphs of about 3–5 min duration. In *conversational speech* mode, the participant speaks freely about any topic for 10 to 15 min. In order to create replay recordings, we prefer latter mode due to the following reasons. The speaker characteristics including behavioral traits are relatively better manifested in conversational speech, and in general the replay attacker preferably acquires the speech samples (secretly) while the target is in conversation with others.

The Phase-I and Phase-II datasets of IITG-MV database contain 148 (112 males and 36 females) non-native English speakers speech samples, recorded at the rate of 16000 samples/second. The duration of the speech samples per speaker varies from 10 to 15 min. For this experimental study, we consider 81 (45 males and 31 females) speakers speech data and segregate into two groups: Dataset-I and Dataset-II. Dataset-I includes 5 male and 6 female speakers speech data amounting to 1 h from each gender for building gender-dependent UBM models. The Dataset-II is developed with 65 speakers speech data (comprising 40 males and 25 females) for evaluation purpose. Each speaker's first 2 min speech data are used for enrollment. The remaining data are converted into several segments of 30 s duration and used for test trials. Each test segment of each speaker is used as a genuine trial for the same target model and an impostor trial against other speakers model of the same gender. This resulted into a huge number of trials. The detail statistics are summarized in Table 2. Altogether, there are 42,440 trials that include 1274 genuine and 41,166 impostor trials. Spoofing an ASV system via replay attempt requires speech recordings from the target claimants only. Hence, number of replay trials are equal to number of target genuine trials.

The replay speech samples are generated manually by replaying the original data through a high quality CREATIVE-SBS-A35 loudspeaker (frequency response 100–15000Hz) almost in acoustically controlled environment (i.e. inside closed room with no fan and air condition noise) and re-recorded through an in-built microphone of HD Webcam C270-Logitech at the sampling rate of 16000 samples/s. We put very careful effort in acquiring the good quality replay speech samples in order to provide more challenging scenario. To verify the quality of the replay data, the original and replay recordings were played in front of few participants. They could hardly differentiating between them. This ensures that the collected replay recordings are of good quality.

The quality of the replay recordings can also be verified by estimating the distortion between actual and corresponding replay recordings using cepstral distance method (Nocerino et al. 1985). Cepstral distance (CSD) is an estimate of measuring distortion in the replay recordings from their corresponding actual signals. It represents the average Euclidean distance between the two recordings and is estimated using standard short-term cepstral analysis with hamming window of duration 20ms and 10ms overlap. The DC coefficient 'c_0' is omitted. Low CSD values characterize high-quality replay recordings. The mean and standard deviation of CSD values, estimated for whole 1274 trials (males and

Table 2 Summary of the dataset used in this work for genuine, impostor and replay trials

Statistics	Male	Female	Total
Background speakers	05	06	11
Target speakers	40	25	65
Genuine trials	706	568	1274
Impostor trials	27534	13632	41166
Replay trials (target claimants)	706	568	1274

The speech samples are taken from IITG-MV database (Haris et al. 2012)

Table 3 Quality verification of developed IITG-MV replay database with respect to standard ASVspoof 2017 database using following parameters: number of genuine and replay trials, mean and standard deviation of cepstral distance (CSD) between actual and replay recordings, and quality of playback and recording device (L = low, M = medium, H = high)

Parameters	Replay database		
	IITG-MV	ASVspoof2017	
		C1	C3
#Genuine trials	1274	1438	2363
#Replay trials	1274	1438	2363
CSD (μ)	0.80	0.79	0.77
CSD (σ)	0.16	0.16	0.28
Playback device quality	H	L	L/M
Recording device quality	M	L/M	L/M

females) are given in Table 3. It also contains two additional columns, representing the CSD values for C1 and C3 out of six evaluation conditions (C1–C6) of ASVspoof2017 database (in (Kinnunen et al. 2017), please refer Table 5 and Fig. 2). Conditions C1 and C3 represent comparatively low category replay trials with substantial spectral distortion. It can be observed that CSD values for the trials of IITG-MV database are in closed matching with CSD values of the trials under either C1 or C3 evaluation conditions of ASVspoof 2017 database. Although, both C1 and C3 are of low category but show wide variations in replay detection performance among top ten systems, thereby simulates challenging evaluation conditions. From this aspect, the developed IITG-MV replay database provides relatively homogeneous but challenging evaluation condition similar to either C1 or C3 category of ASVspoof 2017 database. Hence, it can be considered as useful database for ASV as well as replay spoofing and countermeasure studies. In addition, it facilitates the vulnerability study of ASV systems to replay attacks gender-wise.

3.1.2 ASVspoof 2017

The ASVspoof 2017 database involves speech files from the original *RedDots* corpus and corresponding replayed versions. The original speech corpus is replayed and re-recorded through various different kind of playback devices and microphones, and in different acoustic environments. Thus, provides highly varying acoustics replay attacks scenario. The database consists of three non-overlapping subsets: train, development and evaluation, the details are given in Table 4. The speech files and corresponding replay recordings are collected at 16 kHz with resolution 16-bit per sample.

Table 4 Details of the ASVspoof2017 database used for replay detection experiments

Database statistics	Number of speakers	Speech files	
		Actual	Replay
Train-set	10	1508	1508
Development-set	8	760	950
Evaluation-set	24	1298	12,008

3.2 Experimental setup

This section describes the details of features, classifiers and experimental setups used for SV and spoof detection experiments in this study.

3.2.1 Features

Three LP residual based features *RMFCC*, *LPRHEMFCC* and *RPCC* feature has been extracted using short-term mel-cepstral analysis as shown in Fig. 2. Thus, all three features are derived from all types of speech frames as unvoiced frames also contribute towards replay detection task along with voiced frames (Font et al. 2017; Hanili 2017). *MFCC* feature has been extracted from all speech frames (voiced or unvoiced) using standard short term mel-cepstral analysis and 24 channel triangular filterbanks. *CQCC* (Todisco et al. 2017) feature has been extracted by keeping the same baseline configuration provided in ASVspoof2017 challenge. It is based on the constant-Q transform (CQT), which results in variable time-frequency resolution across entire speech spectrum and thus captures useful informative characteristics for spoof detection. The CQCC features is post processed by cepstral mean variance normalization (CVMN) technique.

3.2.2 Classifiers

In this work, classical GMM-UBM is used at model level for ASV experiments on IITG-MV database. State-of-the-art modelling techniques such as, i-vector and related frameworks are available but requires large amount of data for training. In contrast, GMM-UBM is standard and popular modelling approach, and works satisfactorily at relatively small amount of training data (Reynolds et al. 2000). GMM-UBM outperformed i-vector particularly for unknown types of spoof attacks as reported in the study (Hanili et al. 2015). Further, present work focuses on exploration of discriminatory evidences more at feature level rather than model level. From these aspects, use of GMM-UBM classifier at model level is seems more suitable for this study. For Spoof detection experiments on ASVspoof2017 database, a two-class

GMM-classier is used to discriminate between actual and replay speech samples.

3.2.3 Setups for ASV and spoof detection experiments

From the available comprehensive literature on ASV studies, we have observed that processing speech files at sampling frequency (Fs) 8 kHz to perform SV experiments is very common and still followed mostly in speaker verification studies. Likewise, in the present study we have performed ASV experiments on IITG-MV database at Fs = 8kHz by down-sampling the 16 kHz speech files to 8 kHz speech files. Further for SV experiments, tenth order LP analysis is very common in literature, and adapted by most studies for extracting LP residual signal from speech signal sampled at 8kHz (Makhoul 1975; Prasanna et al. 2006; Raykar et al. 2005). In general, rule (Fs + 2, where 'Fs' is in kHz) is followed in the literature for doing the LP analysis (Makhoul 1975; Prasanna et al. 2006). Also, in Prasanna et al. (2006) the SV experiments were conducted over different LP order (8–20) and it was found that the performance is relatively better around LP order 10, indicating significant speaker-specific excitation information content in tenth order LP residual. Accordingly, 10-th order LP analysis is used to extract LP residual from speech signal sampled at 8 kHz. As per our experience on SV studies, 39-dimensional feature vectors are sufficient to providing significant speaker recognition accuracy and hence considered for ASV experiments, in this study.

The current stand-alone spoof detection methodology involves processing of speech files at their original Fs = 16 kHz for spoof detection experiments using ASVspoof2017 database. Similarly, here also, we have performed spoof detection experiments with Fs = 16kHz and 18-th order LP analysis. This facilitates comparison of our results with the recent prior works related to replay detection task. Further, we experimentally experience that the improvements are comparatively significant for 57-dimensional features and considered in most of the spoof detection studies. Accordingly, 57-dimensional feature vectors are used to conduct spoof detection experiments. Table 5 provides a summary of individual systems and their configurations used for ASV and spoof detection experiments in this study.

3.3 Evaluation process

The following evaluation metrics have been used for evaluating the features performances in this study.

3.3.1 EER and detection error tradeoff (DET)

The speaker verification performance is generally shown by detection error tradeoff (DET) curve and measured in terms of EER, where the false rejection rate (FRR) and false acceptance rate (FAR) are equal (Martin et al. 1997). In false rejection, a genuine is classified as an impostor while in false acceptance, an impostor is accepted as genuine speaker. Under replay spoofing scenario, the replay attackers usually aim at specified (target) speakers and thereby increases the FAR of the system. Thus, under replay attacks, the FAR is more relevant measuring parameter for evaluating the system performance. Accordingly, we have used two metrics to evaluate the system performance under both the baseline and spoofing test conditions: zero-effort false acceptance rate (ZFAR) and replay attack false acceptance rate (RFAR). ZFAR and RFAR is related to zero-effort impostor trials and replay trials, respectively.

The baseline ZFAR or equivalently the EER performance is computed by pooling all genuine and zero-effort impostor trials together. The RFAR performance is computed using the replay trials only, under replay attacks. The computation of RFAR is based on the fixed decision threshold (at EER point) of the baseline system as shown in Fig. 4. As same baseline ASV system is used for both ZFAR and RFAR computation, the difference '*RFAR* − *ZFAR*' directly indicates system vulnerability to replay attacks (Wu et al. 2015a). In positive sense, it represents ASV systems' capability to resist spoof attacks. A smaller value of '*RFAR* − *ZFAR*' indicates

Table 5 Summary of the experimental setups of the systems for ASV and spoof detection experiments. S, D and DD stands for static (excluding first energy term c_0), delta and delta-delta coefficients, respectively

Databases	Systems (classifiers)	Features	Dimension
IITG-MV	ASV (GMM-UBM) 256 Gaussian components	MFCC RMFCC LPRHEMFCC RPCC	(13-S + 13-D + 13-DD)
ASVspoof 2017	Spoof detection (GMM) 512 Gaussian components	MFCC RMFCC LPRHEMFCC RPCC CQCC	(19-S + 19-D + 19-DD)

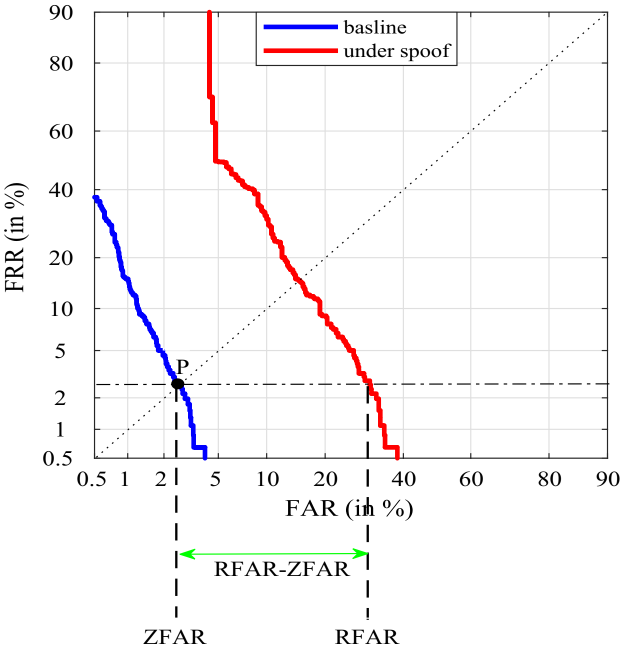


Fig. 4 Synthetic example shows computation of ZFAR and RFAR using decision threshold fixed at EER point (P) of the baseline system

better replay detection accuracy. For a foolproof spoof resistant ($RFAR = 0$) ASV system the $RFAR - ZFAR$ will approaches to ($-ZFAR$). Moreover, since same ASV system is used for both baseline and spoofing tests, the scores and decisions for all genuine trials will remain unaffected. Consequently, the FRR will remain constant, under both test conditions. Altogether, ZFAR, RFAR and their difference '$RFAR - ZFAR$' can be used as evaluation metrics to compare the different ASV systems performance under replay spoofing scenario.

3.3.2 Tandem-detection cost function (t-DCF)

t-DCF is a recently proposed performance evaluation scheme (Kinnunen et al. 2018). It is extension of conventional DCF used in ASV research to scenarios involving spoofing attacks. The present work is related to detection of replay trials directly on *ASV system without any countermeasures*. Accordingly, the suitable empirical formula for t-DCF can be expressed as,

$$\begin{aligned}\text{t-DCF(t)} &= C^{ASV}_{miss}.\pi_{tar}.P^{ASV}_{miss}(t) + C^{ASV}_{fa}.\pi_{non}.P^{ASV}_{fa}(t)\\ &\quad + C^{ASV}_{fa,spoof}.\pi_{spoof}.(1 - P^{ASV}_{miss,spoof}(t))\end{aligned} \tag{9}$$

where

$$\pi_{tar} + \pi_{non} + \pi_{spoof} = 1 \tag{10}$$

C^{ASV}_{miss} cost of ASV system rejecting a target trial, C^{ASV}_{fa}-cost of ASV system accepting a nontarget trial, $C^{ASV}_{fa,spoof}$ cost of ASV system accepting a spoof trial, $P^{ASV}_{miss}(t)$ probability that a target trial is missed by ASV system, $P^{ASV}_{fa}(t)$ probability that a nontarget trial is accepted by ASV system. (1-$P^{ASV}_{miss,spoof}(t)$) probability that a spoof trial is not missed by ASV system, π_{tar}, π_{non} and π_{spoof} are the target, nontarget and spoof a priori parameters, respectively. The π_{spoof} is typically set to 0.05 for computing t-DCF performance.

Equation 9 represents that the t-DCF performance is computed by pooling all genuine, zero-effort impostor trials and replay trials together. Hence, t-DCF given in Eq. 9 can be used as standard measure to evaluate the ASV performance involving replay attacks. This is very much consistent with the work method presented in this paper.

3.4 Experimental results and discussion

This section presents the experimental results for IITG-MV replay database and standard ASVspoof 2017 database, and followed by useful discussions based on the obtained results.

3.4.1 Results on IITG-MV database

Table 6 shows the results of ASV experiments conducted on IITG-MV database using different features and their combinations under both the baseline and replay spoofing test conditions. In case of zero-effort impostor trials the ASV performance is expressed in terms of ZFAR. Under replay attacks the performance is expressed in terms of RFAR. The t-DCF performance considers pooling of all genuine, zero-effort impostor and replay trials together. With reference to ZFAR values, the corresponding RFAR values are higher for MFCC and RMFCC features in all cases. In contrast, opposite patterns are obtained for LPRHEMFCC and RPCC features particularly in male speakers case. Usually, an ASV system shows higher RFAR value (under replay attacks) as compared respective baseline ZFAR value, which is not the case with LPRHEMFCC and RPCC based ASV systems. For a foolproof spoof resistant ASV system the RFAR value should be zero (refer Sect. 3.3.1). In this regard, smaller value of RFAR with respect to corresponding ZFAR for these features shows their strong caliber towards detection of replay attacks. This can also be validated through comparing respective computed t-DCF values across different features given in the last column of the Table 6.

The contribution of this work is better reflected as (source features + MFCC) combinations. From the fourth column of the Table 6, the relative (RFAR-ZFAR) improvements of 86.10% (males), 27.45% (females) and 54.14% (whole-speaker-set) have been achieved for (LPRHEMFCC + RPCC) + MFCC combination as compared to RMFCC +

Table 6 ASV results for different features and their combinations using GMM-UBM classifier. The features are combined at score level using simple weighted linear combination scheme with equal weights

Features	ZFAR(%)	RFAR(%)	Difference (RFAR-ZFAR)	t-DCF ($\pi_{spoof} = 0.05$)
Male				
MFCC	2.97	38.81	35.84	0.225
RMFCC	5.38	15.72	10.34	0.136
LPRHEMFCC	12.62	7.93	− 4.69	0.170
RPCC	12.74	7.51	− 5.23	0.170
LPRHEMFCC + RPCC	8.22	7.79	− 0.43	0.124
RMFCC + MFCC	2.97	29.46	**26.49**	**0.158**
(LPRHEMFCC + RPCC) + MFCC	4.82	8.50	**3.68**	**0.094**
Female				
MFCC	3.69	65.14	61.45	0.364
RMFCC	5.46	51.40	45.94	0.314
LPRHEMFCC	10.78	22.00	11.22	0.220
RPCC	9.68	56.34	46.66	0.382
LPRHEMFCC + RPCC	8.45	34.86	26.41	0.262
RMFCC + MFCC	3.69	61.44	**57.75**	**0.335**
(LPRHEMFCC + RPCC) + MFCC	5.28	47.18	**41.90**	**0.291**
Whole-set				
MFCC	4.24	54.08	49.84	0.314
RMFCC	5.65	31.16	25.51	0.214
LPRHEMFCC	13.20	8.00	− 5.20	0.177
RPCC	12.23	27.63	15.40	0.265
LPRHEMFCC + RPCC	9.81	18.52	8.71	0.194
RMFCC + MFCC	3.80	41.20	**37.51**	**0.237**
(LPRHEMFCC + RPCC) + MFCC	5.80	23.00	**17.20**	**0.175**

Bold results to highlight the significance of processing LP residual signal in terms of its components i.e., Hilbert envelope (LPRHEMFCC) and residual phase (RPCC) rather than direct processing of LP residual signal (RMFCC)

MFCC combination. Whereas in terms of t-DCF metric, the obtained relative improvements are 40.50% (males), 13.13% (females) and 26.16% (whole-set). This indicates the usefulness of processing LP residual in terms of its components (Hilbert envelope and residual phase) rather direct processing of raw LP residual signal for replay detection task. Further, the performance for females speakers is more worsen under replay attacks as compared to male speakers. This may be due to less spectral (harmonics) distortion caused by playback devices in case of female speech as they have lesser number of harmonics than male speech (Pépiot 2014). The DET plots shown in Fig. 5 shows ASV performance of (LPRHEMFCC + RPCC) + MFCC combination and stand-alone MFCC feature under baseline and spoof attacks scenario. Small gap in the baseline performance and large gap in the performance under spoof attacks reflect the significance of the fusion of source features along with MFCC in developing spoof resistant ASV system without using any stand-alone classifier for replay detection.

3.4.2 Results on ASVspoof2017 database

The obtained SV results on IITG-MV database show the usefulness of the Hilbert envelope of the LP residual and Residual phase in the development of spoof resistant ASV system. Nevertheless, efficacy of the proposed system should be validated under replay attacks in practical scenario. In this direction, the experiments for the proposed features set are repeated using pooled ASVspoof2017 database. The systems are trained on train-set and (train + developement) set and tested using evaluation-set. The spoof detection results are presented in Table 7. As expected, the (LPRHEMFCC + RPCC) + MFCC system gives better performance as compared to RMFCC + MFCC system. Likewise, (LPRHEMFCC + RPCC) + CQCC combination outperforms RMFCC + CQCC combination. These observations confirm joint use of Hilbert envelope and residual phase in rejecting replay spoofing trials over the raw LP residual signal. The DET plots corresponding to best EER performances are shown in Fig. 6. Moreover, 8.86% EER is achieved for score level fusion of features LPRHEMFCC, RPCC and CQCC, and

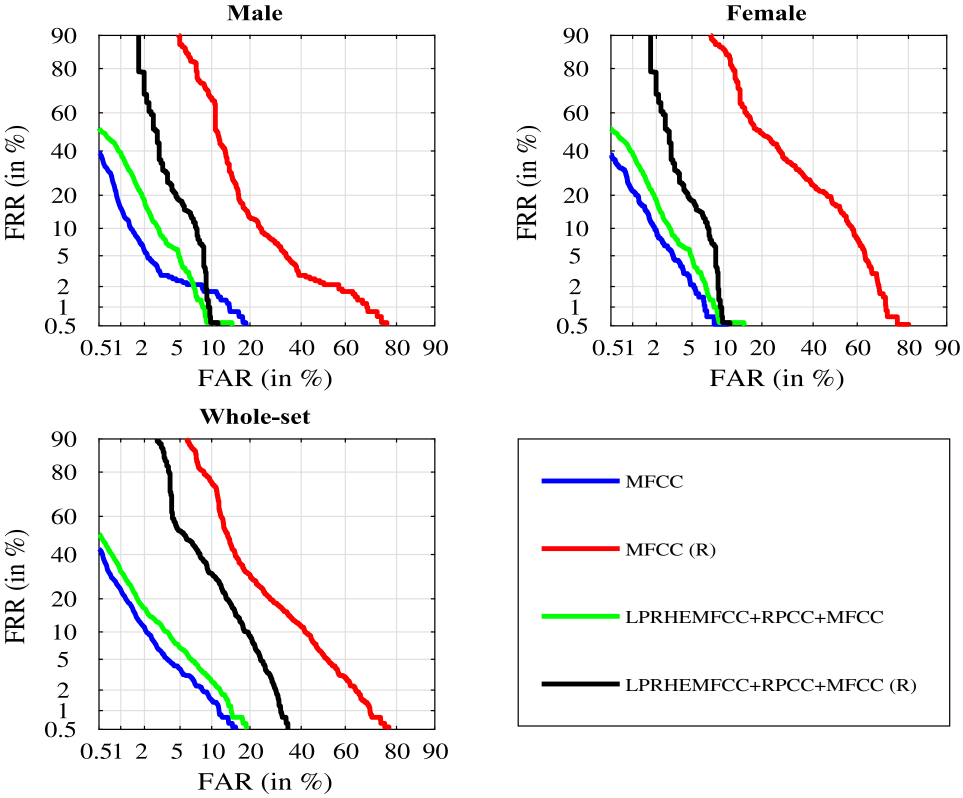


Fig. 5 DET plots showing ASV performance under baseline and replay attacks ('R') test scenario for (LPRHEMFCC + RPCC) + MFCC combination and stand-alone MFCC feature in all cases, i.e., male, female and whole-set. Here, baseline evaluation involves pooling of genuine trials and zero-effort impostor trials while evaluation in spoofing scenario involves pooling of genuine trials and replay trials (as impostor trials)

Table 7 Spoof detection results for different features and their combinations using 2-class GMM classifier

Features	EER (%)	
	Train	(Train + dev)
MFCC	33.62	22.73
RMFCC	27.07	20.88
LPRHEMFCC	25.26	23.65
RPCC	23.52	19.00
CQCC	18.83	15.16
LPRHEMFCC + RPCC	20.20	17.48
RMFCC + MFCC	26.04	16.80
(LPRHEMFCC + RPCC) + MFCC	20.22	14.83
RMFCC + CQCC	11.79	9.50
(LPRHEMFCC + RPCC) + CQCC	11.71	**8.86**
EF (Jelil et al. 2017)	NA	28.66
PSRMS (Jelil et al. 2017)	NA	28.90
(EF + PSRMS) + CQCC (Jelil et al. 2017)	NA	17.61
CILPR (Jelil et al. 2018)	20.66	15.76
CILPR + CQCC (Jelil et al. 2018)	9.77	9.41

The features are combined at score level using bosaris_toolkit (The Bosaris toolkit 2013)

Bold values indicate the minimum best EER (8.86%) has been achieved for the combination LPRHEMFCC+RPCC+CQCC in this work on standard ASVspoof 2017 database

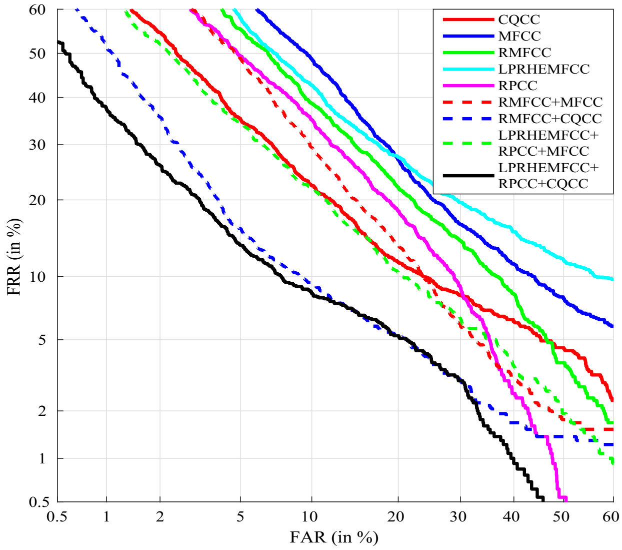


Fig. 6 DET plots representing EER performances of individual features and combinations

thus approaches to results in the state-of-the-art (refer Table 1). Further, Table 7 compares the proposed source features set in this work with the prior works (Jelil et al.

2017, 2018). The (LPRHEMFCC + RPCC) + CQCC (EER = 8.86%) combination outperforms with respect to (EF + PSRMS) + CQCC (EER = 17.61%) and CILPR + CQCC (EER = 9.41%) combinations.

3.4.3 Discussions

The useful remarks from the experimental outcomes are as follows:

- Combining evidences from the source features along with MFCC improves RFAR-ZFAR performance by considerable amount with respect to stand-alone MFCC feature in all cases, especially in case of male dataset (t-DCF = 0.094 and RFAR-ZFAR = 3.68%).
- For female dataset, the RFAR-ZFAR value is very high (41.90%) thereby still vulnerable to replay attacks, and hence requires further enhancement.
- It is easy to fool female speakers as compared to male speakers.
- Score level combination of independently processed LP residual components (i.e. Hilbert envelope and residual phase) performs better than stand-alone raw LP residual signal.
- The obtained t-DCF (IITG-MV whole-set) and EER (ASVspoof 2017) results show that the combination (LPRHEMFCC + RPCC) + MFCC performs better than the MFCC and combination (LPRHEMFCC + RPCC). This represents the consistency of proposed features combination across both the databases.
- Combination of independently processed LP residual components i.e. Hilbert envelope and residual phase is more beneficial as compared to directly processed raw LP residual, in replay detection context.

4 Summary and future scopes

This work demonstrates the usefulness of processing LP residual signal in the form of combination (Hilbert envelope + residual phase) over direct processing of the raw LP residual samples in rejecting replay spoofing trials. The ASV results obtained on IITG-MV database show that the combination LPRHEMFCC + RPCC outperforms RMFCC feature in rejecting replay trials in all cases i.e., male, female and whole-set. The LPRHEMFCC and RPCC in together with MFCC feature reduce the 'RFAR-ZFAR' performance by considerable amount with respect to stand-alone MFCC feature especially in case of male dataset (RFAR-ZFAR = 3.68%). This reflects that joint use of source features set and system feature (such as MFCC and CQCC) may be proved as a reliable candidate for the development of spoof resistant ASV system. The outcomes of ASV experiments on IITG-MV database have been validated through spoof detection experiments on standard ASVspoof 2017 database. The LRHEMFCC + RPCC (EER = 17.48%) outperforms RMFCC feature (EER = 20.88%) by 3.40%.

The presented work in this paper, provides a motivation towards the development of spoof resistance ASV system which will handle the spoofing and ASV problem together. Further, self developed IITG-MV replay database can be a useful contribution to speech research community towards the development of spoof resistance ASV system. Compared to conventional MFCC features extraction steps, feature extraction techniques used here require only few additional signal processing steps, which are computationally inexpensive. Hence, the proposed method in this work can easily be adapted in the hand-held mobile devices for real time applications. Future plan is to search of explicit excitation source characteristics such as fundamental frequency (F0) and strength of excitation (SoE) for further enhancement in the ASV performance under replay attacks scenario.

Acknowledgements This research work is funded by Ministry of Electronics and Information Technology (MeitY), Govt. of India through the project "Development of Excitation Source Features Based Spoof Resistant and Robust Audio-Visual Person Identification System". The research work is carried out in Speech Processing and Pattern Recognition (SPARC) laboratory at National Institute of Technology Nagaland, Dimapur, India.

References

Bonastre, J. F., Matrouf, D., & Fredouille, C. (2007). Artificial impostor voice transformation effects on false acceptance rates. In: *Proceedings of interspeech*, pp 2053–2056

Campbell, J. P, Jr. (1997). Speaker recognition: A tutorial. *Proceedings on IEEE*, *85*(9), 1437–1462.

Das, R. K., & Prasanna, S. M. (2016). Exploring different attributes of source information for speaker verification with limited test data. *The Journal of the Acoustical Society of America*, *140*(1), 184–190.

De Leon, P. L., Apsingekar, V. R., Pucher, M., & Yamagishi, J. (2010a). Revisiting the security of speaker verification systems against imposture using synthetic speech. In: *Proceedings of ICASSP*, pp 1798–1801

De Leon, P. L., Pucher, M., & Yamagishi, J. (2010b). Evaluation of the vulnerability of speaker verification to synthetic speech. In: *Proceeding of Odyssey: The Speaker and Language Recognition Workshop* p 28

Evans, N., Kinnunen, T., & Yamagishi, J. (2013). Spoofing and countermeasures for automatic speaker verification. In: *Proceedings of interspeech*, pp 925–929

Font, R., Espın, J. M., & Cano, M. J. (2017). Experimental analysis of features for replay attack detection—Results on the ASVspoof 2017 challenge. In: *Proceedings of interspeech* pp 7–11

Hanilçi, C. (2017). Linear prediction residual features for automatic speaker verification anti-spoofing. *Multimedia Tools and Applications* pp 1–13

Hanilçi, C., Kinnunen T, Tomi., Sahidullah, M., & Sizov, A. (2015). Classifiers for synthetic speech detection: A comparison. In: *Proceeding of interspeech*, pp 2057–2061

Haris, B. C., Pradhan, G., Prasanna, S. R. M., Das, R. K., & Sinha, R. (2012). Multivaribility speaker recognition database in Indian scenario. *International Journal of Speech Technology (Springer)*, *15*(4), 441–453.

Hautamäki, R. G., Kinnunen, T., Hautamäki, V., Leino, T., & Laukkanen, A. M. (2013). I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In: *Proceeding of interspeech*, pp 930–934

Hautamäki, R. G., Kinnunen, T., Hautamäki, V., & Laukkanen, A. M. (2015). Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication*, *72*, 13–31.

Jelil, S., Das, R. K., Prasanna, S. M., & Sinha, R. (2017). Spoof detection using source, instantaneous frequency and cepstral features. In: *Proceedings on interspeech* pp 22–26

Jelil, S., Kalita, S., Prasanna, S. R. M., & Sinha, R. (2018). Exploration of compressed ILPR features for replay attack detection. In: *Proceedings on interspeech*, pp 631–635

Ji, Z., Li, Z. Y., Li, P., An, M., Gao, S., Wu, D., & Zhao, F. (2017). Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017. In: *Proceedings of interspeech*, pp 87–91

Kamble, M., Tak, H., & Patil, H. (2018). Effectiveness of speech demodulation-based features for replay detection. In: *Proceeding of interspeech*, pp 641–645

Kinnunen, T., Lee, K. A., Delgado, H., Evans, N., Todisco, M., Sahidullah, M., Yamagishi, J., & Reynolds, D. A. (2018). t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. In: *Proceeding of Odyssey the speaker and language recognition workshop*, pp 312–319

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, *52*, 12–40.

Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., & Lee, K. A. (2017). The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In: *Proceeding of interspeech*, pp 2–6

Kinnunen, T., Wu, Z. Z., Lee, K. A., Sedlak, F., Chng, E. S., & Li, H. (2012). Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In: *Proceeding of ICASSP*, pp 4401–4404

Larcher, A., Lee, K. A., Ma, B., & Li, H. (2012). RSR2015: Database for text-dependent speaker verification using multiple passphrases. In: *Proceeding of interspeech*, pp 1580–1583

Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., & Shchemelinin, V. (2017), Audio replay attack detection with deep learning frameworks. In: *Proceeding of interspeech*, pp 82–86

Li, D., Wang, L., Dang, J., Liu, M., Oo, Z., Nakagawa, S., Guan, H., & Li, X. (2018). Multiple phase information combination for replay attacks detection. In: *Proceeding of interspeech*, pp 656–660

Lindberg, J., & Blomberg, M. (1999). Vulnerability in speaker verification: A study of technical impostor techniques. In: *Proceeding of EUROSPEECH*, pp 5–9

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceeding of IEEE*, *63*(4), 561–580.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. In: *Proceeding on European conference on speech communication technology*, Rhodes, Greece, *4*, pp 1895–1898

Murty, K. S. R., & Yegnanarayana, B. (2006). Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Process Letter*, *13*(1), 52–55.

Nagarsheth, P., Khoury, E., Patil, K., & Garland, M. (2017). Replay attack detection using DNN for channel discrimination. In: *Proceeding of interspeech*, pp 97–101

Nocerino, N., Soong, F., Rabiner, L., & Klatt, D. (1985). Comparative study of several distortion measures for speech recognition. *Proceeding of ICASSP*, *10*, 25–28.

Pépiot, E. (2014). Male and female speech: A study of mean F0, F0 range, phonation type and speech rate in parisian french and American English speakers. *Speech Prosody*, *7*, 305–309.

Prasanna, S. R. M., Gupta, C. S., & Yegnanarayana, B. (2006). Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication*, *48*, 1243–1261.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Englewood Cliffs: Prentice-Hall.

Raju Alluri, K., & Gangashetty, A. K. V. (2017). SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017. In: *Proceeding of interspeech*, pp 107–111

Raykar, V. C., Yegnanarayana, B., Prasanna, S. M., & Duraiswami, R. (2005). Speaker localization using excitation source information in speech. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 751–761.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*, 19–41.

Sailor, H., Kamble, M., & Patil, H. (2018). Auditory filterbank learning for temporal modulation features in replay spoof speech detection. In: *Proceeding of interspeech*, pp 666–670

Singh, M., & Pati, D. (2018). Linear prediction residual based short-term cepstral features for replay attacks detection. *Proceeding of interspeech*, *2018*, 751–755.

Suthokumar, G., Sethu, V., Wijenayake, C., & Ambikairajah, E. (2018). Modulation dynamic features for the detection of replay attacks. In: *Proceeding of interspeech*, pp 691–695

Tak, H., & Patil, H. (2018). Novel linear frequency residual cepstral features for replay attack detection. In: *Proceeding of interspeech*, pp 726–730

Tapkir, P., & Patil, H. (2018). Novel empirical mode decomposition cepstral features for replay spoof detection. In: *Proceeding of interspeech*, pp 721–725

The Bosaris toolkit [software package]. Retrieved from https://sites.google.com/site/bosaristoolkit

Todisco, M., Delgado, H., & Evans, N. (2017). Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech and Language*, *45*, 516–535.

Villalba, J., & Lleida, E. (2010). Speaker verification performance degradation against spoofing and tampering attacks. In: *FALA 10 workshop*, pp 131–134

Villaba, J., & Lieida, E. (2011). Preventing replay attacks on speaker verification systems. In: *Proceeding of International carnahan conference on security technology (ICCST)*, pp 1–8

Wang, J., & Johnson, M. (2012). Residual phase cepstrum coefficients with application to cross-lingual speaker verification. In: *Interspeech*

Wang, Z., Wei, G., & He, Q. H. (2011). Channel pattern noise based playback attack detection algorithm for speaker recognition. In: *Proceeding of IEEE Int conference of the biometrics special interest Group (BIOSIG) on machine learning and cybernetics*, pp 1708–1713

Witkowski, M., Kacprzak, S., Zelasko, P., Kowalczyk, K., & Gałka, J. (2017). Audio replay attack detection using high-frequency features. In: *Proceeding of interspeech*, pp 27–31

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., & Li, H. (2015a). Spoofing and counter measures for speaker verification: A survey. *Speech Communication*, *66*, 130–153.

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., & Sizov, A. (2015b). ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In: *Proceeding of interspeech*, pp 2037–2041