Multimed Tools Appl
DOI 10.1007/s11042-017-5181-0

# Linear prediction residual features for automatic speaker verification anti-spoofing

**Cemal Hanilçi**[1]

Received: 15 February 2017 / Revised: 15 July 2017 / Accepted: 30 August 2017
© Springer Science+Business Media, LLC 2017

**Abstract** Automatic speaker verification (ASV) systems are highly vulnerable against spoofing attacks. Anti-spoofing, determining whether a speech signal is natural/genuine or spoofed, is very important for improving the reliability of the ASV systems. Spoofing attacks using the speech signals generated using speech synthesis and voice conversion have recently received great interest due to the 2015 edition of Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015). In this paper, we propose to use linear prediction (LP) residual based features for anti-spoofing. Three different features extracted from LP residual signal were compared using the ASVspoof 2015 database. Experimental results indicate that LP residual phase cepstral coefficients (LPRPC) and LP residual Hilbert envelope cepstral coefficients (LPRHEC) obtained from the analytic signal of the LP residual yield promising results for anti-spoofing. The proposed features are found to outperform standard Mel-frequency cepstral coefficients (MFCC) and Cosine Phase (CosPhase) features. LPRPC and LPRHEC features give the smallest equal error rates (EER) for eight spoofing methods out of ten spoofing attacks in comparison to MFCC and CosPhase features.

**Keywords** Speaker verification · Anti-spoofing · Countermeasure · Linear prediction residual

## 1 Introduction

Automatic speaker verification (ASV) aims at accepting or rejecting an identity claim given a speech signal [13]. Recent developments in ASV technology has led to considerable performance improvements which makes the voice one of the most important biometric modality for person authentication.

✉ Cemal Hanilçi
cemal.hanilci@btu.edu.tr

[1] Department of Electrical and Electronics Engineering, Bursa Technical University, Bursa, Turkey

As in the case for other biometric modalities such as face and fingerprint [26], spoofing attacks are one of the most important security concern for ASV systems [9]. Spoofing attack refers to an attack where an attacker masquerades herself as a target user in order to gain unauthorized access to the system. Spoofing attacks against a biometric person authentication system can be categorized into two groups [26]: *indirect attacks* where attacks are evaluated within the system such as software and *direct attacks* (also known as presentation attacks) where attack is evaluated at the sensor level by presenting a forged biometric data [9, 26] and they have been under active research for all types of biometric systems.

ASV systems can *directly* be spoofed by mainly four types of attacks [36]: (i) *impersonation* where attacker mimics the target speaker's voice [7, 8], (ii) *replay* [3, 31] – presentation of target speaker's prerecorded voice sample–, (iii) *speech synthesis* (SS) [17, 18] – generating target speaker's voice by speech synthesis technique given an input text – and (iv) *voice conversion* (VC) [2, 15] – modifying attacker's voice signal towards the target speaker's speech. For more details about the spoofing attacks against speaker recognition systems, readers are referred to [4, 34, 36].

Countermeasures, determining whether a speech signal is natural or spoofed, play an important role for the reliability of ASV systems against spoofing attacks. Among the four attack types, detecting SS and VC spoofing attacks have gained more attention over the impersonation and replay attacks. This is because of the recently organized *Automatic Speaker Verification Spoofing and Countermeasures Challenge* (ASVspoof 2015) [37]. In ASVspoof challenge, natural (genuine) and synthetic/converted speech signals generated from ten different SS and VC techniques were shared with the participants and asked to determine whether a speech signal is genuine or spoofed.

Various countermeasures have been proposed for spoofing detection. In [35], authors compared Mel-frequency cepstral coefficients (MFCC), modified group delay (MGD) and cosine phase (CosPhase) features for detecting the synthetic speech signals and CosPhase features were found to outperform MFCC and MGD. For the ASVspoof 2015 challenge, phase based features (e.g. CosPhase [35], MGD [35] and relative phase shift (RPS) [29]) were found to be superior to magnitude based features, in general [23, 32, 33, 39]. In [10], linear prediction (LP) residual signal obtained from the LP analysis followed by long-term prediction (LTP) is used to extract audio quality features (e.g. average energy of the LP residual, maximum energy of LTP residual, average and maximum of the LTP gain) for spoofing detection and it was shown that residual features give encouraging results for spoofing detection. Similar to [10], in [1], LP residual is computed by LP analysis followed by non-linear prediction (NLP) and similar audio quality features were extracted from LP-NLP residual for spoofing detection.

Intuitively, spoofing detection requires the features capturing the imperfections produced by VC and SS techniques. For instance, well-known assumption for human speech production model uses a time-varying filter driven by impulse train excitation. The voice coders (vocoders) used for speech waveform generation in VC and SS systems generally uses such simple models [25]. Therefore, source excitation parametrization as features would intuitively be beneficial for spoofing detection. To this end, we propose to use both magnitude and phase based features extracted from the analytic signal of LP residual for spoofing detection. In this study, we propose to use cepstral coefficients extracted from the phase function of the analytic signal computed from the LP residual for spoofing detection and systematically analyse the performance of the proposed features.

## 2 Linear prediction residual features

Linear prediction (LP) analysis assumes that a speech sample, $x[n]$, can be estimated as a linear combination of its $p$ previous samples by, $\hat{x}[n] = -\sum_{k=1}^{p} \alpha_k x[n-k]$ [19]. Here $x[n]$ is the original speech sample, $\hat{x}[n]$ is its predicted counterpart, $p$ is the prediction order and $\{\alpha_k\}_{k=1}^{p}$ are the predictor coefficients. The predictor coefficients mostly represent the information related to vocal tract [19] and removing this information from the speech signal results a residual signal which is roughly an approximation of the source excitation [19, 25]. The LP residual signal (prediction error) is defined as the difference between the actual speech sample $x[n]$ and the predicted sample $\hat{x}[n]$:

$$e[n] = x[n] - \hat{x}[n] = x[n] + \sum_{k=1}^{p} \alpha_k x[n-k]. \tag{1}$$

Features extracted from the LP residual signal have successfully been used for speaker and language recognition in previous studies [21, 22, 24] and it was shown that these features convey complementary information and improve the speaker and language recognition performance when they are combined with standard MFCC features.

Since the values of the LP residual signal samples change rapidly, its amplitude variation is considerably high, it is difficult to extract useful information from LP residual signal using short-term analysis [21]. Thus, the features are extracted from analytic signal derived from the LP residual [21, 24]:

$$e_a[n] = e[n] + j e_h[n] \tag{2}$$

where $e_h[n]$ is the Hilbert transform of the $e[n]$.

Figure 1 shows an example of genuine and spoofed speech frames with their LP residual counterparts and Hilbert envelopes (absolute value of the analytic signal given in (2)). From the figure, it can be seen that it is difficult to discriminate natural speech signal from the spoofed speech using the original time-domain signals (the first row of the figure). However, LP residual of the genuine speech segment differs from that of spoofed segments in some manners. First, the large fluctuations occur in LP residual of genuine segment whereas the LP residual of spoofed segments are more like impulse train shape and the fluctuations occur only around the glottal closure instants within each pitch period. Therefore, signal-to-noise ratio around the glottal closure instants are higher than that of genuine speech. The similar observations hold for the Hilbert envelopes of the LP residual signal (the last row of the figure). Second, the amplitude variations, i.e. the fluctuations occur in signal amplitude over time, in the genuine speech segments are higher than that of the spoofed segments which can be seen from the Hilbert envelope graphs. These variations and differences between the natural and spoofed signals would be important features for the spoofing detection.

In this study we extract three different features derived from the LP residual signal. The first set of features are LP residual Hilbert envelope cepstral coefficients (LPRHEC). LPRHEC features are obtained by applying discrete cosine transform (DCT) to the logarithm of the magnitude of the analytic signal given in (2).

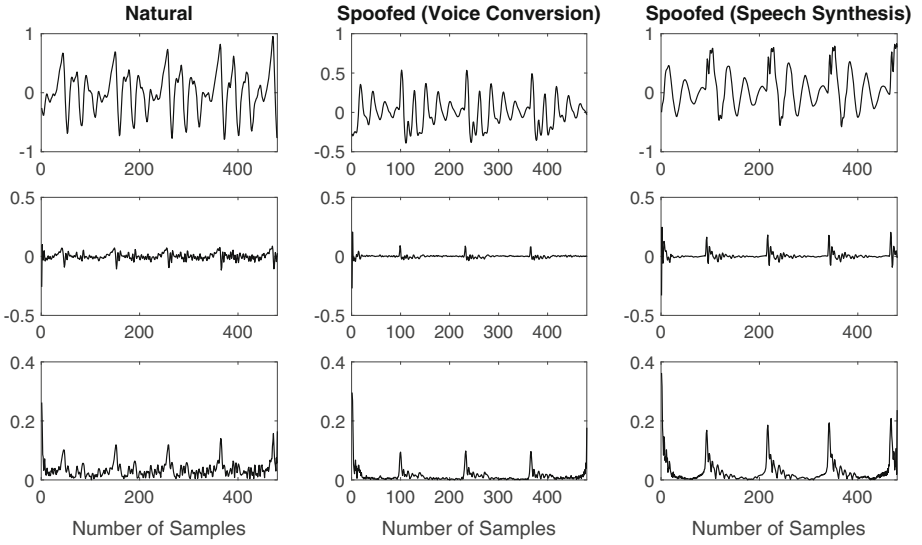


**Fig. 1** Genuine and spoofed voiced speech segments (first row), corresponding LP residual signals (second row) and their Hilbert envelopes (last row)

The second feature representation is LP residual phase (LPResPhase) features. LPResPhase features are defined as the cosine of the analytic signal phase function and obtained by [21, 22, 24]:

$$\cos(\theta[n]) = \frac{e[n]}{\sqrt{e^2[n] + e_h^2[n]}}. \tag{3}$$

LPResPhase features previously used in different recognition tasks based on speech signals such as speaker and language recognition [21, 22, 24]. Besides LPRHEC and LPResPhase features, we propose to use a modified form of the LPResPhase which we refer to as LP residual phase cepstral coefficients (LPRPC). Since the values of the phase function obtained from the LP residual signal mostly consist of the correlated samples, in order to reduce the redundancy and improve the relevance, LPRPC features are obtained by applying discrete cosine transform to the LP residual phase function given in (3). The extraction of LP residual features used in this study are summarized in Fig. 2. Note that, in Fig. 2, the first step of the feature extraction process is the use of pre-emphasis filter. The pre-emphasis filter is used as a pre-processing step because according to previous results reported in [28] and our initial experiments on ASVspoof 2015 database, the high frequency components

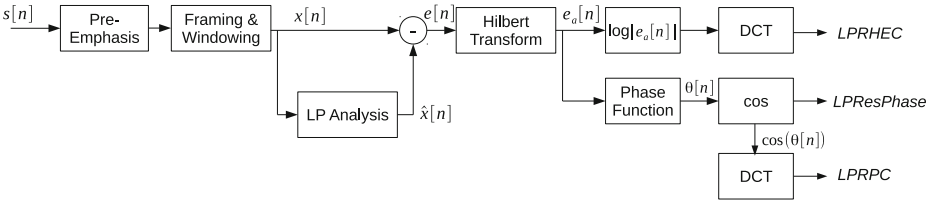


**Fig. 2** Block diagram of LP residual based feature extraction techniques used in this work

were found to be more useful than low-frequency regions for spoofing detection. Therefore high frequency components of the signal are emphasized by applying pre-emphasis filter.

## 3 Anti-spoofing setup

### 3.1 Database

ASVspoof 2015 database [37, 38] is used in the experiments. ASVspoof database consists of natural and synthetic speech signals from 45 male and 61 female speakers. The database is partitioned into three disjoint subsets without any speaker overlap across the subsets [37, 38]:

- **Training Set**: Training set includes 3750 natural and 12625 synthetic speech signals from 10 male and 25 female speakers. Synthetic speech signals are generated by one of the five easily implemented SS or VC algorithms (S1-S5). S1, S2 and S5 are the VC algorithms whereas S3 and S4 are the speech synthesis algorithms. Among the five spoofing algorithms, S1, S2, S3 and S4 all use the STRAIGHT vocoder [12] for waveform generation. However, S5 uses the MLSA vocoder [5]. The training subset is used to train the natural and spoofed speech classes for spoofing detection.
- **Development Set**: Development set consists of both natural and spoofed speech signals from 15 male and 20 female speakers. Total number of natural signals in development set is 3497 and the total number of spoofed signals is 49875. Each of the synthetic signals in the development set is originated from one of the same five spoofing algorithms used to generate training set (S1-S5). The development set is used to optimize the spoofing detection systems.
- **Evaluation Set**: Evaluation set consists of 9404 natural and 184000 synthetic speech signals from 46 speakers (20 male and 26 female). Spoofed speech signals are generated using ten different SS and VC algorithms. The spoofing algorithms include the same five SS/VC techniques that take part in training and development sets (S1-S5) referred to as *known attacks* since the same algorithms used to generate the spoofed speech signals in training set and additional five spoofing algorithms (S6-S10) which are referred to as *unknown attacks*. S6, S7, S8 and S9 are all the VC algorithms using STRAIGHT vocoder [12] whereas S10 is the unit selection based speech synthesis technique implemented with MARY Text-to-Speech Synthesis system[1] that does not use any vocoder for waveform generation. The evaluation subset is used to evaluate the final spoofing detection performance of the system.

The reader is referred to [37, 38] for a detailed description of the ASVspoof 2015 database.

### 3.2 Feature extraction

In the experiments, standard Mel-frequency cepstral coefficients (MFCC) and cosine phase (CosPhase) features are used as baseline countermeasures for spoofing detection. MFCC features are extracted from 20 ms Hamming windowed frames in every 10 ms. Magnitude spectrum of each frame is computed using discrete Fourier transform (DFT). Magnitude

[1] http://mary.dfki.de/

spectrum is then processed through a 30 channel triangular filterbank spaced in Mel-scale. Discrete cosine transform of logarithmic filterbank outputs is used to obtain MFCC features. In the experiments, 60 dimensional MFCC feature vectors (the first 20 static coefficients, $c_1 - c_{20}$, with their first and second order derivatives) are used.

CosPhase features [35] in turn, are extracted from the DFT of Hamming windowed frames with the duration of 20 ms in every 10 ms. The phase spectrum of the windowed frame is first unwrapped. The cosine function of the unwrapped phase is computed to normalize the phase spectrum into the range $[-1, 1]$. The normalized phase spectrum is then converted to cepstral coefficients by discrete cosine transform. The first 20 coefficients (excluding $c_0$) are kept as the feature vectors.

Linear prediction (LP) based features are extracted using the Hamming windowed frames with the same frame length and frame shift durations as in MFCC and CosPhase features. The dimensionality of the feature vectors are fixed to 20. The number of feature coefficiens are selected based on the initial experiments and 20 coefficients ($c_1 - c_{20}$) were found to give the best performance. The prediction order in turn are optimized using development set and the effect of prediction order is analyzed in Section 4.1.

### 3.3 Classifier

Although, there are various classifiers that can be used for anti-spoofing, in [6], different classification techniques were compared for spoofing detection and Gaussian mixture model (GMM) trained using maximum likelihood (ML) criterion [27] was found to yield the best performance. Therefore, we use the GMM classifier for spoofing detection in the experiments.

GMMs consisting of 512 Gaussian components for each class (natural and spoofed), $\lambda_{\text{natural}}$ and $\lambda_{\text{spoofed}}$, are trained using maximum likelihood criterion with 5 expectation maximization iterations [27]. Given a feature vector sequence, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$, extracted from a test utterance and GMM models $\lambda_{\text{natural}}$ and $\lambda_{\text{spoofed}}$, the detection score is computed by the logarithmic likelihood ratio as:

$$\Lambda(\mathbf{Y}) = \log p(\mathbf{Y}|\lambda_{\text{natural}}) - \log p(\mathbf{Y}|\lambda_{\text{spoofed}}), \tag{4}$$

where, $\log p(\mathbf{Y}|\lambda)$ is the average log-likelihood of feature set $\mathbf{Y}$ given GMM model and it is computed by $\lambda$ [27]:

$$\log p(\mathbf{Y}|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log p(\mathbf{y}_t|\lambda). \tag{5}$$

### 3.4 Performance criterion

Equal error rate (EER) is used as the performance criterion. EER is the error rate at the decision threshold where false acceptance ($P_{\text{FA}}$) and false rejection ($P_{\text{FR}}$) rates are equal. $P_{\text{FA}}$ is the ratio of the number of spoofed trials determined as natural to the total number of spoofed trials. Similarly, $P_{\text{FR}}$ is the ratio of the number of natural/genuine trials detected as spoofed to the total number of genuine trials.

The EERs are estimated using the Bosaris toolkit[2] which computes the EER using the receiver operating characterisics convex hull (ROCCH). Although, the EERs for each

[2] https://sites.google.com/site/bosaristoolkit/

spoofing methods are estimated independently, the average EER computed over all spoofing algorithms is used as the performance criterion in the ASVspoof challenge [37]. In addition to EER values, detection error trade-off (DET) curves [20] are also shown in order to analyze the relation between the false alarm and miss detection rates.

## 4 Results

### 4.1 Results on development set

Experiments are first conducted on development set of ASVspoof 2015 database for parameter optimization. We first study effect of the prediction order, $p$, on the spoofing detection performance. The average EERs (%) for different values of $p$ using three LP residual based features are shown in Fig. 3. From the figure, increasing the prediction order yields higher EERs for LPRHEC and LPResPhase features. However, for LPRPC features, EER considerably reduces as $p$ increases. The optimum values of prediction orders are determined as $p = 4$ for LPRHEC and LPResPhase features and $p = 28$ for LPRPC features. The EERs of 0.257%, 3.565% and 0.007% are obtained with LPRHEC, LPResPhase and LPRPC features, respectively. Comparing the proposed features on development set indicates that LPRPC is superior to LPRHEC and LPResPhase features.

Next, we analyze the effect of voice activity detection (VAD) on spoofing detection. To this end, we applied energy based VAD [14] to detect and drop non-speech frames. Results are summarized in Table 1. Similar to observations reported in previous studies [16, 28], removing the non-speech frames significantly reduces the spoofing detection performance independent of the features. This reveals that SS and VC algorithms used to generate spoofed signals may not model the non-speech portions of the signal well. Therefore, non-speech portions contain relevant information for discriminating natural signals from spoofed speech.

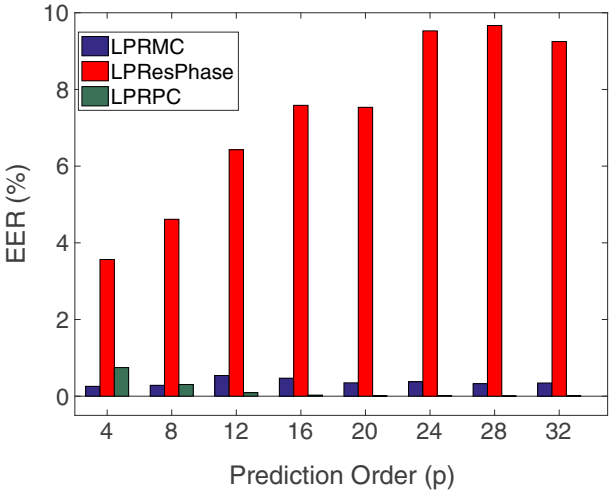


**Fig. 3** Average EERs (in %) for different prediction order on development set

**Table 1** Average EERs (%) with and without VAD on development set

| VAD | LPRHEC | LPResPhase | LPRPC |
|---|---|---|---|
| ✓ | 1.015 | 8.619 | 0.009 |
| ✗ | 0.257 | 3.565 | 0.007 |

We next study how appending the first and second order derivatives ($\Delta$ and $\Delta\Delta$), also known as dynamic features, to the static features affect the spoofing detection accuracies on development set. Table 2 shows the performance for seven different combinations of static ($S$) and dynamic ($\Delta$ and $\Delta\Delta$) features. First, in contrast to previous studies [28, 30] where dynamic features were found to outperform static coefficients for magnitude spectrum based features, the static features gives better performance than dynamic features for LP residual based features. This observation indicates that the spoofing algorithms in ASVspoof database may not model the dynamic information conveyed in LP residual signal well in contrast to magnitude spectrum based features. Second, appending the dynamic coefficients to the static features increases the average EERs in general except for LPRHEC features. For LPRHEC features, the first order derivative coefficients appended to static features ($S+\Delta$) yield approximately 67% performance improvement over the static features (EER reduces from 0.257% to 0.102%). However, appending the second order derivatives does not bring any performance improvement over the $S + \Delta$ combination.

In the last set of experiments on development set, we compare the results obtained using LP residual features with the baseline MFCC and CosPhase features. The EERs for each individual attack in development set and their average values obtained using different features are given in Table 3. First, irrespective of the features, speech synthesis attacks (S3 and S4) are easier to detect than voice conversion attacks (S1, S2 and S5). Second, the EERs for synthesis attacks (S3 and S4) are similar to each other. This is because of the fact that S3 and S4 are both hidden Markov model (HMM) based speech synthesis with different number of training utterances. Thus, both are in fact the same spoofing attack. Independent of the spoofing algorithms, LPRHEC and LPRPC show superior performance in comparison to the baseline MFCC and CosPhase features. LPResPhase yields the highest average EER among the five feature extraction techniques. However, they show encouraging performance against speech synthesis attacks (S3 and S4) and outperforms CosPhase features for S3 and S4 attacks. LPRHEC and LPRPC are the best two methods in terms of average EERs among the five techniques.

**Table 2** Average EERs (%) for different combinations of static and dynamic features on ASVspoof 2015 development set

| Features | LPRHEC | LPResPhase | LPRPC |
|---|---|---|---|
| $S$ | 0.257 | **3.565** | **0.007** |
| $\Delta$ | 0.346 | 10.920 | 0.681 |
| $\Delta\Delta$ | 0.469 | 10.794 | 0.801 |
| $S + \Delta$ | **0.102** | 5.191 | 0.028 |
| $S + \Delta\Delta$ | 0.145 | 5.505 | 0.038 |
| $\Delta + \Delta\Delta$ | 0.220 | 10.333 | 0.458 |
| $S + \Delta + \Delta\Delta$ | 0.173 | 5.510 | 0.057 |

$S$: static, $\Delta$: First order derivative, $\Delta\Delta$: second order derivative coefficients

**Table 3** EERs (%) for each spoofing attack and their average EERs using different feature extraction methods on development set

| Features | S1 | S2 | S3 | S4 | S5 | Average |
|---|---|---|---|---|---|---|
| MFCC | 0.157 | 4.232 | 0.000 | 0.000 | 2.027 | 1.283 |
| CosPhase | 0.170 | 0.985 | 0.237 | 0.219 | 2.700 | 0.862 |
| LPRHEC | 0.041 | 0.363 | 0.000 | 0.000 | 0.107 | 0.102 |
| LPResPhase | 3.110 | 11.822 | 0.029 | 0.050 | 2.814 | 3.565 |
| LPRPC | **0.000** | **0.017** | 0.000 | 0.000 | **0.018** | **0.007** |

## 4.2 Results on evaluation set

In the light of the observations on development set, we carry out the experiments on evaluation set with optimum set of parameters. These are: (i) predictor order is set to $p = 4$ for LPRHEC and LPResPhase and $p = 28$ for LPRPC features. (ii) all the features are extracted from the whole signal using both speech and non-speech frames (without any VAD), (iii) only static coefficients are used for LPResPhase and LPRPC features whereas, static coefficients and their first order derivatives ($S + \Delta$) are used as feature vectors for LPRHEC features. Table 4 shows performance of different features for each spoofing attack in evaluation set. In table, the smallest EER of each spoofing attack (each column) is bolded and the second smallest values are underlined.

First, similar to observations on development set, LPRPC features show excellent performance in comparison to standard MFCC and CosPhase countermeasures on known attacks (S1-S5). Similarly, the LPRHEC is the second best method among the five feature extraction techniques. Voice conversion (VC) attacks (S2 and S5) are the most difficult attack type to detect independent of the features. However, the proposed LPRHEC and LPRPC features yield the EERs lower than 0.5% on detecting VC attacks. LPRHEC shows approximately 90% and 97% better performance than the MFCC features on detecting S2 and S5 attacks, respectively. The LPRPC features in turn, yields 97% and 99% relative improvement over MFCC for S2 and S5 attacks, respectively. This implies that LP residual features capture relavent information for discriminating natural speech from spoofed speech originated by VC techniques. In contrast to VC attacks, speech synthesis attacks (S1, S3 and S4) are easier to detect and almost every feature set gives favourable results except for LPResPhase. However, the proposed LPRHEC and LPRPC features yield much lower EERs than MFCC and CosPhase features for S1 attack. The LPRHEC features show approximately six times better performance than the MFCC (EERs of 0.075% vs. 0.012%) and LPRPC gives fifteen times lower EER than MFCC features. S3 and S4 attacks in turn, are detected perfectly in general.

**Table 4** EERs (%) for each spoofing attack on evaluation set

| Features | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 |
|---|---|---|---|---|---|---|---|---|---|---|
| MFCC | 0.075 | 3.090 | 0.000 | 0.000 | 1.579 | 1.507 | <u>0.259</u> | 0.000 | 0.334 | **18.927** |
| CosPhase | 0.083 | 0.686 | 0.064 | 0.064 | 2.041 | 2.832 | **0.138** | 0.326 | 0.332 | 34.748 |
| LPRHEC | <u>0.012</u> | <u>0.298</u> | <u>0.000</u> | <u>0.000</u> | <u>0.043</u> | <u>0.749</u> | 0.951 | <u>0.080</u> | <u>0.183</u> | 30.613 |
| LPResPhase | 2.591 | 11.778 | 0.029 | 0.025 | 2.672 | 11.158 | 14.749 | 14.772 | 15.965 | <u>27.033</u> |
| LPRPC | **0.005** | **0.071** | **0.000** | **0.000** | **0.010** | **0.073** | 2.313 | **0.000** | **0.077** | 49.948 |

For unknown attacks (S6-S10), spoofing detection performance degrades considerably in comparison to known attacks independent of the features. Interestingly, in contrast to case of known attacks, better performance is achieved on VC attacks (S6, S7, S8 and S9) in comparison to SS attack (S10) for unknown attacks. The proposed LP residual features give encouraging and mostly better performance than MFCC and CosPhase features on VC attacks. For example, 1.507% EER is obtained for S6 attack using MFCC features whereas EER reduces to 0.073% with LPRPC features. However, the performance of S10 attack, the speech synthesis algorithm based on unit selection, significantly differs from VC algorithms and varies considerably. The EERs of S10 attack are much higher than that of other attacks. This is possibly because of the fact that S10 attack does not utilize any vocoder to generate waveform but most of the spoofed speech signals used to train spoof model in training set are generated by the spoofing algorithms utilizing STRAIGHT vocoder. Therefore, this induces a vocoder mismatch between the training set and S10 attack.

Figure 4 shows the DET curves for different features on evaluation set. In contrast to EER, DET curves help to see the performance of a detection system by showing the full trade-off curve between the false alarm and the miss rates. Note that all the scores for each spoofing attack are pooled together while producing the DET curves, although the evaluation metric is the EER for each individual attack. Since the performance of the evaluation set is highly dependent on S10 attack and it yields much higher EER than other nine attacks (S1-S9), the S10 attack is excluded while generating the DET curves. From the figure, proposed LP residual features show better anti-spoofing performance than MFCC and CosPhase features in general. The smallest EER of 0.392% is obtained with LPRHEC features. Similar to previous results, LPResPhase gives the highest EER.

Finally, we compare the results obtained in this study with the results reported in other studies utilizing various LP based features [1, 11]. The results are compared in terms of average of known and unknown attacks seperately and the average of all spoofing attacks on

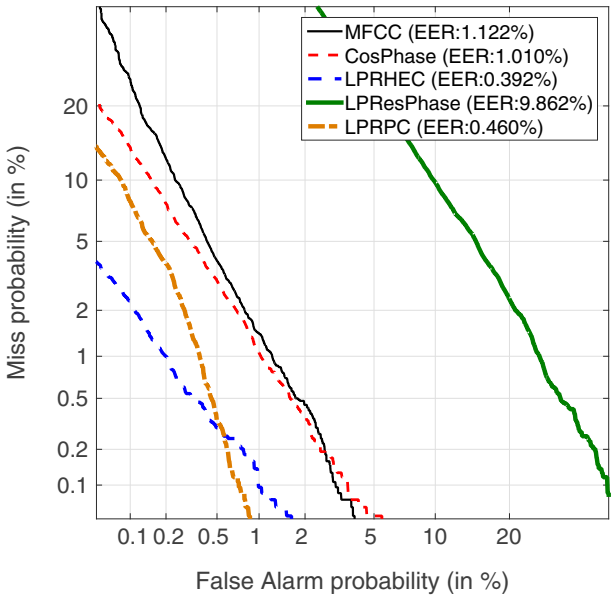


**Fig. 4** DET curves for different features on evaluation set

**Table 5** Comparison of the proposed features (in terms of EER) with the recently proposed LP-LTP [11] and LP-NLP [1] features on evaluation set

| Features | Known (S1-S5) | Unknown (S6-S10) | Average (S1-S10) |
|---|---|---|---|
| LP-LTP [1] | 2.270 | 19.540 | 10.910 |
| LP-NLP [1] | 4.890 | 23.400 | 14.150 |
| LP-LTP [11] | 1.217 | 10.206 | 6.611 |
| LPRHEC | 0.070 | **6.515** | **3.292** |
| LPResPhase | 3.419 | 16.735 | 10.077 |
| LPRPC | **0.017** | 10.482 | 5.249 |

evaluation set. Table 5 summarizes the average EERs for known, unknown and all attacks obtained in this study and other recent studies for comparison. The results given in the table indicates that the proposed LP residual features yield better performance than LP-LTP and LP-NLP methods for both known and unknown attacks.

## 5 Conclusion

In this study, we proposed to use linear prediction (LP) residual based features for automatic speaker verification (ASV) anti-spoofing. Unlike the previous studies where LP residual is estimated from LP followed by long-term prediction (LTP) [10, 11] and non-linear prediction (NLP) [1], we used analytic signal obtained from the LP residual signal for deriving the proposed features. The phase function of the analytic signal computed from LP residual signal (LPResPhase) and two proposed features so-called LP residual Hilbert envelope cepstral coefficiens (LPRHEC) and LP residual phase cepstral coefficients (LPRPC) features were compared for anti-spoofing. First of all, the results indicated that proposed LP residual features show promising results on spoofing detection in comparison to MFCC and CosPhase features.

From the results on development set, the prediction order of $p = 4$ was found to be the optimum value for LPRHEC and LPResPhase features whereas $p = 28$ is the best value for LPRPC features. Similar to previous studies on ASVspoof 2015 database, applying voice activity detection (VAD) increased the EER for all types of features used in this study. Static features were found to give the best performance in comparison to dynamic features and their combinations. Among the three LP residual based features, LPResPhase gave the highest EER.

For evaluation set, as previously shown in other studies, S10 was found to be the most difficult attack type for detection. LPRPC features yields the smallest EERs for eight spoofing attacks among the ten spoofing methods. For S10, MFCC and LPResPhase features are the best two methods in terms of EER. The proposed features showed better performance than recently proposed LP-LTP [11] and LP-NLP methods [1].

Since the proposed feature extraction techniques require only determining the LP coefficients which is computationally inexpensive [25], the proposed methods can easily be adapted to use in real time applications using mobile devices. The application of the proposed feature extraction techniques based on LP analysis to the replay attack detection would be interesting for future studies.

**Acknowledgements** This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) (project #115E916).

## References

1. Bhavsar HN, Patel TB, Patil HA (2016) Novel nonlinear prediction based features for spoofed speech detection. In: Proceedings of INTERSPEECH, pp 155–159
2. Bonastre J, Matrouf D, Fredouille C (2007) Artificial impostor voice transformation effects on false acceptance rates. In: Proceedings of INTERSPEECH, pp 2053–2056
3. Ergünay SK, Khoury E, Lazaridis A, Marcel S (2015) On the vulnerability of speaker verification to realistic voice spoofing. In: Proceedings of BTAS, pp 1–6
4. Evans NWD, Kinnunen T, Yamagishi J, Wu Z, Alegre F, Leon PLD (2014) Speaker recognition anti-spoofing. In: Handbook of biometric anti-spoofing - trusted biometrics under spoofing attacks, pp 125–146
5. Fukada T, Tokuda K, Kobayashi T, Imai S (1992) An adaptive algorithm for mel-cepstral analysis of speech. In: Proceedings of ICASSP, vol 1, pp 137–140
6. Hanilçi C, Kinnunen T, Sahidullah M (2015) Classifiers for synthetic speech detection: a comparison. In: Proceedings of INTERSPEECH, pp 2057–2061
7. Hautamàki RG, Kinnunen T, Hautamàki V, Leino T, Laukkanen A (2013) I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In: Proceedings of INTERSPEECH, pp 930–934
8. Hautamàki RG, Kinnunen T, Hautamàki V, Leino T, Laukkanen A (2015) Automatic versus human speaker verification: the case of voice mimicry. Speech Comm 72:13–31
9. Jain AK, Ross A, Pankanti S (2006) Biometrics: a tool for information security. IEEE, Transactions on Information Forensics and Security 1(2):125–143
10. Janicki A (2015) Spoofing countermeasure based on analysis of linear prediction error. In: Proceedings of INTERSPEECH, pp 2077–2081
11. Janicki A (2017) Increasing anti-spoofing protection in speaker verification using linear prediction. Multimedia Tools and Applications 76(6):9017–9032
12. Kawahara H, Masuda-Katsuse I, de Cheveigne A (1999) Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based {F0} extraction: Possible role of a repetitive structure in sounds. Speech Communication 27(3–4):187–207
13. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. Speech Comm 52(1):12–40
14. Kinnunen T, Saastamoinen J, Hautamàki V, Vinni M, Fränti P (2009) Comparative evaluation of maximum a posteriori vector quantization and gaussian mixture models in speaker verification. Pattern Recogn Lett 30(4):341–347
15. Kinnunen T, Wu ZZ, Lee KA, Sedlak F, Chng ES, Li H (2012) Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech. In: Proceedings of ICASSP, pp 4401–4404
16. Lavrentyeva G, Novoselov S, Simonchik K (2017) Anti-spoofing methods for automatic speaker verification system. Springer International Publishing, Cham, pp 172–184
17. Leon PLD, Apsingekar VR, Pucher M, Yamagishi J (2010) Revisiting the security of speaker verification systems against imposture using synthetic speech. In: Proceedings of ICASSP, pp 1798–1801
18. Leon PLD, Pucher M, Yamagishi J (2010) Evaluation of the vulnerability of speaker verification to synthetic speech. In: Proceedings of Odyssey, p 28
19. Makhoul J (1975) Linear prediction: a tutorial review. Proc IEEE 63(4):561–580
20. Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M (1997) The det curve in assessment of detection task performance. In: Proceedings of EUROSPEECH, pp 1895–1898
21. Murty KSR, Yegnanarayana B (2006) Combining evidence from residual phase and mfcc features for speaker recognition. IEEE Signal Process Lett 13(1):52–55
22. Nandi D, Pati D, Rao KS (2017) Implicit processing of lp residual for language identification. Comput Speech Lang 41(C):68–87
23. Novoselov S, Kozlov A, Lavrentyeva G, Simonchik K, Shchemelinin V (2016) STC anti-spoofing systems for the ASVspoof 2015 challenge, pp 5475–5479
24. Pati D, Prasanna SRM (2011) Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information. Int J Speech Technol 14(1):49–64

25. Rabiner L, Schafer R (2010) Theory and applications of digital speech processing, 1st edn. Prentice Hall Press, Upper Saddle River
26. Ratha NK, Connell JH, Bolle RM (2001) Enhancing security and privacy in biometrics-based authentication systems. IBM, Syst J 40(3):614–634
27. Reynolds DA, Rose RC (1995) Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing 3(1):72–83
28. Sahidullah M, Kinnunen T, Hanilçi C (2015) A comparison of features for synthetic speech detection. In: Proceedings of INTERSPEECH, pp 2087–2091
29. Sânchez J, Saratxaga I, Hernàez I, Navas E, Erro D, Raitio T (2015) Toward a universal synthetic speech spoofing detection using phase information. IEEE, Transactions on Information Forensics and Security 10(4):810–820
30. Todisco M, Delgado H, Evans N (2016) A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients. In: Proceedings of Odyssey, pp 283–290
31. Villalba JA, Lleida E (2010) Speaker verification performance degradation against spoofing and tampering attacks. In: Proceedings of FALA, pp 131–134
32. Villalba JA, Miguel A, Ortega A, Lleida E (2015) Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge. In: Proceedings of INTERSPEECH, pp 2067–2071
33. Wang L, Yoshida Y, Kawakami Y, Nakagawa S (2015) Relative phase information for detecting human speech and spoofed speech. In: Proceedings of INTERSPEECH, pp 2092–2096
34. Wu Z, Li H (2014) Voice conversion versus speaker verification: an overview. APSIPA Transactions on Audio Signal and Information Processing 3(e17)
35. Wu Z, Siong CE, Li H (2012) Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: Proceedings of INTERSPEECH, pp 1700–1703
36. Wu Z, Evans N, Kinnunen T, Yamagishi J, Alegre F, Li H (2015) Spoofing and countermeasures for speaker verification: a survey. Speech Comm 66:130–153
37. Wu Z, Kinnunen T, Evans N, Yamagishi J, Hanilçi C, Sahidullah M, Sizov A (2015) ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In: Proceedings of INTERSPEECH, pp 2037–2041
38. Wu Z, Yamagishi J, Kinnunen T, Hanilçi C, Sahidullah M, Sizov A, Evans N, Todisco M (2017) Asvspoof: The automatic speaker verification spoofing and countermeasures challenge. IEEE Journal of Selected Topics in Signal Processing 11(4):588–604
39. Xiao X, Tian X, Du S, Xu H, Chng ES, Li H (2015) Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge. In: Proceedings of INTERSPEECH, pp 2052–2056

**Cemal Hanilçi** received B.Sc., M.Sc. and Ph.D. degrees from Uludağ University in 2005, 2007 and 2013, respectively, all in Electronic Engineering. From March to December 2011, he was a visiting researcher at the School of Computing, University of Eastern Finland. From 2014 to 2015, he was a post-doctoral researcher at the same school. Currently he is an assistant professor at the Bursa Technical University, Department of Electrical & Electronic Engineering, in Turkey. His research interests include speaker recognition, anti-spoofing, audio forensics.