Speaker Verification Anti-Spoofing Using Linear Prediction Residual Phase Features

Cemal Hanilçi

Department of Electrical-Electronics Engineering
Bursa Technical University, Bursa, Turkey
Email: cemal.hanilci@btu.edu.tr

Abstract—The vulnerability of automatic speaker verification (ASV) systems against spoofing attacks is an important security concern about the reliability of ASV technology. Recently, various countermeasures have been developed for spoofing detection. In this paper, we propose to use features derived from linear prediction (LP) residual signal for spoofing detection using simple Gaussian mixture model (GMM) classifier. Experiments conducted on recently released ASVspoof 2015 database show that LP residual phase cepstral coefficients (LPRPC) outperforms standard MFCC features and considerably improves the spoofing detection performance. With the LPRPC features 97% relative improvement is observed over standard MFCC features on known attacks.

I. INTRODUCTION

Automatic speaker verification (ASV) is the task of accepting or rejecting an identity claim given a speech signal [1]. Recent developments on ASV technology which yields to achieve low equal error rates (EER), has led to an increasing potential use of ASV systems in real case scenarios such as online banking and call centers thereby triggering the adoption by the mass-market. However, as in the case for other biometric modalities (e.g. face and fingerprint), spoofing attacks are one of the most important security concern for ASV systems [2], [3]. With spoofing attack (also known as presentation attack), an attacker aims to gain illegitimate access to the system by presenting a forged biometric data at the sensor level (e.g. camera and microphone) [2] and the vulnerability of ASV systems against spoofing attacks have been confirmed independently in many studies [4], [5], [6], [7].

For the ASV systems, *impersonation* [8], *replay* [9], *speech synthesis* (SS) [6] and *voice conversion* (VC) [10] are the four major direct spoofing attack types against ASV systems [11]. Among these four attack types, impersonation is less likely since it requires a professional skill to mimic a target speaker's voice. Replay attacks in turn, are the most likely attack type because it is caused by presenting a pre-recorded speech signal of the target speaker. The SS refers to synthesize target speaker's voice given a text input whereas VC is the modification of source speaker's (attacker's) voice towards that of target speaker's voice. SS and VC are easily accessible and important attack types due to two main reasons. First, there exists freely available open source toolkits that can be used by non-experts without any background information on SS and VC. Second, state-of-the-art SS and VC techniques produce speech signals of high quality even using small amount of training data. Therefore, SS and VC attacks are potential threats for falsifying ASV systems. For a detailed review and general information on spoofing attacks against ASV systems, the reader is refered to [11].

Spoofing countermeasures, determining whether a speech signal is natural or spoofed, play an important role to cope with spoofing attacks against ASV systems. Detection of speech synthesis (SS) and voice conversion (VC) attacks have gained great interest by the community due to recently organized *ASVspoof Automatic Speaker Verification Spoofing and Countermeasures Challenge* (ASVspoof) [12]. In the challenge, a dataset consisting of natural and spoofed speech signals has been generated by various SS and VC techniques. One of the aims of the challenge was to develop a common dataset and to define a standard evaluation metric for stand-alone spoofing detection. The evaluation dataset in ASVspoof 2015 challenge, contained spoofed speech utterances generated with 10 different SS and VC techniques and developing a generalized countermeasure to detect both known and unknown attacks was another aim of the challenge.

Various countermeasures were proposed for the ASVspoof 2015 challenge dataset for spoofing detection with varying performance [12]. From the evaluation results, phase based features were found to outperform magnitude features in general [12]. For example in [13], amplitude and phase based features were compared for spoofing detection and simple cosine phase features [14] were found to outperform standard amplitude based Mel-frequency cepstral coefficients (MFCC). Seven different (two magnitude and five phase-based) features extraction techniques were compared in [15] and group delay features were reported to give smallest EER on development set of ASVspoof challenge data.

In [16], linear prediction (LP) residual signal obtained from the LP analysis followed by long-term prediction (LTP) block is used to extract audio quality based features (e.g. mean energy of LP residual, maximum energy of LTP residual, mean and maximum of LTP gain) for spoofing detection and it was shown that proposed features yield encouraging results.

Inspired from the success of promising results reported in [16], in this work, we propose to use phase based features extracted from LP residual of the speech signal for spoofing detection. Since the LP residual signal conveys information about the excitation source, intiutively one would expect that

it may capture relevant information to discriminate natural speech from spoofed speech. To this end, besides phase based features, magnitude features are also employed for comparison. To compare the performance of the proposed features, well known features used in spoofing detection, MFCC and cosine phase (CosPhase), are selected as baseline countermeasures.

II. SPOOFING DETECTION

In spoofing detection, we use Gaussian mixture model (GMM) classifier which has been succesfully used for speaker recognition [17] and spoofing detection [18]. In GMM, each class is represented by a weighted sum of M Gaussian densities, $p(\mathbf{x}|\lambda) = \sum_{m=1}^{M} w_m p_m(\mathbf{x})$, where w_i is the mixture weight of the ith Gaussian component and $p_i(\mathbf{x})$ is the multivariate Gaussian density function. Since spoofing detection is a binary classification task, one GMM is trained for natural class using the natural training utterances and another is trained using the training utterances of spoofed class. λ_{natural} and λ_{spoofed} denote the GMMs for natural and spoofed classes, respectively. GMM is trained with expectation maximization (EM) algorithm using maximum likelihood (ML) criterion [17].

During the test phase, features, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, are extracted from the test speech signal and then logarithmic likelihood is computed using the GMM of each class:

$$\mathcal{L}(\mathbf{X}|\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(\mathbf{x}_t|\lambda). \tag{1}$$

Finally, the log-likelihood ratio (LLR) score is computed as the detection score and it is defines as:

$$\Lambda(\mathbf{X}) = \mathcal{L}(\mathbf{X}|\lambda_{\text{natural}}) - \mathcal{L}(\mathbf{X}|\lambda_{\text{spoofed}}). \tag{2}$$

In the experiments, GMMs for natural and spoof classes consisting of 512 Gaussians are trained with 5 EM iterations.

III. LINEAR PREDICTION RESIDUAL BASED FEATURES

Linear prediction (LP) analysis assumes that a speech sample, $x[n]$, can be estimated as a weighted sum of its p previous samples, $\hat{x}[n] = -\sum_{k=1}^{p} \alpha_k x[n-k]$ [19]. Here $x[n]$ is the original speech sample, $\hat{x}[n]$ is its predicted counterpart, p is the predictor order and $\{\alpha_k\}_{k=1}^{p}$ are the predictor coefficients. LP residual (prediction error) is defined as the difference between the actual speech sample $x[n]$ and the predicted sample $\hat{x}[n]$

$$e[n] = x[n] - \hat{x}[n] = x[n] + \sum_{k=1}^{p} \alpha_k x[n-k] \tag{3}$$

Previously, it was shown that LP residual signal, $e[n]$, contain relevant information for speaker recognition [20]. Since LP residual convey information about the excitation source, the input of the speech production system, by intuition, the features derived from LP residual may convey useful information for spoofing detection. Thus, we extract features from LP residual signal for spoofing detection.

A. LP Residual Magnitude and Phase Features

Since the values of the LP residual signal are relatively large, it is difficult to extract useful information from short-term analysis. Therefore, we extract the residual magnitude features from the analytic signal derived from the LP residual signal [20], [21]:

$$e_a[n] = e[n] + je_h[n] \tag{4}$$

where $e_h[n]$ is the Hilbert transform of the $e[n]$. LP residual magnitude cepstral coefficients (LPRMC) are obtained by applying discrete cosine transform (DCT) to the logarithm of the magnitude of the analytic signal given in (4).

LP residual phase features (LPRP) are defined as the cosine of the analytic signal phase function

$$\cos(\theta[n]) = \frac{e[n]}{\sqrt{e^2[n] + e_h^2[n]}} \tag{5}$$

LPRMC and LPRP features were previously used in different recognition tasks based on speech signals such as speaker and language recognition [20], [21], [22]. In addition to these two known feature sets, we propose to use a modified form of LP residual phase which we refer to as LP residual phase cepstral coefficients (LPRPC). LPRPC features are obtained by applying discrete cosine transform to the LP residual phase function given in (5). The extraction process of the features derived from the LP residual signal is summarized in Fig. 1.

IV. EXPERIMENTAL SETUP

A. Database

ASVspoof 2015 database [12] consisting of natural and spoofed speech signals generated by various speech synthesis (SS) and voice conversion (VC) algorithms is used in the experiments. ASVspoof 2015 database is composed from three disjoint subsets: *training*, *development* and *evaluation*:

- **Training set** consists of 3750 natural and 12625 spoofed speech signals. Spoofed signals are generated using three VC algorithms (S1, S2 and S5) and two SS techniques (S3 and S4). Training set is used to train natural and spoof acoustic models for classifier.
- **Development set** includes both natural and spoofed signals from 35 speakers (15 male and 20 female). Spoofed signals are originated from the same five spoofing techniques (S1-S5) used to generate the training set. The development set is used for parameter tuning and optimisation of developed countermeasures.
- **Evaluation set** includes 9404 natural and 184000 spoofed utterances from 46 speakers (20 male and 26 female). Spoofed signals generated using the same five techniques (S1-S5) that take part in training and development sets which are referred to as *known* attacks and five additional spoofing algorithms referred to as *unknown* attacks (S6-S10).

Since evaluation set consists of both known and unknown attacks, it is used to test the generalization capability of developed countermeasures. Known attacks (S1-S5) are expected to
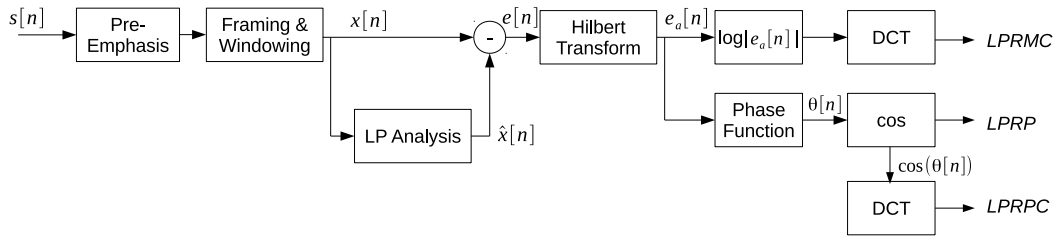
Fig. 1. Block diagram of LP residual based feature extraction techniques used in this work.

yield better performance than the unknown attacks since the same techniques are used to train the classifier.

B. Features

In the experiments, we use the MFCC features as the baseline system. MFCC features are extracted from 20 ms frames in every 10 ms. Power spectrum computed using discrete Fourier transform (DFT) of Hamming windowed speech frames are processed through a 27-channel triangular filterbank. MFCC features are obtained by applying discrete cosine transform (DCT) to the logarithmic filterbank outputs. The first 20 MFCC coefficients ($c_0 - c_{19}$) with their first and second order derivatives (Δ and $\Delta\Delta$) yielding a total of 60 dimensional feature vectors are used in the experiments.

Cosine phase (CosPhase) features [14] are used as the second baseline countermeasure. They are extracted from the DFT phase spectrum of Hamming windowed speech frames. Cosine function is applied to the unwrapped phase spectrum to normalize the unwrapped phase. Then DCT is applied on normalized phase and the first 20 coefficients are retained.

The LP residual features are extracted from the speech frames with the same duration and frame shift lengths used to extract MFCC and CosPhase features. The predictor order is fixed to $p = 24$ and the first 20 coefficients are used as the features.

C. Performance Criterion

Equal error rate (EER) is used as the performance criterion of spoofing detection. EER is the threshold value that false acceptance rate (P_{fa}) and miss rate (P_{miss}) are equal. P_{fa} is the ratio of number of spoofed trials classified as natural to the total number of spoofed trials. P_{miss} in turn, is the ratio of number of natural trials classified as spoofed to the total number of natural trials. As suggested in the ASVspoof 2015 challenge evaluation plan [12], EERs are computed using the BOSARIS toolkit[1].

Apart from reporting the EER values of each individual feature set, in order to find out whether the proposed features contain complementary information over the baseline MFCC and CosPhase features, two different score fusion strategy is considered in this study: (i) **Fusion 1:** is the score fusion based on the logistic regression where fusion weights are trained using the BOSARIS toolkit. Here, we use the development

data to train the fusion paremeters. (ii) **Fusion 2:** is the simple score averaging technique.

V. RESULTS

In the experiments, we first report the results obtained on the development set of ASVspoof 2015 database. The EERs for each individual attack in development set (S1-S5) obtained with different features are summarized in Table I. In the table, the best numbers for each attack are shown in boldface. It can be seen that LP residual magnitude and phase cepstrum features (LPRMC and LPRPC) yield considerably better performance than standard MFCC features independent from the attack type. They outperform CosPhase features, as well, except for the S2 attack. CosPhase features show better performance than LP residual features in detecting S2 attacks. However, LPRMC and LPRPC are superior to CosPhase features in terms of average EER. From the table, LPRP features yield relatively high EERs in comparison to other four feature types. This is possibly because in LPRP, we use the first 20 raw phase values as the features which are known to be correlated and dependent. However applying DCT, results uncorrelated feature coefficients and boosts the spoofing detection performance considerably (average EER reduces from 9.258% to 0.016%). The best performance is achieved with the LPRPC features on development set which implies that LP residual phase based features are potentially a good candidate for spoofing detection.

TABLE I
RESULTS ON DEVELOPMENT SET.

Features	S1	S2	S3	S4	S5	Avg.
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
LPRMC	0.136	1.220	0.000	0.000	0.532	0.377
LPRP	9.330	19.017	0.024	0.037	19.234	9.528
LPRPC	**0.007**	**0.040**	**0.000**	**0.000**	**0.033**	**0.016**
Fusion 1	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Fusion 2	0.007	0.188	0.000	0.000	0.153	0.069

From the last two rows of the Table I, logistic regression based score fusion considerably improves the spoofing detection performance in comparison to best performing system (LPRPC). Score averaging technique in turn, does not bring any performance improvement but slightly increases the EER.

TABLE II
EERs (%) FOR EACH INDIVIDUAL ATTACK ON EVALUATION SET.

Features	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
MFCC	0.075	3.090	0.000	0.000	1.579	1.507	**0.259**	0.000	0.334	**18.927**	**2.577**
CosPhase	0.083	0.686	0.064	0.064	2.041	2.832	0.138	0.326	0.332	34.748	4.131
LPRMC	0.049	1.072	0.000	0.000	0.424	1.283	2.584	0.088	0.968	30.089	3.656
LPRP	8.537	16.983	0.016	0.025	17.398	18.116	17.021	7.973	15.293	29.312	13.067
LPRPC	**0.006**	**0.070**	**0.000**	**0.000**	**0.021**	**0.111**	2.439	**0.000**	**0.062**	49.931	5.264
Fusion 1	**0.000**	**0.017**	**0.000**	**0.000**	**0.003**	**0.046**	**0.236**	**0.000**	**0.005**	40.711	4.102
Fusion 2	0.000	0.256	0.000	0.000	0.101	0.235	0.113	0.000	0.037	**25.363**	**2.610**

Next, we study the spoofing detection performance of the proposed features on evaluation set. The EERs for each individual spoofing attack technique are given in Table II. Independent of the features, S10 attack, unit selection based speech synthesis technique, is the most difficult attack type to detect in comparison to remaining nine techniques. MFCC features yield the smallest EER for S10 attack. Similar to the results on development set, LPRPC features are superior to other features for detecting the eight attacks among the ten techniques. For S7 and S10 attacks, MFCC features show better performance than other features. Since the EERs of S10 attack are much higher than that of other spoofing techniques, the average EERs (the last column in Table II) become highly dependent on the performance of detecting the S10 attack. Therefore, we summarize the results on evaluation set in Table III in a different way. In the table, we report the average EERs for known and unknown attacks, seperately. While computing the average EER of unknown attacks, we excluded S10 and report the EER of S10 attack separately.

As in the case for development set, LP residual phase cepstrum (LPRPC) features yields the smallest average EER for known attacks on evaluation set. LPRPC gives approximately 97% and 96% better performance than MFCC (EERs of 0.019% vs. 0.949%) and CosPhase features (EERs of 0.019% vs. 0.588%) on known attacks, respectively. However, on unknown attacks, MFCC features slightly outperform LPRPC. The performance differences between MFCC and LPRPC further increase on S10 attack. MFCC yields the smallest EER on S10 attack among the five feature extraction techniques. Comparing three LP residual variants, LPRPC features are superior to LPRP and LPRMC for known and unknown attacks. However, interestingly, LPRP gives the smallest EER on S10 attack among the three LP features.

Similar to the observations on development set, applying score fusion considerably reduces the EER of known and unknown attacks, in general. However, the EER of S10 attack after score fusion is not better than the best performing system (MFCCs) for both logistic regression and score averaging fusion strategies. For known and unknown attacks (except S10) Fusion 1 considerably outperforms Fusion 2. However, for S10 attack, the simple score averaging method, Fusion 2, yields approximately 38% smaller EER than Fusion 1 method. This is possibly because of the training the linear fusion weights.

TABLE III
RESULTS ON EVALUATION SET

Features	Known (S1-S5)	Unknown (S6-S9)	S10	Avg. (S1-S10)
MFCC	0.949	**0.525**	**18.927**	**2.577**
CosPhase	0.588	0.907	34.748	4.131
LPRMC	0.309	1.231	30.089	3.656
LPRP	8.592	14.601	29.312	13.067
LPRPC	**0.019**	0.653	49.931	5.264
Fusion 1	**0.004**	**0.072**	40.711	4.102
Fusion 2	0.071	0.096	**25.363**	**2.610**

Since the fusion weights for logistic regression based method (Fusion 1) were trained using the score files on development set, applying the same weights to the evaluation scores may fail to improve the performance because of the existence of unforeseen attacks appear in evaluation set.

Figure 2 shows the DET curves for MFCC and LPRPC features on evaluation set. Note that, although in the ASVspoof evaluation, the results are reported as the average EER over different attacks (Tables I and III), here we pooled the scores of nine attacks (S1-S9) to produce DET curves and to compute the EERs. Similar to previous results, we excluded the S10 attack while generating the DET curves. From the DET curves, it can be seen that MFCC features give almost two times higher EER than LPRPC features which shows the importance and the superior performance of the proposed features on spoofing detection.

VI. CONCLUSION

In this work, we proposed to use various features extracted from LP residual signal for spoofing detection and compared their performances with MFCC and simple but powerful CosPhase features. Experiments on ASVspoof 2015 challenge database revealed that phase features extracted from LP residual signal (LPRPC) conveys relevant information for spoofing detection. The results on development set of ASVspoof database showed that LPRPC features considerably improves the spoofing detection performance in comparison to standard MFCC and CosPhase features. Approximately 97% relative improvement was observed using LPRPC features over MFCC. Similarly, another type of LP residual features, LP
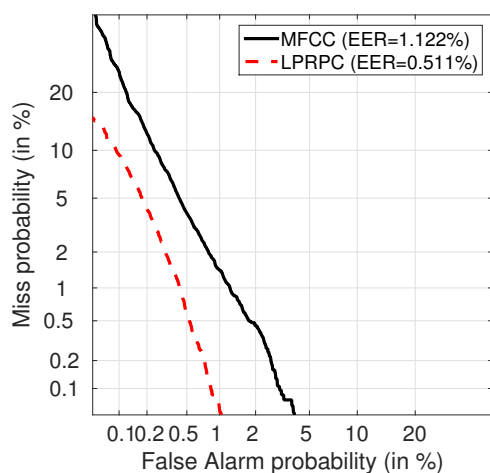
Fig. 2. DET plots for evaluation set. All the scores (except S10) in the evaluation set are pooled to produce DET curves.

residual magnitude cepstral features (LPRMC) were found to be superior to MFCC features on development set.

For the evaluation set, LPRPC yields better performance than other feature sets on unknown attacks. However, for S10 attack, the most difficult attack type in ASVspoof database [12], MFCC features yield better performance than LP residual phase features.

Studying the LP residual features for the case of replay and impersonation attacks against ASV systems would be interesting as a future work.

ACKNOWLEDGMENT

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010.

[2] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, March 2001.

[3] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: A tool for information security," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, November 2006.

[4] L. W. Chen, W. Guo, and L. R. Dai, "Speaker verification against synthetic speech," in *2010 7th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, November 2010, pp. 309–312.

[5] P. L. D. Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 1798–1801.

[6] P. L. D. Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. Odyssey 2010: The Speaker and Language Recognition Workshop*, 2010, p. 28.

[7] F. Alegre, R. Vipperla, N. W. D. Evans, and B. G. B. Fauve, "On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals," in *Proc. European Signal Processing Conference, EUSIPCO*, 2012, pp. 36–40.

[8] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13–31, 2015.

[9] A. Janicki, F. Alegre, and N. W. D. Evans, "An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks," *Security and Communication Networks*, vol. 9, no. 15, pp. 3030–3044, 2016.

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

[11] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification," *Speech Communication*, vol. 66, no. C, pp. 130–153, Feb. 2015.

[12] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH*, 2015, pp. 2037–2041.

[13] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5475–5479.

[14] Z. Wu, C. E. Siong, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. INTERSPEECH*, 2012, pp. 1700–1703.

[15] X. Xiao, X. Tian, S. Du, H. Xu, E. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, 2015, pp. 2052–2056.

[16] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Proc. INTERSPEECH*, 2015, pp. 2077–2081.

[17] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan 1995.

[18] C. Hanilçi, T. Kinnunen, and M. Sahidullah, "Classifiers for synthetic speech detection: A comparison," in *Proc. INTERSPEECH*, 2015, pp. 2057–2061.

[19] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[20] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, Jan 2006.

[21] D. Pati and S. R. M. Prasanna, "Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information," *International Journal of Speech Technology*, vol. 14, no. 1, pp. 49–64, 2011.

[22] D. Nandi, D. Pati, and K. S. Rao, "Implicit processing of LP residual for language identification," *Computer Speech and Language*, vol. 41, no. C, pp. 68–87, Jan. 2017.