INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Novel Nonlinear Prediction Based Features for Spoofed Speech Detection

Himanshu N. Bhavsar, Tanvina B. Patel and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)
Gandhinagar-382007, India
{himanshu_bhavsar, tanvina_bhupendrabhai_patel, hemant_patil}@daiict.ac.in

Abstract

Several speech synthesis and voice conversion techniques can easily generate or manipulate speech to deceive the speaker verification (SV) systems. Hence, there is a need to develop spoofing countermeasures to detect the human speech from spoofed speech. System-based features have been known to contribute significantly to this task. In this paper, we extend a recent study of Linear Prediction (LP) and Long-Term Prediction (LTP)-based features to LP and Nonlinear Prediction (NLP)-based features. To evaluate the effectiveness of the proposed countermeasure, we use the corpora provided at the ASVspoof 2015 challenge. A Gaussian Mixture Model (GMM)-based classifier is used and the % Equal Error Rate (EER) is used as a performance measure. On the development set, it is found that LP-LTP and LP-NLP features gave an average EER of *4.78* % and *9.18* %, respectively. Score-level fusion of LP-LTP (and LP-NLP) with Mel Frequency Cepstral Coefficients (MFCC) gave an EER of *0.8* % (and *1.37* %), respectively. After score-level fusion of LP-LTP, LP-NLP and MFCC features, the EER is significantly reduced to *0.57* %. The LP-LTP and LP-NLP features have found to work well even for Blizzard Challenge 2012 speech database.

Index Terms: Speaker verification, spoof detection, linear prediction, long-term prediction, nonlinear prediction.

1. Introduction

Recently, Automatic Speaker Verification systems (ASVs) are becoming more popular and are widely used. The main goal of ASV systems is to accept the claimed identity of the genuine speaker and reject the claim of an impostor. ASV systems are becoming very popular due to the less Equal Error Rate (EER) achieved. However, they are found to be highly vulnerable to spoofing attacks. These attacks can be due to impersonation, replay, speech synthesis (SS) and voice conversion (VC). Impersonation refers to the attacks using humans to alter their voice (such as mimicking) [1], [2]. Replay attack is caused by reusing pre-recorded speech signal of the genuine or target speaker [3]. The SS technique refers to text-to-speech (TTS) synthesis systems (generally Hidden Markov Model (HMM)-based TTS systems (HTS) and adapted HMM-based systems [4], [5]). The VC technique refers to modifying the source speaker's speech to make it sound-like the genuine target speaker [6], [7]. A review of various spoofing attacks and their countermeasures are provided in [8]. It has been known that SS and VC are easily accessible than impersonation and replay attacks. In addition, SS and VC speech can be generated for any speaker. Thus, we concentrate on SS and VC spoofing attacks. Recently, the ASVspoof challenge had been organized as a special session of INTERSPEECH 2015 [9]. The aim of the ASVspoof 2015 challenge was to design a robust detector which could classify natural and spoofed speech for both *known* and *unknown* attacks. At the ASVspoof 2015 challenge, various countermeasures were proposed which includes modified group delay [10], local binary patterns [11], relative phase shift (RPS) [12], wavelet-based features [13], cochlear filter cepstral coefficients and change in instantaneous frequency (CFCCIF) [14], phonetic-level phoneme posterior probability (PPP) and *i*-vector subsystem-based features [15]. As the SS and VC speech is synthesized using a vocoder, the phase information is lost in synthesized or voice converted speech. Therefore, many of these features use phase-based approaches which may not work for non-vocoder or unit-selection-based speech. In addition to system-based features, several excitation source-based features have been explored using the challenge database. Previous work on the excitation source-based features includes pitch pattern-based features [16], Linear Prediction (LP)-based features [17]- [18] and very recently proposed fundamental frequency (F_0) contour and its variations along with the strength of excitation (SoE) feature for vocoded speech detection in [19].

Our present work is directed towards using excitation source-based features for spoof detection task. In particular, extending the recently reported work in [17], where LP and Long-Term Prediction (LTP)-based analysis was carried out based on the fact that the SS and VC speech are quite likely to be very easily predicted (if it is generated with a simplified acoustic model) or very difficult to predict (if any artifacts are present in the speech signal). Here, we explore the fact that the speech production mechanism is a nonlinear phenomenon and prediction of spoofed speech by nonlinear prediction (NLP) may provide complementary information to that obtained by LP-LTP analysis. The source-based features derived from LP-LTP and LP-NLP analysis are combined at score-level and also with the state-of-the-art Mel Frequency cepstral coefficients (MFCC) spectral features. It is observed that using LP-NLP-based features with the LP-LTP and MFCC features, the % EER for the detection system decreases more than using LP-LTP, LP-NLP or MFCC alone.

2. Prediction Techniques

2.1. Linear Prediction (LP)

The effectiveness of LP analysis is due to its ability to capture implicitly the frequency response of the time-varying vocal tract area function [20]. The speech signal at n^{th} instant can be expressed as a linear combination of previous 'p' samples, i.e.,

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

where $\hat{x}(n)$ is the predicted signal for $x(n)$. The difference between the original signal $x(n)$ and predicted signal $\hat{x}(n)$ is

http://dx.doi.org/10.21437/Interspeech.2016-1002

called LP error or LP residual which is denoted by $e(n)$. It can be said that the prediction coefficients $\{a_k\}$ where $k \in [1,p]$ are able to efficiently model the speech signal within a particular frame based on the prediction gain G_p, defined as [17],

$$G_p = \frac{E_x}{E_e}, \tag{2}$$

where E_x and E_e are the energies of original speech signal and predicted error signal, respectively. If G_p is high the prediction is better. The prediction coefficients are estimated by minimizing the l^2 energy of LP residual and it is given by [20],

$$E = \sum_{n=0}^{N-1} |e(n)|^2. \tag{3}$$

2.2. Long-Term Prediction (LTP)

The LTP technique is widely used in speech coding, e.g., in GSM 06.10 or in narrowband and wideband adaptive multi-rate coders [21]. LP is the short-term correlation of each sample with p immediate preceding samples while LTP represents the long-term correlation of sample $x(n)$ with $2Q+1$ similar samples which are a pitch period T away from sample $x(n)$ (Chap. 8, [22]) as shown in Figure 1. Thus, LTP operates on vectors rather than on individual samples. In this case, a vector of samples can be predicted using another vector of samples from the signal's history. The best matching vector is subtracted from LP residual error $e(n)$ resulting in LTP residual signal $e'(n)$. The LTP works efficiently with quasi-periodic voiced speech signals. The prediction error and prediction gain are the same as that of LP analysis. A schematic diagram of LP-LTP-based countermeasure for spoofing is shown in [17].

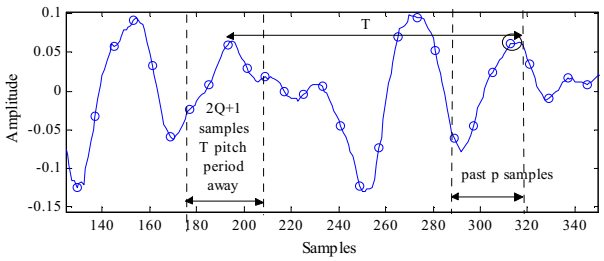

Figure 1: *Schematic of the short-term correlation of a sample with 'p' immediate past samples and the long-term correlation with the samples which are a pitch period 'T' away. After* [22].

2.3. Nonlinear Prediction (NLP)

In practice, our speech production mechanism is a nonlinear phenomenon. This is because of the nonlinear interaction or *coupling* of the excitation source and the system. A nonlinear system with k memory terms, represented by the Volterra series expansion (which relates the input and output of the system) is used. For a dynamic system, a closed-loop version of the Volterra series is used in which the output $x(n)$ is fed back as a delayed input (i.e., $x(n) \equiv y(n)$). Therefore, we analyze the univariate time series by using a discrete Volterra-Wiener (VW) series of degree d and predictor memory k to calculate the predicted series $\hat{x}(n)$ is given by [23]:

$$\hat{x}(n) = a_0 + a_1 x(n-1) + a_2 x(n-2) + \ldots + a_k x(n-k) + a_{k+1} x(n-1)^2 + a_{k+2} x(n-1) \times x(n-2) + \ldots + a_{M-1} x(n-k)^d, \tag{4}$$

$$\therefore \quad \hat{x}(n) = \sum_{m=0}^{M-1} a_m z_m(n), \tag{5}$$

where the functional basis $\{z_m(n)\}$ consists of all the distinct combinations of the embedding space coordinates up to degree d with a total dimension $M=(k+d)!/(k!d!)$. Thus, each model is parameterized by k and d corresponding to the *predictor memory* and the *degree* of nonlinearity in the model, respectively. The coefficients $\{a_m\}_{m\in[1,M]}$ in eq.(5) are estimated by Korenberg's fast algorithm using Gram-Schmidt procedure from the linear and nonlinear autocorrelation of the data series itself [24]. The difference between $x(n)$ and $\hat{x}(n)$ is referred to as the NLP residual. The schematic diagram of LP-NLP-based proposed countermeasure is shown in Figure 2.

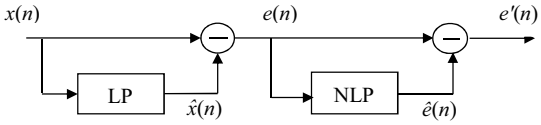

Figure 2: *Schematic diagram of LP-NLP approach as the proposed countermeasure.*

To show how NLP is efficient over the LP and LTP analysis, we compare the LP, LTP and NLP residual as shown in Figure 3. For a voiced region of the natural speech signal, the LP residual, LTP residual and NLP residual is observed. It is observed that the NLP residual has less relatively energy than LP and LTP residual. This was observed over several such voiced regions.

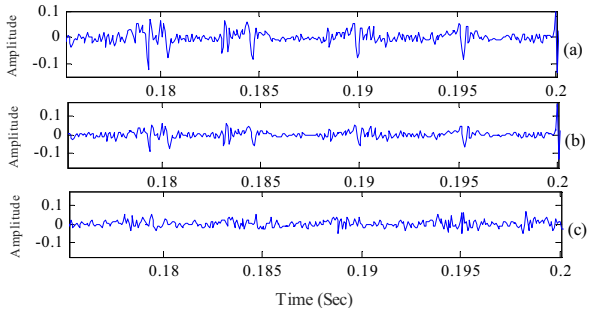

Figure 3: *Diagram of comparison among LP, LTP and NLP (a) LP, (b) LTP and (c) NLP residual for voiced speech.*

To further statistically quantify it, the average l^2 norm is estimated of all residuals for *100* natural utterances of the *D1* speaker of the ASVspoof challenge database. For LP analysis, *d=1* and for NLP, *d=2* is considered. To keep the number of coefficients constant for LP and NLP analysis, *p =2*, *14*, *20* are considered and corresponding to *k =1*, *4*, *5* giving a total of *3*, *15* and *21* coefficients, respectively. The average l^2 energy using eq. (3) are shown in Table 1. From this result, it is observed that the energy of prediction error decreases for NLP as compared to LP and LTP for same number of coefficients.

Table 1. *Average l^2 energy of prediction error signal over 100 utterances*

	Total number of coefficients		
Features	3	15	21
LP residual	12.3	6.42	6.18
LTP residual	10.5	5.52	5.26
NLP residual	10.5	**3.63**	**3.16**

3. Proposed Countermeasures

While considering the prediction error-based countermeasures, we assume that synthetic or converted speech signals will be predicted too well or inefficiently predicted. If it will be predicted too well, the prediction gain G_p in eq. (2) is high, else if it will not be predicted efficiently, the prediction gain is lower than the usual [17]. Similarly, several such measures can

be used as proposed for LP-LTP analysis in [17]. Our proposed countermeasures using LP-NLP approach (as shown in Figure 2) consist of the following features.

- MeanLPerr- mean energy of the LP error, i.e., mean energy of $e(n)$,
- MeanNLPerr- mean energy of the NLP error, i.e., energy of $e'(n)$,
- MaxNLPerr- maximum energy of the NLP error,
- MeanNLPgain- mean NLP gain (i.e., mean ratio between energies of the LP and NLP residual, mean G_p as defined in eq. (2)),
- MaxNLPgain- maximum of the G_p for NLP, MeanErrLen- mean length of segments with the NLP error above the threshold θ,
- MaxErrLen- maximum length of segments with the LTP error above the threshold θ,
- MeanNoErrLen- mean length of segments with the NLP error equal to or below the threshold θ,
- MaxNoErrLen- maximum length of segments with the NLP error equal to or below the threshold θ,
- EnergyLP- total energy of LP residual $e(n)$,
- EnergyNLP- total energy of NLP residual $e'(n)$,
- ErrChangeRate- the NLP threshold crossing rate (counted per *20 ms* frame).

Figure 4 shows the histogram of two selected features, i.e., the MeanNoErrLen for LP-LTP-based and ErrChangeRate for proposed LP-NLP-based features. It is observed that the MeanNoErrLen for LP-LTP analysis has significantly different distribution for human and spoofed speech. For LP-NLP features, the ErrChangeRate is more for natural than for spoofed speech. Thus, spoof-specific differences exist in countermeasures from LP-LTP and LP-NLP analysis.

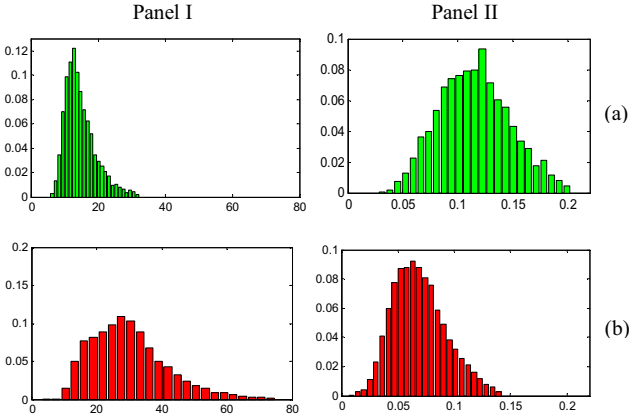

Figure 4: *Histogram of selected countermeasures. Panel I: MeanNoErrLen for LP-LTP and Panel II: ErrChangeRate for LP-NLP for (a) natural speech (b) spoofed speech.*

4. Experimental Setup

4.1. Databases

In this paper, the experiments were conducted on the ASVspoof 2015 challenge database [9]. This dataset is divided into three categories, namely, training, development and evaluation sets. The details of the speakers and number of utterances of natural and spoof speech signals are shown in Table 2. The spoofed speech was generated using *10* different spoofing algorithms (S1, S2, …, S10). These algorithms are either based on the SS or VC technique. The training and development set consists of S1 to S5 spoof and evaluation set consists of S1 to S10 (i.e., testing on unknown attacks).

Table 2. *Statistics of the ASVspoof 2015 challenge dataset*

	No. of speakers		No. of utterances	
Dataset	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	184000

In addition, experiments were also done on the Blizzard Challenge 2012 database. In this database, the synthetic speech utterances were generated from various systems, A- K. In particular, system A contains natural speech signals whereas systems B, G, F and I were built using unit-selection method. Systems E, H, K were built using statistical methods. Systems C and D were built using hybrid approach whereas diphone-based method is used to build system J. Each of the system has two categories, namely, paragraphs and sentences, consisting *60* and *100* speech signals, respectively.

4.2. Feature Extraction

In this paper, for LP analysis $p=20$ is considered for every *25 ms* frame length (due to the relationship between sampling frequency *Fs* and length of vocal tract [25]). The LP residual is calculated after subtracting the predicted signal (obtained using the LP coefficients) from the original speech signal. For LTP and NLP operation, the LP residual is framed using *5 ms* window. Furthermore, for NLP, $d=2$ and $k=5$ is considered in eq. (4). For ErrChangeRate feature in LP-LTP and LP-NLP, the threshold value is set to $\theta=0.02$. The entire analysis is done for voiced regions. The LP-LTP and LP-NLP features form a *12-dimensional* (*12-D*) feature vector for the entire speech signal. In addition to residual-based features, *12-D* MFCC feature vectors (excluding 0^{th} coefficient) were extracted using *28* subband filters for every *25 ms* frame length with a frame shift of *12.5 ms*. In addition to the static features, the Δ and the $\Delta\Delta$ features are considered to obtain *36-D* MFCC features for each frame of the speech signal. For representation, the LP-LTP and LP-NLP approaches are abbreviated as M1 and M2, respectively. In [17], the features are extracted by processing sample by sample on the speech signal, whereas in our work, we consider the frame-level processing for fast processing and to facilitate comparison with the frame-based MFCC.

4.3. Spoof Detection System

For spoofed speech detection, Gaussian Mixture Model (GMM)-based classifier with *128* mixture components was considered. The *3750* training utterances were taken to build GMM *Model1* for natural speech signals whereas *12625* training utterances were taken to build GMM *Model2* for spoofed speech signals. Final scores are represented in terms of log-likelihood ratio (LLR). The decision of the test speech being natural or spoofed is based on the LLR, i.e.,

$$LLR = log\left(LLk\_Model1\right) - log(LLk\_Model2), \qquad (6)$$

where *LLk_Model1* and *LLk_Model2* are the likelihood scores from the *Model1* and *Model2*, respectively. In this work, a score-level fusion is also considered, which is a weighted average of log-likelihood scores as follows,

$$LLK_{combine} = (1-\alpha_f)LLK_{feature1} + \alpha_f LLK_{feature2}\,, \qquad (7)$$

where $LLK_{feature1}$ and $LLK_{feature2}$ are log-likelihood scores of *feature1* and *feature2*, respectively. The weights of the scores are decided by the fusion parameter α_f.

Table 3. *The EER (%) on the development set for score-level fusion between M1, M2 and 36-D MFCC features and the fusion of best combination of M1-M2 (best_M1-M2) with 36-D MFCC features*

Feature1	Fusion Factor α_f											Feature2
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
M1	4.78	4.14	3.71	3.60	**3.60**	3.83	4.28	4.97	5.77	7.32	9.18	M2
M1	4.78	4.52	4.20	3.83	3.40	2.95	2.46	1.89	1.32	**0.80**	1.60	36-D MFCC
M2	9.18	8.64	8.06	7.46	6.66	5.92	4.86	3.95	2.60	**1.37**	1.60	36-D MFCC
best_M1-M2	3.60	3.32	3.03	2.77	2.46	2.14	1.72	1.34	0.92	**0.57**	1.60	36-D MFCC

Table 4. *The average EER (%) on the evaluation set for individual attacks for M1, M2, a fusion of M1 and M2 (best_M1-M2), 36-D MFCC and score-level fusion of best_M1-M2 fusion with 36-D MFCC features*

Features	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Known	Unknown	Average (w/o S10)	Average (with S10)
M1	5.74	4.13	0.02	0.02	1.45	2.55	7.49	0.04	0.95	86.66	2.27	19.54	2.49	10.91
M2	9.07	7.70	1.67	1.45	4.55	10.68	24.50	0.97	3.27	77.59	4.89	23.40	7.10	14.15
best_M1-M2 (α_f=0.4)	2.67	1.82	0.00	0.00	0.48	0.97	6.99	0.02	0.48	83.87	0.99	18.47	1.49	9.73
36-D MFCC	0.01	0.99	0.00	0.00	0.83	0.90	0.05	0.00	0.08	39.72	0.37	**8.15**	0.32	**4.26**
best_M1-M2 + (36-D MFCC)	0.00	0.04	0.00	0.00	0.02	0.02	0.01	0.00	0.01	51.11	**0.01**	10.23	**0.01**	5.12

The Detection Error Tradeoff (DET) curve is used to measure the performance of various features [26]. It gives uniform treatment to both False Acceptance Rate (FAR) and Miss Rejection Rate (MRR) for evaluation of system performance. In DET curve, the operating point where FAR and MRR becomes equal is referred as the EER.

5. Experimental Results

The results in EER on the development set for the M1, M2 and MFCC features are shown in Table 3. The % EER of M1 and M2 on development set are *4.78* and *9.18*, respectively. It is observed that the best score-level fusion of M1 and M2 is obtained for α_f *=0.4* (*best_M1-M2*) for which the EER is *3.6* %. For MFCC, we obtain % EER as *3.26* %, *2.17* % and *1.60* % for *12-D*, *24-D* and *36-D* feature sets, respectively. Therefore, only *36-D* MFCC feature vector is considered hereafter. To explore possible complementary information in source and systems features, the M1 and M2 features are fused with MFCC at various α_f. It is observed from Table 3 that, at score-level fusion of M1 with MFCC and M2 with MFCC, the EER reduces to *0.80* % and *1.37* %, respectively. Furthermore, the *best_M1-M2* scores are further combined at score-level with *36-D* MFCC. This fusion at α_f *=0.1*, reduces the % EER of MFCC from *1.6* to *0.57* (i.e., *65* % decrement). Hence, M1 and M2 extracted additional information which MFCC is unable to extract on the same data. The DET curve on development set is shown in Figure 5. It is observed that the MRR for the *best_M1-M2* and MFCC (at α_f *=0.1)* reduces significantly, as compared to M1, M2 and MFCC used alone.

Thereafter, for the evaluation set, the % EER of the individual attack is reported in Table 4. As the evaluation set has vocoder-based speech (S1-S9) and vocoder-independent speech (S10), we report the average % EER with and without S10. Considering the % EER for only vocoder-based speech, the % EER using M1 and M2 is *2.49* and *7.10*, respectively. For the best fusion of M1 and M2 the EER reduces to *1.49* %. The *36-D* MFCC features on its own gave *0.32* % EER which significantly reduced to *0.01*% after fusion of *best_M1-M2* and MFCC (α_f *=0.1*). Thus, including the LP-NLP-based features with the LP-LTP-based features and MFCC, the % EER is significantly improved for vocoded speech. The average % EER could not improve due to the high % EER obtained for the vocoder-independent S10 spoof. This is because the models were trained on vocoder-based spoofs and the testing was carried on vocoder-independent spoof. Thus, there is a need for more generalized countermeasures.

On testing with the sentence category of Blizzard Challenge 2012 database, it is observed from Table 5 that % EER of speech signals generated using system E, K, B, I and J is less for M1 and M2 than *36-D* MFCC.

Table 5. *The average EER (%) on the Blizzard Challenge 2012 database for M1, M2 and 36-D MFCC features*

System Name	Synthesis Technique	Features		
		M1	M2	36-D MFCC
E	Statistical	**10**	**41**	61
H	Statistical	42	42	3
K	Statistical	**43**	**51**	73
F	Unit selection	63	46	15
G	Unit selection	73	39	27
B	Unit selection	**56**	**53**	67
I	Unit selection	**49**	**28**	69
J	Diphone	**27**	**45**	69
C	Hybrid	54	49	47
D	Hybrid	55	37	42

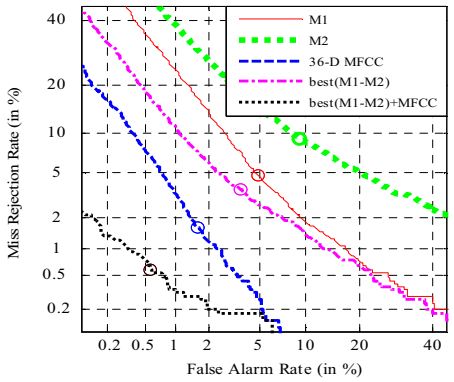

Figure 5: *DET curves on development set*

6. Summary and Conclusions

In this paper, we propose a countermeasure to detect the spoofed speech using LP and NLP-based approach. On the development set, for the LP-LTP approach using sample-by-sample processing [17], the EER obtained is *8.90* % which decreases to *4.78* % by frame-by-frame processing. The % EER further reduces to *3.6* on score-level fusion LP-LTP and LP-NLP-based countermeasures. The countermeasures are tested on the ASVspoof 2015 and the Blizzard Challenge 2012 databases. Our future research work will be focused on the frequency-domain LP analysis for the spoof detection task.

7. References

[1] P. Perrot and G. Chollet, "Helping the forensic research institute of the French Gendarmerie to identify a suspect in the presence of voice surgery," in *Forensic Speaker Recognition,* A. Neustein and H. A. Patil, (Eds.), *New York:* Springer, pp. 469-503, 2012.

[2] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. Int. Symp. on Intell. Multimedia, Video, Speech Process.*, Hong kong, 2004, pp. 145-148.

[3] A. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Proc. Int. Conf. of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2014, pp. 157-168.

[4] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Workshop on Speech Synthesis (SSW)*, Santa Monica, CA, 2002, pp. 227-230.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039-1064, 2009.

[6] Y. Stylianou, "Voice transformation: A survey," in *Proc. Int. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, Taipei, Taiwan, 2009, pp. 3585-3588.

[7] J-F. Bonastre, D. Matrouf, and C. Fredouille, "Transfer function-based voice transformation for speaker recognition," in *Proc. IEEE Speaker Lang. Recogn. Workshop (Odyssey)*, Toledo, 2006, pp. 1-6.

[8] Z. Wu, *et al.*, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130-153, 2015.

[9] Z. Wu, *et al.*, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2037-2041.

[10] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM System for the automatic speaker verification spoofing and countermeasures challenge 2015," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2072-2076.

[11] Y. Li, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2082-2086.

[12] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2042-2046.

[13] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," in *Proc. Int. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5475-5479.

[14] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural *vs.* spoofed speech," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2062-2066.

[15] S. Weng, *et al.*, "The SYSU system for the Interspeech 2015 automatic speaker verification spoofing and countermeasures challenge," in Cornell University Library,arXiv:1507.06711, 2015.

[16] P. L. De Leon, B. Stewar, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Portland, 2012, pp. 370-373.

[17] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, 2015, pp. 2077-2081.

[18] X. Xiao, *et al.*, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2052-2056.

[19] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (F_0) and strength of excitation (SoE) and their dynamics in spoofed speech detection," in *Proc. Int. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, Shanghai, 2015, pp. 5105-5109.

[20] J. Makhoul, "Linear Prediction: A Tutorial Review," in *Proc. IEEE*, vol. 63, no. 4, pp. 581-580, 1975.

[21] "Digital Cellular Telecommunication System (Phase 2+) (GSM); Full rate speech," 1999.

[22] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 2nd ed. Wiley & Sons, 2008.

[23] H. A. Patil and T. B. Patel, "Nonlinear Prediction of speech signal using Volterra-Wiener series," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Lyon, France, 2013, pp. 1687-1691.

[24] M. J. Korenberg, "Identifying nonlinear difference equation and functional expansion representation: The fast orthogonal algorithm," *Annals of Biomedical Engineering*, vol. 16, pp. 123-142, 1988.

[25] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637-655, 1971.

[26] A. Martin, G. Doddington, T. Kamm, and M. Ordowski, "The DET curve in assessment of detection task performance," in *Proc. European Conf. on Speech Comm. and Technol. (EUROSPEECH)*, Rhodes, Greece, 1997, pp. 1895-1898.