EFFECTIVENESS OF FUNDAMENTAL FREQUENCY (F_0) AND STRENGTH OF EXCITATION (SOE) FOR SPOOFED SPEECH DETECTION

Tanvina B. Patel and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, India

{tanvina_bhupendrabhai_patel, hemant_patil}@daiict.ac.in

ABSTRACT

Current countermeasures used in spoof detectors (for speech synthesis (SS) and voice conversion (VC)) are generally phase-based (as vocoders in SS and VC systems lack phase-information). These approaches may possibly fail for non-vocoder or unit-selection-based spoofs. In this work, we explore excitation source-based features, i.e., *fundamental frequency* (F_0) contour and *strength of excitation* (*SoE*) at the glottis as discriminative features using GMM-based classification system. We use F_0 and *SoE1* estimated from speech signal through zero frequency (ZF) filtering method. Further, *SoE2* is estimated from negative peaks of derivative of glottal flow waveform (dGFW) at glottal closure instants (GCIs). On the evaluation set of ASVspoof 2015 challenge database, the F_0 and *SoEs* features along with its dynamic variations achieve an Equal Error Rate (EER) of *12.41%*. The source features are fused at score-level with MFCC and recently proposed cochlear filter cepstral coefficients and instantaneous frequency (CFCCIF) features. On fusion with MFCC (CFCCIF), the EER decreases from *4.08%* to *3.26%* (*2.07%* to *1.72%*). The decrease in EER was evident on both known and unknown vocoder-based attacks. When MFCC, CFCCIF and source features are combined, the EER further decreased to *1.61%*. Thus, source features captures *complementary* information than MFCC and CFCCIF used alone.

Index Terms—*F_0*, SoE, MFCC, CFCCIF, anti-spoofing

1. INTRODUCTION

In voice biometrics or Automatic Speaker Verification (ASV) task, the speaker-specific information from the speech signal is used for authentication purpose with the help of machines. The characteristics of the speech signal being natural to produce makes it an adequate, easily accessible and convenient biometric modality. Current ASV systems with high accuracy and significantly low Equal Error Rates (EER) are still open to *spoofing* scenarios. Spoofing attacks can be due to impersonation (mimicking), replay, speech synthesis (SS) and voice conversion (VC). The effect of these attacks on the *%* EER of ASV systems is reported in [1]. The ease of access of online sources to generate synthetic speech (i.e., by Hidden Markov Model (HMM)-based Text-to-Speech (TTS) systems and adapted HMM systems [2]- [3]) and voice converted speech [4]- [5] make them vulnerable to ASV systems.

Previously proposed countermeasures were generally phase-based, including relative phase shift (RPS) [6]- [7], modified group delay [8], temporal modulation [9], etc. Most of these use Gaussian Mixture Model (GMM)-based classifier. Additional studies are also based on improving the back-end models [10]. Generally, these approaches were based on *known* attacks (i.e., using prior information of spoofing algorithm) instead of the real-case scenarios of *unknown* attacks or mismatched conditions. To address this, very recently, the ASVspoof 2015 challenge has been organized as a special session at INTERSPEECH 2015 [11]. Here, the task was to design an ASV-independent detector to classify natural *vs.* spoofed speech for both *known* and *unknown* attacks [11]. A generalized dataset was provided by the organizers and results in *%* EER were returned based on the scores submitted. The various approaches proposed at the challenge used features based on magnitude and phase spectrum of group delay [12]- [13], relative phase [14]- [15] and exploring back-end of spoofing detectors [16]- [17]. Efforts had also been made to exploit new features such as Linear Prediction (LP) residual [18]- [19] and wavelet-based features [20] for spoof detection systems (SDS). For the challenge, the authors proposed a Cochlear Filter Cepstral Coefficients and Instantaneous Frequency (CFCCIF) feature which was relatively the best performing system [21]. The CFCCIF features use envelope at output of each cochlear subband filter, i.e., CFCC [22] and its IF information.

In this work, we explore excitation source-based features to improve performance of SDS. Earlier in [23], [24], [25], pitch, pitch patterns and its variability were used to detect SS spoof. Not much work is reported for VC spoof. In this work, we explore F_0 contour and strength of excitation (*SoE*) features at the glottal closure instants (GCIs) in the voiced regions to detect spoofed speech. We use F_0 and measures of *SoE1* estimated from the speech signal through zero frequency (ZF) filtering method [26]. The ZF filtering is used due to its effectiveness to estimate both F_0 and

The authors would like to thank Department of Electronics and Information Technology (DeitY), New Delhi, India for TTS (Phase-II) and ASR (Phase-II) consortium projects and the authorities of DA-IICT.

excitation strength from speech. The *SoE2* is estimated from negative peaks of derivative of Glottal Flow Waveform (dGFW). The GFW is estimated by Iterative Adaptive Inverse Filtering (IAIF) method [27]. Humans vary their vocal fold movements and *SoE* at the glottis depending on the type of utterance which can affect the F_0 contour and the *SoE* of the speech. However, there is *no* true glottal closure phenomenon during generation of spoofed speech. Thus, relationship between F_0 and *SoEs* of natural and spoofed speech should be different. In addition, it was observed that their dynamic information provided good discrimination even with much less feature dimensions. Individually, MFCC and CFCCIF features gave low *%* EER on known and unknown attacks, respectively. When information from source-based features (i.e., F_0, *SoE1* and *SoE2* and their dynamics) were fused with MFCC and CFCCIF features at score-level, the EER decreased further. Thus, improvement in performance of SDS on using *source-based* information demonstrate that natural and spoofed speech vary in terms of excitation source characteristics which should be further explored as discriminative features in spoof detection.

2. DESCRIPTION OF SOURCE FEATURES

2.1. Estimation of F_0 and SoE from the Speech Signal

The ZF filtering method, also known as *0-Hz* resonator [26] is used here to obtain the excitation source information. The idea is effect due to an impulse is uniformly spread across *all* frequency regions including *zero* frequency. When speech is passed through a ZF filter, the vocal tract information from speech signal is separated. The negative-to-positive zero-crossings of the filtered signal provide an estimate of the GCIs and hence, the F_0 contour is estimated. The slope of ZF filtered signal at negative-to-positive zero-crossings gives a measure of the strength of glottal closure, i.e., *SoE* [28]. The slope (or derivative) is a *point* property of a system which at the negative-to-positive zero-crossings indicates the strength of abruptness of glottal closure. Thus, both F_0 contour and *SoE* is estimated from *0-Hz* resonator.

2.2. Estimation of SoE from dGFW

The GFW results from the movement of slit-like opening between the glottal folds called *glottis*. The glottal flow may be gradual or sudden depending on the movement of the glottis. Using IAIF method, the effect of the vocal tract system and lip radiation is cancelled from speech to estimate the GFW [27]. Thereafter, its derivative is computed and negative peak of the dGFW at GCIs gives the *SoE* at GCI. The strength of the negative peaks of dGFW indicates the strength/force with which the glottis closes. Fig. 1 shows the F_0 and *SoE* derived from speech (*SoE1*) and dGFW (*SoE2*) for a natural speech (Panel I) and HMM-based SS spoof from the SAS database (Panel II) [29]. In Fig. 1 (d), only the negative part of the dGFW is plotted and the magnitude of dGFW at the GCI is indicated as *SoE2* in Fig 1(d).

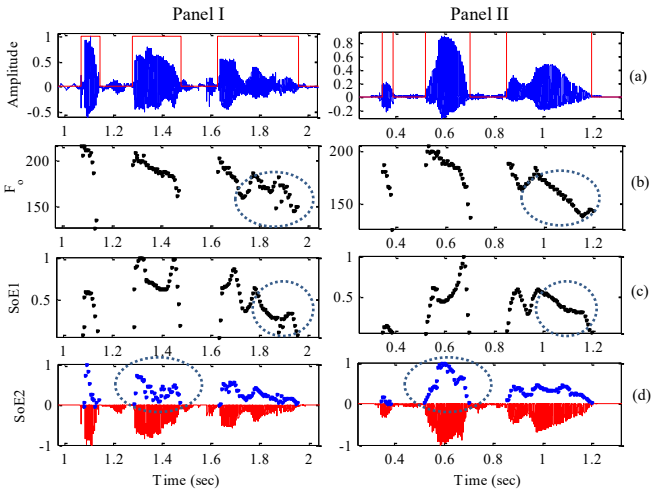


Fig. 1. Panel I: Natural speech and Panel II: Spoofed speech (SS) (a) speech signal \it's nice to hear\, (b) F_0 contour estimated by ZF filtering method (c) normalized *SoE1* at GCIs estimated by ZF method and (d) the dGFW (red) and normalized *SoE2* estimated from dGFW at GCI's from ZFF method (dotted blue).

The relation between source-based features for natural and SS spoof in Fig 1 is shown by scatter plot of F_0, *SoE1* and *SoE2* at GCIs in Fig. 2. The correlation coefficients between: F_0 *vs. SoE1, SoE1 vs. SoE2* and *SoE2 vs.* F_0 are *0.51*, *0.73* and *0.51* for natural speech and *0.34*, *0.645* and *0.45* for SS speech, respectively. Thus, it is observed that *correlations* vary for natural and SS speech. In addition, as shown by dotted regions in Fig. 1, there exist variations in excitation source features for natural and SS speech. Such variations were found over several utterances for SS and VC spoof. Although a direct relationship amongst F_0, *SoE1* and *SoE2* cannot be specified for different spoofing algorithms, there do exist differences in natural and spoofed speech due to excitation source characteristics. This is verified by using F_0, *SoE1* and *SoE2* and their dynamics as discriminative features for proposed spoof detection task in Section 4.

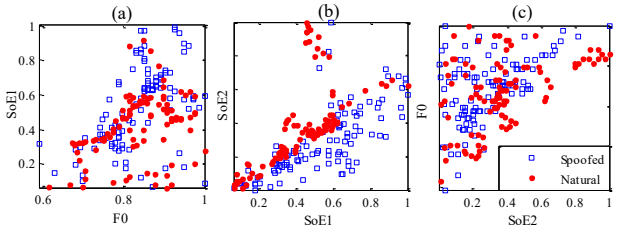


Fig. 2. Scatter plots (a) F_0 *vs. SoE1* (b) *SoE1 vs. SoE2* and (c) *SoE2 vs.* F_0 for the natural and spoofed (SS) utterances in Panel I and Panel II (from Fig. 1), respectively.

3. EXPERIMENTAL SETUP

3.1. Database

The database provided for the ASVspoof 2015 challenge is used here for experiment purpose. Details of spoofing (S) algorithms are provided in [11]. The training and development dataset consists of spoofed utterance generated by

five spoofing algorithms while evaluation data was based on ten spoofs, i.e., both *known* and *unknown* attacks. The *S3*, *S4* and *S10* are SS spoof and remaining are VC spoofs. The *S5* VC spoof uses Mel Log Spectrum Approximation (MLSA) filter [30] and *S10* SS spoof is vocoder-independent based on Modular Architecture for Research on speech sYnthesis (MARY) TTS system [31] that uses FESTIVAL framework [32]. Other spoofs are STRAIGHT vocoder-based [33].

Table 1. Summary of utterances used in training, development and evaluation sets of the ASVspoof 2015 challenge database [11]

	No. of speakers		No. of utterances	
Dataset	**Male**	**Female**	**Genuine**	**Spoofed**
Training (S1-S5)	10	15	3750	12625
Development (S1-S5)	15	20	3497	49875
Evaluation (S1-S10)	20	26	9404	184000

3.2. Classifier Details and Performance Evaluation

Gaussian Mixture Model (GMM) is used to model classes corresponding to natural and spoofed speech (using speech from training dataset). Scores are represented in terms of log-likelihood ratio (LLR). The decision of the test speech being human or spoofed is based on the LLR, i.e.,

$$LLR = \log(llk_{human}) - \log(llk_{spoof}), \quad (1)$$

where llk_{human} and llk_{spoof} are the likelihood scores from the GMM of human and spoofed speech, respectively. To utilize possible complementary information in source-based features as compared to acoustic features (such as MFCC and recently proposed CFCCIF features), we use their score-level fusion, i.e.,

$$LLk_{combine} = (1-\alpha_f)LLk_{feature1} + \alpha_f LLk_{feature2}, \quad (2)$$

where $LLk_{feature1}$ and $LLk_{feature2}$ are log-likelihood score of MFCC/CFCCIF and excitation source features, respectively. Parameter α_f decides the weights for fusion. Detection Error Tradeoff (DET) curve is used to measure the performance of SDS [34]. The operating point where the false acceptance rate (FAR) and false rejection rate (FRR) becomes equal is called as EER and is used as performance measure [35].

3.3. Feature Sets

The source features, i.e., F_0, *SoE1* and *SoE2* are extracted at GCIs estimated by ZF and IAIF method using a frame size of *25 ms* and with a shift of *50%*. The F_0, *SoE1* and *SoE2* gives a *3-dimensional (3-D)* static feature vector i.e., *Ds* for each GCI location. The dynamics of the F_0, *SoE1* and *SoE2* features are also considered by taking their first derivative, i.e., velocity, (*d1*: ΔF_0, $\Delta SoE1$ and $\Delta SoE2$) and appended to the *Ds* to get *6-D* feature vector (*D1=Ds+d1*). This was done till 5^{th} order derivative (i.e., acceleration, jerk, jounce, crackle) to get *D2*, *D3*, *D4* and *D5*, corresponding to *9-D*, *12-D*, *15-D* and *18-D* feature vectors, respectively. For score-level fusion, *36-D* MFCC and *36-D* CFCCIF feature vectors comprising of static and dynamic (i.e., *12*static+*12*Δ +*12*ΔΔ) are used in addition to source-based features.

4. EXPERIMENTAL RESULTS

4.1. Results on Development Set

4.1.1. Effect of source features and their dynamics

The effect of source features and their dynamics is studied by evaluating the *%* EER of the detector on the development set when trained on the training data for various number of mixture components in GMM (as in Fig. 3). It is observed from Fig. 3 that the *%*EER on the development set decreases significantly when the dynamic information is added to the static features. The decrease is significant with the increase in number of mixture models. The *%*EER with *128* mixtures for *Ds*, *D1*, *D2*, *D3*, *D4* and *D5* are *24.8*%, *16.1*%, *13.6*%, *12.7*%, *12.8*%, and *13.4*%, respectively. With higher-order derivative than the jerk (*D3*), the decrease is not significant and also increases slightly. Thus, throughout this work, *D3* feature vector with *128* mixtures GMM will be considered.

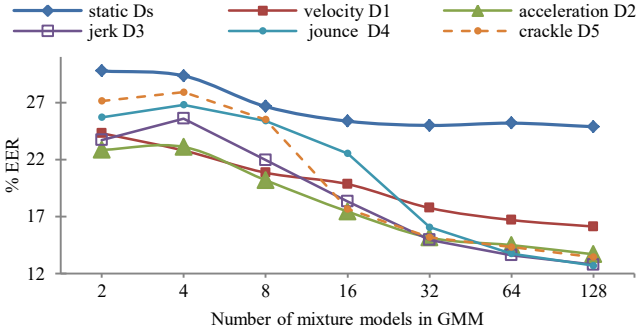


Fig. 3: The *%* EER obtained on the development set when the static and various dynamics, i.e., velocity, acceleration, jerk, jounce and crackle of F_0, *SoE1* and *SoE2* are considered.

To observe the effect of F_0, *SoE1* and *SoE2*, the *%* EER with F_0, *SoE1* and *SoE2* used individually up to third order derivative (i.e., *12-D*) was estimated. Next, the performance on using only two features at a time was also studied. It was observed from Table 2 that individually for F_0, *SoE1* and *SoE2* features, the *%* EER is very high *~27%*. Surprisingly, on combining F_0 features with two *SoEs*, the *%* EER increased, while on combining the two *SoEs* the *%* EER decreased (indicating that the *SoEs* capture complementary information). However, the *%* EER is not less than *12.7%* that was obtained when all three features are used (Fig. 3). Thus, all F_0, *SoE1* and *SoE2* features are essential for SDS.

Table 2. The *%* EER of F_0, *SoE1* and *SoE2* features used alone and when combined with each other using *D3* feature set

Individual	**% EER**	**Combined**	**% EER**
D3: F_0	27.94	**D3: F_0 & SoE1**	45.98
D3: SoE1	25.54	**D3: F_0 & SoE2**	43.92
D3: SoE2	27.68	**D3: SoE1 & SoE2**	**18.82**

4.1.2. Fusion of source-based features

The source-based features, (*D3* feature vector) were fused at score-level with MFCC and CFCCIF features (as in eq. (2)). It was observed that for $\alpha_f = 0.3$ for MFCC and $\alpha_f = 0.2$ for

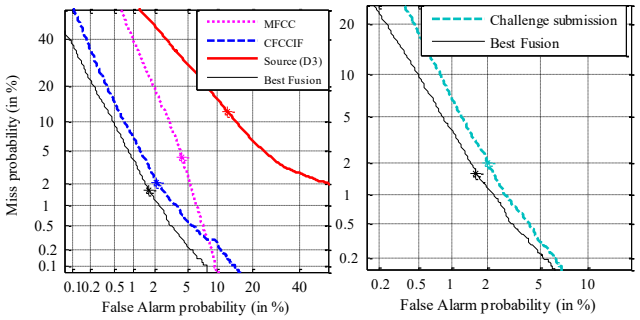
Table 3. The *%* EER of F_0, *SoE1* and *SoE2*, MFCC and CFCCIF along with their score-level fusion on the evaluation set.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Known	Unknown	Average
Ds	16.4	57.46	25.3	24.1	10.1	19.8	16.5	13.84	24.13	56.63	26.66	26.17	26.41
D1	2.66	55.76	11.4	10.8	2.82	8.59	7.11	5.86	10.25	61.29	16.68	18.62	17.65
D2	0.07	54.96	8.13	7.95	0.90	3.40	2.68	1.82	4.23	56.55	14.40	13.74	14.07
D3	0.01	53.90	6.35	6.34	0.23	1.58	0.86	0.58	2.99	51.25	13.37	11.45	12.41
MFCC	0.01	1.04	0.00	0.00	0.86	0.94	0.05	0.00	0.09	37.80	0.38	7.78	4.08
CFCCIF	0.03	0.72	0.00	0.00	2.24	0.98	0.16	0.88	0.29	15.42	0.60	3.55	2.07
MFCC+Ds (α_f=0.3)	0.00	1.11	0.00	0.00	0.43	0.54	0.03	0.00	0.07	36.13	0.31	7.35	3.83
MFCC+D1 (α_f=0.3)	0.00	0.88	0.00	0.00	0.22	0.38	0.03	0.00	0.04	34.39	0.22	6.97	3.59
MFCC+D2 (α_f=0.3)	0.00	0.73	0.00	0.00	0.12	0.22	0.02	0.00	0.03	32.27	0.17	6.51	3.34
MFCC+D3 (α_f=0.3)	0.00	0.76	0.00	0.00	0.08	0.18	0.02	0.00	0.03	31.58	0.17	6.36	3.26
CFCCIF+Ds (α_f=0.2)	0.02	0.82	0.00	0.00	1.28	0.74	0.11	0.71	0.25	15.32	0.42	3.43	1.92
CFCCIF+D1 (α_f=0.2)	0.01	0.68	0.00	0.00	0.83	0.51	0.08	0.58	0.17	15.12	0.30	3.29	1.80
CFCCIF+D2 (α_f=0.2)	0.00	0.67	0.00	0.00	0.55	0.34	0.06	0.46	0.08	14.75	0.24	3.14	1.69
CFCCIF+D3 (α_f=0.2)	0.00	0.74	0.00	0.00	0.40	0.33	0.05	0.37	0.08	15.26	0.23	3.22	1.72
D3+ MFCC+CFCCIF	0.00	0.375	0.00	0.00	0.18	0.16	0.02	0.087	0.022	15.30	**0.11**	**3.12**	**1.61**

CFCCIF the EER on the development set decreased from *1.6%* to *1.02%* and *1.5%* to *0.71%,* respectively. This EER achieved on the development set is less than that submitted by the authors in their work for the ASVspoof 2015 challenge (i.e., *0.83%* with fusion of MFCC and CFCCIF). Thus, from the development set, $\alpha_f = 0.3$ and $\alpha_f = 0.2$ is decided for fusion of source-based features with MFCC and CFCCIF on the evaluation set, respectively. The lesser contribution of source features when fused with CFCCIF is justified due to its embedded additional IF information.

4.2. Results on Evaluation Set

The results on the evaluation set for source-based features, MFCC and CFCCIF with *%* EER for individual known and unknown spoofs are shown in Table 3. It is seen that on an average, CFCCIF features gives the best *%* EER amongst all features considered. The average *%* EER of source-based features decreases with increase in dynamic information, i.e., *Ds* to *D3* for F_0, *SoE1* and *SoE2* features. It is observed that individually MFCC and CFCCIF work well for known and unknown attacks, respectively. However, when source information is fused at score-level with MFCC and CFCCIF (as in eq. (2)), the average EER decreases from *4.08%* to *3.26%* for MFCC and *2.07%* to *1.69%* for CFCCIF. Thus, F_0, *SoE1* and *SoE2* features contribute to decrease in *%* EER. Furthermore, on fusing F_0, *SoE1* and *SoE2* (*D3* source features), with MFCC (spectral) and CFCCIF (envelope in time-domain and IF) at all possible α_f, a weight factor of *0.2*, *0.1* and *0.7*, respectively, gave the best EER of *0.11%* for known attacks and *3.12%* for unknown attacks. This is considered as the *best fusion*. The decrease is more evident on vocoder-based attacks than on non-vocoder attacks (*S10*). The DET curves for MFCC, CFCCIF and source-features individually and the best fusion is shown in Fig. 4 (a). It is seen that MFCC and CFCCIF had high *%* FRR and *%* FAR than each other, respectively, which decreases when both are fused (this system was submitted by the authors for the ASVspoof 2015 challenge submission (Fig. 4 (b)). However, when the source features were added, the *%* EER further decreases. This best fusion performs relatively best at almost *all* operating points of the DET curve (Fig. 4 (b)).



(a) (b)

Fig. 4: (a) DET curve for MFCC (magenta), CFCCIF (blue), source features with their dynamics (*D3*) (red) and proposed *best fusion* (black), (b) DET curve for the relatively best system at the ASVspoof 2015 challenge (cyan) and proposed best fusion (black).

5. SUMMARY AND CONCLUSIONS

In this work, we explore F_0 contour and *SoE* as source-based features from speech signal for spoof detection. A simple GMM classifier with only *12-D* feature vector of F_0, *SoE1* and *SoE2* with its dynamics is used. The EER reduced significantly for vocoder-based attacks when fused at score-level with *36-D* MFCC and *36-D* CFCCIF. At the ASV spoof 2015 challenge, generally phase-based methods were used as vocoders in state-of-the-art SS and VC techniques lack phase information. These countermeasures gave less *%* EER for known attacks. However, they performed poorly for non-vocoder MARY TTS spoof (*S10*) with around *20-40 %* EER. In future, we plan to explore additional excitation source features, especially for non-vocoder attacks. Various aspects of the source-related information are essential to develop generalized countermeasures and to precisely detect wide range of spoofing attacks against ASV systems.

6. REFERENCES

[1] Z. Wu, et. al, "Spoofing and countermeasures for speaker verification: A survey," *Speech Comm.*, vol. 66, pp. 130-153, Feb. 2015.

[2] K. Tokuda, H. Zen and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE workshop on Speech Synthesis*, 2002, pp. 227-230.

[3] J. Yamagishi, et. al, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 17, no. 1, pp. 66-83, Jan. 2009.

[4] Y. Stylianou, O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.,* vol. 6, no. 2, pp. 131-142, Mar. 1998.

[5] 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.

[6] P. L. De Leon, et. al, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Audio, Speech, & Lang. Process.,* vol. 20, no. 8, pp. 2280-2290, Oct. 2012.

[7] J. Sanchez, et. al, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Trans. Info. Foren. and Sec.,* vol. 10, no. 4, pp. 810-820, April 2015.

[8] Z. Wu, E. S. Chng and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. INTERSPEECH*, Portland, USA, 2012, pp. 1700-1703.

[9] Z. Wu, et. al, "Synthetic speech detection using temporal modulation feature," in *Proc. IEEE ICASSP*, Vancouver, BC, Canada, 2013, pp. 7234-7238.

[10] A. Sizov, et. al, "Joint speaker verification and anti-spoofing in the i-vector space," *IEEE Trans. Info. Foren. and Sec.,* vol. 10, no. 4, pp. 821- 832, Feb. 2015.

[11] Z. Wu, et. al, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2037-2041. URL: http://www.spoofingchallenge.org/.

[12] Y. Liu, et. al, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2082-2086.

[13] X. Xiao, et. al, "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2052-2056.

[14] J. Sanchez, et. al, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2042-2046.

[15] L. Wang, et. al, "Relative phase information for detecting human speech and spoofed speech," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2092-2096.

[16] N. Chen, et. al, "Robust deep feature for spoofing detection-The SJTU system for ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2097-2101.

[17] J. Villalba, et. al, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2064-2071.

[18] M. J. Alam, P. Kenny, G. Bhattacharya and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasure challenge 2015," in *Proc. INTERSPEECH*, Dresden, Germany,2015, pp.2072-2076

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

[20] S. Novoselov, et. al, "STC anti-spoofing systems for the ASVspoof 2015 challenge," arXiv:1507.08074, 2015.

[21] 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. INTERSPEECH*, Dresden, Germany, pp. 2062-2066, 2015.

[22] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Trans. on Audio, Speech and Lang. Process.,* vol. 19, no. 6, pp. 1791-1801, 2011.

[23] T. Masuko, K. Tokuda and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. ICSLP*, Beijing, 2000, pp. 302-305.

[24] A. Ogihara, H. Unno and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Trans. on Fundamentals of Elect. Comm. and Computer Sciences,* vol. 88-A, pp. 280-286, 2005.

[25] B. Steward, P. L. De Leon and J. Yamagishi, "Synthetic speech discrimination using pitch patter statistics derived from image analysis," in *Proc. INTERSPEECH 2012*, Portland, Oregon, USA, pp. 370-373, 2012.

[26] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. on Speech and Audio Process.,* vol. 16, no. 8, pp. 1602-1613, Nov. 2008.

[27] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Comm.,* vol. 11, no. 2-3, pp. 109-118, 1992.

[28] K. S. R. Murty, B. Yegnanarayana and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Sig. Process. Letters,* vol. 16, no. 9, pp. 469-472, 2009.

[29] Z. Wu, et. al, "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE ICASSP*, Brisbane, Australia, 2015, pp. 4440-4444.

[30] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. IEEE ICASSP*, San Francisco, CA, pp. 137-140, 1992.

[31] "MARY Text-to-Speech System (MaryTTS)," [Available Online]: http://mary.dfki.de/ {Last accessed: 24th Aug. 2015}.

[32] A. Black, P. Taylor and R. Caley, "The Festival speech synthesis system," 1988. [Available Online].: http://festvox.org/festival/ {Last accesssed: 24th Aug. 2015}.

[33] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Comm.,*vol.27,no.3-4, pp.187-207,April 1999.

[34] A. Martin, G. Doddington, T. Kamm and M. Ordowski, "The DET curve in assessment of detection task performance," in *Proc. EUROSPEECH*, Rhodes, 1997, pp.1895-98.

[35] "DET-Curve Plotting (MATLAB)," [Available Online]: http:// www.itl.nist.gov/iad/mig/tools/{Last accessed: 24th Sept. '15}.