A Comparison of Features for Synthetic Speech Detection

Md Sahidullah, Tomi Kinnunen, Cemal Hanilçi

School of Computing, University of Eastern Finland, Finland
sahid@cs.uef.fi, tkinnu@cs.joensuu.fi, chanil@cs.uef.fi

Abstract

The performance of biometric systems based on automatic speaker recognition technology is severely degraded due to spoofing attacks with synthetic speech generated using different voice conversion (VC) and speech synthesis (SS) techniques. Various countermeasures are proposed to detect this type of attack, and in this context, choosing an appropriate feature extraction technique for capturing relevant information from speech is an important issue. This paper presents a concise experimental review of different features for synthetic speech detection task. A wide variety of features considered in this study include previously investigated features as well as some other potentially useful features for characterizing real and synthetic speech. The experiments are conducted on recently released ASVspoof 2015 corpus containing speech data from a large number of VC and SS technique. Comparative results using two different classifiers indicate that features representing spectral information in high-frequency region, dynamic information of speech, and detailed information related to subband characteristics are considerably more useful in detecting synthetic speech.
Index Terms: anti-spoofing, ASVspoof 2015, feature extraction, countermeasures

1. Introduction

Synthetic speech signal created using different voice conversion (VC) and speech synthesis (SS) techniques can be used to *spoof* biometric systems based on automatic speaker recognition technology [1, 2, 3, 4]. Over the past few years, considerable research effort has been devoted to protect the speaker recognition systems by developing various *countermeasures* [3]. Countermeasures consists of two parts: front-end for parameterizing the speech signal and back-end to determine whether it is a natural or synthetic speech. The front-end or feature extraction unit should capture relevant information from speech signal that reflects artifacts related to conversion or synthesis process. The other part includes a modeling technique to effectively represent those speech features. A number of techniques have been proposed for both parts to improve the spoofing detection performance. For example, mel-frequency cepstral coefficients (MFCCs), cosine phase, and modified group delay features were investigated in [5] for VC-based synthetic speech detection using a Gaussian mixture model (GMM) as back-end. Phase information obtained from *relative phase shift* (RPS) is also used in SS-based synthetic speech detection with high recognition accuracy as compared to MFCCs [6, 7]. The authors of [8], in turn, proposed to use one-class approach using *local binary pattern* [9] of linear frequency cepstral coefficients (LFCCs) followed by support vector machine (SVM) for voice conversion, speech synthesis and artificial signal detection [8]. A good overview of various countermeasures techniques is given in [3].

But most of the prior investigations are restricted to a certain type of spoofing technique, and only a limited number of countermeasures are studied. It is also not possible to compare the reported results across different studies since the experiments are conducted on different databases with varying configuration of features, classifiers and evaluation metrics. As a result for an end-user (e.g. administrator of an ASV system), it is difficult to choose one technique over another for his/her applications. A systematic benchmarking of the different proposed techniques in presence of various spoofing attacks is highly demanding. Further, it is crucial to know which kind of technique is more useful for a certain kind of spoofing attack.

In this paper, we experimentally compare 19 speech front-end features for spoofing attack detection, and compare their relative performances. We not only evaluate the performance of previously investigated features for spoofing detection, but include other features also which are successfully used in speaker verification task and have a potential for robust detection of spoofing attacks. The performances are separately evaluated with Gaussian mixture model (GMM) and support vector machine (SVM) based classifiers that are successfully employed in detecting synthetic speech. We report our results on the ASVspoof 2015 corpus which is provided with *First Automatic Speaker Verification Spoofing and Countermeasure Challenge* [10]. As far as we are aware, our study is the most extensive comparative evaluation of features in spoofing detection.

2. Feature Extraction Techniques

Here we describe the compared features briefly. We divide all the methods into three categories as shown in Table 1: short-term power spectrum features, short-term phase features, and feature involving long-term processing steps.

2.1. Short-Term Power Spectrum Features

Log-spectrum: The logarithm of power spectrum contains useful information related to the speech signal [16]. We have used raw log-spectrum (Spec) computed directly from speech frames as features.

Cepstrum: Cepstral coefficients (Cep) are computed from the power spectrum by applying discrete cosine transform (DCT) [17]. Usually, only the lower-order coefficients are retained in speech processing front-ends. Here, however, we retain *all* the coefficients since especially the higher-order coefficients could be useful for characterizing synthetic speech [18].

Δ-Cepstrum and Δ^2-Cepstrum: Traditional dynamic coefficients, i.e. deltas and double-deltas [19], are useful for speech and speaker recognition. Most of the synthetic speech generation techniques do not fully model temporal characteristics of speech. Therefore, intuitively deltas and double-deltas could be useful in detecting synthetic speech.

Filter bank based cepstral features: The main issue with

Table 1: *Summary of the evaluated features evaluated in this paper with the values of required control parameters/implementation details and references related to their earlier studies in spoofing detection.*

Type	Name (dim.)	Configuration Parameter(s)/Implementation Details	Used for Spoofing Detection in
Short-term power spectrum features	Spec/Cep (257)	Number of DFT bins = 512	—
	Δ-Spec/Δ-Cep (257)	Computed with three frames using differentiation	—
	Δ^2-Spec/Δ^2-Cep (257)	Computed with three frames using differentiation	—
	LFCC/MFCC (60)	No. of filter=20	[4, 11, 12]
	RFCC/IMFCC (60)	No. of filter=20	—
	LPCC (60)	LP Order=20	[13]
	PLPCC (63)	No. of filters in Bark scale=21	—
	SSFC (60)	No. of Subbands=20, rectangular window	—
	SCFC/SCMC (60)	No. of Subbands=20, rectangular window	—
Short-term phase features	MGDF (60)	$\alpha = 0.4$, $\gamma = 1.2$, First 20 coefficients are retained after DCT	[5]
	APGDF (60)	LP Order=20	—
	CosPhase (60)	First 20 coefficients are retained after DCT	[5]
	RPS (60)	RPS computed with COVAREP tool [14], 20 filters in mel filter bank	[6, 7]
Spectral features with long-term processing	SDC (56)	From MFCC with N=7, d=1, P=3, K=7	—
	Mod-Spec (60)	From 20 mel filter log-energies using window of 510 ms with shift of 10 ms	[15]
	FDLP (60)	FDLP package[1]	—
	MHEC (60)	No. of filters in Gammatone filter bank=20, f_c of LPF=30Hz	—

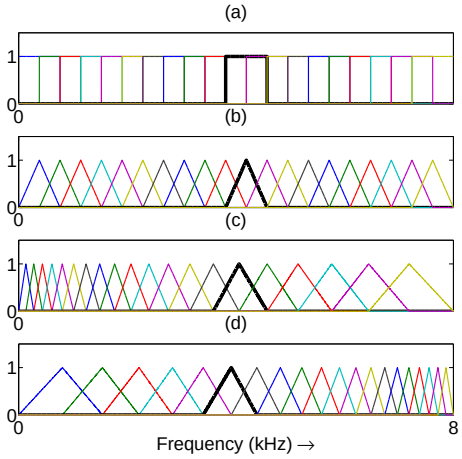

Figure 1: *Figure showing filter bank used in the computation of (a) RFCC, (b) LFCC, (c) MFCC, and (d) IMFCC.*

Spec and Cep features is high their dimensionality. This drawback is addressed by a filter bank. The power spectrum is first integrated using overlapping band-pass filters and logarithmic compression followed by DCT is performed to produce the cepstral coefficients. We consider four types of filter bank cepstral features as illustrated in Fig 1. In rectangular filter cepstral coefficients (RFCCs), integration is performed using a rectangular window [20] and the filters spaced in linear scale. Linear frequency cepstral coefficients (LFCCs) are extracted the same way but the filters are triangular in shape [8]. In MFCC, the filters are placed in mel scale, having denser spacing in the low-frequency region [21]. Finally, *inverted* mel frequency cepstral coefficient (IMFCC) uses filters that are linearly spaced on "inverted-mel" scale, giving higher emphasis to the high frequency region [22].

All-pole modeling based cepstral features: Cepstral coefficients are also derived from all-pole modeling representation of signal where linear prediction coefficients (LPC) are converted to linear prediction cepstral coefficients (LPCC) [23]. Another all-pole representation of speech called perceptual linear prediction cepstral coefficients (PLPCC) is also computed by first performing a series of perceptual processing prior to LP analysis [24].

Spectral flux based feature: Spectral flux measures the frame-by-frame change in the power spectrum [25]. It is computed as the Euclidean distance between normalized power spectrum of consecutive frames. We investigate a new feature that we term *subband spectral flux coefficient* (SSFC). First, we compute the subband spectral flux (SSF) of the i-th subband of the t-th speech frame as, $\text{SSF}_t^i = \sum_{k=1}^{M/2+1} \|\bar{S}_t(k) - \bar{S}_{t-1}(k)\|^2 w_i(k)$, where $\bar{S}_t(k)$ is the magnitude of k-th frequency component of normalized power spectrum of t-th frame, $w_i(k)$ is the spectral window function to obtain the frequency response of the i-th subband, and M is the number of bins in discrete Fourier transform (DFT). SSFCs are then obtained by performing logarithm and DCT on SSFs.

Subband spectral centroid based feature: Spectral subband centroids represent centroid frequencies of subbands, and they have properties similar to formant frequencies [26]. In [27], spectral centroid magnitude (SCM) is investigated along with subband centroid frequency (SCF) for speaker recognition. For the i-th subband of the t-th speech frame, they are defined as, $\text{SCF}_t^i = \sum_{k=1}^{M/2+1} f(k) S_t(k) w_i(k) / \sum_{k=1}^{M/2+1} S_t(k) w_i(k)$ and $\text{SCM}_t^i = \sum_{k=1}^{M/2+1} f(k) S_t(k) w_i(k) / \sum_{k=1}^{M/2+1} f(k) w_i(k)$, where $S_t(k)$ and $f(k)$ represent the power spectrum magnitude of t-th frame and normalized frequency ($0 \leq f(k) \leq 1$) corresponding to k-th frequency component. Both SCF and SCM contain complementary information related to subbands, not captured in cepstral features. The finer details of speech spectrum are not preserved in synthetic speech as VC and SS techniques mostly focus on producing identical overall envelope of the speech spectrum. Therefore, speech features representing SCF and SCM could be useful in detecting synthetic speech. We convert them to feature vectors following the process described in [27]. SCFs are directly used to create SCF coefficients (SCFCs) feature while log and DCT operations are performed on SCM to get SCM coefficients (SCMCs).

[1] `http://www.clsp.jhu.edu/~sriram/research/fdlp/feat_extract.tar.gz`

2.2. Short-Term Phase Features

Modified group delay function (MGDF): Modified group delay function was proposed to represent the phase information of a signal [28]. It is defined as, $\tau_t(k) = \text{sgn} \times \left|[X_{\text{R}}(k)Y_{\text{R}}(k) + X_{\text{I}}(k)Y_{\text{I}}(k)]/H(k)^{2\gamma}\right|^{\alpha}$, where sgn is the sign of $X_{\text{R}}(k)Y_{\text{R}}(k) + X_{\text{I}}(k)Y_{\text{I}}(k)$, $X_{\text{R}}(k)$ and $X_{\text{I}}(k)$ represent real and imaginary part of DFT for a speech frame $x(n)$ of L samples (for $n = 0, 1, 2, ..., L-1$), $Y_{\text{R}}(k)$ and $Y_{\text{I}}(k)$ represent the real and the imaginary parts of DFT for $nx(n)$, $H(k)$ is the speech spectrum after cepstral smoothing, while α and γ are two control parameters. Cepstral like features are formulated from MGDF by processing with logarithm followed by DCT. This feature was used for detecting synthetic speech in [5].

All-pole group delay function (APGDF): Recently, a phase-based feature using all-pole modeling is investigated in speaker recognition [29]. The advantage over MGDF is fewer parameters: only the all-pole predictor order needs to be optimized.

Cosine-phase function (CosPhase): Phase spectrum obtained during short-term speech analysis is used for synthetic speech detection [5]. Features are created from unwrapped phase by cosine normalization followed by DCT.

Relative phase shift (RPS): In the context of harmonic speech models, RPS describes the "phase shift"" of the harmonic components with respect to the fundamental frequency [30]. Features are computed from raw RPS by performing phase-unwrapping and differentiation followed by mel-scale integration and DCT. It was used in [6, 7] for detecting synthetic speech.

2.3. Spectral Features with Long-term Processing

Modulation spectrum (ModSpec): Modulation spectrum contains long-term temporal characteristics of speech signal [31]. It is computed by performing DFT in temporal domain on each dimension of feature vector. Non-linear processing, such as logarithmic compression on both the power spectrum, i.e. short-term and modulation, are often used in computing modulation spectrum based features [32]. In [15], modulation spectrum from MFCCs was used for synthetic speech detection, where feature vector is obtained by performing principal component analysis (PCA) on stacked modulation spectra.

Shifted delta coefficients (SDCs): SDC which also captures long-term speech information and was originally used for language recognition [33]. It is computed by augmenting delta coefficients of near-by frames. SDCs are specified by four parameters N, d, P, and k, where N is the number of cepstral coefficients, d is the number of frames for delta computation, P is the gap between the blocks of delta, and k is the number of blocks.

Frequency domain linear prediction (FDLP): In FDLP, LP analysis is performed in different subbands obtained by performing DCT on speech signal. FDLP features were recently studied in speaker recognition with promising results in both clean and noisy conditions [34].

Mean Hilbert envelope coefficients (MHECs): In MHEC, the speech signal is passed through a Gammatone filter bank. Then Hilbert envelope is computed from each filter output and they are processed using a low-pass filter for smoothing. Finally, MHEC features are derived by dividing the subband signals into sub-frames and computing the mean [35].

[2] http://www.spoofingchallenge.org/

3. Experimental Setup and Results

3.1. Database Description

The accuracy of different features for synthetic speech detection is evaluated on ASVspoof 2015 corpus distributed with First Automatic Speaker Verification Spoofing and Countermeasure Challenge[2]. A detailed description about the challenge and the corpus is available in [10]. The database has its own training segments from natural speech and synthetic speech. The synthetic speech data contains speech signals from five types of spoofing attacks (*known attacks*). The development section includes trials from natural speech and trials from synthetic speech of known attacks. On the other hand, the evaluation section contains trials from some additional spoofing techniques (*unknown attacks*) which are not included in training.

3.2. Classifier Description

In a different study with classifiers, we have shown that GMM-based technique yields reasonably good accuracy in ASVspoof 2015 corpus [36]. So, we choose this classifier for benchmarking of various features. We have also evaluated the performance with recently proposed SVM-based approach for detecting synthetic speech.

GMM-ML: Two separate GMMs are trained first using *maximum-likelihood* (ML) criteria from natural and synthetic speech-data. Then likelihood of test-segment is computed as, $\Lambda(\mathbf{X}) = \log p(\mathbf{X}|\lambda_n) - \log p(\mathbf{X}|\lambda_s)$, where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ represents the feature vectors of the test-segment containing T frames while λ_n and λ_s are the GMMs for natural and synthetic speech, respectively. We train GMMs with 512 mixtures and 10 EM iterations.

LBP-SVM: In LBP-SVM, first a *textrogram* is computed from feature-matrix using LBP analysis followed by one-dimensional histogram computation as detailed in [8]. Since seven out of ten spoofing techniques of ASVspoof 2015 are based on VC, we consider two-class SVM as back-end which gives best recognition accuracy for this type of spoofing attack [8]. We use linear kernel SVM from *LIBSVM*[3] package.

3.3. Performance Evaluation

Spoofing detection accuracy is measured by computing equal error rate (EER) [10]. We use Bosaris[4] toolkit to calculate the EER using receiver operating characteristics convex hull (ROCCH) method. Here, we report average EER by computing them separately for each spoofing technique.

3.4. Feature Extraction Parameters

Short-term features are extracted from speech frames with frame size 20 ms and of overlap 50%. The main control parameters and other implementation details of feature extraction techniques are given in Table 1. We have also included the energy coefficients when applicable. For meaningful comparison of performances, we choose the number of base coefficients such that the final feature dimensions, after adding Δ and Δ^2, are comparable as shown in the second column of Table 1. However, for Spec and Cep, the dimensionality is considerably high (257). Based on observations from preliminary experiments, we have not applied any voice activity detector (VAD) except for RPS as it requires only voiced frames [6].

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

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

Table 2: *Comparative accuracy (Avg. EER in %) using Spec, Cep, and RFCC features for static and dynamic coefficients on development set using GMM-ML classifier.*

	Static	Δ	Δ^2	Static+$\Delta\Delta^2$	$\Delta\Delta^2$
Spec	0.24	0.11	0.07	N/A	N/A
Cep	0.02	0.13	0.18	N/A	N/A
RFCC	2.41	0.34	0.35	0.75	**0.21**

Table 3: *Comparative accuracy (Avg. EER in %) of different features on the development set for both the classifiers.*

Feature	GMM-ML			LBP-SVM		
	Static	Static+ $\Delta\Delta^2$	$\Delta\Delta^2$	Static	Static+ $\Delta\Delta^2$	$\Delta\Delta^2$
RFCC	2.41	0.75	0.21	6.25	2.12	3.38
LFCC	2.46	0.66	0.12	5.05	1.56	2.37
MFCC	3.46	1.09	0.64	7.71	4.78	7.99
IMFCC	1.33	0.48	0.20	6.03	**1.50**	2.10
LPCC	2.44	0.68	0.14	5.44	2.47	3.94
PLPCC	2.95	1.61	1.51	9.09	5.48	8.07
SSFC	0.96	0.60	0.49	4.57	2.80	5.82
SCFC	1.77	**0.25**	**0.05**	23.43	1.87	**1.91**
SCMC	2.76	0.95	0.20	5.62	1.85	2.85
MGDF	4.71	2.24	2.69	7.15	3.81	7.41
APGDF	2.44	0.75	0.19	5.67	2.42	4.20
CosPhase	0.82	1.11	1.89	15.45	10.83	13.30
RPS	**0.21**	0.37	6.44	**2.45**	1.80	13.21
FDLP	5.71	2.18	1.99	12.17	6.50	9.44
MHEC	7.69	3.30	2.01	11.88	6.54	8.09
SDC-MFCC	4.37	-	-	-	7.06	-
ModSpec	4.41	-	-	-	5.92	-

Table 4: *Comparative accuracy (Avg. EER in %) of different features on the evaluation set for both the classifiers.*

	GMM-ML		LBP-SVM	
	Known	Unknown	Known	Unknown
MFCC (Static-$\Delta\Delta^2$)	0.83	5.17	4.35	17.18
RPS (Static)	0.10	10.51	**1.66**	20.04
RFCC ($\Delta\Delta^2$)	0.12	1.92	3.20	19.96
LFCC ($\Delta\Delta^2$)	0.11	**1.67**	2.13	19.45
MFCC ($\Delta\Delta^2$)	0.39	3.84	7.78	19.22
IMFCC ($\Delta\Delta^2$)	0.15	1.86	1.96	**9.97**
LPCC ($\Delta\Delta^2$)	0.11	2.31	3.54	13.90
SSFC ($\Delta\Delta^2$)	0.30	1.96	5.22	14.91
SCFC ($\Delta\Delta^2$)	**0.07**	8.84	1.81	17.54
SCMC ($\Delta\Delta^2$)	0.17	1.71	2.36	19.10
APGDF ($\Delta\Delta^2$)	0.16	2.34	3.74	13.10

3.5. Results

We first perform experiments on the development set for comparing the performance of the full spectrum (Spec and Cep) and RFCC feature. From the results in Table 2, we find that Spec and Cep lead to promising recognition accuracy can be obtained by compromising computational cost. Importantly, the dynamic coefficients of Spec and RFCC are more useful than static coefficients. This is reasonable since dynamic characteristics of spectral content are not well-modeled in most VC and SS techniques.

Motivated by this preliminary observations, we perform further experiments for both back-end, separately for static, dynamic, and combined coefficients with all the features described in Section 2 (except for ModSpec and SDC which already contain contextual information in their design). The results are shown in Table 3. For both the short-term power spectrum features as well as features involving long-term processing (i.e. FDLP and MHEC), it is clear that the dynamic coefficients outperform the static coefficients in almost all cases. Regarding the filter bank features, LFCC which uses triangular filter for local integration of the power spectrum outperforms RFCC where rectangular filter is used. Further, IMFCC, a feature set which emphasizes high-frequency spectral information beats MFCCs that emphasize the low-frequency region. Filter bank features and LPCC, giving equal emphasis to all frequencies, also outperform MFCCs and PLPCCs. Note that in PLPCC, low-frequency region is given more importance, too. SSFCs carry information related to spectral flux in different subbands is also found useful in comparison to other spectral features. Centroid frequency and magnitude features also perform well. The overall best recognition accuracy on development set (EER of 0.05%) is obtained with SCFC features and GMM-ML back-end.

We also observe high recognition accuracy with short-term phase based features. However, in contrast to the power spectrum features, dynamic coefficients are not always better than their static counterpart. For instance, for RPS features with GMM-ML back-end, EERs of static and dynamic coefficients are 0.21% and 6.44%, respectively. Perhaps the dynamic coefficients of phase are sensitive to small variations in signal. However, for MGDF and APGDF, Δ and Δ^2 are useful, possibly because of their resemblance with spectral characteristics [29, Fig.1]. Finally, somewhat different to what the authors initially assumed, for features with long-term processing, the recognition accuracy is low. This might be because long-term features have been found useful in mismatched conditions. But in ASVspoof 2015, there is no channel or environment mismatch and signals are already available with good quality.

The results on *evaluation set* are shown in Table 4 for top 11 features on the development set. Here, also, we find that dynamic coefficients and high-frequency information are useful. RPS feature performs well for known attacks, but for unknown attacks its performance is worst among all other features. The highest recognition accuracy for known attacks (EER of 0.07%) is obtained with SCFC features and GMM-ML classifier. However, for the unknown attacks, dynamics of cepstral features are better, and $\Delta\Delta^2$ of LFCCs gives the highest recognition accuracy (EER of 1.67%).

4. Conclusion

We have performed an extensive study with different feature extraction techniques for synthetic speech detection. Our results indicate that features conveying information related to high-frequency region, dynamic characteristic and detailed spectral information are useful. Those details are not accurately modeled during voice conversion or speech synthesis process.

5. Acknowledgements

This work was funded from Academy of Finland (proj. no. 253120 and 283256).

6. References

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, Mar 1998.

[2] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1, May 1998, pp. 285–288 vol.1.

[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, no. 0, pp. 130 – 153, 2015.

[4] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 4401–4404.

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

[6] P. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2280–2290, Oct 2012.

[7] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *Information Forensics and Security, IEEE Transactions on*, vol. 10, no. 4, pp. 810–820, April 2015.

[8] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*. IEEE, 2013, pp. 1–8.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, Jul 2002.

[10] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH*, 2015, (Accepted).

[11] Z. Wu and H. Li, "Voice conversion and spoofing attack on speaker verification systems," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, Oct 2013, pp. 1–9.

[12] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *INTERSPEECH*, 2013.

[13] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 3068–3072.

[14] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep - a collaborative voice analysis repository for speech technologies," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 960–964.

[15] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7234–7238.

[16] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993.

[17] D. G. Childers, D. Skinner, and R. Kemerait, "The cepstrum: A guide to processing," *Proceedings of the IEEE*, vol. 65, no. 10, pp. 1428–1443, Oct 1977.

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

[19] F. Soong and A. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 6, pp. 871–879, Jun 1988.

[20] T. Hasan, S. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 6783–6787.

[21] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[22] S. Chakroborty, A. Roy, and G. Saha, "Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks," *International Journal of Signal Processing*, vol. 4, no. 2, pp. 114–122, 2007.

[23] S. Furui, "Cepstral analysis technique for automatic speaker verification," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 2, pp. 254–272, Apr 1981.

[24] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[25] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2, Apr 1997, pp. 1331–1334.

[26] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, May 1998, pp. 617–620.

[27] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition," in *Odyssey*, 2010, p. 7.

[28] H. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 1, April 2003, pp. I–68–71.

[29] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alku, "Using group delay functions from all-pole models for speaker recognition," in *INTERSPEECH*, 2013, pp. 2489–2493.

[30] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 45, no. 7, pp. 381–383, March 2009.

[31] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 43 – 55, 1999.

[32] T. Kinnunen, K.-A. Lee, and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification." in *Odyssey*, 2008, p. 30.

[33] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *INTERSPEECH*, 2002.

[34] S. Ganapathy, S. Mallidi, and H. Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 8, pp. 1285–1295, Aug 2014.

[35] S. O. Sadjadi, T. Hasan, and J. H. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition," in *INTERSPEECH*, 2012.

[36] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: A comparison," in *INTERSPEECH*, 2015, (Accepted).