INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features

Sarfaraz Jelil, Rohan Kumar Das, S. R. M. Prasanna and Rohit Sinha

Department of Electronics and Electrical Engineering,
Indian Institute of Technology Guwahati, Guwahati-781039, India
{sarfaraz, rohankd, prasanna, rsinha}@iitg.ernet.in

Abstract

This work describes the techniques used for spoofed speech detection for the ASVspoof 2017 challenge. The main focus of this work is on exploiting the differences in the speech-specific nature of genuine speech signals and spoofed speech signals generated by replay attacks. This is achieved using glottal closure instants, epoch strength, and the peak to side lobe ratio of the Hilbert envelope of linear prediction residual. Apart from these source features, the instantaneous frequency cosine coefficient feature, and two cepstral features namely, constant Q cepstral coefficients and mel frequency cepstral coefficients are used. A combination of all these features is performed to obtain a high degree of accuracy for spoof detection. Initially, efficacy of these features are tested on the development set of the ASVspoof 2017 database with Gaussian mixture model based systems. The systems are then fused at score level which acts as the final combined system for the challenge. The combined system is able to outperform the individual systems by a significant margin. Finally, the experiments are repeated on the evaluation set of the database and the combined system results in an equal error rate of 13.95%.

Index Terms: anti-spoofing, ASVspoof 2017, epochs, peak to side lobe ratio, IFCC, CQCC.

1. Introduction

Speaker verification (SV) systems are highly vulnerable to spoofing attacks and it has been observed that their performances are severely degraded when subjected to these attacks [1, 2]. Hence in recent years, detecting spoofing attacks has become an integral part of SV systems and it is an area of active research. Spoofing attacks can be broadly classified into four categories: impersonation, replay, text-to-speech (TTS) synthesis and voice conversion [3]. The first ASVspoof challenge was organized in Interspeech 2015 after the need to have common dataset, protocols and metrics was realized in Interspeech 2013 ASVspoof special session [4]. ASVspoof 2017 is the second edition of the challenges on spoofing and countermeasures for automatic SV and it focuses on the detection of only replay attacks [5, 6]. These attacks are very simple to achieve since they require no prior knowledge of speech processing technologies and high quality recording devices are available at very nominal costs [3].

In [7], channel noise pattern is determined from denoising filter and statistical frames that are trained by support vector machine based classifier to identify whether the input speech is authentic or replayed. The authors of [8] have used spectral bitmaps to identify whether a speech is genuine or replayed for text-dependent SV. For text-independent SV, a similar technique of average spectral bitmaps has been used as the low frequency contents of a replayed speech gets suppressed and it can be considered to be a discriminative feature [9]. Both score normalization and spectral features are utilized to design a playback attack detection algorithm that is robust to adverse acoustic environment [10]. The work in [11] describes the spoofing competition organized by the biometric group at Idiap Research Institute which also included replay attacks. The countermeasures for TTS and voice conversion attacks are studied in [12, 13, 14]. The authors of [15] have addressed the issues involved in impersonation by analyzing glottal and vocal tract parameters.

In this present work, the task of detecting replay attacks is dealt with. The work addresses the problems posed by the ASVspoof 2017 challenge and uses two speech source features that can be used as countermeasures. One of these features is a two dimensional feature containing the epoch intervals and the corresponding strength of excitation. The mean and the skewness of the peak to side lobe ratio (PSR) of Hilbert envelope (HE) of linear prediction (LP) residual is taken as the second source feature. Along with these two source features, the work also explores the impact of instantaneous frequency cosine coefficient (IFCC) feature [16], and cepstral features constant-Q cepstral coefficient (CQCC) [17] and mel frequency cepstral coefficient (MFCC) on the problem at hand. The work advocates the use of a combination of all these features since they contain complementary information and their fusion is likely to result in an improved performance of the spoof detection system. To measure the effectiveness of the individual features, the stand-alone systems are created on the development set of the ASVspoof 2017 database. These systems are built using Gaussian mixture model (GMM) based framework [18]. Subsequently, score level fusion of all these systems is carried out. It is expected that this fusion should enhance the system performance due to different front-end processing and features involved. The experiments are repeated on the evaluation set of the database but the difference in these experiments lies in the use of both the train and the development set to build the genuine and the spoofed speech models using GMM. The major contributions of this paper are the use of speech signal specific knowledge and exploiting the differences in the nature of genuine and spoofed speech signal to classify them accordingly.

The remainder of this paper is organized as follows. In Section 2, the features used for classifying genuine and spoofed speech are explained in detail. Section 3 describes the experiments conducted for this work and describes the process of development of the spoof detection systems and subsequent fusion of these systems. The results of these experiments and a following discussion are reported in Section 4. Conclusions of the work are presented in Section 5.

2. Features Used for Spoof Detection

This section provides an explanation of the source, instantaneous frequency and cepstral features used for building the spoof detection systems and the motivation behind using them.

http://dx.doi.org/10.21437/Interspeech.2017-930

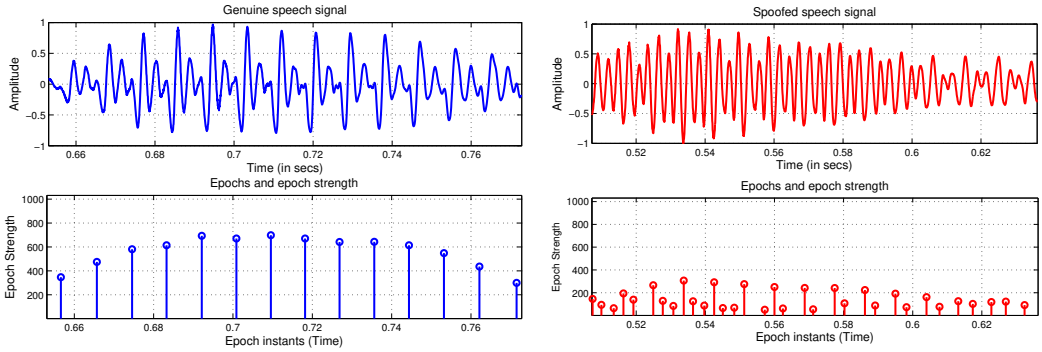

Figure 1: *Epochs and its strength for a segment of genuine and spoofed speech signals taken from ASVspoof 2017 database.*

2.1. Epoch and Epoch Strength Feature

The main focus of this work is to understand the dissimilarities in the a genuine and a spoofed speech signal. To this end, the first source feature used is epoch instants and their corresponding epoch strength. Epochs are defined as the instants where significant excitation is present during speech production [19]. Glottal closure instants are the regions around which most significant excitations occur for voiced speech. In this work, the epochs are extracted using zero frequency filter (ZFF) method [19]. The following steps are used to the determine epochs and epoch strengths from the speech signal $s[n]$ [20].

- Difference the speech signal

$$x[n] = s[n] - s[n-1] \tag{1}$$

- Pass $x[n]$ twice through the zero frequency resonator.

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n], \tag{2}$$

and

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n], \tag{3}$$

where $a_1 = -2$ and $a_2 = 1$. This is equivalent to integrating four times successively.

- Remove the trend by subtracting $y_2[n]$ with the average value of $y_2[n]$ calculated over window length of average pitch period.

$$y[n] = y_2[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2[n+m], \tag{4}$$

where $2N+1$ is the number of samples in the average pitch period. This trend removed signal is called the ZFF signal.

- The positive crossings of the ZFF signal are taken as the epochs.
- Slope of the ZFF signal is called the epoch strength or strength of excitation $S_e(el)$ [21].

$$S_e(el) = ||y[el+1] - y[el-1]|| \tag{5}$$

Figure 1 shows the plot of epoch and epoch strength for a speech segment. It is observed that the epochs are equally spaced for the genuine speech while in the spoofed speech this property is lost due to spurious detection of epochs. These additional epochs detected are errors produced by the algorithm because of the degraded spoofed speech signal. However, they are still useful for the discrimination between genuine and spoofed signals. Further, the epoch strengths are also higher for the genuine speech as compared to the spoofed speech. Thus, a two dimensional feature containing the difference of epochs (epoch interval) as one dimension and the corresponding epoch strength as another dimension is created which is referred to as the epoch feature (EF). In Figure 2, the distribution of EF for a genuine speech and its corresponding spoofed speech are shown that highlights their discrimination.

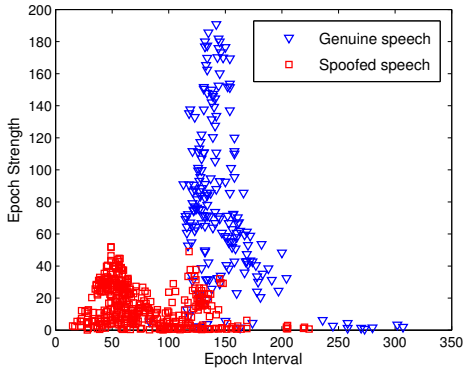

Figure 2: *Discrimination of EF for a genuine speech signal and its replayed version taken from ASVspoof 2017 database.*

2.2. Peak to Side Lobe Ratio (PSR) of the Hilbert Envelope (HE) of Linear Prediction (LP) Residual

LP residual gives information about the excitation source information, most importantly the epoch sequence for a segment of voiced speech. The residual error is large around the epochs and the prediction is poor [22]. However, since the residual signal amplitudes depend on the phase of the signal it may cause ambiguity in determining the epochs. Thus, instead of using the LP residual directly, the HE of the LP residual signal is used which helps in reducing the ambiguity about the peaks [22]. Hilbert envelope of LP residual $r(n)$ is computed using the following equation.

$$h(n) = \sqrt{r^2(n) + r_h^2(n)}, \tag{6}$$

where $r_h^2(n)$ is the Hilbert transform of $r(n)$.

Figure 3 shows the HE of LP residual for a segment of a genuine and a spoofed speech. It is observed that the peaks in the genuine speech are more well defined than the peaks in the

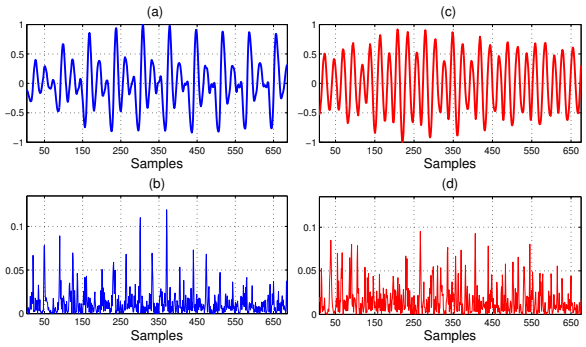

Figure 3: *(a) Segment of a genuine speech example. (b) HE of LP residual of the genuine speech segment. (c) Segment of a spoofed speech example. (d) HE of LP residual of the spoofed speech segment. Examples taken from ASVspoof 2017 database.*

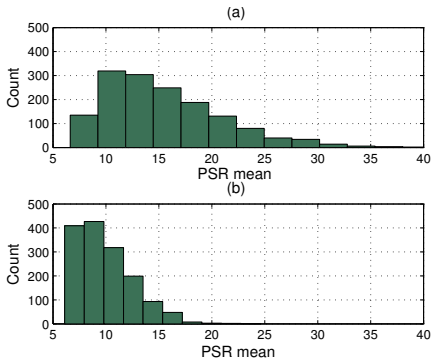

Figure 4: *Histogram of the PSR mean values of (a) genuine utterances and (b) spoofed utterances on train set of ASVspoof 2017 database.*

spoofed speech and are less affected by the side lobes. Hence, we use the parameter PSR of the HE of LP residual. The peaks are first computed by considering both sides of epoch locations acquired from ZFF signal. These peaks are searched within a window of 3 ms around the epoch locations. The maximum peak value within that window is taken as the peak of the HE of LP residual of speech signal. For calculating side-lobe values, the mean of sample values 1.5 ms to the right and 1.5 ms to the left of the peak value is taken. PSR is calculated by dividing the peak of HE of LP residual by the side-lobe value [23]. The histogram of the PSR mean for the train set of the ASVspoof 2017 database is shown in Figure 4. From this figure, it can be observed that the mean of the genuine speech is much higher than that of the spoofed speech. The distribution is also more skewed for the spoofed speech as compared to that of the genuine speech. Taking these factors into consideration, a two-dimensional feature vector consisting of the mean and skewness of the PSR values of a signal is created and is referred to as PSRMS.

2.3. Instantaneous Frequency Cosine Coefficient

The instantaneous frequency cosine coefficient (IFCC) is an attempt to extract features from the analytic phase of speech signal for speaker verification [16]. In order to overcome the problem of phase warping, the instantaneous frequency (IF) is computed with the help of Fourier transform properties without explicit involvement of computation of analytic phase. The narrow-band components of speech are taken to compute IF in the following way,

$$\theta'[n] = \frac{2\pi}{N} Re\left\{\frac{F_d^{-1} kZ[k]}{F_d^{-1} Z[k]}\right\}, \quad (7)$$

where F_d^{-1} denotes inverse discrete Fourier transform (IDFT), N being the length of the narrow band signal and $Z[k]$ is the DFT of the analytic signal $z[n]$, obtained from the narrow-band component of speech signal as explained in [24].

The computation of IF is followed by discrete cosine transform (DCT) on deviations in IF computed from narrow-band components of speech to extract IFCC features as a compact representation [16].

2.4. Cepstral Features

Recently, CQCC features have been proposed as countermeasure for spoofing and they are found to work extremely well [17]. It uses constant-Q transform instead of short time Fourier transform for spectral analysis [25]. On a speech signal $x(n)$, constant-Q transform is first applied to obtain $X^{CQ}(k)$ and then the power spectrum $|X^{CQ}(k)|^2$ is calculated. Then logarithm is taken on this to find the log power spectrum $\log|X^{CQ}(k)|^2$ and subsequently uniform resampling and discrete cosine transform are done to determine the CQCCs [17]. Apart from this, MFCC features are also used in this work due to its ubiquitous nature in the field of speech processing [26].

Table 1: *ASVspoof 2017 database details*

Database Subset	Number of Speakers	Utterances	
		Genuine	Spoofed
Train	10	1,508	1,508
Development	8	760	950
Evaluation	24	1,298	12,008

3. Development of Spoof Detection Systems

This section first describes the details of the ASVspoof 2017 database used for the experiments. It then explains the experimental setups of the systems developed for the challenge.

3.1. Database

The ASVspoof 2017 database is derived from the RedDots corpus where the original corpus acts as the genuine recordings and its replayed version serves as the spoofed recordings [27, 28]. It consists of three subsets: train, development and evaluation, the details of which are given in Table 1. All utterances have a sampling rate of 16 kHz and resolution of 16 bits per sample. For tuning the systems, the train set is used to build the genuine and the spoof speech models and the development set is used for testing. The systems submitted to the challenge evaluation uses only the train set for building the models and evaluation set as the test set. Post evaluation, the models are rebuilt using both the train and the development set utterances. The results presented in this paper for the evaluation set are the ones obtained using the latter configuration.

3.2. Experimental Setup

There are five stand-alone systems that are developed using the five features discussed in the previous section. For the source features, first glottal activity regions (GAR) are detected. System-1 is built with the EF features calculated within the GARs. Delta and delta-delta coefficients are extracted on which mean and variance normalization is performed. System-2 uses PSRMS features where from each signal the features are calculated within the GARs.

Table 2: *Summary of the experimental setups of the systems on ASVspoof 2017 database*

System	Features	Classifier
System-1 (S1)	EF (2-static + 2-Δ + 2-$\Delta\Delta$)	GMM (128 components)
System-2 (S2)	PSRMS	GMM (16 components)
System-3 (S3)	IFCC (20-static + 20-Δ + 20-$\Delta\Delta$)	GMM (512 components)
System-4 (S4)	CQCC (30-$\Delta\Delta$)	GMM (512 components)
System-5 (S5)	MFCC (13-static + 13-Δ + 13-$\Delta\Delta$)	GMM (512 components)

System-3 is based on the IFCC feature. To build this system, short-time analysis of speech is done using Hamming window of duration 20 ms with a frame shift of 10 ms. Energy based voice activity detection (VAD) is performed to separate the speech regions from the silence regions. IFCC of 20-dimensions are extracted as explained in Section 2.3. The delta and delta-delta features are then calculated to obtain 60-dimensional features.

System-4 uses 90-dimensional CQCC features that includes the zeroth coefficient, 29 static coefficients, 30 delta and 30 delta-delta coefficients. However, only the 30 delta-delta coefficients are used as they are more discriminative as reported in [25]. System-5 is built using 39-dimensional MFCC features having 13 static, 13 delta and 13 delta-delta coefficients. The preprocessing steps involved while computing these features are similar to that of System-3. All the systems use GMM of appropriate sizes as a classifier. Table 2 provides a summary of the individual systems and their configurations. Different combinations of the systems are fused at score level using Bosaris toolkit [29]. The final system submitted to the challenge is the score level fusion of all the five systems.

4. Results and Discussions

The results of the experiments conducted on the development set are presented in Table 3. It is observed that although the source features do not perform well on their own but when they are combined with all the other features the system performance increases sufficiently. This confirms that the source features carry information that is complementary to that carried by the features IFCC, CQCC and MFCC. For the evaluation set, the results of two baseline systems (B01 and B02) are reported in [6] as a summary to the ASVspoof 2017 challenge. The systems B01 and B02 are built using CQCC of 90 dimensions (30-static + 30-Δ + 30-$\Delta\Delta$) with GMM classifier. B01 is trained using both train and development set, whereas B02 is trained using only the train set. The equal error rate (EER) of B01 and B02 are 24.77% and 30.6% respectively. Compared to these, the final fused system submitted to the challenge produced an EER of 24.88%. As discussed earlier, the submitted system was developed with only the train set. Post evaluation, the systems are retrained using both train and development set as significant improvement of performance is observed for baseline system B01 over B02 as mentioned in [6]. The results of these experiments for individual as well as combined systems are shown in Table 4. The trend in performances for individual systems S1, S2 and S3 is inverted as compared to that in the development set while it remains the same for S4, S5 and the combined systems. The EER of the final fused system reduces to 13.95% compared to the submitted system having an EER of 24.88% with the use of additional data for learning the genuine and the spoofed speech models. This drop in EER points out the fact that to make the system robust to these attacks the knowledge of different recording and playback devices is important.

Table 3: *Results of the stand-alone spoof detection systems and their fusion on the development set of ASVspoof 2017 database*

System	EER (%)
S1	36.29
S2	31.60
S3	24.81
S4	9.79
S5	18.90
S4 + S5	6.82
S3 + S5	7.22
S3 + S4 + S5	6.39
S1 + S3 + S4 + S5	6.00
S1 + S2 + S3 + S4 + S5	**5.31**

Table 4: *Results of the stand-alone spoof detection systems and their fusion on the evaluation set of ASVspoof 2017 database*

System	EER (%)
S1	28.66
S2	28.90
S3	35.19
S4	19.58
S5	23.55
S3 + S4 + S5	15.31
S1 + S2 + S4	17.61
S1 + S2 + S3 + S4	15.15
S1 + S2 + S3 + S5	14.16
S1 + S2 + S3 + S4 + S5	**13.95**

5. Conclusion

This work presents different features to tackle the problem of replay attacks for the ASVspoof 2017 challenge. The main objective is to study the nature of genuine and spoofed speech signals to exploit their characteristics to design countermeasures for these attacks. Two source features EF and PSRMS are proposed for spoof detection. The work then explores the effectiveness of IFCC, CQCC and MFCC features to detect spoofing attacks. The individual systems developed using GMM based approach for each of these features are fused at the score level. The combined system is able to provide an improved spoof detection showing discriminative nature of the features considered for this work. This improvement is reflected in terms of performance achieved for both development as well as evaluation set while performing fusion of all the subsystems. In future, other characteristics of source can be studied to characterize genuine and spoofed speech. Deep neural network based classification can also be performed to further enhance the performance.

6. Acknowledgement

The authors thank the organizers of ASVspoof 2017 challenge for scoring the additional set of score files post evaluation.

7. References

[1] Y. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the YOHO speaker verification corpus," in *Knowledge-Based Intelligent Information and Engineering Systems, Springer*, 2005.

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

[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, pp. 130 – 153, 2015.

[4] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, no. 99, pp. 1–1, 2017.

[5] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "ASVspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," December 23, 2016.

[6] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Interspeech 2017 (Submitted)*, 2017.

[7] Z. F. Wang, G. Wei, and Q. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *ICMLC*, 2011.

[8] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec 2014, pp. 1–5.

[9] A. Paul, R. K. Das, R. Sinha, and S. R. M. Prasanna, "Countermeasure to handle replay attacks in practical speaker verification systems," in *2016 International Conference on Signal Processing and Communications (SPCOM)*, June 2016, pp. 1–5.

[10] J. Galka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, pp. 143 – 153, 2015.

[11] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonalves, A. G. S. Mello, R. P. V. Violato, F. O. Simoes, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi, H. Dinkel, N. Chen, Y. Qian, D. Paul, G. Saha, and M. Sahidullah, "Overview of btas 2016 speaker anti-spoofing competition," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sept 2016, pp. 1–6.

[12] 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 *Interspeech*, 2015.

[13] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Interspeech*, 2015.

[14] M. J. Alam, P. Kenny, V. Gupta, and T. Stafylakis, "Spoofing detection on the ASVspoof 2015 challenge corpus employing deep neural networks," in *Odyssey*, 2016.

[15] T. Amin, P. Marziliano, and J. German, "Glottal and vocal tract characteristics of voice impersonators," *IEEE Transactions on Multimedia*, vol. 16, pp. 668–678, April 2014.

[16] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54 – 71, 2016, phase-Aware Signal Processing in Speech Communication.

[17] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech and Language*, 2017.

[18] 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.

[19] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, Nov 2008.

[20] N. Adiga and S. R. M. Prasanna, "Significance of instants of significant excitation for source modeling," in *INTERSPEECH*, 2013.

[21] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469–472, June 2009.

[22] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 751 – 761, 2005.

[23] B. Sharma and S. R. M. Prasanna, "Sonority measurement using system, source, and suprasegmental information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 505 – 518, 2017.

[24] S. L. Marple, "Computing the discrete-time analytic signal via fft," in *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers (Cat. No.97CB36136)*, vol. 2, Nov 1997, pp. 1322–1325.

[25] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Odyssey*, 2016.

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

[27] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brummer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, M. J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Interspeech, Annual Conf. of the Int. Speech Comm. Assoc*, 2015.

[28] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamaki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamaki, and K. A. Lee, "Red-dots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *ICASSP*, 2016.

[29] The BOSARIS toolkit, (accessed on 10th Dec. 2013). [Online]. Available: www.sites.google.com/ site/bosaristoolkit/