Interspeech 2018
2-6 September 2018, Hyderabad

Exploration of Compressed ILPR Features for Replay Attack Detection

Sarfaraz Jelil[1], *Sishir Kalita*[1], *S. R. Mahadeva Prasanna*[1,2] *and Rohit Sinha*[1]

[1]Department of Electronics and Electrical Engineering,
Indian Institute of Technology Guwahati, Guwahati-781039, India
[2]Department of Electrical Engineering,
Indian Institute of Technology Dharwad, Dharwad-580011, India
{sarfaraz, sishir, prasanna, rsinha}@iitg.ernet.in, prasanna@iitdh.ac.in

Abstract

This paper deals with the problem of detecting replay attacks on speaker verification systems. In literature, apart from the acoustic features, source features have also been successfully used for this task. In existing source features, only the information around glottal closure instants (GCIs) have been utilized. We hypothesize that the feature derived by capturing the temporal dynamics between two GCIs would be more discriminative for such task. Motivated by that, in this work we explore the use of discrete cosine transform compressed integrated linear prediction residual (ILPR) features for discriminating between genuine and replayed signals. A spoof detection system is built using the compressed ILPR feature and a Gaussian mixture model (GMM) classifier. A baseline system is also built using constant-Q cepstral coefficient feature with GMM backend. These systems are tested on the ASVSpoof 2017 Version 2.0 database. On fusing the systems developed using acoustic and proposed source features an equal error rate of 9.41% is achieved on the evaluation set.

Index Terms: spoof detection, replay attacks, playback attacks, CILPR, CQCC.

1. Introduction

Automatic speaker verification is defined as the task of accepting or rejecting the identity claim of a speaker [1, 2, 3]. In recent years, the adoption of speaker verification technologies for commercial applications has seen a sharp increase. As these speaker verification systems are vulnerable to different kinds of spoofing attacks, research in the area of detecting and preventing spoofing attacks has steadily gained momentum. Spoofing attacks are classified into four different types: replay, voice conversion, text-to-speech (TTS) and impersonation [4]. This work deals exclusively with replay attacks.

The earliest work on studying the vulnerabilities of a speaker verification system to replay attacks was done in [5]. It performs the replay attacks by concatenating isolated digits from recordings of a genuine user and playing it back to the SV system and it showed a marked increase in both equal error rates (EER) and false acceptance rates (FAR) in the presence of replay attacks. Similar trends were reported in [6] which showed that EER of a joint factor analysis (JFA) based SV system was 0.71% when only non-spoofing impostor trials were used but when this EER operating point was selected as the decision threshold, 68% of the replayed trials were falsely accepted by the system. Recently, constant-Q cepstral coefficients (CQCC) features have been successfully used as a countermeasure for replay attacks [7]. The ASVSpoof 2017 challenge helped to increase awareness and the need for research in the area of replay attack detection [8]. In [9], replay attacks were tackled with the help of high frequency cepstral coefficients features at the signal level and a deep neural network (DNN) based feature extractor at the modeling level which is trained to distinguish between different playback, recording and environmental conditions. An ensemble learning technique is proposed in [10]. It uses a combination of acoustic features and different GMM based classifiers. The authors of [11] have explored the efficacy of standalone convolutional neural networks (CNN) and a combination of CNN and recurrent neural networks (RNN).

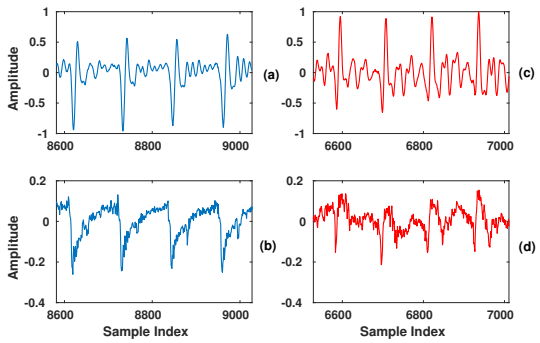

Figure 1: *ILPR signals for segments of genuine and spoofed speech signals. (a)-(b) and (c)-(d) represent the speech signal and its corresponding ILPR signal for genuine and spoofed signals, respectively.*

In our previous work, two source features namely epoch feature (EF) and mean and skewness of peak to side lobe ratio of the Hilbert envelope of linear prediction residual (PSRMS) were explored that characterize the excitation source behavior around the glottal closure instants (GCIs) [12]. However, they do not capture the dynamic characteristics of the source signal between two GCIs. This information can be extracted with the help of the integrated linear prediction residual (ILPR) signal which models the temporal shape of voice source signal between two GCIs. Figure 1 shows four glottal cycles of a speech signal and the corresponding ILPR for original and spoofed signal. It can be clearly seen from the figure that the dynamics of the ILPR signal between two GCIs is totally distorted as that of original. It is expected that characterizing the source temporal dynamics between two GCIs may give improvement in the spoof detection system.

In this present work, the task of discriminating between genuine and replayed speech signals is performed by utilizing the temporal dynamics of a signal between two GCIs. Since this dynamics is captured by ILPR, its use as a feature for spoof detection is proposed. However, the ILPR between two GCIs

10.21437/Interspeech.2018-1297

does not yield fixed dimensional vectors as the number of samples between two GCIs is not constant. In order to solve this problem, the discrete cosine transform (DCT) is applied on the ILPR in a pitch synchronous fashion. On account of the energy compaction achieved with DCT, a fixed dimension representation can be obtained. So derived compressed excitation source features are called as CILPR feature in this work.

It has also been established that combining source features with system level features helps in enhancing the performance of a replay attack detection system. CQCCs have proven to be a robust spoof detection feature. Thus, in this work a combination of CILPR with CQCC is proposed as a countermeasure for replay attack detection. Initially, a spoof detection system is developed using CILPR feature and Gaussian mixture model (GMM) classifier. Two GMMs are trained with these features: one for genuine class and one for spoof class. A CQCC based front-end and GMM based back-end baseline system is also developed. The efficacy of these systems is tested on the development set of the ASVSpoof 2017 Version 2.0 database. Since the two systems are built with features having complementary information, it is expected that their fusion will result in improved performance. The two systems are hence fused at the score level with the help of Bosaris toolkit [13]. The same systems are then built on the evaluation set of the database. Two different sets of experiments are performed for the evaluation set. The first set of experiments are conducted using only the train set of the database to learn the GMMs. In the second set, data from both train and development sets are taken to build the GMMs. The main contribution of this work is the use of ILPR to characterize the voice source of genuine and replay signals and the proposal to apply CILPR for replay attack detection.

The remainder of the paper is organized in the following way: Section 2 explains the method of extraction of CILPR feature in detail. In Section 3, the process of development of proposed spoof detection system using CILPR is described. Section 4 details the baseline system built using CQCC. Experimental results and discussions are provided in Section 5. Finally, the conclusions are presented in Section 6.

2. CILPR for excitation source characterization

CILPR is computed from the ILPR based voice source representation and captures the temporal shape of voice source signal between two GCIs. This feature has also been explored for speaker identification in [14]. ILPR is estimated by passing a non pre-emphasized version of speech signal through an LP inverse filter, the LP coefficients of the inverse filters are obtained from the corresponding pre-emphasized speech signal. The LP order is considered to be $f_s/1000 + 4$, where, f_s is sampling frequency. CILPR feature is computed pitch synchronously and requires GCIs to mark the pitch period. Let, $r_i(n)$ be a pitch synchronous segment of ILPR signal between i^{th} and $(i+1)^{th}$ GCIs. Then, DCT-II of $r_i(n)$ segment is computed by projecting it in to the discrete-cosine basis, as given below,

$$c(k) = \sum_{n=0}^{N-1} r_i(n) cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], \qquad (1)$$

where, $c(k)$, $k = 0, 1, 2, ...N-1$ are the DCT coefficients, N is the number of DCT coefficients. Further, the lower order DCT coefficients are considered and resultant feature vector is referred to as CILPR. It provides a compact representation of

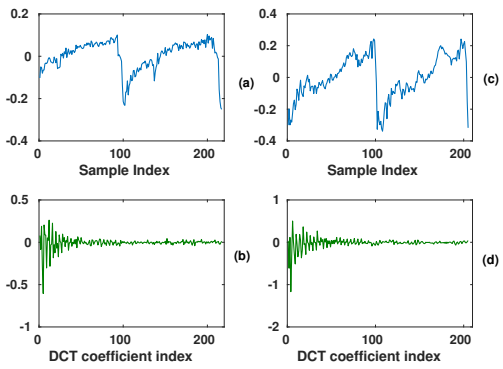

Figure 2: *(a)-(b) and (c)-(d) represent the ILPR signals and their corresponding non-truncated CILPR features for two different genuine speech signals, respectively.*

Table 2: *Details of the ASVSpoof 2017 Version 2.0 database*

Database Subset	Number of Speakers	Number of utterances	
		Genuine	Replayed
Train	10	1,507	1,507
Development	8	760	950
Evaluation	24	1,298	12,008

the ILPR signal between two GCIs and captures the dynamic characteristics between them.

To illustrate the compaction property of DCT and to show that the lower order DCT coefficients encapsulate the information contained in the ILPR, the ILPR of two pitch periods and the corresponding non-truncated CILPR for two different speech segments are depicted in Fig 2. From the figure it can be clearly noticed that the CILPR is a compressed version of the ILPR and that most of the information is contained in the first few coefficients.

3. CILPR based spoof detection system

In this section the process of development of the proposed spoof detection system using CILPR is explained. First a description of the ASVSpoof 2017 Version 2.0 database is presented. Next the experimental setup of this spoof detection system is described explaining the details of the CILPR based system.

3.1. Database

The experiments in this work are conducted on the ASVSpoof 2017 Version 2.0 database [15]. The original ASVSpoof 2017 database contained some anomalies which were removed in version 2.0. The database consists of genuine and replayed speech signals and is designed specifically for replay attack detection. The genuine signals are taken from the RedDots corpus and the replay attacks are made using different configurations [16, 17]. There are three subsets in this database: train, development and evaluation. The sampling rate of the signals is 16 kHz and the resolution is 16 bits per sample. The details of the database are provided in Table 2.

3.2. Development of spoof detection system using CILPR

CILPR feature is calculated from the ILPR in temporal domain in pitch synchronous manner [18]. Since 16 kHz sampling frequency is considered in this work, 20 order of LP filter is used.

Table 1: *Tuning of the dimensionality of the proposed CILPR features on development set of ASVSpoof 2017 Version 2.0 database*

CILPR dimensionality	4	8	12	16	20	24	28	32	36
EER (%)	31.06	25.55	21.52	20.34	20.0	**19.68**	20.11	19.95	20.06

As only voiced regions are considered to compute the CILPR, a glottal activity detection algorithm is needed. Zero-frequency filtering (ZFF) based method is applied to detect the glottal activity and to estimate the GCI locations in the speech signal [19, 20]. In this case, initially, differenced version of speech signal is passed through a zero frequency resonator (ZFR) and the output of the ZFR is exponentially growing or decaying in nature depending on the signal polarity. The trend of the exponential signal is removed by a moving average filter with size of approximately two pitch periods and the resultant signal is termed as zero frequency filtered (ZFF) signal. The positive to negative zero crossings are considered as the estimated GCIs of the signal. At every GCI, the positive to negative slope is termed as strength of excitation (SoE). The SoE is used to detect the glottal activity regions of the genuine and spoofed signals. The detected glottal activity regions are further used to extract the CILPR features pitch synchronously. At each GCI, ILPR segment from current GCI to next GCI is considered and normalized by the norm of the segment before applying DCT-II computation. The DCT-II of pitch synchronous ILPR segment is computed for both the genuine and spoofed signals and first few coefficients are considered, excluding the first coefficient. The low order DCT coefficients are termed as CILPR in this work. Initially, an experiment is performed to obtain the optimum number of DCT coefficients to develop the spoof detection system. The lower order DCT coefficients are varied from 4 to 36 with an increment of 4 to perform the experiment. GMM based models of 512 mixtures are built from the train subset of the database using each extracted CILPR features. The testing is done on the development data and the results are shown in the Table 1. From the table it can be observed that the best performance in terms of EER is obtained with 24 dimensional CILPR. After tuning the parameters of the CILPR on the development set, the system is tested on the evaluation set.

4. Constant-Q cepstral coefficients based baseline spoof detection system

CQCCs have recently become very popular as a countermeasure against replay attacks. This section describes the method of extracting CQCCs from a speech signal. The experimental setup of the baseline spoof detection system developed using CQCC is then discussed.

4.1. Method of extracting CQCC

CQCCs are calculated from the constant-Q transform (CQT) instead of the conventional Fourier transform [21]. In the calculation of Fourier transform, regularly spaced frequency bins are used which leads to poor frequency resolution at lower frequencies and poor temporal resolution at higher frequencies. CQT, on the other hand, applies geometrically spaced frequency bins which ensures higher frequency resolution at lower frequencies and higher temporal resolution at higher frequencies. This kind of frequency spacing resembles the human perception system more closely [7].

The CQT $X^{CQ}(k,n)$ of a discrete signal $x(n)$ is calculated in the following way [21]:

$$X^{CQ}(k,n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j-n+N_k/2) \quad (2)$$

where, $k = 1, 2, ..., K$ is the index of the frequency bin, a_k^* denotes the complex conjugate of a_k and N_k represents variable window lengths.

The CQT is then converted to a linear space and conventional cepstral analysis is performed to extract CQCC features.

$$CQCC(p) = \sum_{l=1}^{L} \log |X^{CQ}(l)|^2 \cos\left[\frac{p(l-\frac{1}{2})\pi}{L}\right] \quad (3)$$

where $p = 0, 1, 2, ..., L-1$ and l are newly resampled frequency bins [21].

4.2. Experimental setup

The baseline spoof detection system in this work is developed using CQCC features. They are calculated as described in Section 4. The system uses 19 static CQCC coefficients plus the log-energy coefficient to which delta (Δ) and delta-delta ($\Delta\Delta$) features are appended making it a 60 dimensional feature vector. Cepstral mean variance normalization (CMVN) is applied upon these features. From these features, two GMMs of 512 mixtures each are learned for the genuine and spoof classes.

5. Experimental Results and Discussion

The performances of the systems developed on the ASVSpoof 2017 Version 2.0 database using the proposed CILPR features and the baseline CQCC features are presented in Table 3 in terms of EER and minimum detection cost function (min. DCF). For the experiments on the development set, the train set is used to learn the GMMs of the two classes. The experiments on the evaluation set are conducted with two different training configurations. The first configuration uses only the train set to learn the GMMs while for the second configuration, data from the train and development sets are pooled to train the GMMs. These two configurations are called C1 and C2, respectively. From Table 3, it can be seen that the baseline system gives an EER of 9.19% for the development set. EERs for C1 and C2 configurations of the evaluation set are 13.84% and 12.58%, respectively. The proposed system results in an EER of 19.68% for the development set. Its EER for C1 configuration is 20.66% and 15.76% for C2. The baseline and the proposed system are then fused at score level. The fused system results in the best EER of 9.41% and minimum DCF of 0.474 for the C2 configuration of evaluation set. For the development set, the fused system produces an EER of 5.89% and minimum DCF of 0.338. This proves that combining source and acoustic features for replay attack detection can lead to significant performance enhancement. The detection error trade-off (DET) curves for the different spoof detection systems are given in Figure 3 which show similar performance trends.

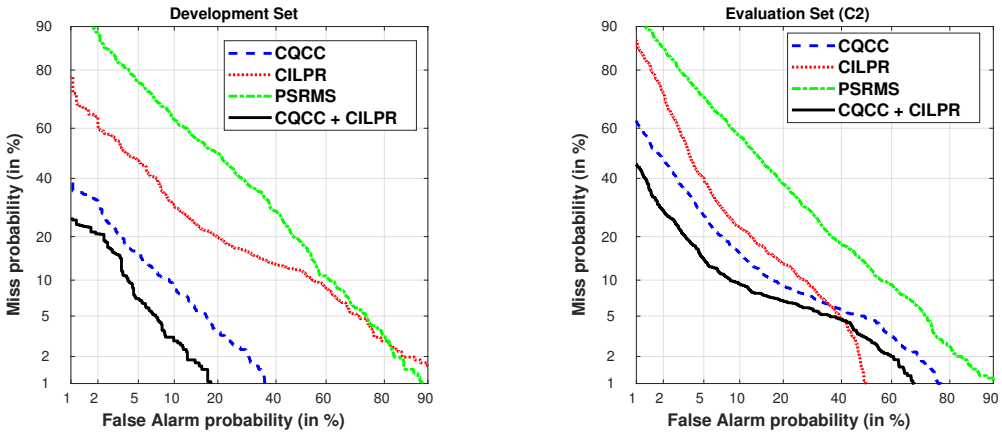

Figure 3: *DET curves for the spoof detection systems developed using different kinds of features and their fusion. These curves are plotted for development set and configuration C2 of evaluation set*

Table 3: *Performance comparison (in terms of %EER and minimum DCF) of different spoof detection systems and their fusion*

System	Development Set		Evaluation Set			
	Train		Train (C1)		Train + Development (C2)	
	EER	Min. DCF	EER	Min. DCF	EER	Min. DCF
Baseline: CQCC	9.19	0.455	13.84	0.739	12.58	0.662
Proposed: CILPR	19.68	0.799	20.66	0.947	15.76	0.847
Contrast: PSRMS	33.38	0.952	28.16	0.996	27.81	0.991
Fusion: CQCC + CILPR	**5.89**	**0.338**	**9.77**	**0.552**	**9.41**	**0.474**

For contrast purpose, a system is developed using PSRMS source feature proposed in our earlier work [12]. For calculating the PSRMS, first the linear prediction (LP) residual is estimated from the speech signal. From the LP residual a smoothed Hilbert envelope (HE) is obtained. The peaks in the HE correspond to the GCI locations. The side lobes around each peak is considered to measure the peak to side-lobe ratios. The mean and the skewness of these ratios form a 2-dimensional PSRMS features. GMM of 16 components is learned from these features. From Table 3, it can be noted that the PSRMS based contrast system produces an EER of 33.38% for the development set which is about double that for the CILPR system. The contrast system results in an EER of 28.16% and 27.81% on the evaluation set for configurations C1 and C2, respectively.

Another experiment is conducted to support our hypothesis that the information extracted between two GCIs is more useful than that obtained around a GCI for replay attack detection. In this experiment, all the genuine and spoofed signals in the development set are considered. To find the separation between the two classes in the case of CILPR and PSRMS features, the Bhattacharya distances have been computed and are shown in Figure 4. It is observed that the Bhattacharya distance for CILPR is significantly greater than that of PSRMS which confirms our hypothesis.

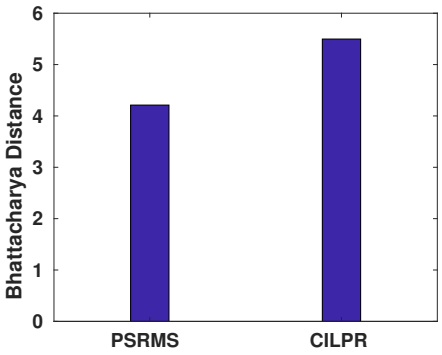

Figure 4: *Bhattacharya distance between the genuine and spoof classes for PSRMS and CILPR features. This supports enhanced detectability achieved with proposed CILPR features.*

6. Conclusion

This work explores the recently proposed ILPR voice source feature for the task of detecting replay attacks. The ILPR feature captures the dynamic characteristics of a signal between two GCIs and hence it has the potential to be more sensitive to the differences in genuine and replayed signals. However, the ILPR feature is calculated pitch synchronously, thus it does not produce fixed dimensional representation. On applying pitch synchronous DCT to the ILPR, a fixed dimensional feature is derived and referred to as the CILPR feature. First, a baseline spoof detection system employing the CQCC feature and GMM classifier is built and evaluated on the ASVSpoof2017 Version 2.0 database. For evaluating the CILPR feature another spoof detection system is created. On score-level combination of the acoustic and the proposed source feature based spoof detection systems, the EERs of 5.89% and 9.41% are obtained for the development set and the evaluation set with pooled data training of the GMMs, respectively. In the future, other speech source features can be explored for detecting replay attacks. DNN based classifiers can also be utilized to further increase the performance of the system.

7. References

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep 1997.

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

[3] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56 – 77, 2014.

[4] 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.

[5] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification : A study of technical impostor techniques," in *EUROSPEECH 1999, 6th European Conference on Speech Communication and Technology*, 1999.

[6] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA*, 2010.

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

[8] 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 *Proc. Interspeech 2017*, 2017, pp. 2–6. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1111

[9] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," in *Proc. Interspeech 2017*, 2017, pp. 97–101. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1377

[10] Z. Ji, Z. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in ASVSpoof2017," in *Proc. Interspeech 2017*, 2017, pp. 87–91. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1246

[11] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Proc. Interspeech 2017*, 2017, pp. 82–86. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-360

[12] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 22–26.

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

[14] A. G. Ramakrishnan, B. Abhiram, and S. R. M. Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *JASA EL*, 2015.

[15] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVSpoof 2017 version 2.0: meta-data analysis and baseline enhancements," in *Submitted to Odyssey 2018*, 2018.

[16] K. 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.

[17] 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.

[18] R. K. Das and S. R. M. Prasanna, "Exploring different attributes of source information for speaker verification with limited test data," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 184–190, 2016. [Online]. Available: https://doi.org/10.1121/1.4954653

[19] 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.

[20] 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.

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