Interspeech 2018
2-6 September 2018, Hyderabad

Linear Prediction Residual based Short-term Cepstral Features for Replay Attacks Detection

Madhusudan Singh and Debadatta Pati

Department of Electronics and Communication Engineering
National Institute of Technology Nagaland, Chumukedima, Dimapur-797103, India
madhusudan_niit@yahoo.co.in, debapati2003@yahoo.com

Abstract

Modern automatic speaker verification (ASV) systems are highly vulnerable to spoof attacks, and developing ASV anti-spoofing algorithms to protect ASV systems form these attacks is currently a part of active research. Contrarily to current trends on development of stand-alone spoof detection system, this work aims detection of replay attacks directly on the ASV system. The claim made through replay spoofing trials is rejected as impostors directly by ASV system. The objective here is to model the changes in the excitation signal characteristics caused by playback devices for replay detection. Accordingly, two linear prediction (LP) residual based source features are proposed for rejecting replay spoofing trials namely, RMFCC (residual mel-frequency cepstral coefficients) and LPRHEMFCC (LP residual Hilbert envelope MFCC). A comparative analysis between these two source features has been performed through speaker verification experiments to evaluate their effectiveness for ASV anti-spoofing applications. The comparison between the two has been made in the form of (source feature + MFCC) combination. The experiments are conducted using self-developed IITG-MV replay database. From the experimental results, it has been observed that 'LPRHEMFCC+MFCC' combination outperforms 'RMFCC+MFCC' combination, under replay attacks. Finally, the experiments are repeated on ASVspoof2017 database to validate the efficacy of proposed work.

Index Terms: Speaker Verification (SV), Replay detection, IITG-MV replay database, Hilbert Envelop, Source+MFCC.

1. Introduction

Automatic speaker verification (ASV) system accepts/rejects a claimed identity on the basis of provided speech samples [1]. In the present scenario where modern ASV systems have achieved state-of-the-art performances, they are highly vulnerable to a variety of spoof attacks [2]. Spoof attacks are classified as impersonation, replay, speech synthesis and voice conversion. Among these attacks, replay attack is very simple, and can be implemented easily using a high quality recording and playback device with little speech processing knowledge. Hence, the development of robust replay detection methods for ASV anti-spoofing is currently in progress.

Mostly, footprints of playback and recording devices in the replay signals were being used for identifying replay speech samples from the corresponding originals in prior works. In [3], increment in noise and reverberations in the replay signals from the surroundings was used for replay attack detection. In [4], the channel pattern noise from original and replay recordings was used as an indicator for detecting replay signals. The high frequency imperfections caused by additional anti-aliasing filtering process during re-recording via microphone were used for detecting replay signals in [5]. In [6], the authors explored the evidences from high frequency regions of speech to identify replay samples. In [7], the channel artifacts present at low signal-to-noise ratio time instants were used for replay detection task. In [8], the variations in the spectral envelope during transmission through recording and playback devices were modelled for replay detection. In [9], inclusion of additional epochs and corresponding strength in the replay signals was used to discriminate between actual and replay speech samples. Moreover, a comprehensive study on a set of different conventional and non-conventional features for the development of replay detection system has been reported in [10]. Conclusively, the exploration of excitation source information towards replay attacks detection has been ignored by the above mentioned prior works with the exception [9]. LP residual signal as excitation source parameterization has been widely explored already in literature such as in studies [11, 12, 13, 14]. Therefore, excitation source parameterization of LP residual would intuitively be useful for replay attacks detection. The preceding statement is also supported by developed replay detection system using epoch (source) feature in the recent prior work [9]. With this motivation, the present study deals with mel-based cepstral domain parameterization of LP residual for ASV anti-replay spoofing.

In this work, changes in the excitation component of the speech signal caused by playback devices specifically loudspeakers have been explored for replay detection task [15, 16]. The LP residual, obtained by LP analysis of speech signal, mostly contains information about the excitation source [17, 18]. Accordingly, the excitation source component is modelled using the LP residual and Hilbert envelope of the LP residual in the form of RMFCC and LPRHEMFCC features to detect replay speech samples, respectively. From the signal point of view, the peaks (epochs) look more significant in the Hilbert envelope of the LP residual in comparison to the LP residual [19]. Thereby, excitation source parameterization of Hilbert envelope may be more effective than direct LP residual parameterization. In this direction, a comparative analysis has been performed for the combinations RMFCC+MFCC and LPRHEMFCC+MFCC to analyse the strength of source features towards replay attacks detection. This is achieved through SV experiments conducted on self developed IITG-MV replay database. As per our knowledge, the feature LPRHEMFCC (mel based parameterization of Hilbert envelope of LP residual) is used for the first time in this study and hence, novel contribution to this work. More on this, contrarily to current trends on development of stand-alone countermeasures, the proposed work aims to reject replay spoofing trials directly on the ASV system. The genuine trials and either zero-effort imposter trials and/or replay trials are classified using EER decision threshold of the ASV system. The significant point of interest here in that the proposed approach does not require a dedicated spoof detection system.

10.21437/Interspeech.2018-1128

The rest of the paper is organized as follows: Section 2 presents description of proposed LP residual features for replay detection. A comparative analysis on proposed features has been performed through SV experiments using IITG-MV database in Section 3. The experimental observations of Section 3 are validated using ASVspoof2017 database in Section 4. The conclusions of the work are reported in Section 5.

2. LP residual based features

In LP model of speech, each speech sample is predicted as a linear combination of past p samples, where p is the order of prediction. Each speech sample $s(n)$ is predicted as,

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

where, $\hat{s}(n)$ is the predicted speech sample and $a_k s$ are LP coefficients (LPCs). The error between original and predicted signal is known as LP residual $r(n)$ and is given by,

$$r\,(n) = s\,(n) - \hat{s}\,(n) = s\,(n) + \sum_{k=1}^{p} a_k s\,(n-k) \tag{2}$$

2.1. RMFCC features

Figure 1 shows features extraction steps to obtained RMFCC feature. Discrete Fourier transform (DFT) is performed to obtained LP residual spectrum. The magnitude of LP residual spectra is passed through a bank of non-uniform triangular band pass filters placed on the mel-frequency scale. At the end, discrete cosine transform (DCT) is applied on the logarithm of the sub-band energies obtained from mel-filters bank to obtained RMFCC features.

If $R(e^{jw})$ is the spectrum of the LP residual $r(n)$, the magnitude of which is passed through mel-filters bank (M_{el}) for sub-band energy calculations. Then RMFCC feature ($R(k)$) is computed as,

$$R\,(k) = DCT[log(M_{el}(|R(e^{jw})|)] \tag{3}$$

2.2. LPRHERMFCC features

The LPRHEMFCC feature involves short-term cepstral processing of Hilbert envelope of LP residual signal in mel-domain. The Hilbert envelope $h(n)$ of LP residual $r(n)$ can be expressed as the magnitude of a complex time function given by,

$$h(n) = sqrt(r^2(n) + r_h^2(n)) \tag{4}$$

where $r_h(n)$ is the Hilbert transform of LP residual $r(n)$.

The feature extraction steps to obtained LPRHEMFCC feature is given in Figure 1. If $H(e^{jw})$ is the spectrum of the Hilbert envelope $h(n)$ of LP residual signal, then similar to RMFCC feature, LPRHEMFCC feature ($H(k)$) is computed in the following way,

$$H\,(k) = DCT[log(M_{el}(|H(e^{jw})|)] \tag{5}$$

The source features RMFCC and LPRHEMFCC involve short-term cepstral domain processing of the LP residual and Hilbert envelope of the LP residual, respectively with $20ms$ framesize and $10ms$ overlap. Hence, they model the glottal information averaged over two to three pitch periods [20]. Accordingly, changes in the excitation source characteristics made by replay attacks would possibly be captured by both features and thereby, ensuring their candidature towards developing replay detection systems.

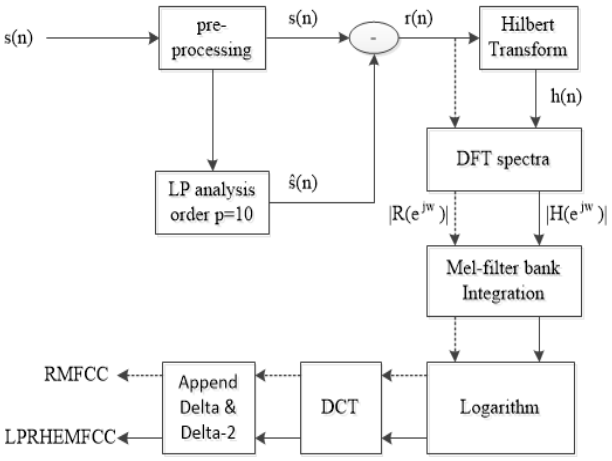

Figure 1: *Block diagram showing various steps of extracting RMFCC and LPRHEMFCC features.*

3. Experimental Study

3.1. Database Design

In this study, the replay database is manually developed by using publicly available Indian Institute of Technology Guwahati Multi-Variability (IITG-MV) speaker recognition database [21]. The Phase-I (office) and Phase-II (laboratory) datasets of IITG-MV database are collected using five different microphone sensors in multiple environment conditions and in different sessions. Therefore suitable for robust speaker verification, to design database for replay attack and anti-spoofing studies like RSR database [22].

The Phase-I and Phase-II datasets of IITG-MV database contain 148 (112 males and 36 females) non-native English speakers speech samples, recorded at the rate of 16000 samples/second. The duration of the speech samples per speaker varies from 10 to 15 minutes. For this experimental study, we consider 81 (45 males and 31 females) speakers speech data and segregate into two groups: Dataset-I and Dataset-II. Dataset-I includes 5 male and 6 female speakers speech data amounting to one hour from each gender for building gender-dependent UBM models. The Dataset-II is developed with 65 speakers speech data (comprising 40 males and 25 females) for evaluation purpose. Each speaker's first two minutes speech data are used for enrollment. The remaining data are converted into several segments of 30 seconds duration and used for test trials. Each test segment of each speaker is used as a genuine trial for the same target model and an impostor trial against other speakers model of the same gender. This resulted into a huge number of trials. The detail statistics are summarized in Table 1. Altogether, there are 42440 trials that include 1274 genuine and 41166 impostor trials. Spoofing an ASV system via replay attempt requires speech recordings from the target claimants only. Hence, number of replay trials are equal to number of target genuine trials.

The replay speech samples are generated manually by replaying the original data through a high quality CREATIVE-SBS-A35 loudspeaker (frequency response 100-15000Hz) al-

Table 1: *Summary of the developed IITG-MV replay dataset used in this work for genuine, impostor and replay trials.*

Statistics	Male	Female	Total
Background speakers	05	06	11
Target Speakers	40	25	65
Genuine trials	706	568	1274
Impostor trials	27534	13632	41166
Replay trials (Targets only)	706	568	1274

most in acoustically controlled environment (i.e. inside closed room with no fan and air condition noise) and re-recorded through an in-built microphone of HD Webcam C270-Logitech at the sampling rate of 16000 samples/second. We put very careful effort in acquiring the good quality replay speech samples in order to provide more challenging scenario. To verify the quality of the replay data, the original and replay recordings are played in front of few participants. They hardly differentiate between them, ensuring the quality of the replay data.

The quality of the replay recordings can also be verified by estimating the distortion between actual and corresponding replay recordings using cepstral distance method [23]. Cepstral distance (CSD) represents the average Euclidean distance between the two recordings and is estimated using standard short-term cepstral analysis with hamming window of duration 20ms and 10ms overlap. The DC coefficient 'c_0' is omitted. Low CSD values characterize high-quality replay recordings. The mean and standard deviation of CSD values, estimated for whole 1274 trials (males and females) are given in Table 2. Table 8 also contains two additional columns, representing the CSD values for C1 and C3 out of six evaluation conditions (C1-C6) of ASVspoof2017 database (in [24], please refer Table 5 and Figure 2). It can be observed that CSD values for the trials of IITG-MV database are in closed matching with CSD values of the trials under either C1 or C3 evaluation conditions of ASVspoof2017 database. Although, both C1 and C3 are of low category but show wide variation in replay detection performance among top ten systems, thereby simulates challenging evaluation conditions. From this aspect, the developed IITG-MV replay database provides relatively homogeneous but challenging evaluation condition similar to either C1 or C3 category of ASVspoof2017 database. Hence, it can be considered as a useful database for spoofing and anti-spoofing studies on ASV systems in the context of replay attacks. In addition, it also facilitates the vulnerability study of ASV systems to replay attacks gender-wise.

3.2. Experimental Setup

Advanced modelling techniques such as, i-vector and DNN frameworks require large amount of data for training. In contrast, classical GMM-UBM [25] works satisfactorily at relatively small amount of training data, and also outperform i-vector particularly for unknown types of spoof attacks as reported in the study [26]. Moreover, present work is more related to exploration of discriminatory evidences at the feature level rather than the model level. Therefore, at this stage GMM-UBM seems to be good choice at model level to examine the strength of proposed features for replay detection task. In this work, a GMM-UBM ASV system is proposed which uses 39-dimensional (13 static, 13 delta and 13 delta-delta coefficients, excluding first energy coefficient) RMFCC and

Table 2: *Verification of developed IITG-MV replay database quality with respect to standard ASVspoof2017 database using following parameters: number of genuine and replay trials, mean and standard deviation of CSD between actual and replay recordings, and quality of playback and recording device (L=low, M=medium, H=high).*

Parameters	IITG-MV	ASVspoof2017 C1	ASVspoof2017 C2
Genuine trials	1274	1438	2363
Replay trials	1274	1438	2363
CSD(μ)	0.80	0.79	0.77
CSD(σ)	0.16	0.16	0.28
Playback device Quality	H	L	L/M
Recording device Quality	M	L/M	L/M

LPRHEMFCC features as a means of rejecting replay spoofing trials. A reference GMM-UBM ASV system is built using 39-dimensional standard MFCC feature to evaluate the robustness of both source features for ASV anti-spoofing in the form of (source + MFCC) feature combination. Some important features extraction parameter details are: #mel-filters $= 24$, sampling frequency $(f_s) = 8kHz$, #DCT coefficients $= 24$, frame-size $= 20ms$, frameshift $= 10ms$, LP order $= 10$.

3.3. Evaluation Process

The SV performance is measured in terms of equal error rate (EER), where the false rejection rate (FRR) and false acceptance rate (FAR) are equal [27]. In false rejection, a genuine speaker is classified as an impostor while in false acceptance, an impostor is accepted as genuine speaker. Replay attackers usually targets the enrolled speakers to spoof ASV system. Thus, under replay attacks scenario FAR is more relevant measuring parameter for evaluating the system performance. Accordingly, we have used two metrics: zero-effort false acceptance rate (ZFAR) and replay attack false acceptance rate (RFAR). ZFAR and RFAR is related to zero-effort impostor trials and replay trials, respectively. The EER or equivalently the ZFAR is computed by pooling all genuine and zero-effort impostor trials together. We call it as the baseline performance of the ASV system. Under replay spoofing, all the target trials by actual and replay speech are considered as genuine speaker trials and impostors, respectively. The RFAR is computed using the target trials by replay speech. The RFAR is measured based on the fixed threshold (at EER point) of the baseline systems. As same baseline ASV system is used for both ZFAR and RFAR computation, the difference '*RFAR-ZFAR*' directly indicates system vulnerability to replay attacks [2]. In positive sense, it represents ASV system capability to resist spoof attacks. A smaller value of '*RFAR-ZFAR*' indicates better replay detection accuracy. Moreover, since same ASV system is used for both baseline and spoofing tests, the scores and decisions for all genuine trials will remain unaffected. Consequently, the FRR will remain constant, under both conditions. Altogether, ZFAR, RFAR and their difference '*RFAR-ZFAR*' can be used as performance metrics to compare different ASV systems under replay attacks.

3.4. Experimental Results and Discussion

Table 3 shows stand-alone performance of MFCC, RMFCC and LPRHEMFCC features based ASV systems as well as joint performance of 'RMFCC+MFCC' and 'LPRHEMFCC+MFCC'

Table 3: *ZFAR(%) and RFAR(%) results for different features based GMM-UBM ASV system. In case of zero-effort imposter trials the performance is expressed in terms of ZFAR. Under replay attacks the performance is expressed in terms of RFAR. The word 'Difference' stands for 'RFAR-ZFAR'.*

1. Male			
Features	**ZFAR**	**RFAR**	**Difference**
MFCC	2.97	38.81	35.84
RMFCC	5.38	15.72	10.34
LPRHEMFCC	12.62	7.93	4.69
RMFCC+MFCC	**2.97**	**29.46**	**26.49**
LPRHEMFCC+MFCC	**2.97**	**16.71**	**13.74**
2. Female			
Features	**ZFAR**	**RFAR**	**Difference**
MFCC	3.69	65.14	61.45
RMFCC	5.46	51.40	45.94
LPRHEMFCC	10.78	22.00	11.22
RMFCC+MFCC	**3.69**	**61.44**	**57.75**
LPRHEMFCC+MFCC	**3.69**	**55.28**	**51.59**
3. Whole-set			
Features	**ZFAR**	**RFAR**	**Difference**
MFCC	4.24	54.08	49.84
RMFCC	5.65	31.16	25.51
LPRHEMFCC	13.20	8.00	5.20
RMFCC+MFCC	**3.80**	**41.20**	**37.51**
LPRHEMFCC+MFCC	**4.47**	**32.03**	**27.56**

combinations, under both baseline and replay spoofing test conditions. With reference to corresponding ZFAR performances, the features MFCC and RMFCC show higher value of RFAR, indicating considerable degradations in their performances under replay attacks in all cases i.e. males, females and whole-set. However, opposite patterns are shown by LPRHEMFCC feature for males and whole-set case, but it can be potentially acceptable from spoofing point of view. Further, degradations are relatively more in case of the female speakers as they may have less spectral distortion than the male speakers.

The contribution of this work is better reflected in case of combination of source features with MFCC. The baseline ZFAR performances are same for both the combinations for males and females cases, and almost same for whole-set case. Therefore, under replay attacks RFAR and 'RFAR-ZFAR' performances for these combinations are directly comparable. Under replay attacks, it is clearly observed that 'LPRHEMFCC+MFCC' combination outperforms 'RMFCC+MFCC'combination by notable margin in all cases. Thus, from the experimental results given in Table 3, it can be easily concluded that Hilbert envelope of the LP residual provides better modelling of excitation source for replay attacks detection than the LP residual.

4. Replay detection experiments using ASVspoof2017 database

In this section, three stand-alone and two fused replay attacks detection systems are developed as shown in Table 4. The developed systems are trained using training-set and tested on development-set and evaluation-set of standard ASVspoof2017 database. ASVspoof2017 database is a sub-part of original *Red-Dots* corpus. Training, development and evaluation sets consist 3016, 1710 and 13306 speech files, respectively. The speech files and corresponding replay recordings are collected at sampling rate 16000 samples per second and 16-bit resolution per sample. The features extraction process is same as discussed in the preceding Section 3.2. However, speech files are processed at their original 16kHz sampling rate without down-sampling. Accordingly, LP order ($p = 18$) is used to get LP residual from speech signal [17]. GMM-classier is used to discriminate between actual and replay speech samples.

Table 4: *EER(%) Results of the stand-alone and fused replay attacks detection systems on pooled ASVspoof2017 database.*

System	**Features**	**EER**	
		dev	**eval**
S1	MFCC	16.32	35.03
S2	RMFCC	20.33	29.33
S3	LPRHEMFCC	10.86	30.59
S1+S2	**RMFCC+MFCC**	**16.05**	**29.29**
S1+S3	**LPRHEMFCC+MFCC**	**7.25**	**28.42**
B02 [24]	CQCC	10.35	30.60

From the EER results shown in Table 4, system fused system (S1+S3) outperforms system (S1+S2). This confirms the higher potential of Hilbert envelope of the LP residual in rejecting replay spoofing trials over the LP residual signal.

5. Conclusions

This work demonstrates the effectiveness of LPRHEMFCC over RMFCC features in rejecting replay spoofing trials through SV experiments conducted on self-developed IITG-MV replay database. It has been observed that the combination of LPRHEMFCC+MFCC provides better results as compared to the combination of RMFCC+MFCC, under replay attacks scenario. Similar pattern of results are also obtain in spoof detection experiment conducted on ASVspoof2017 database, and hence validates the trueness of experimental outcome obtained on IITG-MV database. Significance difference in the replay detection performance has been observed in case of development-set for system 'S1+S3' (EER$= 7.25\%$) over system 'S1+S2' (EER$= 16.05\%$) . However, in case of evaluation-set the difference is relatively very low. Further, fused system S1+S3 shows a relative improvement nearly $\sim 7\%$ over baseline CQCC (B02) system in case of evaluation-set and hence, required further efforts to obtain higher level performance. Future plan is explore Hilbert phase and advance modelling technique such as i-vector and DNN techniques for notable enhancement in the performance under highly varying acoustic replay attacks conditions.

6. Acknowledgements

This work is funded by Department of Electronics and Information Technology (DeitY), Govt. of India under the project title "Development of Excitation Source Features Based Spoof Resistant and Robust Audio-Visual Person Identification System". The research work is carried out in Speech Processing and Pattern Recognition (SPARC) laboratory at National Institute of Technology Nagaland, Dimapur, India.

7. References

[1] J. P. Campbell, Jr., "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sept. 1997.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and counter measures for speaker verification: A survey," *Speech Comm. (Elsevier)*, vol. 66, pp. 130–153, Feb. 2015.

[3] J. Villaba and E. Lieida, "Preventing replay attacks on speaker verification systems," in *Int. carnahan conf. on security technology (ICCST)*, October 2011, pp. 1–8.

[4] Z. Wang, G. Wei, and Q. H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," *proc. IEEE Int. conference of the biometrics special interest Group (BIOSIG) on machine learning and cybernetics*, pp. 1708–1713, 2011.

[5] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," *in proc. Interspeech*, pp. 27–31, 2017.

[6] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," *in proc. Interspeech*, pp. 97–101, 2017.

[7] K. Raju Alluri and A. K. V. Gangashetty, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," pp. 107–111, 2017.

[8] H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, "Novel variable length teager energy separation based instantaneous frequency features for replay detection," *in proc. Interspeech*, pp. 12–16, 2017.

[9] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," *in proc. Interspeech*, pp. 22–26, 2017.

[10] R. Font, J. M. Espın, and M. J. Cano, "Experimental analysis of features for replay attack detection–results on the asvspoof 2017 challenge," *in proc. Interspeech*, pp. 7–11, 2017.

[11] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of crim system for the automatic speaker verification spoofing and countermeasures challenge 2015," *in proc. Interspeech*, pp. 2072–2076, Sept. 2015.

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

[13] H. N. Bhavsar, T. B. Patel, and H. A. Patil, "Novel nonlinear prediction based features for spoofed speech detection." in *INTERSPEECH*, 2016, pp. 155–159.

[14] C. Hanilçi, "Linear prediction residual features for automatic speaker verification anti-spoofing," *Multimedia Tools and Applications*, pp. 1–13, 2017.

[15] J. Eargle, *Loudspeaker handbook*. Springer Science & Business Media, 2013.

[16] F. Rumsey and T. McCormick, *Sound and recording: applications and theory*. CRC Press, 2014.

[17] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[18] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Commun.*, vol. 48, pp. 1243–1261, Jun. 2006.

[19] V. C. Raykar, B. Yegnanarayana, S. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 751–761, 2005.

[20] R. K. Das and S. Mahadeva 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.

[21] Haris B. C., G. Pradhan, S. R. M. Prasanna, R. K. Das, and R. Sinha, "Multivaribility speaker recognition database in indian scenario," *Int. J. of Speech Technology (Springer)*, vol. 15, no. 4, pp. 441–453, March 2012.

[22] A. Larcher, K. A. Lee, B. Ma, and H. Li, "RSR2015: Database for text-dependent speaker verification using multiple pass-phrases," in *proc. of Interspeech*, 2012.

[23] N. Nocerino, F. Soong, L. Rabiner, and D. Klatt, "Comparative study of several distortion measures for speech recognition," in *Proc. ICASSP*, vol. 10, 1985, pp. 25–28.

[24] 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 proceedings of INTERSPEECH*, 2017.

[25] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, vol. 10, 2000, pp. 19–41.

[26] C. Hanilçi, T. Kinnunen, Tomi, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: A comparison," in *proc. Interspeech. ISCA*, 2015, pp. 2057–2061.

[27] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *proc. Eur. conf. on speech communication technology, Rhodes, Greece*, vol. 4, 1997, pp. 1895–1898.