Interspeech 2018
2-6 September 2018, Hyderabad

Novel Linear Frequency Residual Cepstral Feature For Replay Attack Detection

Hemlata Tak and Hemant A. Patil

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India

{hemlata_tak, hemant_patil}@daiict.ac.in

Abstract

Replay attack poses the most difficult challenge for the development of countermeasures for spoofed speech detection (SSD) system. Earlier researchers mainly used vocal tract-based (segmental) information for replay detection. However, during replay, excitation source-based information also gets affected (in particular, degradation in pitch source harmonics at higher frequency regions) due to recording environment and replay devices. Hence, in addition to the vocal tract-based system information, we have also explored the excitation source-based informations for SSD. In particular, we have used Linear Frequency Residual Cepstral Coefficients (LFRCC) for replay detection. The objective of this paper is to explore possible complementary excitation (glottal) source information present in the Linear Prediction residual-based features. Experiments performed on the ASV Spoof 2017 Challenge database with Gaussian Mixture Model (GMM) and Convolutional Neural Network (CNN) classifiers. When we combined the source and system-based information, we obtained on an average 28.77% and 42.72% relative decrease in Equal Error Rate (EER) on development and evaluation set, respectively. Furthermore, when we perform score-level fusion of feature sets (for a fixed classifier) followed by a classifier-level fusion of GMM and CNN (for a fixed feature set), we obtained reduced EER of 2.40% and 9.06% on dev and eval set, respectively.

Index Terms: Replay, Linear Prediction Residual, Convolutional Neural Network.

1. Introduction

Automatic Speaker verification (ASV) system is a biometric system that verifies speaker's claimed identity from his or her voice with the help of machines [1]. We would like ASV system to be robust against various variations (such as microphone and transmission channel, intersession, acoustic noise, etc). This robustness makes ASV system to be more *vulnerable* to various spoofing attacks as it tries to nullify these effects. Hence, we would like the system to be secure against spoofing attacks. The various types of spoofing attacks included in the literature are impersonation, replay, speech synthesis (SS), voice conversion (VC) and twins [1–4]. Replay attack is a pre-recorded speech samples of a target speaker used to get the access of a system [4]. The ASV Spoof 2017 Challenge was mainly focused on the development of robust countermeasure with the capability of detecting various replay spoofing attacks in all the unseen conditions. The challenge organizers provided the baseline system Constant Q Cepstral Coefficients (CQCC) with Gaussian Mixture Model (GMM) as a classifier [5]. Various countermeasures were proposed for detecting the replay spoofed speech in recent ASV spoof 2017 Challenge [6]. Some of the countermeasures focused on the normalization techniques [7], Instantaneous Frequency (IF)-based features were explored in [8, 9]. The high-resolution temporal-based features (such as single frequency filtering (SFF)) were used in [10]. The high frequency-band selection in CQCC feature set also performed better compared to using the full-band CQCC feature set [11]. Some of the deep learning methods were also studied in [12–14]. For replay spoofed speech detection (SSD), phase variations are not as much prominent. The earlier study used high-dimensional feature sets, such as Log Magnitude Spectrum (LMS) and Residual Log Magnitude Spectrum (RLMS) derived from the magnitude spectrum of the speech signal to detect SS and VC spoofing attacks [15]. The LMS captures the information from the magnitude spectrum, such as pitch (F_0), formants and harmonics of a speech signal. The formants (especially lower) are important for Automatic Speech Recognition (ASR), however, not much useful for spoofing detection and hence, RLMS feature set was derived from the Linear Prediction (LP) residual of speech to suppress the effect of formants [15]. The excitation source is found to contain speaker-specific information [16], [17]. Motivated by this study, several methods for modeling the speaker-specific information from the source has been proposed [18–21]. The LP residual is one of the techniques that captures more speaker-specific information from the excitation source for replay SSD task [15, 21, 22].

The key idea in our paper is to exploit flat *vs.* degraded spectral characteristics of LP residual for natural and replay speech, respectively. In particular, due to bandpass nature of frequency response characteristics of replay devices, microphone, loudspeaker and acoustic environment, the LP residual spectrum of replayed speech is expected to experience degradation in specific frequency regions dictated by the replay mechanism w.r.t given hardware. However, along with vocal tract-based system information, we propose the source-based information for the replay SSD. We have proposed the LP residual-based Linear Frequency Residual Cepstral Coefficients (LFRCC) feature set. In cepstral-domain, the excitation source information is obtained from the short-time magnitude spectrum of LP residual magnitude of speech signal [18], [19]. The proposed feature set is extracted by using linearly-spaced triangular filterbank with pre and post-processing techniques (i.e., pre-emphasis filter and Cepstral Mean Normalization (CMN)). For the classification task, we have to used GMM and Convolutional Neural Network (CNN) as a classifier. The proposed feature set is compared with CQCC, MFCC, and Linear Frequency Cepstral Coefficients (LFCC) feature sets and then we combined the excitation source and vocal tract-based system information to improve the performance of SSD system.

2. Linear Prediction (LP) Residual

The LP technique was originally used in system identification literature followed by its novel application to speech cod-

10.21437/Interspeech.2018-1702

ing, where a new technique called as Linear Predictive Coding (LPC) was developed [23]. In the LP analysis, each speech sample is represented by a linear weighted sum of past 'p' speech samples, where p represents the order of linear predictor and weights are called as Linear Prediction coefficients (LPCs) [23]. If $s(n)$ is the current speech sample, then the predicted sample is represented as:

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

where a_k are the LPCs . The difference between the actual speech sample ($s(n)$) and the predicted samples ($\hat{s}(n)$) is known as LP residual, i.e., $r(n)$ and is computed by using Eq. (2) [23]:

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

The LP residual signal is generated by the inverse filtering operation of the speech signal using LP analysis, i.e.,

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}, \quad (3)$$

where $A(z)$ is an inverse filter corresponding to all-pole LP filter $H(z)$, that represents the vocal tract-based system information [24]. The excitation source-based information complements the vocal tract-based system information. The LP residual signal mostly contains the excitation source information. The source information present in the LP residual depends on the prediction order. The earlier study also shows that the LP residual extracted using prediction order in the range of 8-20 (for 8 kHz sampling frequency) is the best choice for proper representation of excitation source information [21]. To extract the excitation source information, the LP residual can be represented in different domains, such as frequency, cepstral or joint time-frequency [25]. In this work, the LP residual is processed in the cepstral-domain for extracting the speaker-specific excitation source information. In the cepstral-domain, the excitation source information is obtained from the short-time LP residual magnitude spectrum of speech signal [18], [19]. LP residual coveys more information about the excitation source and hence, we have used the LP residual-based features for the replay SSD task [26].

3. Proposed Feature Set

The proposed feature extraction framework is shown in Figure1. The speech signal is passed through a pre-emphasis highpass filter to emphasize the high frequency components. The higher formants (F_3 and F_4) explicitly used for speaker discrimination are present in the higher frequency range and hence, features from these high frequency regions are more important for replay SSD task [27]. The pre-emphasis is a technique for balancing the lower and higher frequency components [27]. The system function of pre-emphasis is $H(z) = 1 - \alpha z^{-1}$, where α (i.e., filter coefficient) = 0.97 [27].

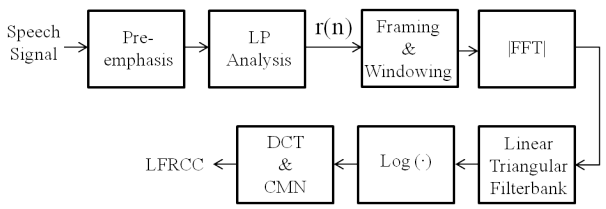


Figure 1: *Schematic block diagram of proposed LFRCC feature extraction.*

The pre-emphasized speech signal is passed through LP analysis block to obtain LP residual (r(n)) waveform. Further, frame blocking and windowing is applied on LP residual waveform over a short duration of 25 ms with 10 ms frame shift. After that, the power spectrum is computed for each LP residual frames. This power spectrum is passed through 40 linearly-spaced triangular filterbanks to obtain the filterbank energies. The linear triangular filterbank, bandwidth is equally distributed throughout the entire available frequency range (for a given sampling frequency) that makes it more reliable to extract the features. To decorrelate the feature set, Discrete Cosine Transform (DCT) is used to compute low-dimensional feature representation. After DCT, we retain only few initial coefficients that are further post-processed with CMN technique to reduce the channel mismatch distortion effects [28]. The basic principle behind CMN is based upon the behavior of the cepstrum under the convolution distortions [29], [30]. It has been observed that by using CMN technique results are improved [7]. CMVN has been also found useful during ASV Spoof 2017 Challenge campaign [31]. Furthermore, to capture *transitional* information across feature vectors, static features are appended with their Δ and $\Delta\Delta$ features.

3.1. Spectrographic Analysis

The spectral energy density of Mel spectrogram (Panel I), RLMS using Mel (Panel II) and linear (Panel III) triangular filterbank for natural and replayed speech is shown in Figure 2(a) and Figure 2(b), respectively. The formants are clearly visible with Mel spectrogram, while they do not appear clearly with RLMS using Mel filterbank. In the LP residual spectrum, the possible speaker-specific excitation source information is represented by the *harmonic* structure and hence, the formants are not visible in the spectral density obtained from RLMS [25]. In this work, linearly-spaced filterbanks are more significant than the nonlinearly-spaced filterbanks (such as Mel triangular filterbank). The difference between Figure 2 (a) and Figure 2 (b) of Panel II is that in the replayed speech, the LP residual spectrum is relatively less dense or lesser spectral energy density than of natural speech signal. Furthermore, when the spectral energies of Panel I are compared the, replayed speech signal have more information in the higher frequency regions. There is also some noticeable difference between the LP residual spectrum in Panel III, for replayed speech. In particular, the LP residual spectrum is relatively more attenuated (less dense or less spectral energy density) in higher frequency regions than the LP residual spectrum of natural speech (highlighted by the dotted circle). This *decay* in a spectrum is mainly due to the bandpass filter characteristics of an impulse response of recording studio and the strong attenuation of the loudspeaker (which is used in the smartphone that can serve as replay device).

4. Experimental Setup

The ASV spoof 2017 Challenge dataset mainly focus on the RedDots corpus, and its replayed speech [32]. The detailed statistics of the database is provided in [6]. The proposed feature set, i.e., LFRCC is extracted from linear triangular filterbank using predictor order '$p = 8$' with 25 ms window duration and 10 ms frame shift. The linear triangular filterbank is computed with the frequency range from $F_{min} = 0$ Hz and $F_{max} = 8000$ Hz with 40 subbands filtered signal. The feature parameters used for LFRCC are *120*-D (static+Δ+$\Delta\Delta$). We have compared our proposed feature set with CQCC, MFCC and LFCC feature sets with the feature dimension of *90*-D, *39*-

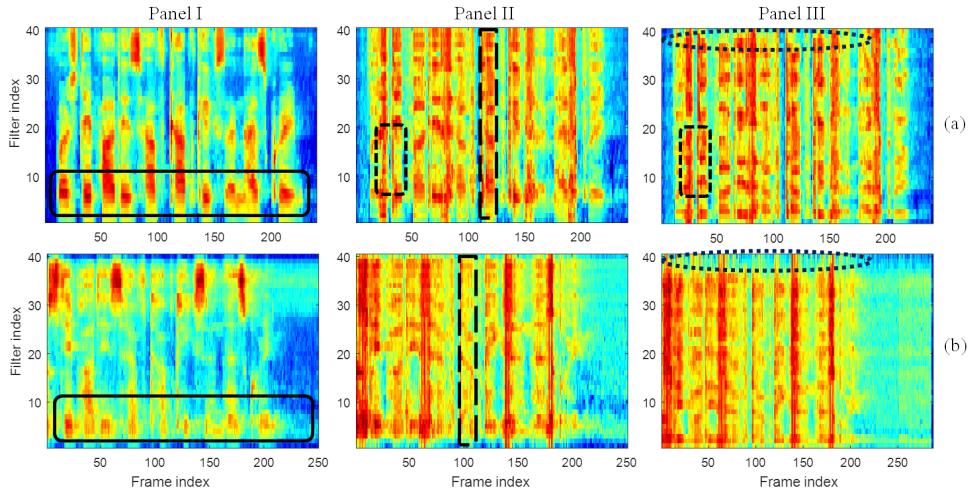


Figure 2: *Comparison of spectral energy density obtained from the Mel spectrogram (Panel I), RLMS using Mel triangular filterbank (Panel II) and RLMS using linear triangular filterbank (Panel III) of an utterance, 'Actions speak louder than words'. Spectral energy density for (a) natural speech signal, and (b) for its replayed version.*

D, and *120*-D, respectively. MFCC and LFCC feature sets are extracted by using 40 subband (Mel and linear triangular) filterbanks. We have used GMM classifier with 512 Gaussian mixture components and CNN as a classifier to obtain the training models to classify the natural and spoofed speech. Final scores are represented by Log-Likelihood Ratio (LLR). To decide whether the test speech is natural or spoofed, LLR scores are used. In particular,

$$LLR = \log \frac{P(X|H_0)}{P(X|H_1)}, \quad (4)$$

where $P(X|H_0)$ and $P(X|H_1)$ are the likelihood scores of natural and replay utterances (with hypothesis H_0 and H_1), respectively. To obtained possible complementary information of the proposed LFRCC feature set, score-level fusion is performed with CQCC, MFCC and LFCC feature sets as per given Eq. (5):

$$LLK_{fused} = \alpha LLK_{feature1} + (1-\alpha) LLK_{feature2}, \quad (5)$$

where $LLK_{feature1}$ is a log-likelihood score of either CQCC or MFCC or LFCC, whereas $LLK_{feature2}$ represent the score of the LFRCC feature set. The fusion parameter (α) lies between $0 < \alpha < 1$ to decide the relative weight of scores. The performance of the system is measured using Equal Error Rate (EER) in % and Detection Error Trade-off (DET) curve based on LLRs of natural and spoofed speech [33].

4.1. Convolutional Neural Network (CNN)

In this work, we have used the CNN as a classifier along with GMM classifier for SSD task. We have used the same CNN architecture proposed in [34]. This architecture consists of the three convolutional layers followed by a max-pooling layer and three fully connected (FC) layers with a softmax layer. A max-pooling layer is used to downsample the data across the spatial dimension. The input given to the CNN architecture is a 2-D image of size $d \times N$, where d represents the dimension of a feature set and N represents the number of frames per second (*N=100*). Similar to proposed CNN architecture, the first three convolutional layers have a filter/kernel size of [$d \times 3$, 1×3, 1×3] dimension, respectively. Each convolutional layer has *128* subband filters. The fourth layer is a max-pooling layer used with 1×2 stride on the output of the third convolutional layer. The last three FC layer with *256* units (neurons) are used for computing the final score. We have used dropout of 0.5 to all the three FC layers to reduce the effect of overfitting. Furthermore, we have used a softmax layer as an output layer for SSD task. We have used 90 % overlapping data, which makes the network to learn the data-dependency accurately. To train the network, we used 64 batch size with ReLU activation function. In addition, Adam optimizer was used to train the network for 100 epochs. The CNN model was implemented using TensorFlow library [35].

5. Experimental Results

5.1. Results with GMM and CNN Classifier

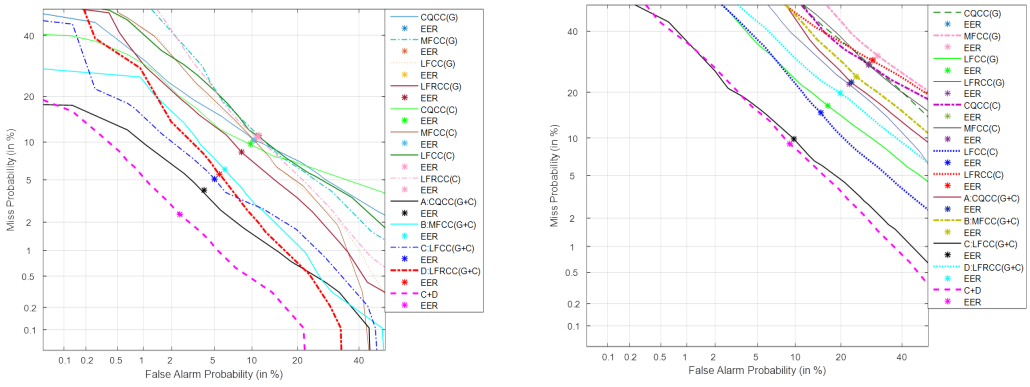
Table 1 shows the results (in % EER) for feature sets with two classifiers, namely, GMM and CNN. We have compared our proposed feature set with the baseline feature, namely, CQCC, LFCC, and MFCC. The baseline system (CQCC) gave an EER of 10.21 % and 28.48 % on dev and eval set, respectively. For MFCC and LFCC feature set, the EER obtain on dev set is 11.21 % and 10.58 %, Whereas 31.30 % and 16.62 % on the eval set. The proposed feature set LFRCC, show the significant reduction in EER (8.38 %) on dev set. However, for eval set the reduction in EER (22.28 %) is not more significant. Similarly, with CNN classifier the baseline gave the lower EER of 10.00 % and 28.42 % on the dev and eval set, respectively. The proposed feature set gave an EER of 10.78 % (dev) and 29.81 % (eval). With CNN classifier, LFCC feature set, show the significant reduction in EER on eval set is 15.10 %.

Table 1: *Results with GMM and CNN classifier*

Feature Set	**GMM**		**CNN**	
	Dev	Eval	Dev	Eval
CQCC	10.35	28.48	**10.00**	28.42
MFCC	11.21	31.30	10.52	28.35
LFCC	10.58	**16.62**	11.05	**15.10**
LFRCC	**8.38**	22.28	10.78	29.81

5.2. Results with Score-level Fusion

To investigate the possible complementary information captured by various feature sets, we have used their score-level



(a) (b)

Figure 3: *DET curves for various replay SSD systems. (a) the individual DET curve of CQCC, MFCC, LFCC and LFRCC, their classifier-level fusion (A, B, C, D) and score-level fusion (C+D) on dev set. (b) DET curves for same feature sets, their similar classifier-level fusion and score-level fusion on eval set.*

fusion as per given Eq. (5) (i.e., for a given classifier, scores from two feature sets are fused). Results of score-level fusion with both GMM and CNN classifiers are shown in Table 2. The fused scores of LFRCC and MFCC have show the significant reduction in % EER on the dev set (6.20 %) with GMM classifier. While with CNN classifier, EER reduced to 6.45 % on dev set. However, for eval set, the reduction in % EER of LFRCC and LFCC is 14.31 % and 15.00 % with GMM and CNN classifiers. Thus, score-level fusion indeed helps to reduce the EER for the proposed feature set.

Table 2: *Results with score-level fusion*

Feature Set	GMM		CNN	
	Dev	Eval	Dev	Eval
LFRCC + CQCC	06.67	19.68	07.52	25.26
LFRCC + MFCC	**06.20**	21.77	**06.45**	26.15
LFRCC + LFCC	06.48	**14.31**	07.02	**15.00**

+ : Score-level fusion

5.3. Results with classifier-level fusion

The results of the classifier-level fusion (i.e., for a given feature set, scores from GMM and CNN classifiers are fused) for CQCC, MFCC, LFCC and proposed feature sets are shown in Table 3. The classifier-level fusion helps to reduce the EER more than the individual systems for all the feature sets. The

Table 3: *Results with classifier-level fusion (GMM+CNN) for a given feature set*

Feature Set	CQCC (A)	MFCC (B)	LFCC (C)	LFRCC (D)
Dev	**04.06**	06.14	05.05	05.57
Eval	22.96	24.46	**09.80**	19.82

EER with classifier-level fusion reduced to 4.06 %, 6.14 %, 5.05 % and 5.57 % for CQCC, MFCC, LFCC and LFRCC, respectively. On eval set, it is reduced to 22.96 % (CQCC), 24.46 % (MFCC), 9.80 % (LFCC) and 19.82 % (LFRCC). With CQCC feature set, we obtained the lower EER of 4.06 % (dev) and 22.96 % (eval) whereas with LFCC feature set, we obtained an EER of 5.05 % (dev) and 9.80 % (eval). For each feature set, scores are obtained using GMM and CNN classifiers followed by their classifier-level fusion. Finally, the resultant relative scores from each feature sets (i.e., A, B, C and D shown in Table 3) are further fused at the score-level in Table 4. The

Table 4: *Results with score-level fusion followed by the classifier-level fusion*

Feature Set	Dev	Eval
A+D	01.83	18.24
B+D	**01.69**	18.01
C+D	02.40	**09.06**

A, B, C and D as per Table 3.

lower EER obtained using score-level fusion (B+D) is 1.69 % (dev) and 18.01 % (eval) and score-level fusion (C+D) gave an EER of 2.40 % (dev) and 9.06 % (eval). The performance is also shown by DET curves of various feature sets, such as CQCC, MFCC, LFCC, and LFRCC with both GMM and CNN classifiers, their classifier-level and score-level fusion in Figure 3 (a) for dev set and in Figure 3 (b) for eval set. On dev set, score-level fusion (C+D) is clearly distinct at *all* the operating points of the DET curve. On eval set, score-level fusion (C+D) have significantly a lower false alarm and miss probabilities in DET curve as compared to the CQCC, MFCC and LFCC feature sets.

6. Summary and Conclusions

In this study, along with vocal tract-based system information, we have proposed the use of excitation source information by processing the LP residual signal using linear triangular filterbank. The objective of this work is to exploit possible complementary information present in the LP residual-based feature set which indeed helps in improving the performance of replay SSD. We have also shown the significance of linear triangular filterbank and Mel triangular filterbank during feature extraction. In linear filterbank, bandwidth is equally distributed throughout all the frequency components that makes it more reliable to extract the features. By combining the system and excitation source-based information, the performance of the combined system is improved over the individual system. Moreover, the results obtained with their classifier-level fusion indeed help to reduce the EER from the individual EER of all the feature sets. Our future work includes exploring Long-Term Prediction (LTP) and Non Linear Prediction (NLP) residual-based feature for replay SSD task.

7. Acknowledgements

The authors would like to thank Ministry of Electronics and Information Technology (MeitY) and authorities of DA-IICT Gandhinagar.

8. References

[1] 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.

[2] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, and S. King, "SAS: A speaker verification spoofing database containing diverse attacks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, 2015, pp. 4440–4444.

[3] M. Pal and G. Saha, "On robustness of speech based biometric systems against voice conversion attack," *Applied Soft Computing*, vol. 30, pp. 214–228, 2015.

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

[5] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[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*, Stockholm, Sweden, 2017, pp. 1–6.

[7] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection results on the ASVspoof 2017 challenge," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 7–11.

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

[9] 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 *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 12–16.

[10] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 107–111.

[11] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high frequency features," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 27–31.

[12] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 82–86.

[13] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 102–106.

[14] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, "Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 17–21.

[15] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASV spoof 2015 challenge," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2052–2056.

[16] B. S. Atal, "Automatic speaker recognition based on pitch contours," *The Journal of the Acoustical Society of America (JASA)*, vol. 52, no. 6B, pp. 1687–1697, 1972.

[17] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, no. 10, pp. 782–796, 2008.

[18] P. Thévenaz and H. Hügli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Communication*, vol. 17, no. 1-2, pp. 145–157, 1995.

[19] S. Hayakawa, K. Takeda, and F. Itakura, "Speaker identification using harmonic structure of LP-residual spectrum," in *International Conference on Audio and Video-Based Biometric Person Authentication (AVBPA)*, Crans-Montana, Switzerland, 1997, pp. 253–260.

[20] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[21] S. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

[22] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *in IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, 1999.

[23] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[24] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America (JASA)*, vol. 55, no. 6, pp. 1304–1312, 1974.

[25] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Salt Lake City, Utah, USA, 2001, pp. 409–412.

[26] C. Hanilçi, "Speaker verification anti-spoofing using linear prediction residual phase features," in *IEEE European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, 2017, pp. 96–100.

[27] L. Deng and D. O'Shaughnessy, *Speech Processing – A Dynamic and Optimization-Oriented Approach*. 1^{st} Edition, Marcel Dekker Inc., June 2003.

[28] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP)*, vol. 1, Hong Kong, china, 2003, pp. 656–659.

[29] A. A. Garcia and R. J. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP)*, vol. 1, Phoenix, Arizona, USA, 1999, pp. 325–328.

[30] M. Westphal, "The use of cepstral means in conversational speech recognition," in *the European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, 1997, pp. 1143–1146.

[31] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans *et al.*, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," *Submitted in The Speaker and Language Recognition Workshop (Speaker Odyssey)*, 2018.

[32] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The RedDots data collection for speaker recognition." in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2996–3000.

[33] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of decision task performance," in *the European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, 1997, pp. 1895–1898.

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

[35] "Tensorflow," URL: https://www.tensorflow.org/, {Avaiable online; Last Accessed: 14 March, 2018}.