Linear Frequency Residual Cepstral Features for Replay Spoof Detection on ASVSpoof 2019

Priyanka Gupta, and Hemant A. Patil
Speech Research Lab
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, India
Email: {priyanka_gupta, hemant_patil}@daiict.ac.in

***Abstract*—Playing a pre-recorded speech to gain illegal access to Automatic Speaker Verification (ASV) system is one of the easiest attacks to execute but difficult to detect. Such attacks are called as replay attacks. Designing robust ASV systems from such attacks motivates to explore signal processing framework for Spoof Speech Detection (SSD). This paper exploits excitation source-based information in the form of Linear Frequency Residual Cepstral Coefficients (LFRCC) feature set for SSD task. In the source-filter model of speech production, the excitation source is also known to contain speaker-specific information. In this context, the residual obtained from Linear Prediction (LP) of speech is exploited in cepstral domain to detect replay attack. Improvements in results are obtained by choosing appropriate order of LP, as the order of LP controls the amount of information carried by the residual. Experiments performed on ASVSpoof 2019 Physical Access (PA) dataset using Gaussian Mixture Model (GMM) and Convolutional Neural Network (CNN) show that the optimal LP order is 8 which gives EER on the evaluation set as 17.30% and 15.21% using GMM and CNN classifiers, respectively.**

***Index Terms*—Automatic Speaker Verification, Replay Attack, Linear Prediction Residual, Linear Prediction Order, Gaussian Mixture Model (GMM).**

I. Introduction

Automatic Speaker Verification (ASV) systems or voice biometric systems deal with verification of claimed identity of speakers (which can be genuine or impostor), with the help of machines [1]. Nevertheless, some impostors, other than zero-effort impostors, deliberately try to fool the ASV system in order to gain an unauthorized access. The deliberate attempts made by the impostor (i.e., attacker) are called attacks. Such illegal attempts which are at the microphone and transmission-level, are called as spoofing attacks on ASV systems [2]. Among all the spoofing attacks (such as speech synthesis, voice conversion, impersonation, twins, and replay [2]–[6]), replay attacks are the simplest to execute but hard to detect. Execution of replay attacks requires only a recording device to record the speech sample of the genuine user. The attacker then replays the pre-recorded speech later to fool the ASV system. Hence, no technical knowledge is required to mount this kind of attack, which makes it a significant threat to the security of ASV system. Therefore, it is important to develop robust Spoofed Speech Detection (SSD) systems that can effectively detect the presence of a spoofed input to the ASV system.

The excitation source-based information is also known to carry speaker-specific information [7]–[11]. The frequency response characteristics of microphone, replay device, and acoustic environment are bandpass in nature. Due to the bandpass nature, the spectrum of the LP residual of replay speech is expected to degrade for high frequency regions. The Linear Prediction (LP) residual is known to capture discriminating information for replay SSD task [12]–[15]. In this context, according to proposition by Mallat [16], a function $s(t)$ is bounded and k times continuously differentiable with bounded derivatives if

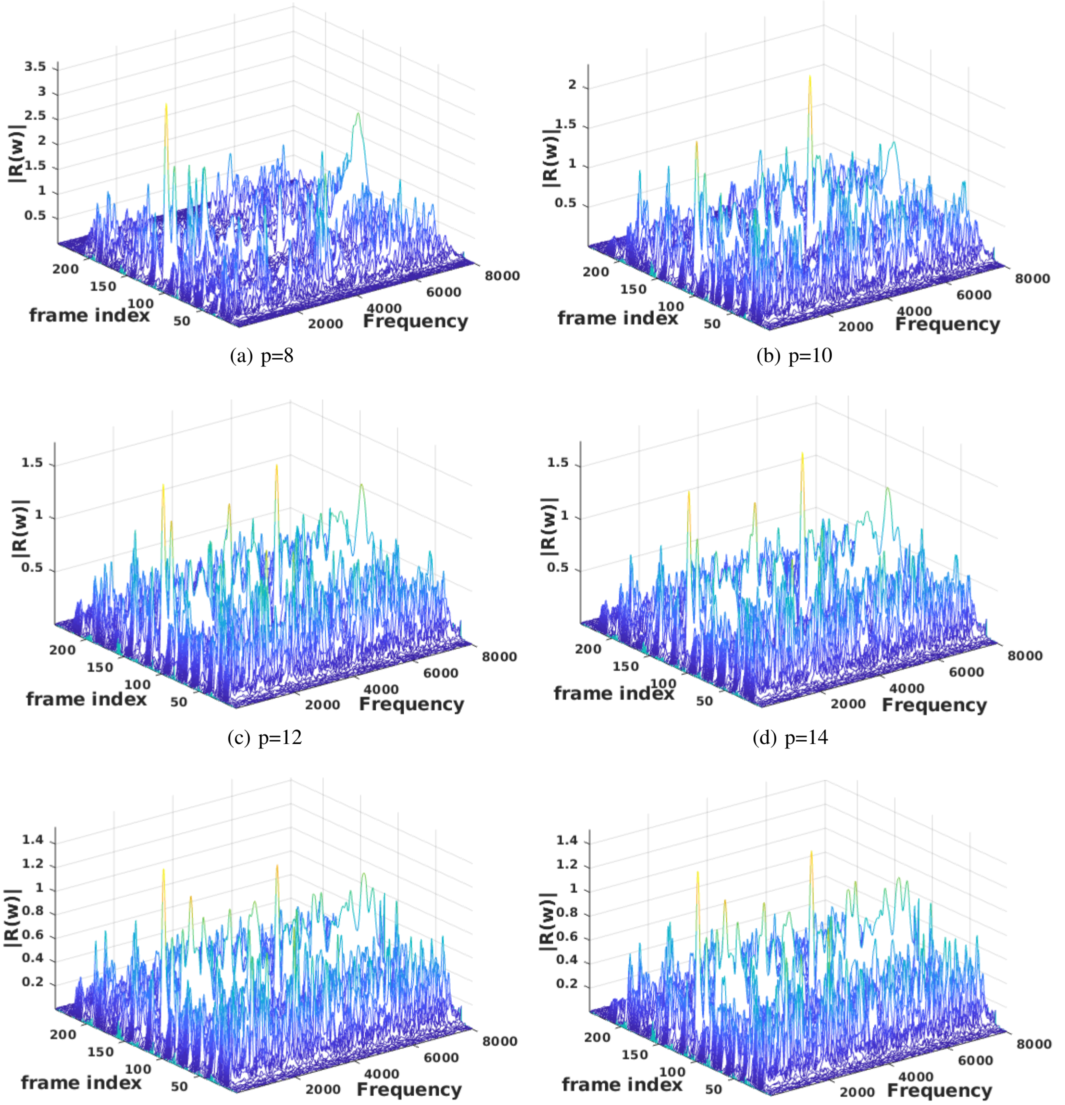
$$\int_{-\infty}^{+\infty} |S(\omega)|(1+|\omega|^k)d\omega < +\infty, \tag{1}$$

where $S(\omega) \in L^1(R) = \mathcal{F}s(t)$. It is known that the decay of spectrum $|S(\omega)|$ of a signal $s(t)$ depends on the worst singular behaviour [16]. For example, in replay speech, the replay noise has sudden discontinuities which are absent in genuine speech. Hence, the spectrum of replay speech is decaying in nature which is the discrimative cue for SSD task by the LFRCC feature set [17].

For SSD task, the frequency spacing at higher frequencies is sparse (such as in Mel-frequency warping). Therefore, to consider the effect of replay mechanism on higher frequency regions, we consider linear frequency scale in this work and exploit linear subband energies in this paper. Furthermore, this paper exploits recently proposed Linear Frequency Residual Cepstral Coefficients (LFRCC) feature set for ASV spoof 2019 PA dataset. Unlike [17], we have analyzed the effect of LP order on the residual. Furthermore, we have optimized the value of LP order to 8. First, the optimization is done empirically by framewise analysis of residual signal, and then followed by experimentation done by varying LP order, using traditional Gaussian Mixture Model (GMM)-based classifier, and state-of-the-art Convolutional Neural Network (CNN).

II. Linear Prediction (LP) Residual

Linear prediction of speech has been widely used in many applications, from speech coding to analyzing excitation source-based information. Each speech sample $\tilde{s}(n)$ is said to be equal to the weighted sum of past p samples, where p is the order of the linear predictor and the weights are called as



(a) p=8

(b) p=10

(c) p=12

(d) p=14

(e) p=16

(f) p=18

Fig. 1: Plots of framewise magnitude spectrum $|R(\omega)|$ of the LP residual of genuine speech, for different values of LP order p.

TABLE I: Log Spectral Distance (LSD) between LP residuals of speech signal (with $F_s = 16$ kHz) with various LP orders (p).

p	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**
2	0	1.35	2.19	2.46	2.64	2.81	2.97	3.06	3.17	3.23
4	1.35	0	1.42	1.74	1.97	2.16	2.34	2.44	2.55	2.61
6	2.19	1.42	0	0.79	1.08	1.31	1.51	1.63	1.75	1.82
8	2.46	1.74	0.79	0	0.63	0.95	1.20	1.33	1.47	1.54
10	2.64	1.97	1.08	0.63	0	0.62	0.94	1.09	1.24	1.32
12	2.81	2.16	1.31	0.95	0.62	0	0.62	0.82	1.00	1.10
14	2.97	2.34	1.51	1.20	0.94	0.62	0	0.47	0.71	0.83
16	3.06	2.44	1.63	1.33	1.09	0.82	0.47	0	0.49	0.65
18	3.17	2.55	1.75	1.47	1.24	1.00	0.71	0.49	0	0.38
20	3.23	2.61	1.82	1.54	1.32	1.10	0.83	0.65	0.38	0

Linear Prediction Coeffcients (LPCs), denoted by $\{\alpha_k\}_{k\in[1,p]}$.

Mathematically, this is represented as [15]:

$$\tilde{s}(n) = -\sum_{k=1}^{p} \alpha_k s(n-k). \tag{2}$$

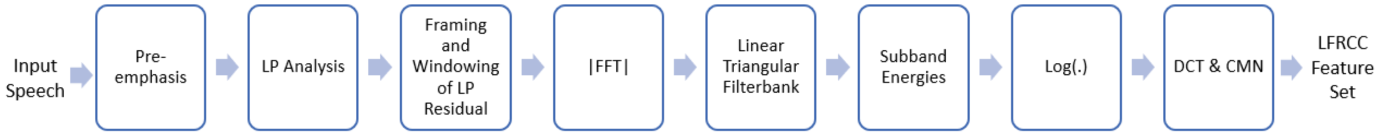


Fig. 2: Block diagram of LFRCC feature extraction. After [17].

The prediction error is called as the LP *residual* as shown in eq. (3). It carries excitation source component of the speech and it is given by [18].

$$r(n) = s(n) - \tilde{s}(n) = s(n) + \sum_{k=1}^{p} \alpha_k s(n-k). \qquad (3)$$

The LP residual is obtained by the all-pole inverse filter A(z) which is mathematically represented in eq. (4).

$$A(z) = 1 + \sum_{k=1}^{p} \alpha_k z^{-k}. \qquad (4)$$

Furthermore, replayed speech signal ($s_r(n)$) can be expressed as distributive property of convolution under the assumption of LTI system, i.e.,

$$s_r(n) = \left[-\sum_{k=1}^{p} \alpha_k s(n-k) + r(n) \right] * h_r(n), \qquad (5)$$

where ’*‘ indicates the convolution operation, and $h_r(n)$ is the impulse response of the playback device used for replay attack. Notably, the information carried by the LP residual also depends on the LP order, p. A large value of order will lead to good prediction of speech and hence, lower error (i.e., residual). However, for SSD task, our aim is not to have a good prediction of speech, rather to exploit the residual at an order optimally suited for SSD task. This is the novel aspect of our work. Figure 1 shows waterfall plot of the magnitude spectrum of the residual for varying order, p. It can been be observed that the plot has highest $|R(\omega)|$ for $p = 8$. For Figure 1, the speech sample taken into consideration had 16 kHz sampling frequency (F_s). This means that the optimum prediction would be achieved at $((F_s/1000)+2)$, i.e., at order p=18 [7]. However, for exploiting source-based information for SSD task, the residual should have more information. Hence, for $p = 8$, we can observe the optimal order for our purpose. In addition, the Table I shows Log Spectral Distance (LSD) between residuals of different LP orders. The LSD is estimated as [19]

$$LSD = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10log_{10}\frac{P(\omega)}{\tilde{P}(\omega)}\right]^2 d\omega}, \qquad (6)$$

where $P(\omega)$ and $\tilde{P}(\omega)$ denote the two power spectra between which LSD is estimated. The diagonal elements of Table I are zero because the LSD between two identical signals is zero. It can be observed that as we move from left to right in the Table I, the LSD keeps on increasing. This also means that the LP order p has a significant effect on the amount of information carried by the LP residual. Furthermore, the experimental results shown in the next Section also confirm our hypothesis that LP order of 8 is *optimal* for SSD task.

III. Experimental Setup

A. Dataset Used

Experiments are performed on the recently released and statistically meaningful ASVspoof 2019 Physical Access (PA) database, which is derived from VCTK corpus [20], [21]. It includes speech data from 107 speakers (46 males, 61 females). The dataset is divided into three subsets, namely, training, development, and evaluation which contain speech from 20 (8 male, 12 female), 10 (4 male, 6 female), and 48 (21 male, 27 female) speakers, respectively. The recording conditions of training, development, and evaluation dataset are identical. The spoof speech in training and development data is from *known* attacks, i.e., generated with identical algorithms. The evaluation set contains attacks generated with different algorithms (designated as unknown attacks). The detailed statistics of the dataset is given in [22].

B. Features and Classifier

In this work, recently proposed LFRCC feature set [17] is used to investigate the effect of LP order. Figure 2 shows the framework for LFRCC feature extraction. Pre-emphasis of the speech signal is done by passing it through a highpass filter, which emphasizes the high frequency regions in short-time speech spectrum. Then, LP analysis is performed to extract the LP residual. The residual signal then undergoes framing and windowing of a short segment of 25 ms with 10 ms frame shift. Furthermore, the power spectrum is estimated frame-wise which is further given to a linear triangular filterbank in order to obtain filterbank energies. Finally, Discrete Cosine Transform (DCT) is applied on the log-filterbank energies to get de-correlated and energy compact cepstral features. To reduce the distortions of the transmission channel, Cepstral Mean Normalization (CMN) is applied on the optimized LFRCC [23].

In order to perform classification for SSD task, a Gaussian Mixture Model (GMM) based Bayesian classifier was used [22]. The performance of the SSD system is measured in terms of % Equal Error Rate (EER) metric. Final scores are represented in terms of the Log-Likelihood Ratio (LLR) given by:

$$LLR = \log \frac{p(X|\lambda_0)}{p(X|\lambda_1)}, \qquad (7)$$

where $p(X|\lambda_0)$, and $p(X|\lambda_1)$ are the likelihood scores from the GMM for the genuine and impostor trials, respectively.

Another classifier used was CNN [24]. We use deep Convolutional Neural Network (CNN) architecture capable of differentiating patterns in features corresponding to genuine and replay speech signals. The CNN architecture used in our experiments has 3 convolutional layers (i.e., Conv1, Conv2, and Conv3). After the convolution operation, in order to introduce non-linearity in the neuron output, an activation function is used. In our CNN architecture, we use the Rectified Linear Unit (ReLU) as the activation function. This operation is followed by a pooling layer of kernel size of 3 × 3 and stride 1 is used. The flattened output is then fed to 2 Fully Connected (i.e., FC1 and FC2) layers. The output of the final FC2 layer gives us a probabilistic output for classification. The loss function used is binary cross-entropy, and the optimization algorithm used is gradient descent.

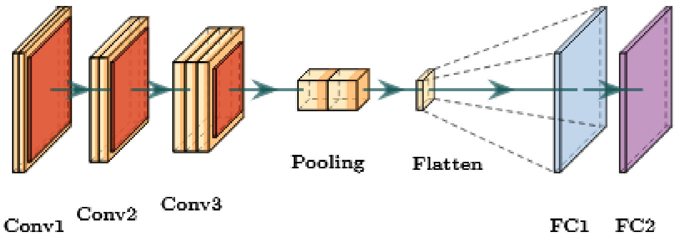


Fig. 3: CNN architecture used for classification of genuine *vs.* replay spoof speech.

C. Results

We present experimental results on ASV spoof 2019 PA dataset using LFRCC feature set for SSD task. We consider the effect on the Equal Error Rate (EER) due to various evaluation factors, such as LP order, and number of subband filters. Table II shows the effect of LP order on the EER for LFRCC feature set. It is observed that the best achieved EER is 15.21% on the evaluation set using CNN. Furthermore, on GMM, the best achieved EER is 17.30%. Both of these results are obtained when LP order is kept 8, which we have hypothesized as optimal (through an analysis as discussed in Section 2). To that effect, the LP order is fixed as 8 for the rest of the experiments in our work.

TABLE II: Effect of LP order on EER for LFRCC Features

Prediction Order (p)	% EER (GMM)		% EER (CNN)	
	Dev.	Eval.	Dev.	Eval.
6	6.84	18.21	6.23	16.17
8	6.77	**17.30**	6.08	**15.21**
10	6.89	19.53	5.35	16.88
12	7.02	19.80	6.72	17.20
14	7.19	20.63	6.96	18.37
16	7.54	20.42	7.34	19.28
18	8.38	21.84	7.94	20.36
20	9.43	23.49	8.86	21.11
24	10.97	24.82	8.93	21.89

Additional experimental results to observe the impact of subband filters, and dimension of feature vector as shown in Table III. While keeping the LP order as 8, the number of subband filters in the filterbank are varied from 40 to 140. We observe that the best performance on evaluation set using GMM as classifier is 16.32%. This is obtained when the number of subband filters is 120. Furthermore, when CNN is used as the classifier, the best performance of 14.83% EER is observed when the number of subband filters is 140. These observations indicate that the optimized LP order for replay spoof detection on ASVSpoof 2019 PA dataset is 8. However, the performance of the countermeasure system is also improved by increasing the number of subband filters, i.e., by increasing the resolution in frequency domain.

TABLE III: Effect of Number of Subband Filters on EER

No. of Subband Filters	%EER (GMM)		%EER (CNN)	
	Dev.	Eval.	Dev.	Eval.
40	6.77	17.30	6.08	15.21
60	6.85	19.01	4.87	17.70
80	7.14	20.47	4.16	18.29
100	7.93	18.28	5.21	17.72
120	8.40	**16.32**	6.78	15.26
140	9.10	16.79	7.53	**14.83**

IV. Summary and Conclusions

In this study, the LFRCC feature set is exploited to analyze the effect of optimal LP order on LFRCC feature set for SSD task. The optimal order of 8 was observed to result in maximum information in the LP residual. Thus, the importance of LP order for SSD task was emphasized, and it was shown that the optimal order for LP w.r.t. speech production mechanism is not the same as for SSD task, which is the novel aspect of this work. To that effect, analysis is shown w.r.t. framewise magnitude spectrum of the LP residuals with varying order. Furthermore, Log Spectral Distance (LSD) between LP residuals of with various LP orders shows that the optimal value of LP order is 8. This analysis is further confirmed by the experimental results obtained. To that effect, the best performance is obtained on LP order 8 with EER of 15.21% achieved on evaluation set using CNN. Furthermore, the effect of number of subband filters is observed. It is observed that the performance of the SSD system further improves on increasing the number of subband filters in the filterbank. However, the underlying assumption of this paper is based on the linearity of source-filter model of speech production, which further leads us to linear production of speech. In future, we would like to exploit the non-linear aspects of speech production (via nonlinear prediction, using Voltera-Weiner series) for SSD task to get better analysis.

V. Acknowledgements

The authors would like to thank the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, for sponsoring consortium project titled ‘Speech Technologies in Indian Languages’ under ‘National Language Translation Mission (NLTM): BHASHINI’, subtitled ‘Building Assistive Speech Technologies for the Challenged’ (Grant

ID: 11(1)2022-HCC (TDIL)). We also thank the consortium leaders Prof. Hema A. Murthy, Prof. S. Umesh, and the authorities of DA-IICT Gandhinagar, India for their support and cooperation to carry out this research work.

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] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[4] Y. Stylianou, "Voice transformation: A survey," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 19-24 April, 2009, pp. 3585–3588.

[5] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, September 2014, pp. 1–6.

[6] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *International Symposium on Intelligent Multimedia, Video, and Speech Processing*, Hong Kong, 20-22 October, 2004, pp. 145–148.

[7] 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.

[8] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *The Journal of the Acoustical Society of America (JASA)*, vol. 45, no. 2, pp. 458–465, 1969.

[9] 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.

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

[11] P. Gupta, G. P. Prajapati, S. Singh, M. R. Kamble, and H. A. Patil, "Design of voice privacy system using linear prediction," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 7-10 December 2020, Auckland, New Zealand, pp. 543–549.

[12] S. Cheedella 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.

[13] J. Mishra, M. Singh, and D. Pati, "Processing linear prediction residual signal to counter replay attacks," in *International Conference on Signal Processing and Communications (SPCOM)*, IISc, Bangaluru, India, 16-19 July, 2018, pp. 95–99.

[14] C. Hanilçi, "Speaker verification anti-spoofing using linear prediction residual phase features," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, August 8 - September 2, 2017, pp. 96–100.

[15] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. 2^{nd} Edition, Pearson Education India, 2004.

[16] S. G. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 2nd Ed. 1999.

[17] H. Tak and H. A. Patil, "Novel Linear Frequency Residual Cepstral Features for Replay Attack Detection," in *INTERSPEECH, Hyderabad, India*, 2-6 September, 2018, pp. 726–730.

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

[19] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[20] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "The ASVSpoof 2019 database," *arXiv preprint arXiv:1911.01601*, 2019, {Last Accessed: 24-10-2020}.

[21] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASV Spoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019, {Last Accessed: 22-10-2020}.

[22] 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," *INTERSPEECH 2017*, Stockholm, Sweden, 20-24 August 2017.

[23] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. Hong Kong, 6-10 April, 2003, pp. I–I.

[24] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT Press Cambridge, 2016, vol. 1, no. 2.