Replay Spoof Detection using Power Function Based Features

Prasad A. Tapkir[1], Madhu R. Kamble[1] Hemant A. Patil[1] and Maulik Madhavi[2]
[1]Speech Research Lab
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India
[2]Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore
E-mail: {prasad_tapkir, madhu_kamble, hemant_patil}@daiict.ac.in and maulik.madhavi@nus.edu.sg

***Abstract*—Among various types of spoofing attacks replay possess a greater threat to the Automatic Speaker Verification (ASV) system. In our previous study, we found that the replay spoof detection is effective when human auditory system is modeled by power law nonlinearity. In this paper, we design the replay spoof detection system using power function-based features, namely, Power Normalized Cepstral Coefficients (PNCC) and Q-Log Normalized Cepstral Coefficients (QLNCC). The PNCC and QLNCC feature sets are noise robust and they are able to capture the speaker-specific information in noisy environments. The PNCC feature set uses power law nonlinearity, however, the QLNCC feature set uses *q*-log nonlinearity. The experiments are performed on ASVspoof 2017 challenge version 2.0 database with Gaussian Mixture Model (GMM) as a classifier. The individual PNCC and QLNCC feature set gives an Equal Error Rate (EER) of 23.02 % and 24.12 % on evaluation set, respectively. Furthermore, to capture the possible complementary information, score-level fusion of PNCC and QLNCC feature sets with Constant Q Cepstral Coefficients (CQCC) feature set was performed resulting in reduced EER of 13.02 % and 13.62 % on evaluation set, respectively.**

I. Introduction

The Automatic Speaker Verification (ASV) system aims to verify the claimed identity of the speaker using speech samples provided by the speaker. To use the ASV system for voice biometric authentication purpose, the ASV system needs to be robust against transmission channel, intersession recordings, speaker health, speaker aging, etc. However, nullifying the effect of these variabilities makes ASV system susceptible to various kinds of spoofing attacks, namely, impersonation, voice conversion, speech synthesis, replay [1].

Replay is one of the most easy way to get spoofed as only simple recorder and playback device is required to generate the spoofed speech from the target speaker and do not need any prior speech processing techniques. Hence, it is more challenging task to detect the replay speech signal. The key objective during replay detection is to emphasize on the characteristics of the intermediate devices [2]. Because replay speech is a convolution of the natural speech signal with the impulse response of the intermediate devices and environmental conditions. Hence, it is a *blind* deconvolution problem that is still a major research issue.

In the past years, many of the researchers have focused on the study of replay spoof speech detection (SSD) task. One of the studies done in the far-field recording condition was reported in [3]. The study was done using the approach of cut and paste to replay speech which was detected with use of pitch (i.e., fundamental frequency, F_0) and Mel Frequency Cepstral Coefficients (MFCC) feature set [3]. The spectral bitmap approach is used to identify live and non-live (recorded) speech [4]. Recently, second ASVspoof 2017 Challenge was organized as a special session focusing exclusively on the replay spoofing attack during INTERSPEECH 2017. The organizers of the challenge provided a common database for the replay SSD task. Many of the countermeasures were explored during this challenge. The research findings of this challenge were, how the replay speech signal gets affected by the intermediate devices and largely it affects the high frequency regions. Hence, one of the findings was the significance of high frequency regions [5]. Moreover, it was found that use of the Cepstral Mean Variance Normalization (CMVN) technique helps to discriminate the natural and replayed speech [5].

In our previous study, we found that the genuine and replay speech is more distinguishable, when the logarithmic nonlinearity in MFCC is replaced with the power law nonlinearity [6]. In this paper, we explore two more feature sets that are based on power function family, namely, Power Normalized Cepstral Coefficients (PNCC) [7] and Q-Log Normalized Cepstral Coefficients (QLNCC) [8]. The PNCC feature uses power law nonlinearity and QLNCC feature uses *q*-log nonlinearity. The *q*-log nonlinearity is the generalization of log function and example of power function. We developed the replay SSD system using Gaussian Mixture Model (GMM) classifier for PNCC, QLNCC feature set. To the best of authors' knowledge, this is the first study reporting use of PNCC and QLNCC for replay SSD task.

II. Power Function Family

At the onset of signal, the average auditory nerve firing rate observes an overshoot, some studies reports that the human auditory system focuses on the onset than the valley of power envelope [9]. Hence, the human auditory system can be mathematically modeled as the function of onset firing rate and sound pressure level [9]. The power law nonlinearity is able to approximate this function [10]. The various studies in speech recognition uses several power function-based features

[7]. In this paper, we explored two power function-based nonlinearities, one is power law nonlinearity given by [7]:

$$y(n) = (s(n))^{\gamma}, \tag{1}$$

where $s(n)$ is input signal, $y(n)$ is output signal and γ is some constant. Some studies suggest that normalizing at q log-domain is efficient than the normal log-domain [11], [12]. Hence, the second nonlinearity we explored is q-log nonlinearity and is given by [8]:

$$y(n) = \log_q(s(n)) = \frac{s(n)^{(1-q)} - 1}{1-q}. \tag{2}$$

Fig. 1 compares the natural and replay speech in power law and q-log domain using their gammatonegrams. The rectangular box shows the distinguishable points in the natural *vs.* replay speech gammatonegrams and the elliptical regions shows the distinguishable points in gammatonegram obtained for power law nonlinearity *vs.* q-log nonlinearity.

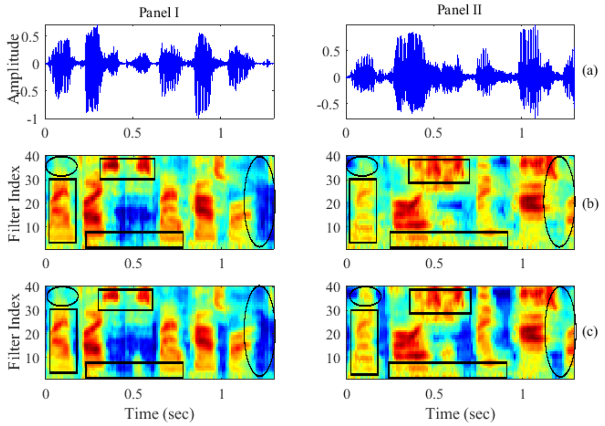

Fig. 1. Comparison of natural *vs.* replay speech in power law and q-log domain. (a) speech signal, gammatonegram using (b) power law nonlinearity, (c) q-log nonlinearity (Panel-I: natural speech, Panel-II replay speech). After [8].

From Fig. 1(b) and Fig. 1(c), it can be observed that the power function nonlinearities captures the effect of noise introduced by the intermediate devices and environmental conditions in replay speech at lower as well as at higher frequency regions. In addition, they also captures the change in energy levels at formant frequencies. It can also be observed from Fig. 1(b) that, the power law nonlinearity shows more distinguishable cues at higher frequency regions in gammatonegram compared to the q-log nonlinearity (elliptically highlighted region).

III. Feature Extraction

The block diagram of the PNCC and QLNCC feature extraction process is shown in Fig. 2.

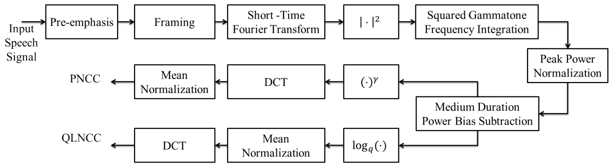

Fig. 2. Functional block diagram of PNCC and QLNCC feature extraction.

A. Power Normalized Cepstral Coefficients (PNCC)

The speech signal is passed through the pre-emphasized filter in order to emphasize high frequency components in the signal. The pre-emphasized signal is segmented into short duration signal with overlapping short-duration windows and Discrete-Time Fourier Transform (DTFT) is applied to obtain Short-Time Fourier Transform (STFT) of signal. Furthermore, the magnitude squared power is passed through gammatone-shaped filterbank, which uses Equivalent Rectangular Bandwidth (ERB) frequency scale. The gammatone filterbank is designed in such a way that the sum of the squared transfer function of each filter is equal to one, i.e., each filter in filterbank must satisfy the following equation [7]:

$$\sum_{k=0}^{K/2-1} \mid G_m\left(e^{j\omega_k}\right) \mid^2 = 1, \tag{3}$$

where $G_m\left(e^{j\omega_k}\right)$ is transfer function of m^{th} filter in gammatone filterbank, ω_k, is discrete frequency and K is DTFT size. After the gammatone filterbank block, the short-time spectral power (P_s) is given by [7]:

$$P_s(i,m) = \sum_{k=0}^{K/2-1} \mid S(i, e^{j\omega_k}) G_m\left(e^{j\omega_k}\right) \mid^2, \tag{4}$$

where $S(i, e^{j\omega_k})$ is STFT of speech signal $s(n)$, i is frame index. Furthermore, the spectral power is normalized by the peak power (P_{pk}) of the signal and scaled by some constant p_0 as follows [7]:

$$P_{norm}(i,m) = p_0 \frac{P_s(i,m)}{P_{pk}}, \tag{5}$$

where P_{norm} is normalized power. In the next step, the median-time power is obtained by computing the moving average of spectral power $P_s(i,m)$. The median-time power in a single analysis frame is given as [7]:

$$P_{med}(i,m) = \frac{1}{2I+1} \sum_{i'=i-I}^{i+I} P_s(i',m), \tag{6}$$

where I is temporal integration factor. This median-time power is used for power bias subtraction in each filter to sharpen the power distribution. Once we get the power bias subtracted signal, the power law nonlinearity is applied to map the information in power law-domain. The Discrete Cosine Transform (DCT) is applied to obtain cepstral coefficients. Only first few coefficients are retained from the decorrelated features. Next, the Cepstral Mean Normalization (CMN) is performed in order to remove the channel effects. In the last step, the first and second-order dynamic coefficients are appended to incorporate the long-range temporal dynamics.

B. Q-Log Normalized Cepstral Coefficients (QLNCC)

The feature extraction process of QLNCC feature set is almost the same as PNCC feature set except last three blocks, namely, nonlinearity block, DCT and mean normalization. The power law nonlinearity in PNCC feature set is replaced with the q-log nonlinearity as per Eq. (2). Furthermore, the order of last two blocks, i.e., mean normalization and DCT is reversed. Unlike other feature sets, in QLNCC feature set, the mean normalization is performed at high frequency resolution (i.e., before applying DCT). This is motivated from the studies reported in [13] which shows that applying mean normalization at low frequency resolution is less effective compared to applying on high frequency resolution. After mean normalization, the DCT is applied to decorrelate the features and first few coefficients are retained. The static co-efficients are appended with first and second-order derivatives to capture the dynamics of signal.

IV. Experimental Setup and Results

All the experiments are performed on the ASVspoof 2017 challenge V2 database [14]. All speech utterances have a resolution of 16-bits per sample and sampling frequency of 16 kHz. The database is based on the RedDots corpus and its replay version. Table I shows the statistics of the ASVspoof 2017 challenge V2 database. All the systems are implemented with GMM classifier with 512 Gaussian mixture components. Two GMMs are trained for genuine and spoof class using only training set of ASVspoof 2017 challenge V2 database.

TABLE I
Statistics of ASVspoof 2017 challenge V2 database. After [14]

Subset	# Speakers	# Utterances	
		Genuine	Spoofed
Training	10	1507	1507
Development	8	760	950
Evaluation	24	1298	12008

A. Effect of Gamma in PNCC

The PNCC features are extracted from pre-emphasized speech using Hamming window of duration 25 *ms* and 10 *ms* window shift. Total 40 number of gammatone filters are used to obtain subband filtered signal. To obtain the medium duration power $I = 5$ is used . First 20 static coefficients are retained after DCT and first and second-order dynamics are appended to obtain high-dimensional feature vector. The experiments are carried out for various gamma (γ) values. Fig. 3 shows the graphical representation of effect of gamma values on system performance. Empirically, we found that, $\gamma = 0.04$ gives the relatively lower Equal Error Rate (EER) of 20.78 % and 23.74 % on development and evaluation set, respectively.

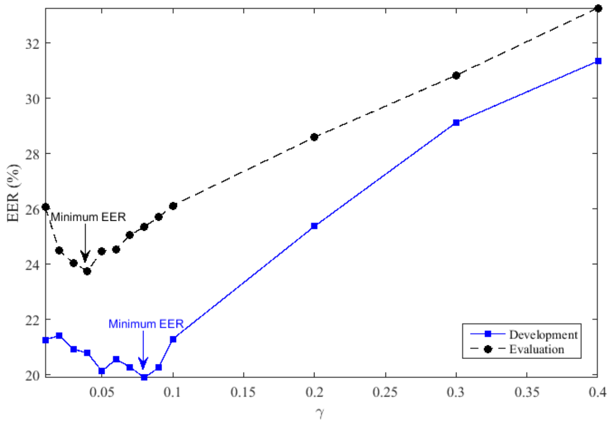

Fig. 3. Graphical representation of effect of γ values on system performance.

B. Effect of q in QLNCC

In this sub-Section, we analyze the effect of q value on system performance developed using QLNCC feature set. The QLNCC features are extracted using the same parameters used for PNCC feature extraction. In this experiment, we examine the effect of q value on system performance by varying q from 0 to 1. Fig. 4 shows the graphical representation of effect of q. Empirically, we found that $q = 0.97$ gives the relatively lower EER of 21.81 % and 24.67 % on development and evaluation set, respectively.

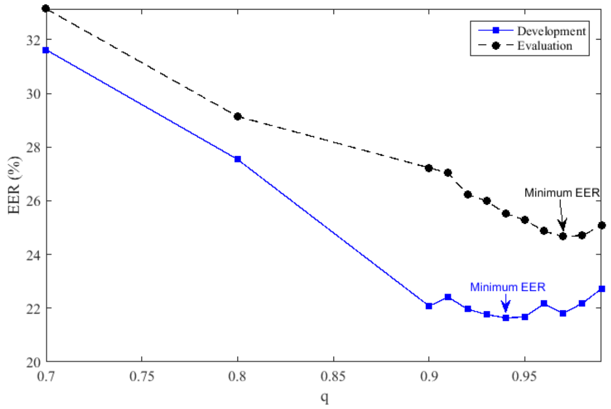

Fig. 4. Graphical representation of effect of 'q' values on system performance.

C. Score-Level Fusion

The final replay SSD system is developed using the best parameters from the above experiment (i.e., q=0.97 and γ=0.04). Table II shows the results (in % EER) on development and evaluation set for individual and fused systems. The PNCC and QLNCC feature sets are fused with CQCC feature set at the score-level. The individual CQCC+GMM system is baseline system provided by the organizers of ASVspoof 2017 challenge having an EER of 19.04 % on evaluation set. When CQCC feature set is fused with PNCC feature set, the % EER

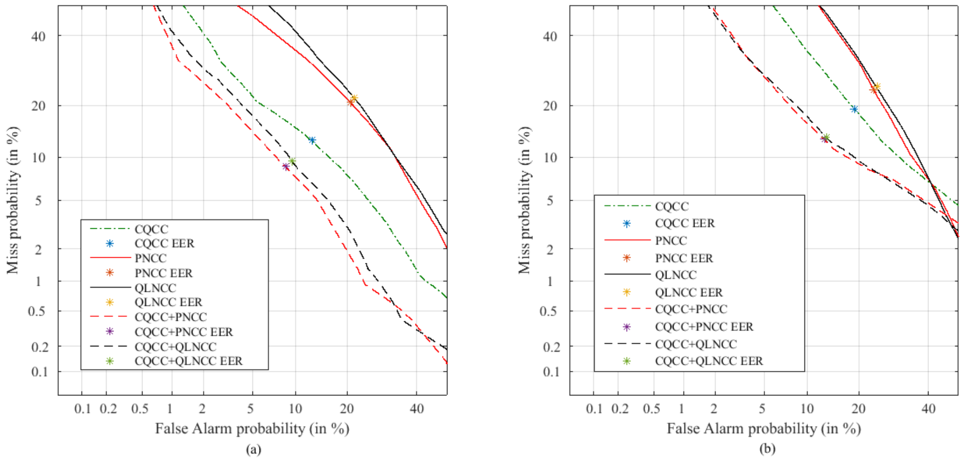

Fig. 5. DET curves for (a) development set, and (b) evaluation set.

gets reduced to 12.98 % from its individual EER of 23.74 %. However, the score-level fusion of CQCC and QLNCC feature sets gives % EER of 13.27 from its individual EER of 24.67 %. Fig. 5(a) and Fig. 5(b) shows the Detection Error Trade-off (DET) curves of individual as well as fused systems for development and evaluation set, respectively.

TABLE II
RESULT ON DEVELOPMENT AND EVALUATION SET

Feature Set	EER (%)	
	Development	Evaluation
CQCC	12.81	19.04
QLNCC ($q = 0.97$)	21.81	24.67
PNCC ($\gamma = 0.04$)	20.78	23.74
CQCC + QLNCC	9.49	13.27
CQCC + PNCC	**8.73**	**12.98**

V. Summary and Conclusions

The replay attack is the simplest and most accessible spoofing attack. Hence, the performance of ASV system degrades to the greater extent in the presence of replay attacks. In this study, we developed the replay spoof detection system using power function-based feature sets, namely, PNCC and QLNCC. It is observed that the genuine and replay speech are more distinguishable after applying power law and q-log nonlinearity. The PNCC and QLNCC feature sets are fused with CQCC feature set at the score-level to develop final spoof detection system. The final system performance is much better than the baseline CQCC system. Our future work will be directed to explore more power function-based features along with some other classifiers, such as CNN, SVM, BLSTM, etc. to detect replay attacks.

Acknowledgments

The authors would like to thank the organizers of the APSIPA 2018 for conducting special session on replay spoof detection. This research is supported by Programmatic Grant No. A1687b0033 from the Singapore governments Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain). The authors would also like to thank authorities of DA-IICT, Gandhinagar to carry out this research work.

References

[1] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *INTERSPEECH*, Lyon, France, 2013, pp. 925–929.

[2] B. S. M. Rafi, K. S. R. Murty, and S. Nayak, "A new approach for robust replay spoof detection in ASV systems," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Montreal, QC, Canada, 2017, pp. 51–55.

[3] Villalba, Jesús and Lleida, Eduardo, "Detecting replay attacks from far-field recordings on speaker verification systems," in *European Workshop on Biometrics and Identity Management*. Roskilde, Denmark: Springer, 2011, pp. 274–285.

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

[5] 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.

[6] P. A. Tapkir, A. T. Patil, N. Shah, and H. A. Patil, "Novel spectral root cepstral features for replay spoof detection," submitted for possible publication in INTERSPEECH, Hyderabad, India, Sept. 02-06, 2018.

[7] C. Kim and R. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4101–4104.

[8] H. F. Pardede, "On noise robust feature for speech recognition based on power function family," in *IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Nusa Dua, Indonesia, 2015, pp. 386–390.

[9] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 556–565, May 2009.

[10] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *The Journal of the Acoustical Society of America (JASA)*, vol. 106, no. 4, pp. 2040–2050, 1999.

[11] J. Lee, S. Baek, and H.-G. Kang, "Signal and feature domain enhancement approaches for robust speech recognition," in *IEEE International Conference on Information, Communications and Signal Processing (ICICS)*, Singapore, 2011, pp. 1–4.

[12] S. Baek and H.-G. Kang, "Mean normalization of power function based cepstral coefficients for robust speech recognition in noisy environment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 1735–1739.

[13] C. Avendano and H. Hermansky, "On the effects of short-term spectrum smoothing in channel normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 4, pp. 372–374, 1997.

[14] H. Delgado *et al.*, "ASVspoof 2017 Version 2.0: Meta-data analysis and baseline enhancements," in *The Speaker and Language Recognition Workshop ODYSSEY*, Les Sables d'Olonne, France, 2018. [Online]. Available: http://www.eurecom.fr/publication/5504, Last Access 28-May-2018