Significance of CMVN for Replay Spoof Detection

Ankur T. Patil and Hemant A. Patil
Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology
(DA-IICT), Gandhinagar-382007, Gujarat, India.
E-mail: {ankur_patil, hemant_patil}@daiict.ac.in

***Abstract*—In this paper, significance of the Cepstral Mean and Variance Normalization (CMVN) is investigated for replay Spoofed Speech Detection (SSD) task. Literature shows that application of the CMVN produces significantly better performance on many feature sets, which is counter-intuitive for replay SSD task. This behaviour is analyzed by performing experiments for environment-independent and dependent cases with % Equal Error Rate (EER) as evaluation metric. Furthermore, analysis is also performed with the help of estimated probability density function (pdf) of the genuine *vs.* spoof speech feature representations. The experiments are performed on the publicly available and statistically meaningful ASVspoof 2017 version-2 dataset using well-known CQCC-GMM and LFCC-GMM SSD systems. This dataset comprised of seven acoustic environments for replay speech. This study reveals that performance of the SSD system is better with application of the CMVN on environment-independent case. Whereas performance degrades drastically on environment-dependent scenario with application of the CMVN. For this scenario, the CMVN suppresses the transmission channel distortion, which is in fact the discriminative cues for the genuine *vs.* replay speech signal. This results in degradation of the performance. However, for environment-independent scenario, CMVN scale down the variability in feature space across the different environment, which improves the performance.**

Keywords: Cepstral Mean and Variance Normalization, replay, spoof, ASVspoof 2017.

I. Introduction

Normalization techniques have been used in various speech applications, such as automatic speech and speaker recognition to improve the performance of the systems [1]–[6]. The literature includes several forms of normalization techniques, which includes normalization w.r.t. n^{th} order expectation of random variable X for each dimension. The first and second-order expectations are known as mean and variance, respectively. If normalization is applied on the cepstral feature representation based on mean and variance, then it is known as Cepstral Mean and Variance Normalization (CMVN). However, if we consider only mean value for normalization, then it is called as Cepstral Mean Normalization (CMN) or Cepstral Mean Subtraction (CMS) [7], [8].

In particular, CMVN is the most common and computationally inexpensive approach of normalization. It reduces the distortion due to transmission channel effects, and improve the recognition performance of the speech and speaker recognition systems. The use of normalization techniques in Automatic Speech Recognition (ASR) for environmental mismatch conditions is well known in the literature [2], [3], [6]. These approaches use maximum likelihood estimates (MLE) to get the mean and variances along the feature dimensions, and then normalize it. In [9], authors propose the Bayesian approach to estimate the mean and variance. In [6], segmental CMVN is used to employ the noise-robust ASR system in real-time, where normalization is performed over short segment of the utterance in order to reduce the *latency* period.

The replay spoof speech signal is formed by convolving the genuine version of speech sample with the impulse responses of the recording and replay environments and devices. In spoof speech detection (SSD) task, we need to identify this additional channel effects present in spoof speech signal. The application of the CMVN/CMN to the speech and speaker recognition system supresses the channel effects. Hence, its use in SSD task seems counter-intuitive. However, among the many countermeasure systems developed on ASVspoof-2017 dataset, it is observed that CMVN/CMN has been effectively utilized for the replay Spoof Speech Detection (SSD) task to give significant improvement in the performance of the SSD system [10]–[18]. This contradictory results motivated us for further investigation over applicability of the CMVN. We performed experiments for environment-independent and environment-dependent scenario. Furthermore, probability density function (pdf) are estimated over several dimensions of feature representations.

II. Cepstral Mean and Variance Normalization (CMVN)

CMN was initially proposed to eliminate the channel distortions that are introduced into the signal by convolving the signal with the impulse response of the transmission channel. In cepstral-domain, convolutional vector space is mapped to the additive vector space. The CMN estimates the mean along every dimension of the cepstral feature representation of the speech sample and this mean value is subtracted from the corresponding dimension to transform the feature representation to zero-mean. Whereas, the CMVN transforms each cepstral feature representation of the speech sample to zero-mean and unit-variance. Mean and variance can be estimated for a segment of the utterance to reduce the latency period [6].

Let x_t denote the d-dimentional feature vector at the frame index t of the utterance, and $x_t(i)$ represent the i^{th} component of x_t. The speech utterance is passed through frame-blocking, denoted as X = [x_1, x_2, ... , x_T], where T denote the number of speech frames. The mean and variance values are estimated for every dimension in maximum likelihood (ML) framework as [9]:

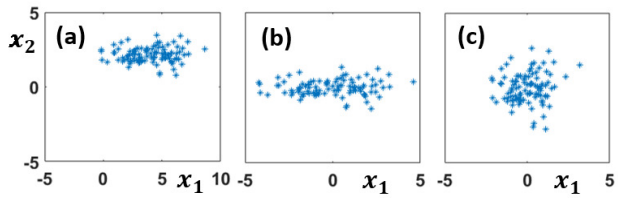

Fig. 1. Scatter plot for (a) the unnormalized data, (b) with CMN normalization, and (c) CMVN normalization. X = [x_1 x_2] denotes the samples drawn from the bivariate Gaussian distribution. Ytick values of Fig. 1(b) and Fig. 1(c) are same as that of Fig. 1(a).

$$\mu_{ML}(i) = \frac{1}{T}\sum_{t=1}^{T} x_t(i), \qquad 1 \leq i \leq d, \tag{1}$$

$$\sigma^2_{ML}(i) = \frac{1}{T-1}\sum_{t=1}^{T}(x_t(i) - \mu_{ML}(i))^2, \quad 1 \leq i \leq d. \tag{2}$$

where μ_{ML} and σ_{ML} corresponds to mean and variance values, estimated in ML framework. The CMVN is applied to obtain normalized cepstrum of the frame as:

$$\hat{x}_t(i) = \frac{x_t(i) - \mu_{ML}(i)}{\sigma_{ML}(i)}, \quad 1 \leq t \leq T, 1 \leq i \leq d. \tag{3}$$

To visualize the effect of the CMN and CMVN, we generated the data samples with the help of two random variables from the normal distributions, $\mathcal{N}(4,4)$, and $\mathcal{N}(2,0.25)$. The scatter plot of the generated data samples is shown in Fig. 1(a). Fig.1(b) and Fig.1(c) shows the scatter plot for CMN and CMVN normalized data samples, respectively. For CMN data samples, it can be observed that the data samples are centered around the origin, and variance is maintained the same as that of the original data samples. With CMVN, the mean and variance are normalized. The spread along both the axes is maintained at unity variance in CMVN as shown in Fig.1(c).

Similar kind of observations regarding normalization can be seen in Fig. 2 which shows the scatter plots for the first two dimensions (D) of the CQCC feature set for genuine *vs.* two spoof speech signals. The data samples for the genuine speech is shown by red '*' symbol, whereas spoof speech data samples for balcony and studio environments are shown by green and blue '*' symbol, respectively. The CQCC feature extraction and dataset details are discussed in Section IV. Fig. 2(a), 2(b) and 2(c) shows the scatter plots for original features (initial 2-D), it's CMN and CMVN normalized versions, respectively. Again, feature representation with CMN is centered around origin with original feature representation variance, whereas feature representation with CMVN is zero-centered with unity variance. The other intuition from this scatter plot is discussed in Section V.

III. Replay Speech Signal Modelling and CMVN

Using linear system theory, the speech signal, $s(n)$ is modelled as the convolution of the glottal airflow, $g(n)$ with the impulse response of the vocal tract system, $v(n)$, i.e.,

$$s(n) = g(n) * v(n). \tag{4}$$

In many speech signal processing applications, speech signal is represented in cepstral-domain. The cepstral representation of the speech signal is obtained as the inverse Fourier transform of the logarithm of the spectrum of the speech signal. This transformation maps the convolutionally-combined vectors to additively combined vectors [19]–[24]. Let $\hat{s}(n)$, $\hat{g}(n)$, and $\hat{v}(n)$ represents the cepstrum of the speech signal, glottal airflow, and vocal tract system, respectively. Cepstral representation of the speech signal in eq. (4) is given as:

$$\hat{s}(n) = \hat{g}(n) + \hat{v}(n). \tag{5}$$

In SSD framework, the signal $s(n)$ is treated as genuine speech signal. The effect of the distortion due to replay mechanism on genuine speech signal, can be modelled by linear filtering. In replay mechanism, the genuine signal is recorded, and again replayed back. In this process, genuine signal is distorted by the impulse responses of the recording environment, $a(n)$, recording device, $b(n)$, playback device, $c(n)$, and playback environment, $d(n)$, respectively. By linear filter theory, the replayed speech is referred to as convolution of the genuine speech signal with this additional components, i.e.,

$$r(n) = s(n) * a(n) * b(n) * c(n) * d(n). \tag{6}$$

Let all these additional elements $(a(n), b(n), c(n),$ and $d(n))$ contribute to the overall impulse response of the replay mechanism system, $h(n)$ (i.e., $h(n) = a(n)*b(n)*c(n)*d(n)$) which will distort the genuine speech signal. Hence, replayed signal is modeled as:

$$r(n) = s(n) * h(n). \tag{7}$$

In cepstral-domain, eq. (7) is written as:

$$\hat{r}(n) = \hat{s}(n) + \hat{h}(n). \tag{8}$$

If an utterance consists of T number of speech frames, then the cepstrum for each frame can be written as:

$$\begin{aligned} \hat{r}_1(n) &= \hat{s}_1(n) + \hat{h}(n), \\ \hat{r}_2(n) &= \hat{s}_2(n) + \hat{h}(n), \\ &\vdots \\ \hat{r}_T(n) &= \hat{s}_T(n) + \hat{h}(n). \end{aligned} \tag{9}$$

Taking average over the T frames, we get,

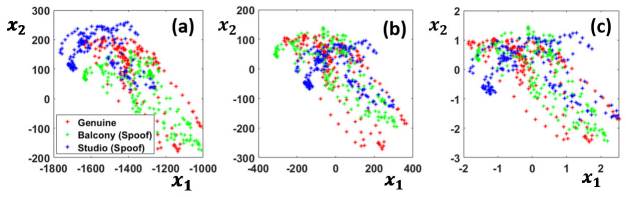

Fig. 2. Scatter plot for (a) the unnormalized data, (b) with CMN normalization, and (c) CMVN normalization. X = [x_1 x_2] denotes the first and second dimension of CQCC feature vector. Legends of Fig. 2(b) and 2(c) are same as that of Fig. 2(a).

$$\frac{1}{T}\sum_{t=1}^{T}\hat{r}_t(n) = \frac{1}{T}\sum_{t=1}^{T}\hat{s}_t(n) + \hat{h}(n). \tag{10}$$

Here, we modeled the effect of distortion by linear filter approach. Then, the distortion from the observed signal is removed by the inverse filtering. The cepstrum is computed as inverse transform of the log of the Fourier transform. Hence, the effect of the distortion can be removed (at least supressed) by subtracting the characteristics of the distortion filter from the cepstrum of the observed signal. In eq. (10), the cepstrum of the distortion filter, $\hat{h}(n)$, can be subtracted to obtain the distortion-less signal. Let us assume that the genuine speech signal is the zero-mean process. Then,

$$\hat{h}(n) = \frac{1}{T}\sum_{t=1}^{T}\hat{r}_t(n) = c_\mu. \tag{11}$$

Then, CMN is supposed to remove or at least supresses the effect of the distortion from the replayed speech signal. Inevitably, the average over the cepstral coefficients include the speech and speaker information, and the effect of the channel distortion.

IV. Experimental Setup

A. Dataset Used

In this study, ASVspoof 2017 challenge version-2 database is used. This standard dataset is designed to develop countermeasure system to protect the ASV systems against the replay spoofing attacks. For the challenge, dataset is partitioned into three subsets, namely, training, development, and evaluation set [10]. The detailed distribution of the dataset is shown in Table I. The dataset consists of bonafide utterances in each subset. Spoofed utterances are created by replaying, and recording bonafide utterances using a variety of heterogeneous devices, and seven acoustic environments. In this study, we aim to investigate the application of the CMVN for spoof detection capability over environment-independent and environment-dependent cases.

In environment-independent case, target environment is unseen by the defense model. To perform the experiments on environment-independent case, the same statistical distribution of the speech samples as provided by the organizers is used. The statistics of the the dataset is shown in Table I. The

TABLE I
Statistics of the ASVspoof 2017 dataset for the environment-independent case. After [10].

Subset	# Spk	Utterances		Environments
		Genuine	Spoof	
Train	10	1507	1507	E3, E6
Dev	8	760	950	E3, E5, E6
Eval	24	1298	12008	E1, E2, E3, E4, E5, E6, E7
Total	42	3565	14465	

E1: Anechoic Room, E2: Analog Wire, E3: Balcony, E4: Canteen, E5: Home, E6: Office, E7: Studio, Spk: Speaker

training subset includes balcony, and office environments. Whereas, development subset includes the balcony, home, and office environments. The spoofed speech samples for evaluation subset are taken from all the environments. Here, spoof speech utterances are taken from the different environments for training set than that of the development and evaluation set. Therefore, this data distribution provided by the organizers is considered as environment-independent case as trained defense model is tested against the unseen environments.

For environment-dependent case, target environment is seen by the defense model. In this case, training and testing is performed on each individual environment. The distribution of the number of spoof speech utterances for each environment is varying and shown in Table II. To develop individual environment-dependent replay spoof speech detection (SSD) system, half of the spoof speech utterances for corresponding environment are chosen for training purpose, and remaining half is used for testing the performance of the model. To train the genuine speech signal model, equal number of genuine utterances are selected as that of spoof speech utterances, used for training in corresponding environment.

TABLE II
Distribution of spoof speech utterances among the environments in ASVspoof 2017 dataset.

Environment	**# Utteartnces**
Anachoic	748
Analog Wire	543
Balcony	1184
Home	570
Canteen	3517
Office	7565
Studio	342

B. Feature Set and Classifiers

The key objective of this paper is to investigate the significance of the CMVN for the replay Spoof Speech Detection (SSD) task. Hence, experiments are performed using CQCC and LFCC feature sets as these feature sets are used as baseline feature sets by the ASVspoof challenge organizers [10], [25]. The SSD systems are trained using GMM [26]–[28]. Two individual GMMs are trained for the genuine, and spoof speech samples of the training data. Each GMM is trained for 512 Gaussian mixtures. The Expectation Maximization (EM)

algorithm is used to update the parameters of the GMMs [29]–[31]. The log-likelihood (llk) score, $s(X)$, is computed as [32]:

$$s(X) = llk(X|\lambda_g) - llk(X|\lambda_s), \tag{12}$$

where λ_g, and λ_s represents the GMM models trained on genuine and spoofed speech samples, respectively. The CQCC and LFCC, both feature sets are designed to be of 90-dimensions (*90-D*) feature set, which includes static (*30-D*), delta (*30-D*), and delta-delta (*30-D*) components.

V. Experimental Results

In this Section, we present results to investigate the issue on application of the CMVN technique for Spoof Speech Detection (SSD) task. To that effect, Table III shows the performance of the environment-indpendent CQCC-GMM and LFCC-GMM SSD systems for ASVspoof 2017 challenge datasets. Performance of the environment-dependent scenario for CQCC and LFCC feature sets using ASVspoof 2017 dataset is displayed in Table IV. The Equal Error Rate (EER) is used as the performance evaluation metric. It is observed that, for environment-independent case in ASVspoof-2017 dataset, CMVN normalization technique works significantly better.

Fig. 3 shows the estimated pdf for the few selected dimensions of the CQCC feature set for genuine speech samples and spoof speech samples for all the individual environments. Fig. 3(a), (b), (c), (g), (h), (i), (m), (n), and (o) shows the estimated pdfs of the CQCC feature set for 1^{st}, 3^{rd}, 5^{th}, 10^{th}, 12^{th}, 15^{th}, 20^{th}, 25^{th}, and 30^{th} dimensions with CMVN normalization, respectively. Whereas, Fig. 3(d), (e), (f), (j), (k), (l), (p), (q), and (r) shows the estimated pdfs for 1^{st}, 3^{rd}, 5^{th}, 10^{th}, 12^{th}, 15^{th}, 20^{th}, 25^{th}, and 30^{th} dimensions for without application of the CMVN, respectively. This figure can be used to analyze the behaviour of the feature distribution in environment-dependent case. As observed from Fig. 3, estimated pdf for the CMVN case, all the environments are seems to be aligned with the pdf of the genuine speech signal. The alignment of the pdf is produced because of CMVN, which further leads to degradation of the results for environment-dependent scenario as distribution of the genuine speech data seems similar to that of the spoof speech data. Observations are also made for other dimensions of the feature vector other than mentioned dimensions but pdfs for few selected dimensions is presented in Fig. 3. It is observed that, after eleventh dimension pdfs of the spoof speech data for the CMVN case, all other environments mostly aligned with the genuine data and almost no difference exists in their pdfs. This fact can be observed from Fig. 3(k), (l), (p), (q), and (r), where all the pdfs are almost merged. Whereas, the pdfs of genuine data is much different to that of individual spoof speech environments in without CMVN case. Furthermore, for without normalization scenario, the distinct difference in pdfs can be observed for almost all the dimensions. For without CMVN case, if GMM parameters (i.e. *mean* and *variance* of the Gaussian mixtures) are estimated for the pdfs of the genuine *vs.* any other environment for spoof speech data, then for most of the environments, we obtained the well distinguishable GMM parameters. In particular, it can be observed that GMM parameters of the genuine data *vs.* spoof speech data from balcony/studio would be well distinguishable. Hence, corresponding SSD systems are producing 0% EER. However, with CMVN applied to feature set, all the pdfs of spoof speech signal representations for individual environment, do not lie on either side of the pdf of genuine signal representations. This fact is also observed from Fig. 2(a), where the data samples for genuine speech signal lie in the middle of the other two spoof speech environments. Hence, cumulative distribution of all the environments (shown in Fig. 4) for spoof speech data, could not produce the distinguishable GMM parameters w.r.t. genuine data.

Fig. 4(a) and (b) shows the estimated pdf of the genuine *vs.* spoof speech signal over first dimension of the CQCC feature set, obtained by application of the CMVN and without CMVN, respectively. Here, spoof speech data is obtained from all the possible environments. In this case, if GMM parameters are estimated from the pdfs, then Fig. 4(a) will have the more distinguishable GMM parameters as the GMM parameters estimated from this pdf would be better separated than the case of without CMVN. It is because of the fact that, pdf maximas of this pdf are well separated For Fig. 4(b), many local maximas are observed and random variable values corresponding to these maximas are mixing with each other. Because of these closely-spaced GMM parameters for genuine and spoof speech signals, classifier model may pose ambiguity when test sample is presented to trained model for the SSD task. This might be the reason for getting better results for environment-independent case with application of the CMVN compared to without CMVN case. Authors believe that pdf corresponding to higher cepstral dimensions show less discrimination between genuine *vs.* spoof due to decay of cepstrum w.r.t. time [20], [21], [23], [23]

TABLE III
Results of CQCC-GMM and LFCC-GMM systems in %EER for environment-independent case on ASVspoof-2017 dataset.

		dev	eval
CQCC	without CMVN	**10.31**	28.02
	CMVN	12.48	**18.17**
LFCC	without CMVN	**7.02**	32.62
	CMVN	14.79	**14.83**

Fig. 5(a) and Fig. 5(b) shows the detection error trade-off (DET) curves for the system developed using the feature set with and without application of the CMVN on ASVspoof 2017 challenge dataset, respectively [33]. These DET curves are shown for environment-dependent scenario. Two systems are showing 0 % EER which cannot be observed on the DET curve. However, these plots are shown by the point at the origin. It can be observed from the DET curves that the performance of environment-dependent case is significantly improved without normalization of the feature set. The false alarm rate is below 50 % for all the environmental cases (in

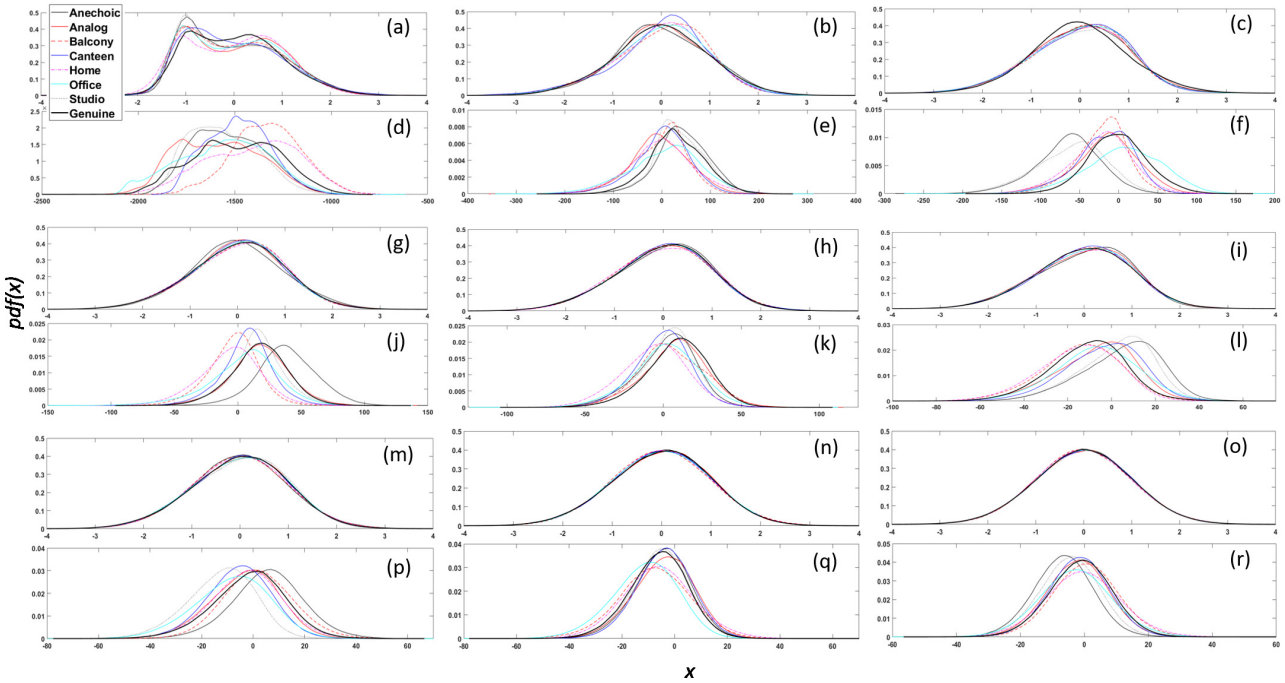

Fig. 3. Estimated *pdf* of genuine and environmentwise spoof speech samples over the (a) 1^{st}, (b) 3^{rd}, (c) 5^{th}, (g) 10^{th}, (h) 12^{th}, (i) 15^{th}, (m) 20^{th}, (n) 25^{th}, and (o) 30^{th} feature dimensions with application of CMVN, whereas Fig. (d), (e), (f), (j), (k), (l), (p), (q), and (r) shows the estimated *pdf*s for without CMVN case with the same sequence of dimensions as that of CMVN case. Legends of all figures are similar as given in Fig. 3(a).

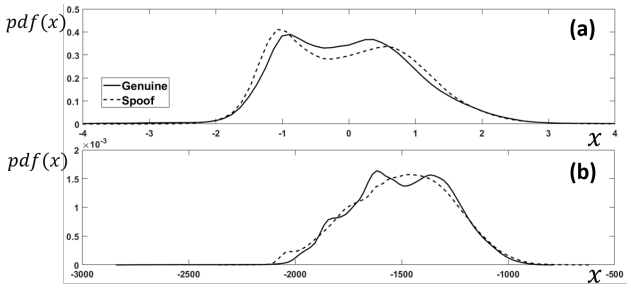

Fig. 4. Estimated *pdf* of genuine and spoof speech samples over the first feature dimension for (a) CMVN and (b) Without CMVN. Legends of Fig. 4(b) are similar to that of Fig. 4(a).

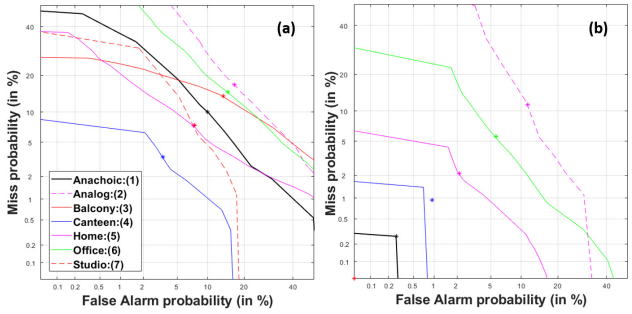

Fig. 5. DET plots for environment-dependent case using ASVspoof-2017 dataset (a) with application of the CMVN, and (b) without application of the CMVN on feature set. Legends for Fig. 5(a) and Fig. 5(b) are the same.

TABLE IV
RESULTS OF CQCC-GMM SYSTEM IN %EER FOR ENVIRONMENT-DEPENDENT CASE ON ASVSPOOF-2017 DATASET

	CQCC		LFCC	
	CMVN	Without CMVN	CMVN	Without CMVN
Anechoic Room	10.02	**0.26**	10.60	**0**
Analog wire	16.99	11.42	22.09	10.89
Balcony	13.81	**0**	9.60	**0.13**
Canteen	3.43	**0.93**	2.73	1.33
Home	7.39	2.12	9.23	2.51
Office	14.99	5.63	17.62	7.22
Studio	7.53	**0**	7.21	**0**

environment-dependent case).

VI. SUMMARY AND CONCLUSIONS

In this study, we performed the experiments for environment-dependent, and environment-independent scenarios with and without application of the channel normalization techniques. We observed that the application of the CMVN to the feature set do not always guarantee better results for the classification task. However, it depends upon the variability of the speech samples in terms of channel noise. If we train the models for environment-independent case, then applying CMVN might be the good choice. It reduces the large channel noise variability among the environments in spoof speech and brings the *pdf*s of the feature representations closely aligned

with each other and hence, with their cumulative distribution. Whereas in environment-dependent case, its better to not apply the CMVN. In this case, SSD is performed for the single environment only. For most of the environments, there is large variation of the distribution is observed without application of the CMVN for genuine *vs.* spoof speech. If we normalize this setting, then variation in distribution between the genuine *vs.* spoof is reduced significantly. Thus, CMVN can be considered as a double-edged sword and hence needs to be applied very carefully based on recording and transmission channel conditions. This fact is also observed in original speaker recognition literature [8,28].

The basis for the analyses presented in this paper are linearity of the overall channel effect in a replay attack scenario. However, the nonlinearities in sound wave propagation model for the speakers, and the environmental background can be analyzed in future study.

REFERENCES

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[2] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[3] F. Hilger and H. Ney, "Quantile-based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 845–854, 2006.

[4] A. De La Torre, A. M. Peinado, J. C. Segura, J. L. Pérez-Córdoba, M. C. Benítez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[5] O. M. Strand and A. Egeberg, "Cepstral mean and variance normalization in the model domain," in *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, Norwich, UK, August 2004.

[6] O. Viikki and K. Laurila, "Cepstral-domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[7] 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*, vol. 55, no. 6, pp. 1304–1312, 1974.

[8] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, p. 58, 1996.

[9] N. V. Prasad and S. Umesh, "Improved cepstral mean and variance normalization using bayesian framework," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, December 2013, pp. 156–161.

[10] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements," in *Odyssey 2018: The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, June 2018, pp. 296–303.

[11] A. T. Patil, R. Acharya, P. A. Sai, and H. A. Patil, "Energy Separation-Based Instantaneous Frequency Estimation for Cochlear Cepstral Feature for Replay Spoof Detection," in *INTERSPEECH*, Graz, Austria, Sept. 2019, pp. 2898–2902.

[12] 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, August 2017, pp. 7–11.

[13] Z. Ji, Z.-Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017," in *INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 87–91.

[14] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection." in *INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 641–645.

[15] H. B. Sailor, M. R. Kamble, and H. A. Patil, "Auditory filterbank learning for temporal modulation features in replay spoof speech detection." in *INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 666–670.

[16] M. Saranya and H. A. Murthy, "Decision-level feature switching as a paradigm for replay attack detection." in *INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 686–690.

[17] H. Tak and H. A. Patil, "Novel linear frequency residual cepstral features for replay attack detection." in *INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 726–730.

[18] M. R. Kamble and H. A. Patil, "Analysis of reverberation via teager energy features for replay spoof speech detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 2607–2611.

[19] R. W. Schafer, "Echo removal by discrete generalized linear filtering," *MIT Research Laboratory of Electronics, Massachusetts, USA*, 1969.

[20] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226, Jun 1968.

[21] 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.

[22] ——, "Superposition in a class of nonlinear systems," in *MIT Research Laboratory of Electronics*, Massachusetts, USA, 1965.

[23] A. V. Oppenheim, R. W. Schafer, and T. Stockham, "Nonlinear filtering of multiplied and convolved signals," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 3, pp. 437–466, 1968.

[24] J. Lim, "Spectral root homomorphic deconvolution system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 3, pp. 223–233, 1979.

[25] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *INTERSPEECH 2019*, Graz, Austria, 2019, pp. 1008–1012.

[26] 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.

[27] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[28] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–32, 1994.

[29] D. Chuong, B and S. Batzoglou, "What is the expectation maximization algorithm?" *Nature Biotechnology*, vol. 26, no. 8, pp. 897–899, 2008.

[30] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[31] D. A. Reynolds, "A Gaussian mixture modeling approach to text-independent speaker identification." Ph.D. dissertation, Georgia Institute of Technology, USA, 1993.

[32] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2^{nd} ed. John Wiley & Sons, 2012.

[33] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH*, Rhodes, Greece, Sept. 1997, pp. 1895–1898.