Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 1581–1590

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

Performance Evaluation of Normalization Techniques in Adverse Conditions

Renu Singh[a]* , Utpal Bhattacharjee[b], Arvind Kumar Singh[a]

[a]*Department of Electrical Engineering, North Eastern Regional Institute of Science and Technology, Itanagar, Arunachal Pradesh,Pin-791109,India*
[b]*Department of Computer Science and Engineering, Rajiv Gandhi University, Doimukh, Arunachal Pradesh, Pin-791112,India*

Abstract

This paper explores the behavior of different normalization techniques viz. cepstral mean normalization, cepstral variance normalization, cepstral mean subtraction. cepstral mean and variance normalization, wiener filter, and spectral subtraction in noisy conditions. The performance parameters viz. EER (Equal Error Rate) and DCF (Detection Cost Function) has been calculated using NIST 2003 SRE and Aurora 2 with the help of various normalization techniques considered in this paper for different noisy backgrounds at 0, 5 and 10 dB signal-to-noise ratio. The experimental results obtained from these techniques reveal that cepstral mean normalization (CVN) normalization method is found to be better when compared to other normalization techniques used in this paper.

Peer-review under responsibility of the scientific committee of the Third International Conference on Computing and Network Communications (CoCoNet'19).

Keywords: CMN; CVN; CMS; CMVN; SS; Wiener-Filtering, signal-to-noise ratio.

* Corresponding author. Tel.: +91-9436252336
E-mail address: renumona08@gmail.com

Peer-review under responsibility of the scientific committee of the Third International Conference on Computing and Network Communications (CoCoNet'19).
10.1016/j.procs.2020.04.169

1. Introduction- Voice biometric is one of the most prominent source of authentication in many areas like banking, online shopping, ATM transaction, access control, database security and applicable in many areas where it used for security purpose. Voice biometric system verifies the identity of a claimed person based on the extracted features of speech by comparing it with the stored voice templates. The performance of speaker recognition systems degrades in presence of noise [1 - 2]. Feature compensation (normalization) techniques are widely and effectively used for speaker recognition task such as speaker verification and speaker identification. Normalization process reduces the effect of noise and alleviates linear and non-linear channel effects. Robustness issue of the system can be improved by applying the normalization techniques [3].

To improve the performance of speaker recognition system various normalization techniques have been proposed which compensate the effects of environmental mismatch [4 – 9]. Robustness issue of speaker verification system has been improved by applying normalization techniques. Chougule and Chavan [4] have used CMN technique to minimize the influence of convolution noise in Hindi language based speech database. Barras and Gouvam [5] have presented some experimental analysis on text independent cellular data with the help of CMS, T-norm, Z-norm normalization techniques. Al-Kaltakchi et. al [6] have studied the speaker identification system which uses PNCCs and MFCC for feature extraction and CMVN as well as FW for the normalization of the system on acoustic model. Grozdic et. al [7] used various normalization techniques such as CVN, CMN, CMVN, CGN in their analysis of whisperd speech recognition system. Hardt and Fellbaum [8] used the spectral subtraction technique for analyzing telephonic based text dependent speaker verification system using different noises. Upadhyay and Jaiswal [9] have carried out the enhancement of single channel speech in stationary environments with the help of Wienner filtering normalization technique.

In this study, various normalization techniques have been considered for different noisy conditions. For the analysis NIST 2003 SRE and Aurora 2 data has been used for the calculation of performance parameters.

2. Normalization techniques

Normalization techniques are implemented to lower the noise impact, speech signal distortion and channel distortion. In paper, we discuss some normalization strategies, which are as follows:

2.1 Cepstral mean normalization (CMN)

CMN is the one of the simplest feature normalization technique to execute and it gives numbers of the advantages compared to advanced algorithms. To obtain normalized vector $\hat{x}_t$ CMN subtracts the mean feature vector μ_x from each vector x_t [10].

$$\mu_x = \frac{1}{T}\sum_t x_t \tag{1}$$

$$\hat{x}_t = x_t - \mu_x \tag{2}$$

2.2 Cepstral variance normalization (CVN)

Cepstral variance normalization (CVN) is a supplement technique of cepstral mean normalization which estimates varianceσ_n, of each cepstral dimension and normalizes it to unity [7].

$$C_{n,t}^{CVN} = \frac{C_{n,t}}{\sigma_n} = \frac{C_{n,t}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(C_{n,t}-\mu)}} \tag{3}$$

where n & t represents the nth cepstral dimension and the index of cepstral samples in the window respectively.

2.3 Cepstral mean subtraction (CMS)

Cepstral mean subtraction is one of the most extensively used normalization methods [11]. Reynolds [12] has reported an utterance X=$\{x_i\}$, i$\in[1, N]$ feature with a feature frame $\{x_i\}$, the mean vector (m) of all the frames for the given utterance, is considered as:

$$\mathrm{m} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{4}$$

The normalized feature $\hat{x}_t$ with CMS is expressed by

$$\hat{x}_t = x_i - m \tag{5}$$

2.4 Cepstral mean and variance normalization (CMVN)

Cepstral mean and variance normalization normalize both mean and variance algorithm.

$$\sigma x^2 = \frac{1}{T} \sum_{t=0}^{T-1} x_t^2 - \mu_x^2 \tag{6}$$

$$x_t = \frac{x_t - \mu_x}{\sigma_x}$$

After normalization, the mean of the cepstral sequence is zero, and it has a variance of one [13].

2.5 Wiener-filtering

The wiener filtering is wavelet-based used to suppress additive noise based on the concept of wiener gains which can be calculated given as,

$$k_m = \frac{S(a^2)m}{S(a^2)m + D(a^2)m} \tag{7}$$

where, $S(a^2)m$ and $D(a^2)m$ represents the speech power and noise power respectively [14].

2.6 Spectral subtraction (SS)

Boll in 1979 introduced spectral subtraction technique of speech enhancement which is used to reduce the additive noise [15-16].

$$y(m) = x(m) + n(m) \tag{8}$$

$y(m)$ represents noisy signal, $x(m)$ is the speech signal and $n(m)$ is the noise.
In frequency domain it can be represents as follow:

$$Y(j\omega) = X(j\omega) + N(j\omega) \tag{9}$$

where, $Y(j\omega), X(j\omega), N(j\omega)$ is fourier transforms of $y(m), x(m), n(m)$, respectively.

3. Baseline System based on GMM-UBM

Fig.1 shows the basic structure of ASR system, which consists following phases: pre-processing, front-end processing/feature extraction, training of model and testing/recognition.

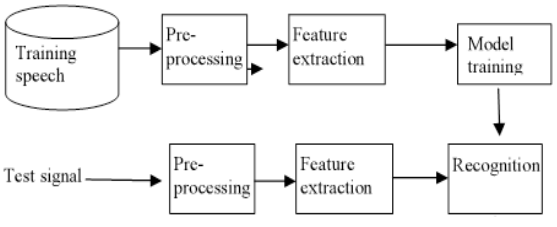


Fig. 1 Basic framework structure of ASR system [17]

3.1 Pre-processing

In pre-processing process speech data has been prepared. Pre-processing includes various tasks like sampling, pre-emphasis, framing (segmenting the speech into frames) and windowing.

3.2 Front end processing / feature extraction

Mel frequency cepstral coefficients, Bark scale filter bank cepstrum coefficients, Linear prediction cepstral coefficients and Perceptual linear prediction cepstral coefficients are the most common used acoustic vectors for speaker verification. All the features based on the spectral information are derived from a short time- windowed segment of speech. MFCC features are derived from the FFT power spectrum whereas LPCC and PLPC use an all-pole model to represent the smoothed spectrum. The proposed normalization technique used Mel-Frequency Cepstral Coefficients (MFCC) feature extraction for further processing [18].

3.3 Training of speaker model

GMM-UBM methodology has been considered as a speaker model for testing the speaker verification system; Fig. 2 shows the GMM-UBM based framework of speaker verification system. Firstly, speech features are extracted using mel-frequency cepstral coefficients after that a gender dependent universal background model is generated which based on Expectation Maximization (EM) algorithm.

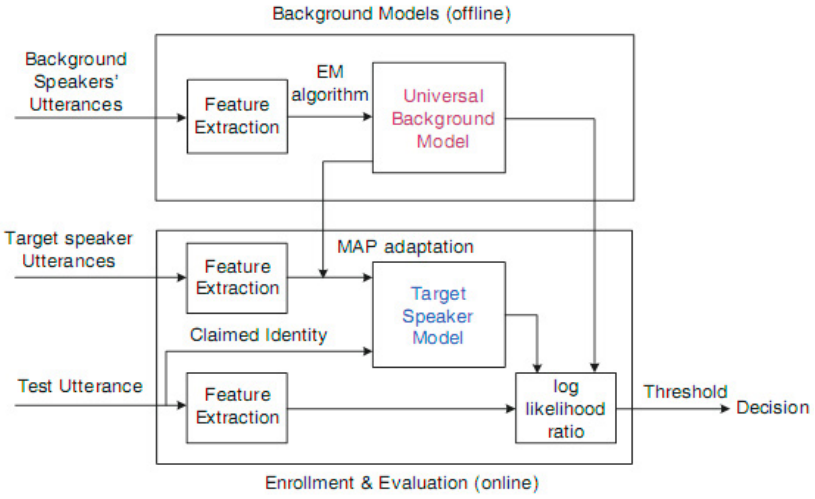


Fig. 2 GMM-UBM framework for speaker verification [19]

The computation of log-likelihood ratio $\wedge\ (X)$ has been done by scoring the test feature vectors against the claimant model and the universal background model by the expression given below:

$$\wedge\,(X) = \log p\left(X|\lambda_{hyp}\right) - \log p\left(X|\lambda_{\overline{hyp}}\right)$$

The claimant speaker is accepted if the value of $\wedge\,(X) \geq \theta$ or otherwise rejected. The substantial concern in speaker verification is to obtain a decision threshold θ for the decision making [20].

4. Experiments and results

In the present work, system performances are analysed in various types of additive noise background such as airport noise, babble noise, car noise and train noise at 0, 5, and 10 dB signal-to-noise ratio (SNR).

4.1 Database

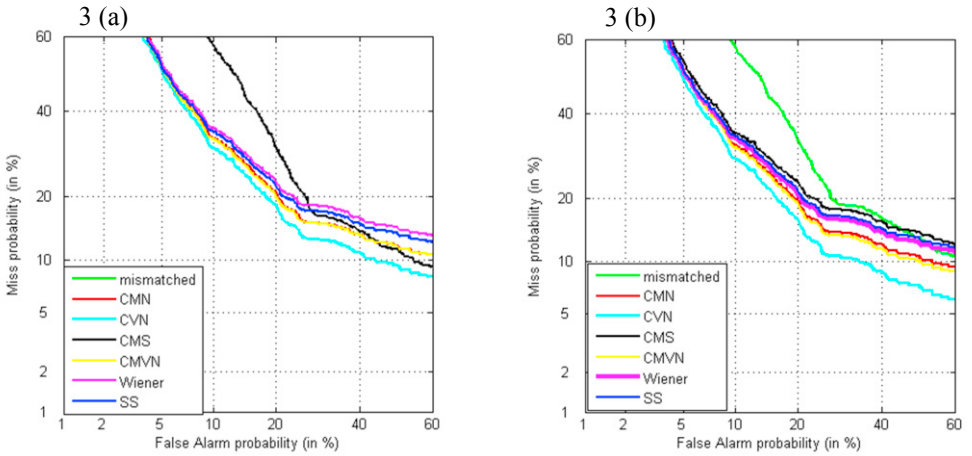
NIST -2003 –SRE database is used for the system development. The database consists of conversational speech of 149 male speakers. Aurora 2 database has been used for artificially simulating noisy environments at different SNR level.

4.2 Performance evaluation

In the present study, equal error rate (EER) has been used to quantify the performance of the system. Detection Error Trade-off (DET) curve is obtained by plotting the 'miss probability' (when a true identity is rejected) false alarm' (when an imposter's claim is accepted). The EER is the operating point in the DET curve where both the miss rates (P_{miss}) and false (P_{fa}) rates become equal, whereas DCF is the base estimation of a weighted cost function which is given by $0.01 * P_{miss} + 0.99 * P_{fa}$.

4.3 Results and discussions

Figures 3 - 5 represent the DET plots of mismatched conditions. The various normalization techniques which have been used in the paper are discussed in the section 2. Each set of curve in a subfigure deviate to a particular type of normalization methods in different noisy environments (airport noise, babble noise, car noise and street noise) at 0 dB, 5 dB and 10 dB SNRs respectively. The DET curve is consistently shifted towards origin with an increase in signal-to-noise ratio inferring performance improvement with reducing the strength of noise.. The performance summary of the normalization techniques is shown in Table 1. The order of precedence in terms of EER value for the system performance accuracy are mismatched, CMN, CVN, CMS, CMVN, Wiener filter and SS. MinDCF follows the same pattern with the exception of the way that they don't show a consistently pattern over the different techniques . The main exemption to this order is found in all noise types except babble noise at at 0 dB and 5 dB signal-to-noise ratios.



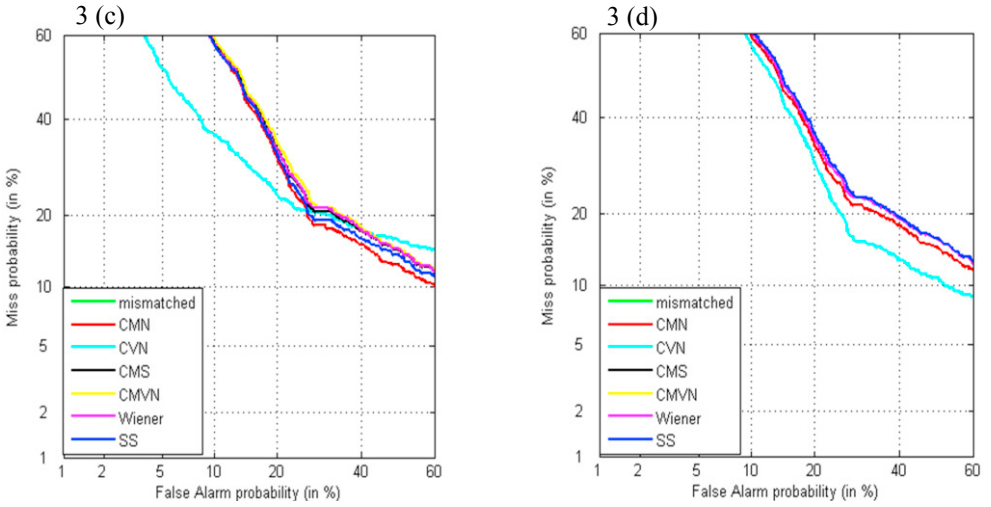
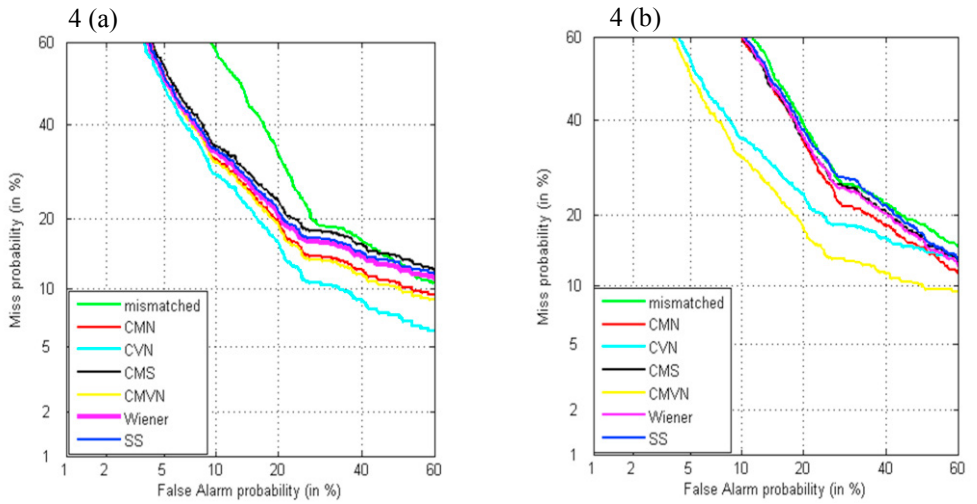


Fig. 3 DET plots of normalization techniques for (a) airport noise (b) babble noise (c) car noise (d) street noise at 0dB SNR



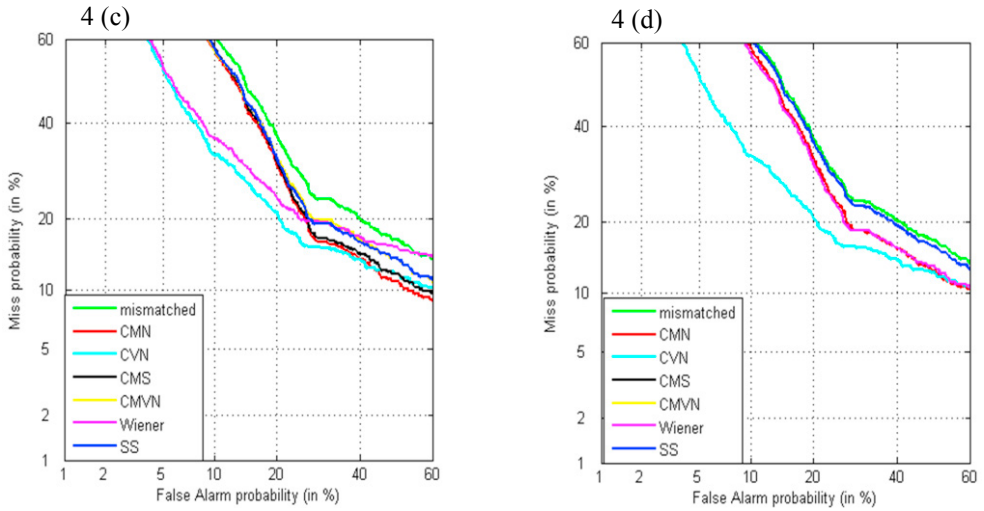
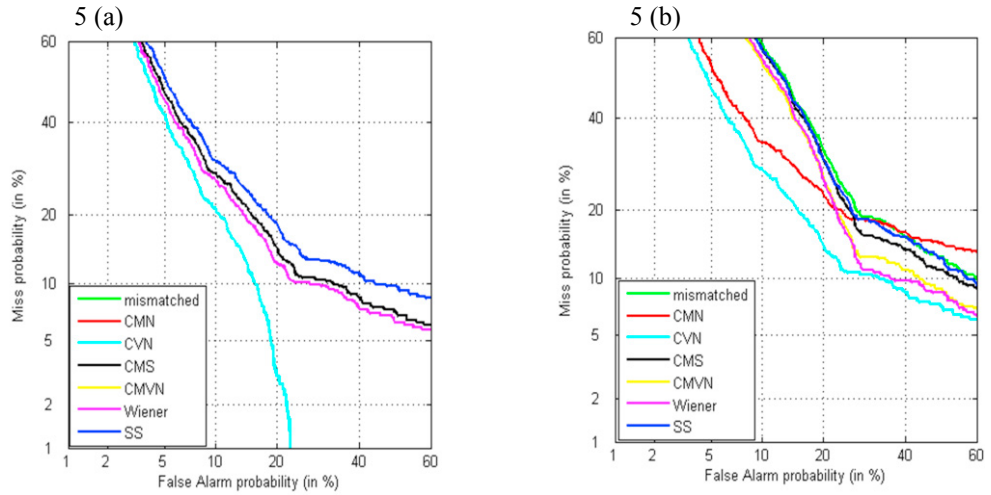


Fig. 4 DET plots of normalization techniques for (a) airport noise (b) babble noise (c) car noise (d) street noise at 5 dB SNR



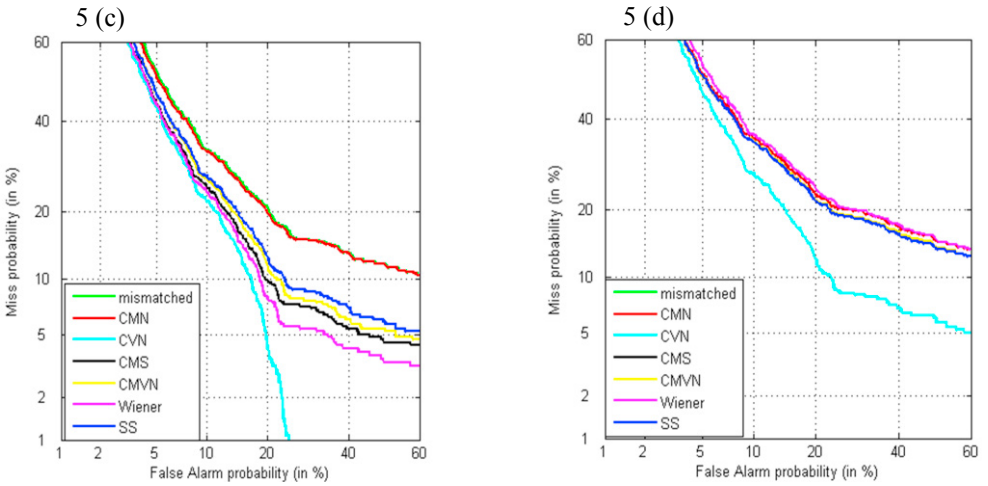


Fig. 5 DET plots of normalization techniques for (a) airport noise (b) babble noise (c) car noise (d) street noise at 10 dB SNR

Table 1 Performance of normalization techniques in different environments

SNR(dB)	Techniques	Airport		Babble		Car		Street	
		EER (%)	MinDCF	EER (%)	MinDCF	EER (%)	MinDCF	EER (%)	MinDCF
0	Mismatch	23.243	0.4204	28.4685	0.5393	25.5856	0.4683	26.3063	0.4853
	CMN	20.180	0.3733	26.8468	0.5015	24.1441	0.4384	25.5856	0.4709
	CVN	19.099	0.3481	23.9640	0.4330	22.7027	0.4238	23.0631	0.4114
	CMS	23.243	0.4204	27.0270	0.5128	24.8649	0.4552	26.3063	0.4114
	CMVN	19.819	0.3698	28.4685	0.5393	25.5856	0.4683	26.3063	0.4853
	WF	21.081	0.3986	27.7477	0.5250	25.0450	0.4590	25.9459	0.4793
	SS	20.721	0.3896	27.3874	0.5190	24.3243	0.4469	26.3063	0.4853
5	Mismatch	24.324	0.4430	27.0270	0.5123	26.4865	0.4899	26.6667	0.4943
	CMN	19.819	0.3607	25.7658	0.4709	23.2432	0.4186	24.3243	0.4420
	CVN	17.838	0.3265	21.2613	0.4022	20.3604	0.3751	20.1802	0.3776
	CMS	21.261	0.3986	26.3063	0.4967	23.4234	0.4240	26.3063	0.4114
	CMVN	19.279	0.3553	19.0991	0.3481	24.5045	0.4469	26.3063	0.4114
	WF	20.180	0.3776	26.3063	0.4962	21.9820	0.4112	23.9640	0.4354
	SS	20.540	0.3842	27.3874	0.5190	24.5045	0.4469	26.3063	0.4853
10	Mismatch	19.279	0.3553	24.3243	0.4424	20.1802	0.3733	22.1622	0.4145
	CMN	14.054	0.2131	20.9009	0.3981	19.8198	0.3688	21.9820	0.4054
	CVN	13.874	0.2093	17.2973	0.3166	14.2342	0.2196	16.7568	0.2943
	CMS	17.477	0.3177	22.8829	0.4132	15.3153	0.2738	20.7207	0.3902
	CMVN	13.874	0.2111	21.9820	0.3759	16.3964	0.2896	21.0811	0.3946
	WF	16.577	0.3008	21.8018	0.3664	14.9550	0.2562	19.8198	0.3632
	SS	14.414	0.2209	23.6036	0.4295	16.9369	0.2997	20.9009	0.3928

Table 2 Performance summary of normalization techniques in different environments

Techniques	EER (%)
Mismatch	24.504
CMN	22.222
CVN	19.219
CMS	23.511
CMVN	21.891
WF	22.312
SS	22.778

The worst case scenario shown by mismatched condition for all types of noisy environment with an average value of EER equals 24.504% over all the noises considered for various SNRs. The MinDCF values varied in the range of 0.2111-0.4853 (Table 1). CMN normalization method shows decrement of 2.282% EER over the mismatch condition. However CVN performs far better than other utilized techniques with an improvement of 5.285% EER for airport, babble, car and street background environments shown in Table 2 with respected to mismatched condition. CMS shows minor improvement of 0.993% decrement of EER over the mismatch condition. Other used methods such as CMVN, Wiener filter and SS shows 2.613, 2.192 and 1.726% respectively decrement in EER with respect to mismatch condition. Wiener filter is seen moderately better than CMN, CMVN and SS normalization methods. Compared to other methods CVN show the highest 7.928% drop in EER in case of babble noise at 5 dB SNR, while reduction of 7.027% in EER for babble noise at 10dB SNR has been noticed.

The comparative improvement of normalization methods performance is visible. The average value for EER of 22.222% for CMN, 19.219% for CVN, 23.511% for CMS, 21.891% for CMVN, 22.312% for wiener filter and 22.778% for SS across all noisy environments is acheived.

5. Conclusions

Normalization techniques have been used to compensate the effects of environmental mismatch. In this study behavior of normalization techniques viz. cepstral mean normalization, cepstral variance normalization, cepstral mean subtraction, cepstral mean and variance normalization, wiener filter and spectral subtraction have been studied under different noisy environments such as airport noise, babble noise, car noise and street noise at 0,5 and 10 dB signal-to-noise ratio. On the basis of experimental results it has been concluded that the cepstral variance normalization (CVN) method is comparatively better as compared to other used normalization methods which have been used in this paper and it reveals 5.285% of EER improvement over mismatch condition across all noisy environments and all SNRs whereas other used methods show improvement of 2.282% for CMN, 0.993% for CMS, 2.613% for CMVN, 2.192% for wiener filter and 1.726% for SS respectively over mismatch condition across all SNRs with respect to mismatch condition.

6. References

[1] W. Kim and J. H. L. Hansen (2009), "Feature compensation in the cepstral domain employing model combination," *Speech Communication*, **51**: 83-96.

[2] S. Furui (1997), "Recent advances in speaker recognition," *Pattern Recognition Letters*, **18**: 859-871.

[3] O. Viikki and K. Laurila (1998), "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, **25**: 133-147.

[4] S. K. Chougule and S. Chavan (2014), "Channel robust MFCCs for continuous speech speaker recognition," *Advances in Signal Processing and Intelligent Recognition Systems,* 557-568 (DOI: 10.1007/978-3-319-04960-1_48).

[5] C. Barras and J. L. Gauvain (2003), "Feature and score normalization for speaker verification of cellular data," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 49-52 **(**DOI: 10.1109/ICASSP.2003.1202291).

[6] M. T. S. Al-Kaltakchi, W. L. Woo, S. Dlay and J. A. Chambers (2017), "Evaluation of a speaker identification system with and without fusion using three databases in the presence of noise and handset effects," *EURASIP Journal on Advances in Signal Processing*, **80**: 2017 (DOI: 10.1186/s13634-017-0515-7).

[7] D. Grozdic, S. Jovicic, D. S. Pavlovic, J. Galic and B. Markovic (2017), "Comparison of cepstral normalization techniques in whispered speech recognition," *Advances in Electrical and Computer Engineering*, 17: 21-26 (DOI: 10.4316/AECE.2017.01004).

[8] D. Hardt and K. Fellbum (1997), "Spectral subtraction and Rasta- filtering in text-dependent HMM-based speaker verification," *IEEE International Conference on Acoustics, Speech and Signal Processing,* 867-870 (DOI: 10.1109/ICASSP.1997.596073).

[9] N. Upadhyay and R.K.Jaiswal (2016), " Signal channel speech enhancement: using Wiener filtering with recursive noise estimation," *Procedia Computer Science,* 22-30 (DOI: 10.1016/j.procs.2016.04.061).

[10] J. Benesty, M. M. Sondhi and Y. Huang (2008), *Springer Handbook of Speech Processing*.

[11] W. Dalei, Li Baojie and J. Hui (2008) "Normalization and transformation techniques for robust speaker verification," *Speech Recognition, Technologies and Applications*, Book edited by: France Mihelič and Janez Žibert, 550, November, I-Tech publisher.

[12] Reynolds, D. A (2002), "An overview of automatic speaker recognition technology," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4072-4075 (DOI: 10.1109/ICASSP.2002.5745552).

[13] A. Drygajlo and M. E. Muliki (1998), "Speaker verification in noisy environements with combined spectral subtraction and missing feature theory," *IEEE International Conference on Acoustics Speech and Signal Processing*, 121-124 (DOI: 10.1109/ICASSP.1998.674382).

[14] R. Gomez and T. Kawahara (2010), "Optimizing spectral subtraction and wiener filtering for robust speech recognition in reverberant and noisy conditions," *IEEE International Conference on Acoustics Speech and Signal Processing*, 4566-4569 (DOI: 10.1109/ICASSP.2010.5495568).

[15] J. Beh and H. Ko (2003), "A novel spectral subtraction scheme for robust speech recogniton: spectral subtraction using spectral harmonics of speech, *IEEE International Conference on Acoustics Speech and Signal Processing*, 648-651 (DOI: 10.1109/ICASSP.2003.1198864).

[16] B. Kumar (2018), "Comparative performance evaluation of MMSE-based speech enhancement techniques through simulation and real - time implementation," *International Journal of Speech Technology*, 735-756.

[17] A. I. Amorous and M. Debyeche (2011), "Robust Arabic speech recognition in noisy environments using prosodic features and formants," *International Journal of Speech Technology*, 351-359.

[18] D. A. Reynolds (1997), "Comparison of background normalization methods for text-independent speaker verification," *EuroSpeech*, 963-966.

[19] K. S. Rao and S. Sarkar (2014), "Robust speaker recognition in noisy environments," *Springer Briefs in Electrical and Computer Engineering* (DOI: 10.1007/978-3-319-07130-5-3).

[20] A. Douglas, F. Thomas, Quitieri and B. B. Robert (2000), "Speaker verification using adapted guassian mixture model," *Digital Signal Processing*, 19-41.