Robotics & Automation Engineering Journal
ISSN: 2577-2899

Juniper Publishers
key to the Researchers

Short Communication
Volume 1 Issue 3 - September 2017
DOI: 10.19080/RAEJ.2017.01.555564

Robot Autom Eng J

Spectral Restoration based Speech Enhancement for Robust Speaker Identification

Nasir Saleem[1]* and Tayyaba Gul Tareen[2]

[1]*Department of electrical Engineering, Gomal University, Pakistan*

[2]*Department of electrical Engineering, Iqra National University, Pakistan*

Submission: August 15, 2017; **Published:** September 21, 2017

***Corresponding author:** Nasir Saleem, Department of electrical Engineering, Gomal University, Pakistan, Tel: 9-20333E-12; Email: nasirsaleem@gu.edu.pk

Abstract

Spectral restoration based speech enhancement algorithms are used to enhance quality of noise masked speech for robust speaker identification (SID). In presence of background noise, the performance of speaker identification systems can be severely deteriorated. The present study employed and evaluated the Minimum Mean-Square-Error Short-Time Spectral Amplitude Estimators (MMSE-STSA) with modified a priori SNR estimate prior to speaker identification to improve performance of the speaker identification systems in presence of background noise. For speaker identification, Mel Frequency Cepstral coefficient (MFCC) and Vector Quantization (VQ) is used to extract the speech features and to model the extracted features respectively. The experimental results showed significant improvement in speaker identification rates when spectral restoration based speech enhancement algorithms are used as a pre-processing step. The identification rates are found to be higher after employing the speech enhancement algorithms.

Keywords: *a priori* SNR; Spectral restoration; speech enhancement; speaker identification; MFCC; VQ

Introduction

Speech enhancement aspires to improve quality by employing a variety of speech processing algorithms. The intention of the enhancement is to improve the speech intelligibility and/or overall perceptual quality of speech noise masked speech. Enhancement of speech degraded by background noise, called noise reduction is significant area of speech enhancement and is considered for diverse applications e.g., mobile phones, speech/speaker recognition/identification and hearing aids. The speech signals are frequently contaminated by the background noise, which affects the performance of speaker identification (SID) systems. The SID systems are used in online banking, voice mail, remote computer access etc. Therefore, for effective use of such systems, a speech enhancement system must be positioned in front-end to improve identification accuracy. Figure 1 shows the procedural block diagram of speech enhancement and speaker identification system. The algorithms for speech enhancement are categorized into three fundamental classes, (i) filtering techniques including spectral subtraction [1-4], Wiener filtering [5-7] and signal subspace techniques [8-9], (ii) Spectral restoration algorithms including Mean-Square-Error Short-Time Spectral Amplitude Estimators [10-12] and (iii) speech-model based algorithms. The systems in [5-7,10-12] principally depend on accurate estimates of signal-to-noise ratio (SNR) in all frequency bands, because gain is computed as function of spectral SNR. A conventional and recognized technique for SNR estimate is decision-directed (DD) method suggested in [10]. The DD technique tails the shape of instantaneous SNR for a priori SNR estimate brings in one-frame delay. Therefore; to avoid one-frame delay, momentum terms are incorporated to get better tracking speed of system and avoid the frame delay problem. All the mentioned systems in [10-12] can significantly improve speech quality. Binary masking [13-18] is another class that increases speech quality and intelligibility simultaneously. This paper presents Mean-Square-Error Short-Time Spectral Amplitude Estimators with modified a priori SNR estimate to reduce background noise and to improve identification rates of speaker identification systems in presence of background noises. The paper is prepared as follows. Section 2 presents the overview of speech enhancement system; section 3 gives speaker identification system; section 4 presents the experimental setup, results

and discussions, and section 5 presents the summary and concluding remarks. The Matlab R2015b is used to construct the algorithms and simulations.

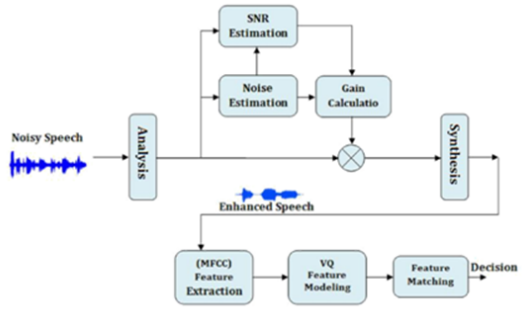

Figure 1: Proposed Speech enhancement and speaker identification system

Spectral Restoration based Speech Enhancement System

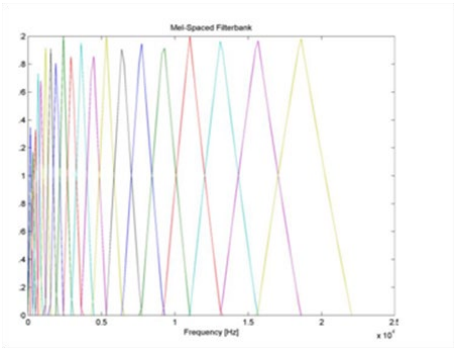

Figure 2: Mel-Spaced Filter bank Response

In classical spectral restoration system based speech enhancement system, the noisy speech is given as; $y(t)=s(t)+n(t)$, where $s(t)$ and $n(t)$ specify clean speech and noise signal respectively. Let $Y(k,\omega k)$, $S(k,\omega k)$ and $N(k,\omega k)$ shows $y(t)$, $s(t)$ and $n(t)$ respectively with spectral element ωk and time frame k. The quasi-stationary nature of speech is considered in frame analysis since noise and speech signals both reveal non-stationary behavior (Figrue 2). A speech enhancement algorithm involves in multiplication of a spectral gain $G(k,\omega k)$ to short-time spectrum $Y(k,\omega k)$ and the computation of spectral gain follows two key parameters, a posteriori SNR and the a priori SNR estimate:

$$\gamma(k,\omega_k)=\frac{|Y(k,\omega_k)|^2}{E\{|N(k,\omega_k)|^2\}}=\frac{|Y(k,\omega_k)|^2}{\sigma_n^2(k,\omega_k)|} \quad (1)$$

$$\xi(k,\omega_k)=\frac{E\{|S(k,\omega_k)|^2\}}{E\{|N(k,\omega_k)|^2\}}=\frac{\sigma_s^2(k,\omega_k)}{\sigma_n^2(k,\omega_k)} \quad (2)$$

Where E{.} shows expectation operator, $\gamma(k,\omega_k)$ and $\xi(k,\omega k)$ presents a posteriori and a priori SNR estimate. In practical implementations of a speech enhancement system, squared power spectrum density of clean speech $|X(k,\omega k)|^2$ and noise $|D(k,\omega k)|^2$ are unrevealed as only noisy speech is available. Therefore; both instantaneous and *a priori* SNR need to be estimated. The noise power spectral density is estimated during speech gaps exploiting standard recursive relation, given as:

$$\hat{\sigma}_n^2\gamma(k,\omega_k)=\beta\hat{\sigma}_n^2(k-1,\omega_k)+(1-\beta)\tilde{\sigma}_Y^2(k-1,\omega_k) \quad (3)$$

Where, β is a smoothing factor and $\tilde{\sigma}_Y^2(k-1,\omega_k)$ is estimate in previous frame. The SNRs can be calculated as:

$$SNR_{INST}(k,\omega_k)=\frac{|S(k,\omega_k)|^2}{|N(k,\omega_k)|^2} \quad (4)$$

$$\xi_{DD}(k,\omega_k)=\alpha\frac{|G(k-1,\omega_k)*Y(k,\omega_k)|^2}{\hat{\sigma}_n^2(k,\omega_k-1)}+(1-\alpha)F\{\gamma(k,\omega_k)-1\} \quad (5)$$

Where α is smoothing factor and has a constant value 0.98, $\xi DD(k,\omega k)$ is a priori noise estimate via decision-direct (DD) method whereas $F\{.\}$ is half-wave rectification. The DD is efficient method and achieve well in speech enhancement applications however; the a priori SNR follows the shape of instantaneous SNR and brings single-frame delay. To overcome the single-frame delay, a modified form of DD approach is used to estimate a priori SNR. The modified a priori SNR can be written as: (6)

$$\xi_{MDD}(k,\omega_K)=\alpha\frac{|G(k,\omega_K)*Y(k,\omega_K)|^2}{\hat{\sigma}_n^2(k,\omega_k-1)}+\mu(k,\omega_K)+(1-\alpha)F\{\gamma(k,\omega_K)-1\}\,\mu(k,\omega_K)=\zeta[\xi_{PRIO}(k,\omega_K)-\xi_{PRIO}(k,\omega_K)]$$

The Eq.6 shows the modified DD (MDD) version used in the speech enhancement system, α is smoothing parameter (α=0.98), ζ is momentum parameter $(\zeta=0.998)$, $\mu(m,\omega k)$ shows momentum terms and $\lambda D(m,\omega k)$ is estimate of background noise variance. The $\xi MDD(k,\omega k)$ shows a priori SNR estimate after modification. The estimated power spectrum of the clean speech magnitude $SEST(k,\omega k)$ is attained by multiplying the gain function with noisy speech $Y(k,\omega k)$ as:

$$|S_{EST}(k,\omega_k)|=|Y(k,\omega_k)|*G(k,\omega_k) \quad (7)$$

The gain function $G(k,\omega k)$ is given as:

$$G(k,\omega_k)=\min\left\{\zeta,\frac{\xi(k,\omega_k)}{1+\xi(k,\omega_k)}\left[\frac{1}{2}\int_{v(k,\omega_k)}^{\infty}\right]\right\} \quad (8)$$

Where, ς ς is used to avoid large gain values at low a posteriori SNR and choose $\varsigma=10$ here.

Speaker Identification System

The intention of a Speaker identification system is to identity information regarding any speaker and categorized into two sub-categories called as Speaker identification (SID) and speaker Verification (SVR). For SID, the Mel Frequency Cepstral coefficient (MFCC) and Vector Quantization (VQ) is used to extract the speech features and to model the extracted features respectively. The speaker identification system drives in two stages, the Training and testing stages. In training mode the system is allowed to create the database of speech signals

How to cite this article: Nasir S, Tayyaba G T. Spectral Restoration based Speech Enhancement for Robust Speaker Identification. Robot Autom Eng J. 2017; 1(3): 555564. DOI: 10.19080/RAEJ.2017.01.555564.

and formulate a feature model of speech utterances. In testing mode, the system used information provided in database and attempt to segregate and identify the speakers. Here, the Mel frequency Cepstral Coefficients (MFCCs) features are used for constructing a SID system. The extracted features of speakers are quantized to a number of centroids employing vector quantization (VQ) K-means algorithm. MFCCs are computed in training as well as in testing stage. The Euclidean distance among MFCCs of all speakers in training stage to centroids of isolated speaker in testing stage is calculated and a particular speaker is identified according to minimum Euclidean distance.

Feature Extraction

The MFCCs are acquired by pre-emphasis [ref] of speech initially to emphasize high frequencies and eliminate glottal and lip radiations. The resulting speech is fragmented, windowed, and FFT is computed to attain spectra. To estimate human auditory system, triangular band-pass filters bank is utilized. For center frequencies lower than 1kHz, a linear relation while beyond 1kHz, a logarithmic relation is assumed. The filter bank response is given in Fig. 2. The Mel-spaced filter bank response is given as:

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \quad (9)$$

The DFT is computed on the log of Mel spectrum to compute Cepstrum as:

$$M_k = \sqrt{\frac{2}{N_f}} \sum_{n=1}^{N_r} \log(\dot{S}(n)) \cos\left(\frac{g\pi}{N_f}(n-0.5)\right) \quad (10)$$

Where Mg shows MFCCs, $\dot{S}$ is nth Mel filter output, K is number of MFCCs chosen between 5 to 26 and N_f is the number of Mel filters. Initial few coefficients are considered since most of the specific information about speakers is present in them [ref].

Vector Quantization

Vector quantization (VQ) is a lossy compression method based on the block coding theory [19]. The purpose of VQ in speaker recognition systems is to create a classification system for every speaker and a large set of acoustic vectors are converted to lesser set that signifies centroids of distribution shown in Figure 2. The VQ is employed since all MFCC generated feature vector cannot be stored and extracted acoustic vectors are clustered into a set of codewords (referred to as codebook) and this clustering is achieved by using the K-Means Algorithm which separates the M feature vectors into K centroids. Initially K cluster-centroids are chosen randomly within M feature vectors and then all feature vectors are allocated to nearby centroid, and the formation of c

entroids new clusters follows this pattern. The process keeps on until a certain condition for stopping is reached, i.e., the mean square error (MSE) among acoustic vector and cluster centroid is lower than a certain predefined threshold or no additional variations in cluster-center task [20-21].

Speaker Identification

The speaker recognition phase is characterized by a set of acoustic feature vectors ,$\{M1, M2,\ldots, Mt\}$ and is judged against codebooks in list. For all codebooks a distortion is calculated, and a speaker with the lowest distortion is selected, and this distortion is sum of squared Euclidean distances among vectors and their centroids. As a result, all feature vectors in M sequence are compared with codebooks, and the codebooks with the minimum average distance are selected. The Euclidean distance between two points, $\lambda = (\lambda 1, \lambda 2 \ldots \lambda n)$ and $\eta = (\eta 1, \eta 2 \ldots \eta n)$ is given by [21]:

$$\sqrt{\left[(\lambda-\gamma)^2 + (\lambda-\gamma)^2 + (\lambda-\gamma)^2 + \ldots\ldots + (\lambda-\gamma)^2\right]} = \sqrt{\sum_{i=1}^{n} (\lambda-\gamma)^2} \quad (11)$$

Results and Discussion

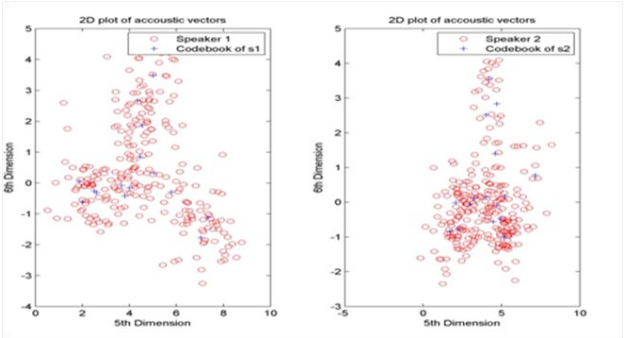

Figure 3: 2D acoustic Vector analysis.

Six different speakers, three male and three female, were selected from Noizeus [22] and TIMIT database respectively. To evaluate the performance of system, four signal-to-noise ratio levels, including 0dB, 5dB, 10dB and 15dB are used. Also three noisy situations including car, street and white noise are used to degrade the Figure 3: 2D acoustic Vector analysis clean speech. The Perceptual evaluation of speech quality (PESQ) [23] and Segmental SNR (SNRSeg) is used to predict the speech quality after speech enhancement. Three sets of experiments are conducted to measure the speaker identification rates including, clean speech with no background noise, speech degraded by background noise and speech processed by the spectral restoration enhancing algorithms. Figure 4 shows PESQ scores obtained after applying Minimum Mean-Square-Error Short-Time Spectral Amplitude Estimators with modified a priori SNR estimate (MMSE-MDD). The modified version offers the best results consistently in all SNR levels and noisy conditions when compared to noisy and speech processed by traditional MMSE-STSA speech enhancement algorithm. Similarly, Figure 5 shows speech quality in terms of segmental SNR (SNRSeg) where highest SNRSeg scores are obtained with MMSE-MDD. The enhanced speech associated with six speakers is tested for speaker identification. Figure 6 shows identification rates, the lowest identification rates are observed in the presence of background noise (Babble, car and street) however, employment of the speech enhancement

How to cite this article: Nasir S, Tayyaba G T. Spectral Restoration based Speech Enhancement for Robust Speaker Identification. Robot Autom Eng J. 2017; 1(3): 555564. DOI: 10.19080/RAEJ.2017.01.555564.

before speaker identification has tremendously increased the identification rates which are evident in Figure 5. The identification rates for MMSE-MDD are higher in all SNR conditions and levels.

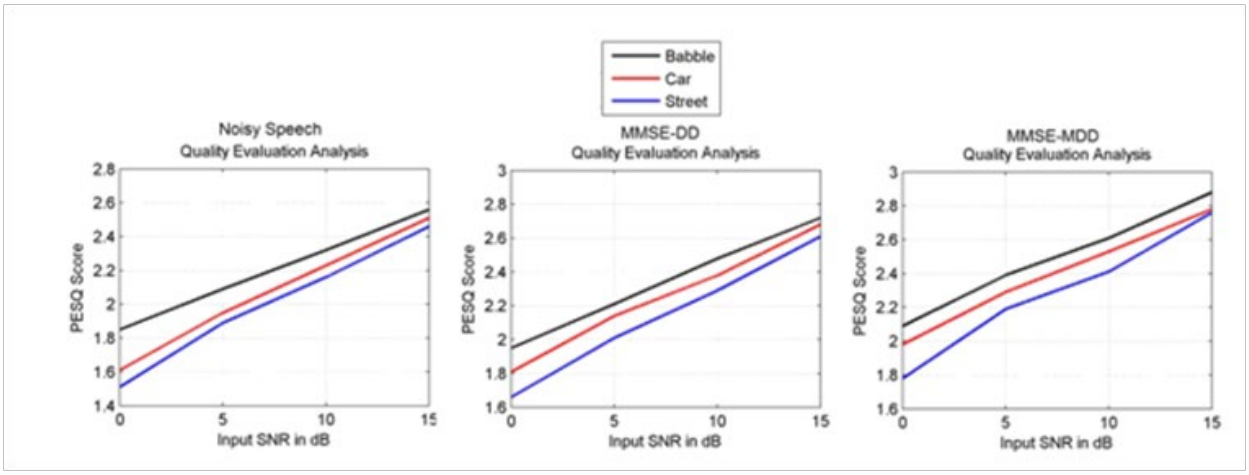

Figure 4: PESQ: Speeh Quality Analysis.

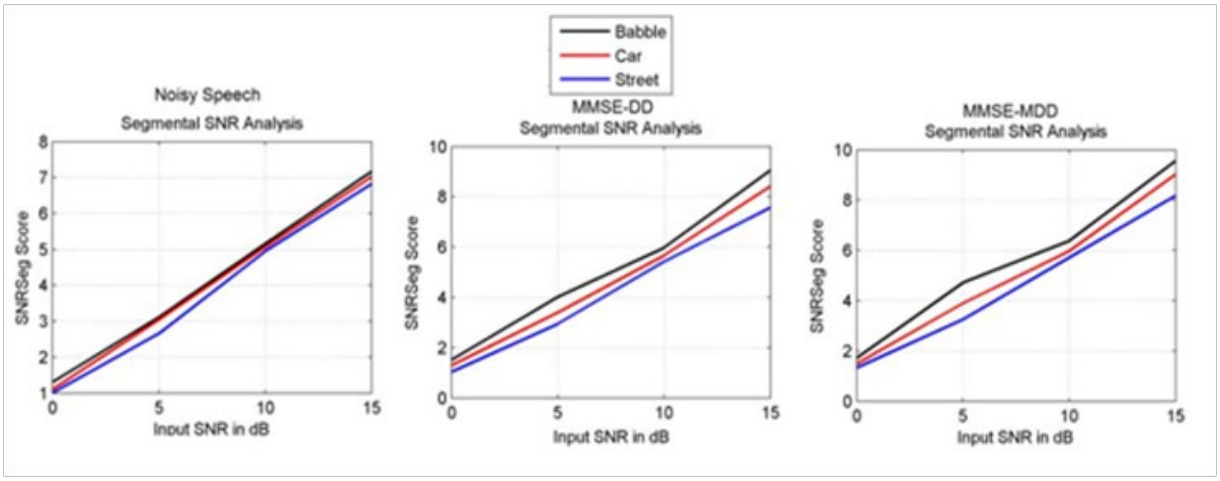

Figure 5: SNRSeg: Segmental SNR Analysis.

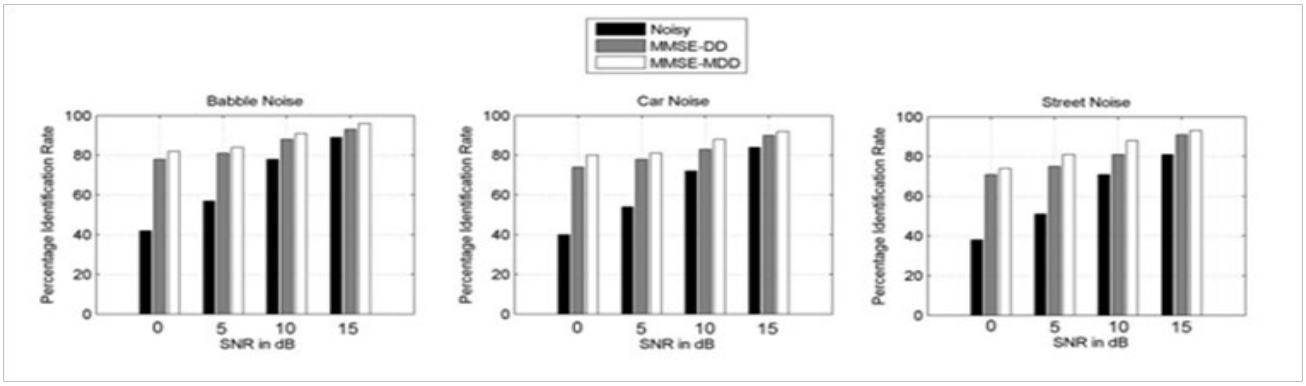

Figure 6: Speaker Identification Rate Analysis.

How to cite this article: Nasir S, Tayyaba G T. Spectral Restoration based Speech Enhancement for Robust Speaker Identification. Robot Autom Eng J. 2017; 1(3): 555564. DOI: 10.19080/RAEJ.2017.01.555564.

Summary and Conclusion

This paper presents Mean-Square-Error Short-Time Spectral Amplitude Estimators with modified *a priori* SNR estimate to reduce the background noise and to improve identification rates of speaker identification systems in presence of background noises. The lowest identification rates are reported when background noises such as Babble, car and street are present however; the use of a speech enhancement system prior to speaker identification remarkably increased the identification rates. On the basis of experimental results, it is concluded and suggested that the use of a speech enhancement system at front-end is necessary when a speaker identification system is working in a noisy environment.

References

1. Berouti M, Schwartz M, Makhoul J (1979) Enhancement of speech corrupted by acoustic noise. Proc IEEE Int Conf Acoust Speech Signal Processing 208-211.
2. Kamath S, Loizou P (2002) A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. Proc. IEEE Int Conf Acoust Speech Signal Processing, Orlando, USA .
3. Gustafsson H, Nordholm S, Claesson I (2001) Spectral subtraction using reduced delay convolution and adaptive averaging. IEEE Trans. on Speech and Audio Processing 9(8): 799-807.
4. Nasir S, Sher A, Usman K, Farman U (2013) Speech Enhancement with Geometric Advent of Spectral Subtraction using Connected Time-Frequency Regions Noise Estimation. Research Journal of Applied Sciences Engineering and Technology 6(06): 1081-1087.
5. Lim J,Oppenheim AV (1978) All-pole modeling of degraded speech. IEEE Trans Acoust Speech Signal Proc 26(3): 197-210.
6. Scalart P, Filho J (1996) Speech enhancement based on a priori signal to noise estimation. Proc IEEE Int Conf Acoust Speech Signal Processing, pp. 629-632.
7. Hu Y, Loizou P (2004) Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans on Speech and Audio Processing 12(1): 59-67.
8. Hu Y, Loizou P (2003) A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans. on Speech and Audio Processing 11: 334-341.
9. Jabloun F, Champagne B (2003) Incorporating the human hearing properties in the signal subspace approach for speech enhancement. IEEE Trans on Speech and Audio Processing, 11(6): 700-708.
10. Ephraim Y, Malah D (1984) Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process ASSP 32(6): 1109-1121.
11. Ephraim Y, Malah D (1985) Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process ASSP 23(2): 443-445.
12. Cohen I (2002) Optimal speech enhancement under signal presence uncertainty using log-spectra amplitude estimator. IEEE Signal Processing Letters 9(4): 113-116.
13. Saleem N, Mustafa E, Nawaz A, Khan A (2015) Ideal binary masking for reducing convolutive noise. International Journal of Speech Technology 18(4): 547-554.
14. Saleem N, Shafi M, Mustafa E, Nawaz A (2015) A novel binary mask estimation based on spectral subtraction gain induced distortions for improved speech intelligibility and quality. Technical Journal UET Taxila 20(4): 35-42.
15. Saleem N (2016) Single channel noise reduction system in low SNR. International Journal of Speech Technology 20(1): 89-98.
16. Boldt JB, Kjems U, Pedersen MS, Lunner T, Wang D (2008) Estimation of the ideal binary mask using directional systems. In Proc int workshop acoust echo and noise control, pp. 1–4.
17. Wang D (2005) On ideal binary mask as the computational goal of auditory scene analysis. In Speech separation by humans and machines, pp.181-197.
18. Wang D (2008) Time-frequency masking for speech separation and its potential for hearing aid design. Trends Amplif 12(4): 332-353.
19. Gray RM (1984) Vector Quantization. IEEE ASSP Magazine, pp. 4-29.
20. Likas A, Vlassis, Verbeek JJ (2003) The global k-means clustering algorithm. Pattern Recognition 36(2): 451-461.
21. Khan SS, Ahmed A (2004) Cluster center initialization for K means algorithm. Pattern Recognition Letters. 25: 11.
22. Hu Y, Loizou P (2007) Subjective evaluation and comparison of speech enhancement algorithms. Speech Commun 49(7-8): 588-601.
23. Rix AW, Beerends JG, Hollier MP, Hekstra AP (2001) Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Acoustics, Speech, and Signal Processing (ICASSP), Pp. 749-752.

DOI: 10.19080/RAEJ.2017.01.555564

How to cite this article: Nasir S, Tayyaba G T. Spectral Restoration based Speech Enhancement for Robust Speaker Identification. Robot Autom Eng J. 2017; 1(3): 555564. DOI: 10.19080/RAEJ.2017.01.555564.