Robust Automatic Speech Recognition System Based on Using Adaptive Time-Frequency Masking

Ahmed Mostafa Gouda, Mohamed Tamazin, Mohamed Khedr
Department of Electronics and Communications Engineering
College of Engineering and Technology
Arab Academy for Science, Technology and Maritime Transport
Alexandria, Egypt
Email: AhmedMost.Gouda@gmail.com

***Abstract*—The Automatic Speech Recognition (ASR) systems suffer from many types of noises in different environments. Nowadays, developing robust ASR system is an attractive research topic due to the high demands in many commercial applications. In this paper, the Mel-Frequency Cepstral Coefficients (MFCC) is modified to robust the noise, where the spectrogram is used as time-frequency analysis tool. The proposed system is designed to remove the energy, which is affected by the noise. It uses an adaptive filtering technique to robust the noise without loss of performance in case of undistorted speech data. The proposed system is evaluated in a noisy environment. The experimental results demonstrated that the proposed MFCC method provides significant improvements in recognition accuracy at low Signal to Noise Ratio (SNR). The average recognition accuracy is improved by 11% and 12.56% compared to standard MFCC and RASTA-PLP, respectively.**

***Keywords*—MFCC, robust speech recognition, feature extraction, time-frequency, mask estimation.**

I. INTRODUCTION

In the recent years, the new technologies, which are serving the humanity, have been expanded. The Automatic Speech Recognition (ASR) system is one of the most important technologies, which is used as a man-machine interface for real-world applications. Nowadays, the ASR system is used in many commercial systems, such as smartphones (e.g., Siri on iPhone, and Google on Android) and personal computers. It is based on getting the corresponding sequence of words using speech signals. Many researchers are motivated to build auditory-based speech processing systems, which have the ability to process speech in the presence of non-stationary disturbances such as noise and background interference. The non-stationary disturbances degrade the performance of the ASR systems, when they are used in several noisy environments.

In the literature, most of speech recognition systems were designed to robust noise in different levels and in various noisy environments. Speech-feature-based systems usually rely on spectral operations techniques such as Linear Predictive Cepestral Coefficients (LPCC) [1], Mel-Frequency Cepstral Coefficients (MFCC) [2] and Perceptual Linear Prediction (PLP) [3]. An enhancement method was used to suppress the background noise using a band-pass filtering like RelAtive SpecTrAl (RASTA) algorithm [4]. Another technique was used to remove the channel instability such as Cepstral Mean and Variance Normalization (CMVN) [5]. There are another advanced noise robust techniques, which include Sparse Perceptual Minimum Variance Distortionless Response (PMVDR) [6], Auditory Reproducing Kernel (SPARK) [7], and Power-Normalized Cepstral Coefficients (PNCC) [8].

The main objective of this paper is to improve the performance of the MFCC system in noisy environments by proposing an adaptive signal-processing technique that can maintain the useful information context in speech signal and eliminate the corrupted information due to noise. In order to mitigate the noise effects, an adaptive unique mask for each word is proposed by taking in concern the useful spectral information located in the spectrogram. The spectrogram is considered as a feature extraction method, which represents the squared magnitude of the time-frequency of speech signal.

This paper is organized as follows: Section II reviews the feature extraction techniques used in the literature; Section III discusses the proposed adaptive method; Section IV shows the experiment and the results; Section V summarizes the outcomes of the paper and future works.

II. METHODLOGY

The well-known speech recognition systems in literature [9] are based on two stages, which are feature extraction stage and classification stage. These two stages are shown in ASR system as in Figure (1).

As shown in Figure (1), Feature extraction is the first stage of speech recognition process. It converts the speech waveform data into reduced feature vectors by retaining the discriminative and non-redundant information in the speech data. These features represent the main characteristics of each word. If these features are extracted correctly, this will lead to high recognition accuracy in the classification stage. The block diagram shown in Figure (2) illustrates the feature extraction stage and explicated as the following:

A. Mel-Frequancy Capstral Coeffients (MFCC)

MFCC is a feature extraction model based on human auditory perception system. It is one of the most popular speech feature extraction methods [9]. The MFCC system diagram is explained as the following:

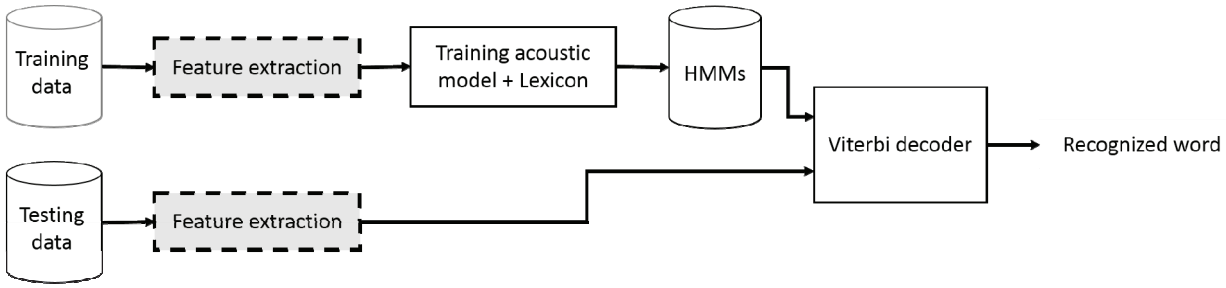


Figure 1: General block diagram of an Automatic Speech Recognition (ASR) system

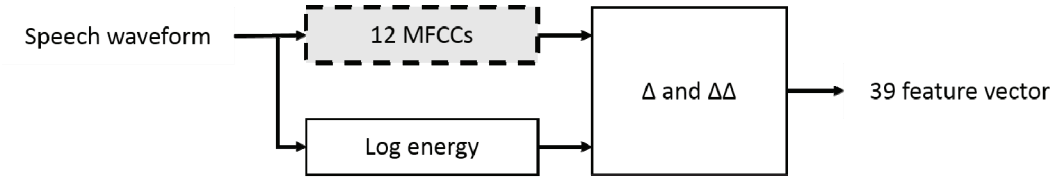


Figure 2: Feature extraction stage

1) Pre-emphasizing

Most of speech energy is concentrated in low frequencies more than middle and high frequency [10]. Pre-emphasizing is a first-order high-pass filter. It is applied to boost the spectrum values in high frequency.

$$y[n] = s[n] - \alpha s[n-1] \quad (1)$$

where $s[n]$ is the input signal, $s[n-1]$ previous sample and α is the filter coefficient in the range of $0.9 \le \alpha \le 1$.

2) Windowing

The speech waveform quasi stationary, therefore it is processed into short overlapped frames. Each frame size is usually between 10 to 25 ms and frame shift between 5 to 10 ms, then each frame is multiplied by Hamming window.

3) Power Spectral Desnsity (PSD)

The Power Spectral Density (PSD) $|Y(k)|^2$ of each frame is then calculated, where it is referred to as the square of Fast Fourier Transform (FFT).

$$|Y(k)|^2 = \left| \sum_{n=1}^{N} y(n) \exp\left(-j \frac{2\pi nk}{N} \right) \right|^2 \quad (2)$$

where *k* is a frequency index, *N* is the total samples and *n* is a time index.

4) Mel-filter banks and frequency warping

The human auditory system can be modeled by a set of overlapped warped triangular shaped filters. Since the relation between Mel scale and frequency scale is nonlinear, the filter banks are warped according to Mel scale [11] as in the relation in equation (3).

$$f_{mel} = 2595 \log_{10}(1 + \frac{f}{700}) \quad (3)$$

where f_{mel} is the Mel-frequency scale and f is the frequency in (Hz).

Then, the power spectral values that is calculated from equation (2) is passed over the triangular filter banks and the output of each filter $E(l)$ is calculated by the summation of the filters banks as shown in equation (4).

$$E(l) = \sum_{k=1}^{M/2} |Y(k)|^2 \Psi_l(k) \quad (4)$$

where l is the filter index, $\Psi_l(k)$ is the function of each filter in frequency domain and *M* is spectrum resolution.

5) Log Discrete Cosine Transform (DCT)

In equation (5), the MFCC feature vector c_i is generated by applying the log operator on equation (4). Then the Discrete Cosine Transform (DCT) is applied.

$$c_i = \sqrt{\frac{2}{L}} \sum_{m=1}^{L} \log(E(l)) \cos\left(\frac{\pi i}{L} (m - 0.5) \right) \quad (5)$$

where *L* is total number MFCC features and *i* feature vector index.

B. Log Energy

The energy feature is calculated by the summation of the power of the samples over the frame from sample at time t_1 to sample at time t_2. Then the energy in log scale is calculated, as shown in equation (6):

$$Energy = \log\left(\sum_{n=t1}^{t2} s^2(n) \right) \quad (6)$$

C. Delta (Δ) and Delta-Delta (ΔΔ) features

Delta and Delta-Delta are applied to the MFCC and log energy features over the frames in order to calculate the

dynamic features [12]. The delta features (Δ) are the velocity variation and the delta-delta features (ΔΔ) are the acceleration variation of the features over the frame.

The final stage is the classification stage, in which the Hidden Markov Model (HMM) is usually used due to the temporal characteristic of speech waveforms [13]. HMM creates a chain of transition probability model for each trained word. The changing of speech signals over time is defined by a stochastic process. In speech recognition systems, words are considered as hidden parameters, while acoustic signals are considered as observed data. The classification stage is divided into two sub stages a training and testing stages a shown in Figure (1).

In training stage the acoustic model is calculated from the extracted features of the training dataset and the number of chains is calculated from the lexicon dictionary. The Hidden Markov Model is constructed for each word then the models are reserved in a database.

In testing stage, feature extraction is applied on the testing dataset. The Viterbi decoder calculates the minimum distance between the extracted features and the reserved models. The system performance is evaluated by scoring the correctly recognized word to calculate the Word Recognition Rate (WRR).

III. PROPOSED METHOD

The new proposed method was developed by a desire to improve the extracted speech features in the presence of Additive White Gaussian Noise (AWGN). The proposed method is designed to robust the noise without loss of the performance in case of undistorted speech wave. Figure (3) illustrates the block diagram of the proposed method.

As discussed in Section II, most of speech energy is concentrated in low frequency. Therefore the pre-emphasizing was removed in the proposed method to prevent boosting the noise amplitude at high frequency value. The uttered word is framed and multiplied by a hamming window. Then, the PSD of each frame is calculated and a time-frequency map for each word is constructed.

The adaptive mask is estimated for each word from the constructed time-frequency map. The block diagram of the proposed adaptive time-frequency masking is shown in Figure (4) and it is explicated as the following:

A. SNR estimation

The SNR estimation technique is applied to measure the noise power level within the speech signal [14]. The Global SNR (*GSNR*) for speech signal is defined as:

$$GSNR = \frac{\sigma_s^2}{\sigma_n^2} \tag{7}$$

where σ_s^2 is the power of signal and σ_n^2 is the power of noise. The standard SNR definition can be evaluated from GNSR, where the speech activity is detected from the speech signal. Therefore, equation (7) can be rewritten as:

$$SNR = 10\log\frac{\sum_{n=0}^{l-1} s^2[n]\cdot vad[n]}{\sum_{n=0}^{l-1} n^2[n].vad[n]} \tag{8}$$

where $s[n]$ is the speech samples, $n[n]$ is the noise samples, and $vad[n]$ is the detected voice activity within the speech waveform. The detection of silence and voice activity is a sensitive stage for correct noise level estimation. There are many approaches to find the voice activity information. One of these approaches is Energy-based algorithm, which is used, in the proposed method:

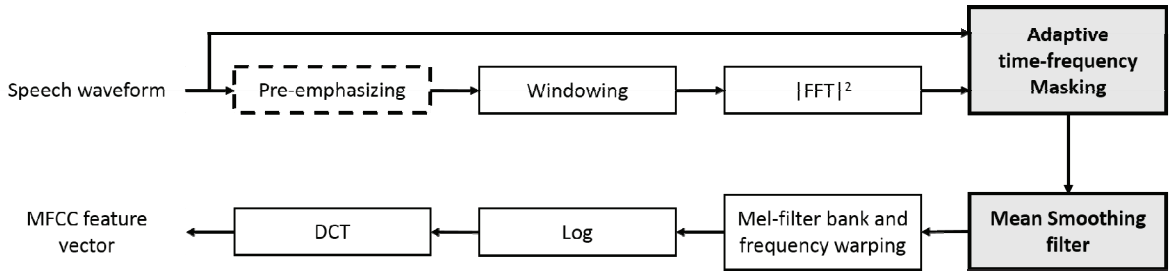


Figure 3: The block diagram of the proposed method MFCC

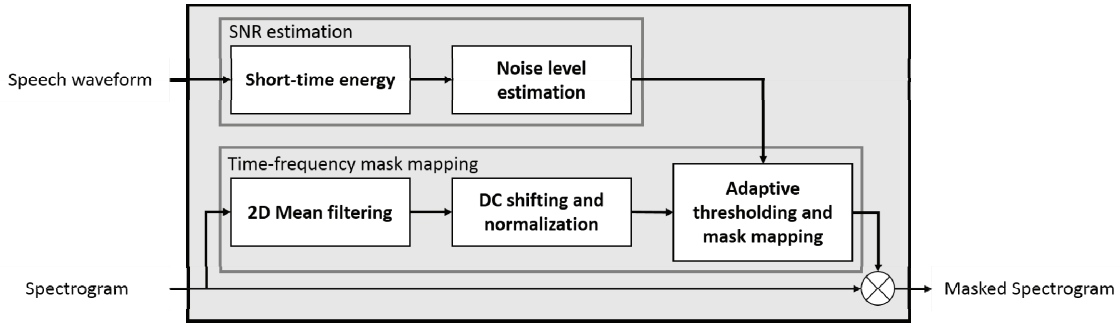


Figure 4: The Adaptive time-frequency masking block diagram

1) Short-time energy

Due to the quasi-stationary property of the speech signal, it is analyzed into short frames. The short time energy STE_i of the noisy speech frames can be defined as:

$$STE_i = \sum_{n=0}^{M-1} x_i[n] = \sum_{n=0}^{M-1} s_i[n] + n_i[n] \quad (9)$$

where n is the sample index, i is the frame index, M is the total number of the samples within the frame, $x_i[n]$ is noisy speech samples, $s_i[n]$ is speech samples and $n_i[n]$ is noise samples.

2) Noise level estimation

In the silence frames, the short time energy is considered as a noise; while the higher energy frames are considered is speech signals with additive noise in the voice activity frames. Consequently, the power of the speech signal to the power of the noise can be rewritten in terms of short time energies as follows:

$$SNR = 20\log \frac{\sum_{i=1}^{k}\sum_{n=0}^{M-1} s_i[n]}{\sum_{i=1}^{k}\sum_{n=0}^{M-1} n_i[n]} \quad (10)$$

$$SNR = 20\log \frac{\sum_{i=1}^{k}\sum_{n=0}^{M-1} x_i[n] - n_i[n]}{\sum_{i=1}^{k}\sum_{n=0}^{M-1} n_i[n]} \quad (11)$$

where k is the total number of frames. The voice short energy varies over the time, while the additive noise energy is almost constant. Estimating the noise frames for the uttered word can be evaluated from the low short energy frames, while the higher short energy frames are considered as the voice activity frame with AWGN; as a result, the Estimated SNR (ESNR) is calculated by using the following relation:

$$ESNR = 20\log \frac{\sum_{i=1}^{k} STE_i - k \cdot \min(STE_i)}{k \cdot \min(STE_i)} \quad (12)$$

The ESNR value is used later in adaptive thersholding and mask-mapping stage.

B. Time-frequency mask mapping

In the speech waveform, high frequencies are more sensitive to the additive noise since the most speech energies are concentrated in low frequencies. Therefore, the adaptive mask is constructed to concentrate the feature values in low frequencies rather than in high frequencies. The following algorithm is applied to extract the less corrupted information from the noisy speech waveform.

1) 2D Average filtering

The aim of this stage is to construct a smooth shape of the spectrogram. It is implemented by convolving 11×11 uniform kernel functions with the time-frequency map. The normalized kernel function equations are given as:

$$I'(m,n) = \sum_{x=-5}^{5}\sum_{y=-5}^{5} 1 \times I(m+x, n+y) \quad (13)$$

$$I'_{norm.}(m,n) = \frac{I'(m,n)}{\sum_{x=-5}^{5}\sum_{y=-5}^{5} 1} \quad (14)$$

where x and y are the filter index and m and n are the spectrogram value indexes.

2) DC shifting and normalization

The scale of the constructed smoothed shape varies for each word. Therefore, it is DC shifted and zero floored by subtracting it from the smallest value, and then dividing it by the maximum number normalizes the total shape.

3) Adaptive thersholding and mask mapping

In this stage, the estimated SNR from equation (12) is used to create an adaptive thresholding. The relation between the threshold and the estimated SNR is demonstrated experimentally by calculating the threshold values that produces the highest recognition rate at different estimated SNRs, then curve fitting is the applied on the obtained points as shown in Figure (5). Equation (15) is the derived formula from curve fitting process where a and b are constants with values 0.047 and 0.8, respectively. The threshold value is calculated to adapt noise level divergence. It is inversely proportional to the Estimated SNR value.

$$Thershold = a.b^{ESNR} \quad (15)$$

The threshold value is applied on the DC shifted and normalized smoothed shape. If the smoothed shape values are greater than the threshold value, the less corrupted speech information is detected and the value of the mask map is equal to 1. Meanwhile, it is equal to 0.1 if the

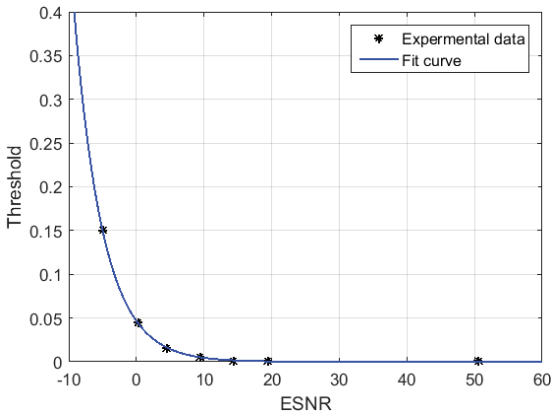


Figure 5: The threshold values at different Estimated Signal to Noise Ratios (ESNR)

smoothed shape values are less than the threshold value, which means highly corrupted speech information with noise.

The time-frequency map is multiplied by constructed mask map, which results the low distorted speech information is multiplied by 1 and highly distorted speech information is multiplied by 0.1. Thus, the speech features are weighted in the less distorted values.

C. *Mean smoothing filter*

Since the power associated with AWGN varies differently from that associated with speech signal, the mean smoothing filter is applied on the adaptive masked time-frequency map to remove the instant variation on the power along the frames. The filter is implemented by calculating the average of three frequency frames. The filter kernel functions are

$$H(z) = \frac{1}{3}(1 + z^{-1} + z^{-2}) \qquad (16)$$

After mean smoothing filtering the log DCT is applied to obtain MFCC features. The proposed method stages are illustrated graphically on the noisy word as shown in Figure (6).

IV. Performance Evaluation

A. *Experiment*

The system is implemented on excerpts of TIDIGITS Database [15]. The complete database consists of 326 speakers (111 men, 114 women, 50 boys and 51 girls), each of them pronounces 22 isolated digits and 55 connected digits and are partitioned into 2 subsets. The first set is a training set and it consists of 55 men, 57 women, 25 boys and 26 girls. The second set is a testing set and it consists of 56 men, 57 women, 25 boys and 25 girls. The words are recorded in 20 kHz.

In the excerpts of database the training set is 37 men and 57 women and the testing set is 56 men and 57 women. The isolated digits for these sets are utilized in this paper to test the performance of the proposed method. The records are down sampled to 8 kHz.

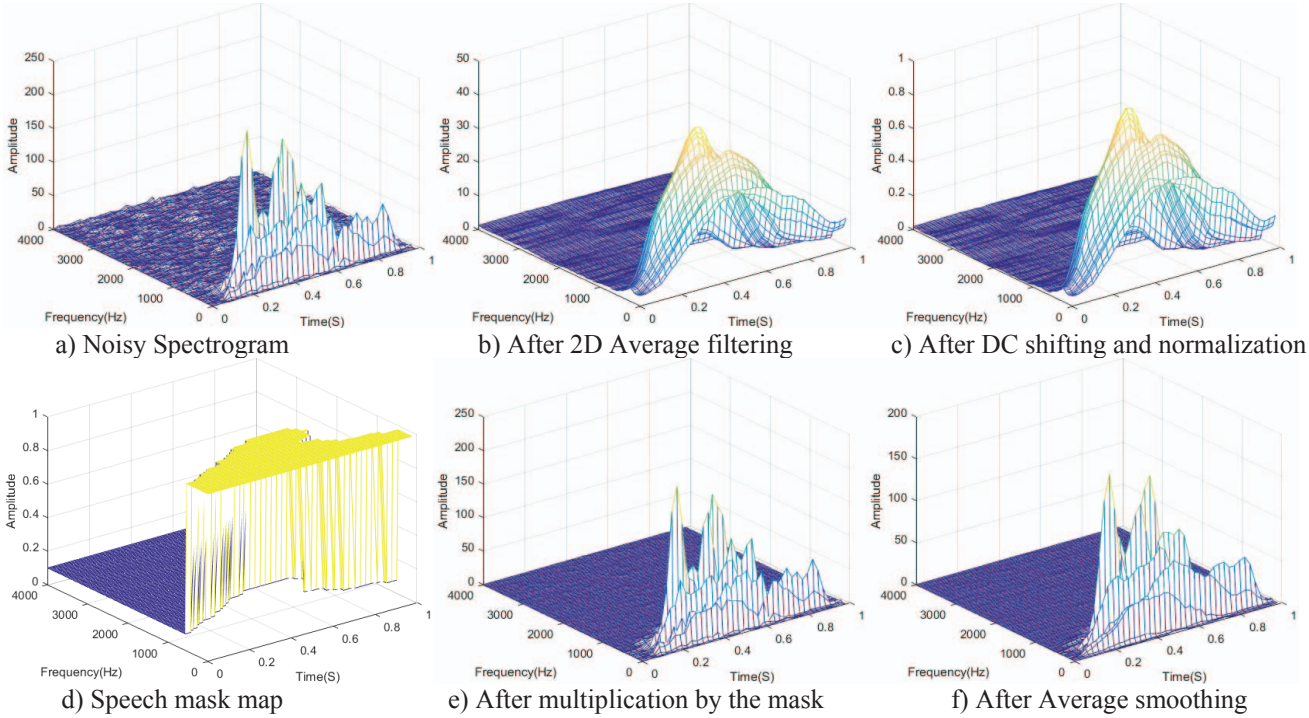
In speech recognition system configuration, each word was framed by 25 ms overlapped Hamming windows, with 10 ms shifting between frames. The FFT resolution was 256 bit and the power spectrum of each frame was calculated and multiplied by 26 overlapped Mel-filter banks and 12 MFCC coefficients were calculated. After calculating log energy, Δ and ΔΔ features were calculated; the total number of extracted features was 39.

The HMM Sphinx CMU lexical dictionary [16] was used. The number of Markov chains was 3 per phoneme. The number of HMM iterations was 10 and WRR was calculated for each iteration. In order to get the total error rate, the average of the last 5 WRR was calculated.

Models were trained with noise-free utterances, while tested with noise-free and noisy utterances. The white noise was generated in a fixed sequence for different SNRs. The experiment was implemented using standard MFCC, RASTA-PLP [17] and the proposed method MFCC without pre-emphasizing. The experiment was executed using desktop with 3.2 GHz Intel Core i5 processor and MATLAB 2015b.

B. *Results*

In this section, the experimental results are presented. The approached MFCC method is evaluated by comparing its recognition accuracy with MFCC and RASTA-PLP system. The performance of MFCC, RASTA-PLP and proposed method is presented in clean data and in different SNRs from -5 dB to 20 dB with step size 5dB. The recognition accuracy is



a) Noisy Spectrogram b) After 2D Average filtering c) After DC shifting and normalization

d) Speech mask map e) After multiplication by the mask f) After Average smoothing

Figure 6: Spectrogram of the uttered word 'one' at SNR = 5 dB

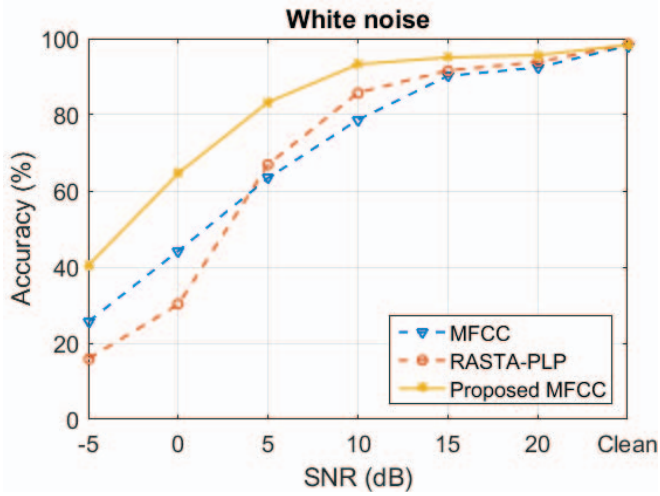


Figure 7: Word Recognition Rate (WRR) at different Signal to Noise Ratios (SNR)

presented graphically as in Figure (7) and numerically as in Table 1.

TABLE1: PERCENTAGE WORD RECOGNITION RATE (WRR)

SNR	White noise		
	MFCC	RASTA-PLP	Proposed MFCC
Clean	98.18	98.48	98.36
20 dB	92.43	93.89	95.67
15 dB	90.23	91.65	95.09
10 dB	78.71	85.83	93.28
5 dB	63.44	66.98	83.36
0 dB	44.27	30.19	64.64
-5 dB	25.70	15.97	40.92
Average	70.42	69.00	81.56

In Table 1, the proposed method outperform other methods in terms of recognition rate at all SNRs. For undistorted data the recognition performance is almost constant. For low distorted data the recognition accuracy is improved by 3.24% and 1.78% in case of 20 dB, 4.87% and 3.44% in the case of 15 dB and 14.57% and 7.45% in the case of 10 dB compared to MFCC and RASTA-PLP, respectively. For the high-distorted data the recognition accuracy is improved by 19.92% and 16.37% in case of 5 dB, 20.37% and 34.45% in the case of 0 dB and 14.79% and 24.52% in the case of -5 dB compared to MFCC and RASTA-PLP, respectively. The average relative improvement is 11% in the case of MFCC method and is 12.56% in the case of RASTA-PLP method. The obtained processing time for the same uttered word "one" is 0.012s, 0.053s and 0.035s for MFCC, RASTA-PLP and proposed MFCC, respectively.

V. CONCULSION AND FUTURE EXTENSIONS

In this paper a proposed method has been developed in order to improve the standard MFCC performance in a noisy environment. It utilized adaptive time-frequency map to estimate which part of the uttered word is highly affected by noise and weighting it. The effect of time-frequency adaptive noise rejection has dramatically improved the recognition at low SNRs. The performance of the proposed system was examined by using TIDIGITS database. The experimental results shows that the proposed method outperforms other methods in terms of average recognition accuracy. It shows improvement of 81.56% compared to 70.42% in the case of standard MFCC system and 69% in the case of RASTA-PLP system. In the future, the performance of the proposed method will be examined with various types of noise and with different database sets.

VI. REFERENCES

[1] L. R. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition. Englewood Cliffs, NJ," Prentice-Hall, 1993.

[2] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Proc. of TASSP*, vol. 28, no. 4, pp. 357–366, 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," in *Proc. of JASA*, vol. 87, no. 4, pp. 1738–1752, 1990.

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, 1994.

[5] O. Viikki, D. Bye, and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," in *Proc. of ICASSP*, pp. 733–736.,1998.

[6] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," in *Proc. of Speech Commun.*, vol. 50, no. 2, pp. 142–152, 2008.

[7] A. Fazel and S. Chakrabartty, "Sparse auditory reproducing kernel (SPARK) features for noise-robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1362–1371, 2012.

[8] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," in *Proc. of TASLP*, vol. 24, no. 7, pp 1315–1329, 2016.

[9] J. Li, D. Li,Y. Gong, Yifan and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition*"* in *Proc. of TASLP*, vol. 22, no. 4, pp 745–777, 2014.

[10] R. Vergin, and O'Shaughnessy, D., "Pre-emphasis and speech recognition" in *Proc. of CCECE*, pp. 1062–1065. , 1995.

[11] S. Molau, M. Pitz, R. Schluter, H. Ney, "Computing Mel-frequency cepstral coefficients on the power spectrum," in *Proc. of ICASSP*, pp. 73–76. , 2001.

[12] B.A. Hanson and T.H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech," in *Proc. of ICASSP*, pp. 857-860, 1990.

[13] V. Diakoloukas and V. Digalakis, "Maximum likelihood stochastic transfor-mation adaptation of hidden Markov models," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 2, pp. 177–187, 2002.

[14] M. Vondrášek and P. Pollak, "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency," in *Proc. of Radioengineering*, vol. 14, no. 1, pp. 6–11, 2005.

[15] R.G. Leonard, "A database for speaker independent digit recognition", in *Proc. of ICASSP*, vol. 3, p.42.11, 1984.

[16] Carnegie Mellon University (2014). CMU dictionary (retrieved Oct. 17, 2016) [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[17] D. Ellis. (2006). PLP and RASTA (and MFCC, and inversion) in MATLAB Using melfcc.m and invmelfcc.m (retrieved Oct. 17, 2016). [Online]. Available: http://labrosa.ee.columbia.edu/matlab/rastamat/