Speech Enhancement Techniques for Noisy Speech in Real World Environments

Htwe Pa Pa Win, Phyo Thu Thu Khine
University of Computer Studies, Yangon
hppwucsy@gmail.com, phyothuthukhine@gmail.com

Abstract

Communication between computer and human has become increasingly popular in today world. Investigation of human emotion importance is also growing in several domains. But under real world condition, speech signal is often, corrupted with several noise types and the accuracy of recognition is degraded from these noisy signal. Therefore this paper focuses on the speech enhancement techniques to develop emotion recognition system for the noisy signal in the real world environment. The various popular enhancement techniques are analyzed by adding the background noise to the clean signal using various SNR. To test the accuracy of the system, the widely used MFCC signal features are against with the SVM classifier. Results after enhancing were compared to that noisy signal and that clean signal to measure the system performance. The experimental results show the best performance algorithm and all enhancement algorithms improve the emotion recognition system performance under various SNRs level of real world background noise.

***Keywords**: Emotion Recognition, Noisy Signal, MFCC, SVM, SNRs*

1. Introduction

Speech is one of the most fundamental means of communication between human to human and human to machine in various fields via automatic speech recognition and speaker identification. The present day speech communication systems are severely degraded due to various types of noises which make the listening task difficult for a direct listener and cause inaccurate transfer of information. Therefore, the noise suppression is one of the main motives of various research endeavors in the field of speech processing over the last few decades. The researchers attempted to suppress the noise level of degraded speech without distorting the speech signal and also tried to make a speech more pleasant and understandable to the listener. The main purpose of speech enhancement research is to minimize the degree of distortion of the desired speech signal and to improve one or more perceptual aspects of speech, such as the speech quality and/or intelligibility. These two measures are uncorrelated and independent of each other. A speech signal may be of high quality and low intelligibility and vice versa [1, 2].

Emotional speech recognition importance is also growing in several domains. Researches have raised the impact of emotion in multidisciplinary applications. Predicting human emotions is catching the attention of many research areas, which demand accurate predictions in uncontrolled scenarios. Psychologists have widely studied the influence of emotional factors, on decision-making. As example, pilots' decision in a flight context may jeopardize several humans' life [3]. The needs for the development of automatic recognizing human emotion in real world environment remain an open research problem.

Speech enhancement algorithms can affect performance of speech recognition and classification systems. The enhancement procedures can significantly modify temporal and spectral characteristics of speech and change affect both linguistic and paralinguistic (emotional) information. These can directly affect the accuracy of speech recognition [4, 5]. Automatic emotion recognition (AER) from speech signals performed under real-life conditions is likely to deal with noisy speech signals. The effects of speech compression on the accuracy of AER from speech signals have not been extensively investigated.

Therefore, this paper presents a system that aims to recognize the emotion from noisy signal like in real world environment. The paper is systemized as follow. Section 2 reviews the previous works related for speech enhancement techniques. Section 3 describes about the Speech Enhancement System and those techniques are defined in section 4. Section 5 and 6 illustrate about the feature extraction and classification method. Experimental results are

discussed in Section 7 and section 8 make conclusion about the enhancement system.

2. Literature Reviews

There are many previous works done for speech enhancement. The researchers in [6-8] made reviews on types of noise and techniques to remove those noise form speech signal and compare them. The extensive work for Spectral-domain is done in [9]. They established a state-of-the-art speech recognition platform for speech enhancement evaluation, and investigated typical spectral-domain enhancement algorithms for different speech recognition decoders under various noise conditions. In paper [10] and [11] proposed the enhancement techniques by modifying the Winner filter and Kalman filter.

3. Speech Enhancement System

Speech enhancement is a step in the digital speech signal processing having an objective of increasing the quality of speech signal i.e. to enhance the clarity, intelligibility, understand ability and comprehensibility of speech signal with the help of some algorithm/filter. There are various reasons which leads to degradation of speech signal due to background noise which are captured during the recording like reverberation, babble etc. For specific type of speech enabled applications clean and noise free speech signals are required. The speech enhancement can be achieved by various methods. According to the type of degradation and the noise in the acquired speech signal the approach to speech enhancement varies. The figure 1 shows the Basic steps of speech enhancement system [6].

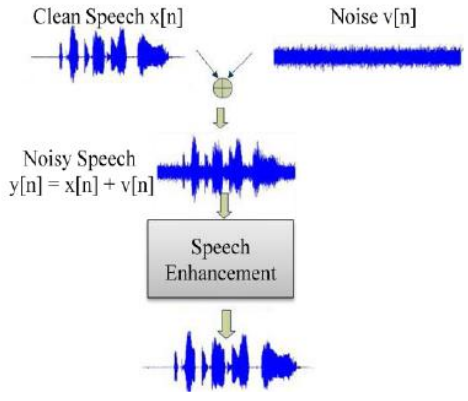


Figure 1. Basic step of Speech Enhancement System

Speech enhancement is a framework that acoustically enhances the desired speech in the captured signals by suppressing the interference, and it has been extensively studied to overcome the above problems. Many speech enhancement techniques have been proposed for noise reduction (denoising), reverberation suppression (dereverberation), and source separation [12].

4. Speech Enhancement Methods

There are different types of speech enhancement techniques which are as follows [8].

4.1. Spectral Subtraction Method

The Spectral subtraction method is most widely used method because of the simplicity of implementation and lower computational load. This approach has some assumption: (i) noise is additive, (ii) signal is absent. It is based on simple principle. Assuming additive noise, one can obtain an estimate of the clean signal spectrum by subtracting an estimate of the noise spectrum from the noisy speech spectrum. Note that the noise spectrum is estimated, and updated, only when signal is absent or when only noise is present.

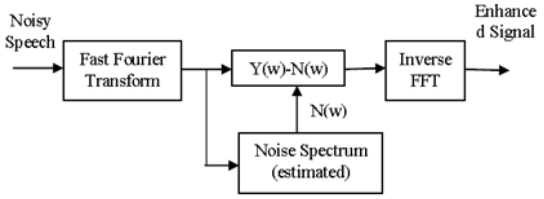


Figure 2. Spectral Subtraction

4.2. Weiner Filtering

The Wiener filter is same as the spectral subtraction in the way that it is derived and makes an attempt to reduce the mean-square error in the frequency domain. It is generally employed in the estimation or prediction of a signal observed in noise. The Wiener filter can also be adaptively estimated used where the surrounding noise has time-varying characteristics. The Filter is used to enhance the quality of speech by removing unwanted noise. The gain function of WF is given by

$$H_{Wiener}(w) = \frac{Ps(w)}{Ps(w) + Pn(w)}$$

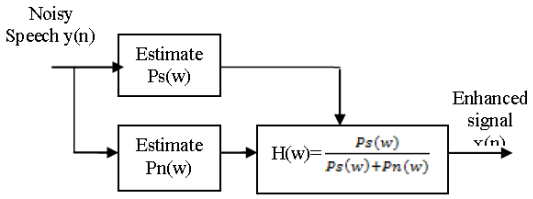


Figure 3. Weiner Filter

4.3. Minimum Mean Square Error

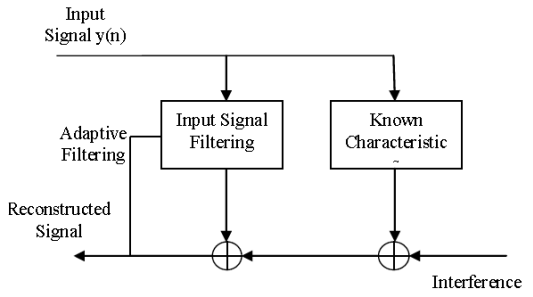


Figure 4. MMSE Filter

MMSE estimation is also known as Ephraim and Malah's estimator used to overcome the problem of the background noise. MMSE method is proposed to minimize the background noise to a considerable amount and thus improved the quality of the resulting enhanced speech. The Minimum mean square error technique is implemented when the input SNR is known. It is an implementation of Weiner Filter [4]. MMSE based algorithms are mainly Minimum Mean Square Error Short-Time Spectral Amplitude (MMSE-STSA) estimator and MMSE Logarithm Spectral Amplitude (MMSE-LSA) estimator. Some power spectrum estimators are used in decision-directed approach for the calculation of *a priori* SNR. The only disadvantage of the MMSE processor is additional complexity in determining the linear estimator.

4.4. Kalman Filtering

The Kalman filter, also called linear Quadratic Estimation (LQE), is a method that uses an arrangement of estimations saw about whether, holding noise (arbitrary varieties) and different mistakes, and produces appraisals of obscure variables that have a tendency to be more exact than those focused around a solitary estimation alone. All the more formally, the Kalman filter works recursively on streams of uproarious data information to generate a measurably ideal appraisal of the underlying system state. The filter is named for Rudolf (Rudy) E. Kálmán, one of the essential designers of its hypothesis.

The method works in a two-stage process. In the prediction step, the Kalman filter produces assessments of the current state variables, alongside their instabilities. Once the result of the following estimation (essentially defiled with some measure of slip, including irregular noise) is watched, these appraisals are overhauled utilizing a weighted normal, with more weight being given to gauges with higher conviction. On account of the calculation's recursive nature, it can run progressively utilizing just the present info estimations and the at one time ascertained state and its instability system; no extra past data is needed [9].

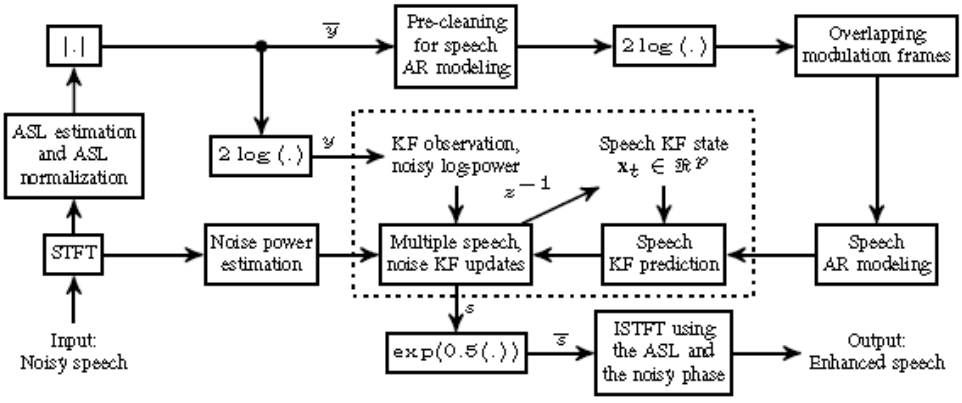


Figure 5: The flowchart diagram of the algorithm. The term z^{-1} refers to one-frame delay. The blocks in the dotted rectangle constitute the KF [13].

5. Feature Extraction

MFCC is the method that is applied for speech parameterization. Human ear has been proven to resolve frequencies non-linearly across the audio spectrum, thus filter bank analysis is more desirable than Linear Predictive Coding (LPC) analysis since it is spectrally based method. Mel-Scale is a frequency binning method based on the human ear's frequency resolution. With the use of frequency bins on the melscale, MFCC is computed and used to parameterize speech data. The mel-scale also attempts to mimic the human ear in terms of the manner with which frequencies are sensed and resolved. The mel-scale is a unit of measurement of perceived frequency (pitch) of a tone.

The MFCC extraction method can be achieved by two ways, either FFT based (Fast Fourier Transform) or LPC based (Linear Predictive Coding). In this paper, MFCC extraction based FFT is used. Generally , MFCC extraction involves several stages, which are pre-emphasis, framing/segmentation windowing, FFT spectrum, mel-spectrum extraction and mel-cepstrum extraction. The general procedure of mel-cepstrum extraction actually involve, dividing the signal into frames, to obtain the power of spectrum, to convert the melspectrum and lastly uses the Discrete Cosines Transform (DCT) to get the cepstrum coefficient [14].

6. Classification

In recent years in speech emotion recognition, researchers proposed many classification algorithms, such as Neural Networks (NN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Maximum Likelihood Bayesian classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN) and Support vector machines (SVM). Support vector machine maps the vector to a higher dimensional vector space, in this space a maximum interval hyper plane to be established. Two parallel hyper planes are established on both sides of the hyper plane. Establish a suitable direction hyper plane to make the distance between the two hyper planes maximization. Its main idea is to use a kernel function to map the original input set to a high dimensional space and then obtain an optimal classification. Since SVM is a simple and efficient computation of machine learning algorithms, and is widely used for pattern recognition and classification problems, and under the conditions of limited training data, it can have a very good classification performance compared to other classifiers [15]. Thus this system adopted the support vector machine to classify the speech emotion.

7. Experimental Results

7.1. Dataset

In this work, we use IEMOCAP database (Interactive Emotional Dyadic Motion Capture), it is an Interactive emotional corpus collected at SAIL lab at USC [16]. This database is composed of recordings in audio, video and motion-capture. Five dyadic sessions of mixed gender pairs lasts approximately a total of twelve hours. Segmentation into utterances was performed manually and annotations was achieved by human annotators in categorical labels :{angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, other} and dimensionally over the axes of :{valence, activation, dominance}. Utterances are recorded in scripts form or in improvisation of hypothetical scenarios. This database contains ten speakers (5 male and 5 female) with five sessions. The audio recordings were sampled at 16 KHz.

This system uses speech information that is available for each utterance specially sentences which are classified in seven emotional states: angry, happy, sad, neural, frustrated, excited, and surprised.

7.2. Experimental setup

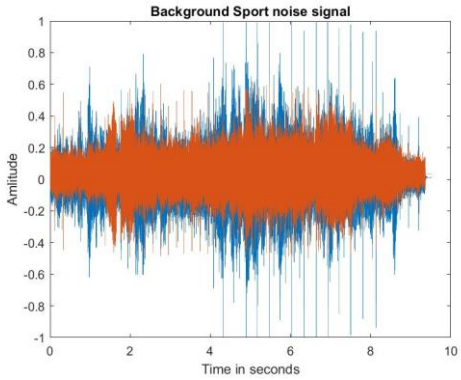
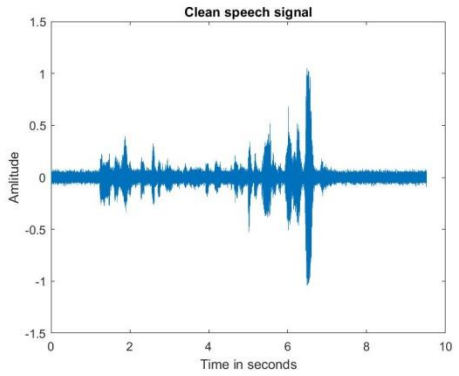
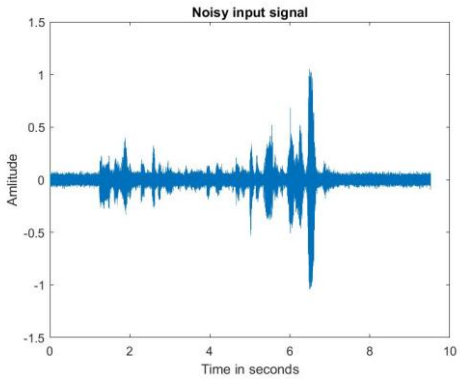


Figure 6: The input sport event background noise signal

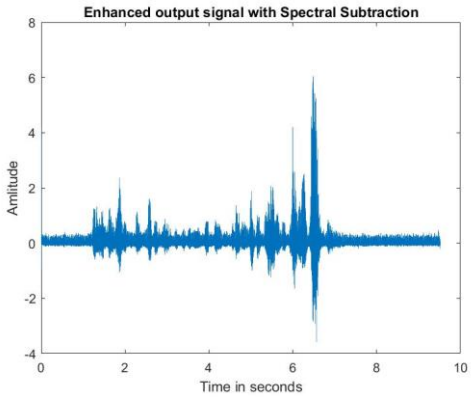
This setup is interested to evaluate the performance of speech emotion recognition system under real world background noise environment. To test the robustness of the system, background noise (sport event), as shown in figure is added to the speech signal respectively at several signal-to-noise ratio (SNR) levels (0dB, 5dB, 10dB and 15dB).



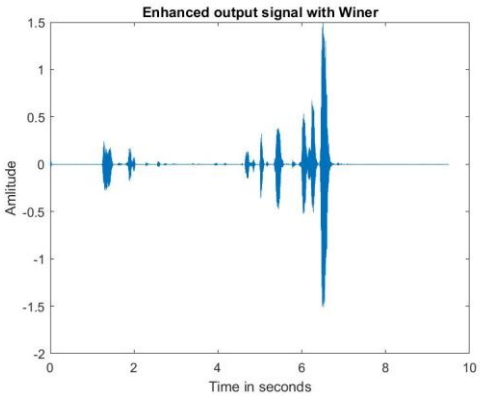
(a) Clean Signal



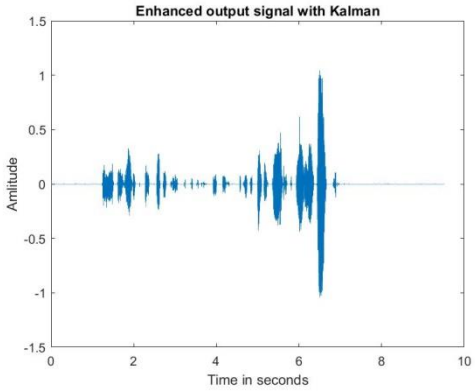
(b) Noisy Signal with SNR 10db



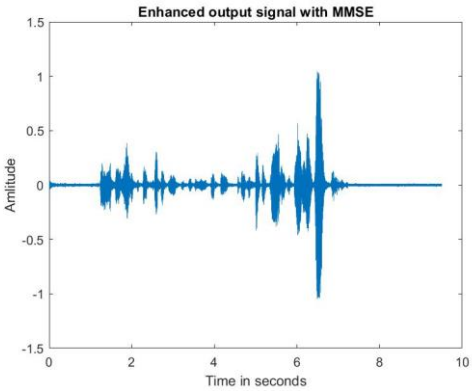
(c) Enhance signal with Spectral Subtraction



(d) Enhance signal with Winer Filter



(e) Enhance signal with Kalman Filter



(f) Enhance signal with MMSE method

Figure 7: Sample Input signal and output signal for angry emotion file

In this analysis, all the available collected utterances which are classified in seven emotional states resulting in a set of 322 sentences. Front-end method based on Mel frequency Cepstral coefficients (MFCC) is performed to extract characteristics of speech signal. For each utterance, the signal was divided into frames of 50ms with 50% overlap between successive frames. Feature vector is composed by 13 coefficients and 13 delta coefficients to get more effective values. First, feature extraction method is performed in clean environment. Then, we explore the robustness of the developed system and compare the performance of speech enhancement algorithms in noisy environment. The emotion classifier is implemented with Support Vector Machine (SVM).

7.3. Classification Results

The following Tables report the recognition performance after adding background noise to the original speech signal using four different SNR (0db, 5db, 10db and 15db) using MFCC as feature method and spectral substraction, wiener, MMSE and Kalman Filter as denoising methods.

Table 1: Recognition Rate in background noise for SNR 0db

Signal Type	Precision	Recall	FMeasure	Accuracy (%)
Noisy Signal	0.101	0.304	0.152	30.4348
Spectral Subtration	0.413	0.307	0.149	30.4348
Winer Filter	0.103	0.320	0.155	31.9876
MMSE	0.242	0.373	0.272	37.2671
Kalman Filter	0.101	0.183	0.115	28.323

Table 2: Recognition Rate in background noise for SNR 5db

Signal Type	Precision	Recall	FMeasure	Accuracy (%)
Noisy	0.101	0.304	0.152	30.4348
Spectral Subtration	0.413	0.307	0.149	30.7453
Winer	0.129	0.323	0.166	32.2981
MMSE	0.263	0.404	0.310	40.3727
Kalman	0.008	0.087	0.014	18.6957

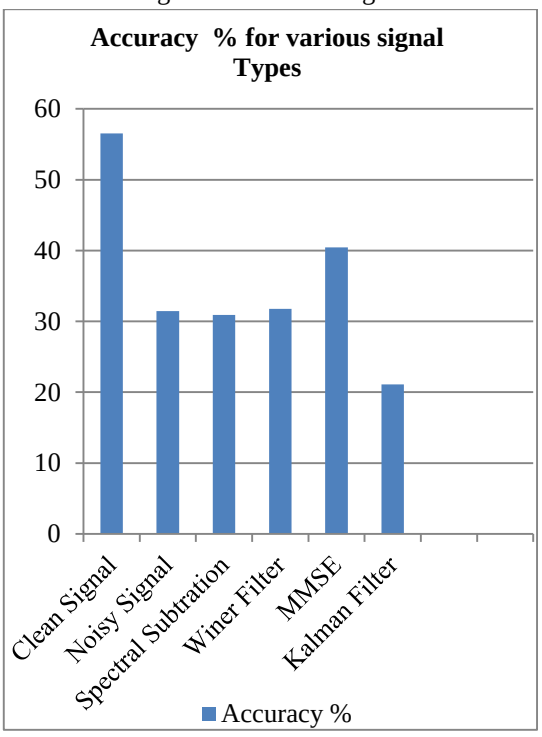
Table 3: Recognition Rate in background noise for SNR 10db

Signal Type	Precision	Recall	FMeasure	Accuracy (%)
Noisy	0.173	0.304	0.148	30.4348
Spectral Subtration	0.413	0.311	0.161	31.0559
Winer	0.140	0.342	0.197	34.1615
MMSE	0.314	0.429	0.328	42.8571
Kalman	0.008	0.087	0.014	18.6957

Table 4: Recognition Rate in background noise for SNR 15db

Signal Type	Precision	Recall	FMeasure	Accuracy (%)
Noisy	0.212	0.345	0.236	34.472
Spectral Subtration	0.272	0.314	0.172	31.3665
Winer	0.325	0.286	0.179	28.5714
MMSE	0.472	0.413	0.317	41.3043
Kalman	0.008	0.087	0.014	18.6957

Results reveal that Simple Kalman Filter is particularly the least efficient when it goes to enhancing emotional speech. It also takes longer time than other algorithms. Therefore, the modification of Kalman Filter is in research trends. MMSE gives the best results than all the methods: it enhances the recognition rate in all SNR level. The table 5 shows also the emotion recognition average rate in noisy environment against with clean signals.



There should be remark that the recognition rate in noisy environment are always lower than those is clean environment. All the methods used in this paper, except Kelman Filter, improve the recognition performance. It is probably due to the fact that the denoising algorithm combined with MFCC is eliminating the noise existing in the original database.

8. Conclusion

In this paper, various types of speech enhancement algorithms are analyzed and tested with IEMOCAP, the standard emotional dataset. Experimental results show that MMSE enhance significantly the recognition rate in sport event in various noise levels. MFCC feature can deal with various types of the signal and SVM can give the robustness classifier concepts for the unseen signals. The speech enhancement techniques are required to get the adaptable emotion recognition system in real world environment.

References

[1] N. Upadhyay and A. Karmakar, "The spectral subtractive-type algorithms for enhancing speech in noisy environments," IEEE International Conference on Recent Advances in Information Technology, ISM Dhanbad, India, March 2012, pp. 841-847.

[2] P. C. Loizou, "Speech Enhancement: Theory and Practice", 2nd edition, CRC Press, 2013.

[3] Causse, M., Dehais, F., Péran, P., Sabatini, U., & Pastor, J., "The effects of emotion on pilot decision-making: A neuroergonomic approach to aviation safety". Transportation research part C: emerging technologies, 33, 272-281, 2013.

[4] He L, "Stress and emotion recognition in natural speech in the work and family environments", Ph.D. thesis, Department of Electrical Engineering, RMIT University, Melbourne, November 2010.

[5] A. Albahri, M. Lech, and E. Cheng, "Effect of speech compression on the automatic recognition of emotions, International Journal of Signal Processing Systems", vol. 4, no. 1, pp. 55–61, 2016.

[6] D. S. Kulkarni , R. R. Deshmukh and P. P. Shrishrimal, "A Review of Speech Signal Enhancement Techniques", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.14, April 2016.

[7] S. Vihari, A. S. Murthy, P. Soni and D. C. Naik, "Comparison of Speech Enhancement Algorithms", Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016).

[8] Ambalika, Er. S. Saini, "A Brief Review on Speech Enhancement Algorithms", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 5, Issue 4, April 2016.

[9] C. H. You and B. MA, "Spectral-domain speech enhancement for speech recognition", Speech Communication 94 (2017) 30–41.

[10] N. Upadhyay and R. K. Jaiswal, "Single Channel Speech Enhancement: using Wiener Filtering with Recursive Noise Estimation", 7th International conference on Intelligent Human Computer Interaction, IHCI 2015.

[11] N. Dionelis and M. Brookes, "Speech Enhancement Using Modulation-Domain Kalman Filtering with Active Speech Level Normalized Log-Spectrum Global Priors", 25th European Signal Processing Conference (EUSIPCO), 2017.

[12] Ogunfunmi, Tokunbo, Togneri, Roberto, Narasimha and M. Sim, "Speech and Audio Processing for Coding, Enhancement and Recognition". ISBN 978-1-4939-1455-5, © Springer Science+Business Media New York 2015

[13] C. Pandey and S. Nemad, "Distinctive Methods for Speech Enhancement using Kalman Filtering", International Journal of Computer Applications (0975 – 8887) Volume 105 – No. 5, November 2014.

[14] C. K. On et.al," Mel-Frequency Cepstral Coefficient Analysis in Speech Recognition", IEEE conference, ICOCI 2006.

[15] P. She, Z. Changjun and X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference on Electronic & Mechanical Engineering and Information Technology, 2011.

[16] C. Busso, M. Bulut, C. Lee, A.Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database", Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.