I.J. Image, Graphics and Signal Processing, **2020, 2, 1-8**
Published Online April 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2020.02.01

Emotion Recognition System of Noisy Speech in Real World Environment

Htwe Pa Pa Win
University of Computer Studies, Hpa-an, Myanmar
Email: hppwucsy@gmail.com

Phyo Thu Thu Khine
University of Computer Studies, Hpa-an, Myanmar
Email: phyothuthukhine@gmail.com

Received: 03 February 2020; Accepted: 15 March 2020; Published: 08 April 2020

Abstract—Speech is one of the most natural and fundamental means of human computer interaction and the state of human emotion is important in various domains. The recognition of human emotion is become essential in real world application, but speed signal is interrupted with various noises from the real world environments and the recognition performance is reduced by these additional signals of noise and emotion. Therefore this paper focuses to develop emotion recognition system for the noisy signal in the real world environment. Minimum Mean Square Error, MMSE is used as the enhancement technique, Mel-frequency Cepstrum Coefficients (MFCC) features are extracted from the speech signals and the state of the arts classifiers used to recognize the emotional state of the signals. To show the robustness of the proposed system, the experimental results are carried out by using the standard speech emotion database, IEMOCAP, under various SNRs level from 0db to 15db of real world background noise. The results are evaluated for seven emotions and the comparisons are prepared and discussed for various classifiers and for various emotions. The results indicate which classifier is the best for which emotion to facilitate in real world environment, especially in noisiest condition like in sport event.

Index Terms—IEMOCAP, MFCC, MMSE, Noisy Signal, SNR, Speech

I. Introduction

The importance of automatic emotional speech recognition is also increasing in several domains of the real world application. Researches have raised the impact of emotion in various types of applications and predicting human emotions, accurate predictions in uncontrolled scenarios, is catching the attention of many research areas. The influence of emotional factors on decision-making and human intelligence are analyzed and clearly stated by many psychologists. The examples can be seen at in pilots' decision in a flight context, at call centers to detect the callers' emotional state in case of emergency, or at the business to identify the level of the customer's satisfaction, at classroom or E-learning to analyses the students' emotion that can provide focus on the enhancement of teaching quality [1-3].

In real world applications, original speech signals are disturbed by real world background noise. These additional signals impact the hearing level and degrade the performance of all the recognition system, especially in emotional system. Therefore for this environment, speech enhancement or removing noise is an essential module. Speech enhancement or de-noising speech is closely related to restoration of the speech because it reconstructs and restores the signal after degradation of the original clean signal [4].

The researchers attempted to overwhelm the noise level of degraded speech without distorting the speech signal and also tried to recognize the emotion of a noisy speech more accurately. However, the performance of the noisy speech emotion recognition is still far from the expectation of researchers because of the enhancement techniques. In noisy speech emotion recognition, there are mainly three difficulties that are how to clean the noisy signal, how to find effective speech emotion features, and how to build a suitable emotion recognition model of speech [5].

The needs for the development of automatic recognizing of human emotion in real world environment that suits for various types of applications remain an open research problem even the existence of large amount of work done. Moreover, analysis of emotions is critical to understand exactly spectators' responses in real world sport events. Therefore, this paper intends to propose the emotion recognition framework in noisy condition with effective enhancement mechanism and effective features. In addition, the system tries to prove the fact that which classifier is most suitable for each data signal and for each emotional state. The experiments are carried out to facilitate the facts for the various types of applications or devices in combination with the recognition of various types of emotions.

The remainder of the paper is organized as follows. In section Ⅱ, a review of the previous system techniques

that are necessary as background required to effectively implement the system is presented. The proposed system is described in Section Ⅲ. After that, experimental results of the proposed system are discussed in section Ⅳ, and finally the conclusion about the proposed system is presented in the last section.

II. Related Works

Although there are many previous works done for speech emotion recognition by using different paradigms, a limited research works have been done in noisy environments.

The related works in [6-9] proposed different mechanism to improve the performance of speech emotion recognition in normal environment. Speech emotion recognition system using CNN with the improvement of CapsNets are proposed in [6] by using IEMOCAP dataset and proved that CapsNets get the better performance than baseline CNNs in building the recognition model. The groups of [7] and [8] also used CNN based classifiers that leads to reliable improvements in accuracy of the speed emotion recognition model and two emotion dataset of IEMOCAP and MSP-IMPROV for unbalanced speed with unsupervised learning and for raw speed. The system used the Bag-of-Visual Words as the classification model on Audio Segment Spectrograms is proposed by the groups [9]. They used SURF features for the SVM recognizer and clearly state that how emotion recognition model is much important in education fields and proved their proposed work improves the accuracy.

The previous works in [10] proposed different types of speech enhancement techniques by using MFCC features and SVM classifier. The authors recommended that for sport noise, Minimum Mean Square Errors, MMSE outperform among the different types of speech enhancement techniques they described.

The authors in [11] proposed the system for Speech emotion recognition in noisy environment. They used three speech enhancement techniques; spectral subtraction, wiener filter and MMSE for noisy signal. MFCC is used for feature extraction and Hidden Markov Model, HMM is used for recognition. They also concluded that although they enhanced the noisy signal, the enhancement techniques are not sufficiently efficient for all types of noise and they remark that recognition rates in both noisy and enhanced environment are independent from various SNR levels.

The researchers in [3] emphasize on automatic speech emotion recognition with the effect of additive noise signals. The white Gaussian noise is added to the original clean signals at several signals to noise ratio (SNR) levels. Gaussian mixture models (GMM) based method is used as the emotion recognizer, the i-vector based method is compared with a baseline conventional method. The results are demonstrated with the superior performance.

The groups of [12] proposed binaural speech system for emotional recognition that is based on the analysis of binaural input signals by using the binaural signal analyzer and on the constructing of speech signal mask in a noisy environment by using emotional mask . They used the emotional speech utterances of Persian database for the experimental purposes. Their simulation results proved that their proposed system achieves higher performance.

III. Proposed Algorithm

This system proposes the processes for the Emotion Recognition system for real world environment as shown in Fig.1. The features of the clean signals in the database are extracted by using MFCC and save at the offline stage of the recognition system. The clean signal is used as the input at the online stage and then sport noise is added to the system. Then the features of the noisy signal are extracted using MFCC and sent to the SVM recognition model. The main processes of the proposed system are as follow;

- Adding noise
- Improving signal quality
- Extracting features
- Recognition model

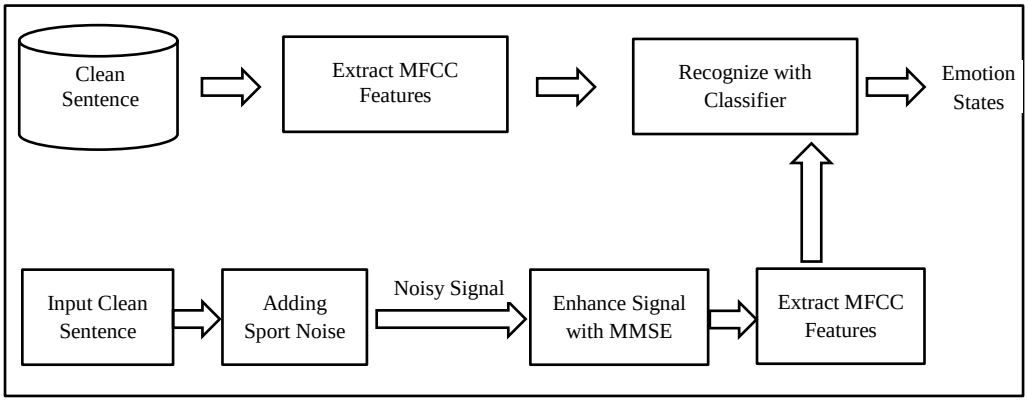

Fig.1. Proposed System Design

A. Adding Noise

This setup is intended to develop a speech emotion recognition system under real world background noise environment. To get the noisy signal, background noise (sport event), as shown in Fig.2 is added to the speech signal respectively at several signal-to-noise ratio (SNR) levels (0dB, 5dB, 10dB and 15dB).

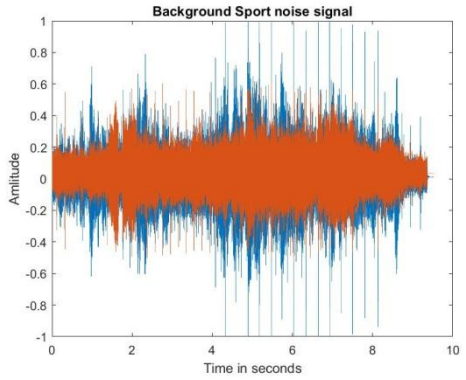

Fig.2. The input sport event background noise signal

B. Improving signal quality

After adding the sport noise to the clean signal, the resulted noisy signal is enhanced by using Minimum Mean Square Error, MMSE.

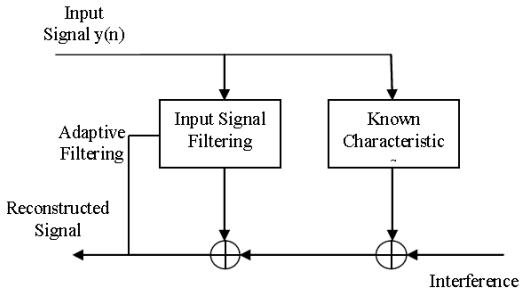

Fig.3. MMSE Filter

MMSE estimation is used to overcome the existence of the background noise. MMSE method has been proposed to minimize the background noise to an acceptable amount and thus improved the quality of the speech. The MMSE technique can be implemented when the input SNR is clearly known [13]. The determination of the linear estimator take additional complexity and it may cause the only disadvantage for MMSE.

The short-time spectral magnitude minimum mean-square error (MMSE) estimation has been largely used in the previous years. The MMSE estimator of the magnitude spectrum of the signal can be calculated as follow:

$$X_k = E\{X_k|Y(w_k)\}, k = 0,1,2,\dots,N-1 \quad (1)$$

$$p(X_k, \theta_k) \quad (2)$$

Where $E\{X_k|Y(w_k)\}$ denotes the expectation operator, $p(X_k, \theta_k)$ is the joint pdf of the magnitude and phase spectra, denotes the noise variance and $p(Y(w_k)|X_k, \theta_k)$ is given by:

$$p(Y(w_k)|X_k, \theta_k) = \frac{1}{\pi\lambda_d(k)}\exp\{-\frac{1}{\lambda_d(k)}|Y(w_k) - X(w_k)|^2\} \quad (3)$$

Where λd(k) denotes the noise variance [11].

C. Extracting Features

The characteristics of the enhanced signal are extracted by using the Mel frequency Cepstral coefficients (MFCC). The Mel-Frequency Cepstral Coefficients (MFCC) features is the most commonly used popular features in speech recognition. It takes the advantages of the cepstrum analysis and a perceptual frequency scale of critical bandwidth. MFCC is based on variation of the human ear's critical bandwidth that cannot perceive frequencies over 1000 Hz. [14]. The steps of the MFCC in speech analysis are as in the following Fig.4. 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 finally the Discrete Cosines Transform (DCT) is used to find the cepstrum coefficient.

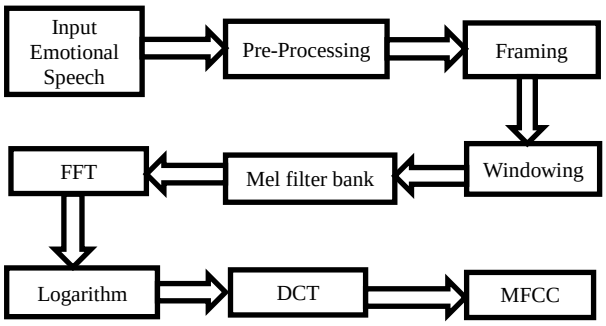

Fig.4. MFCC steps in speech analysis

In this system, all the available collected utterances which are classified in seven emotional states resulting in a set of 322 sentences. The signal was divided into frames of 50ms for each utterance, with 50% overlap between the successive frames. Feature vector is collected by 13 coefficients and 13 delta coefficients to get more effective values.

D. Recogntion model

The researchers proposed many classification algorithms to build the recognition model of the emotion recognition system. Since SVM is a simple and efficient computation of popular machine learning algorithms, and is commonly used for signal classification problems and it can have a very good performance compared to

other classifiers. Thus this system adopted the support vector machine to build the classification model of the speech emotion system.

IV. Experiment simulation and result analysis

A. Dataset

The proposed system is evaluated by using the IEMOCAP database (Interactive Emotional Dyadic Motion Capture), it is an Interactive emotional corpus collected at SAIL lab at USC for research purposes [15]. This database is composed of recordings in audio file, video file and motion-capture file. It takes nearly a total of twelve hours for five dyadic sessions of mixed gender pairs. The manual works are done for segmentation into utterances and annotations in categorical labels of emotions such as angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and other. This database contains ten speakers (5 males and 5 females) with five sessions and the audio recordings files were save with the sampling rate of 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.

B. Accuracy Results

The most popular machine learning methods described above are used to test the performance of the system. The results are carried out by using the matlab 2018a and weka 3.8.1. Firstly the clean signals are tested and the results are shown in the following table. All classification results are obtained by using the clean trained data with the all test data.

Table 1. Emotion recognition results for clean speech signals

Classifier	Accuracy (%)
Deep Learning	54.3478
Neural Network	90.6832
SVM	56.5217
Decision Tree	89.441

As can be seen from the above result, MLP outperform as the best classifier for standard dataset of the clean speech signal. After getting the result of the clean signal, the experiment is carried out for noise signal by adding real world sport noise with various db. The results for each classifier are shown in the following Tables 2 to 5.

Table 2. Emotion recognition results for noisy speech signal using Deep Learning Method

SNR	Precision	Recall	FMeasure	Accuracy (%)
0db	0.139	0.351	0.198	35.0932
5db	0.143	0.270	0.175	27.0186
10db	0.162	0.193	0.150	19.2547
15db	0.162	0.193	0.150	19.2547

Table 3. Emotion recognition results for noisy speech signal using Neural Network

SNR	Precision	Recall	FMeasure	Accuracy (%)
0db	0.001	0.006	0.001	0.6211
5db	0.001	0.009	0.002	0.9317
10db	0.008	0.087	0.014	8.6957
15db	0.008	0.087	0.014	8.6957

Table 4. Emotion recognition results for noisy speech signal using SVM

SNR	Precision	Recall	FMeasure	Accuracy (%)
0db	0.101	0.304	0.152	30.4348
5db	0.101	0.304	0.152	30.4348
10db	0.173	0.304	0.148	30.4348
15db	0.212	0.345	0.236	34.472

Table 5. Emotion recognition results for noisy speech signal using Decision Tree

SNR	Precision	Recall	FMeasure	Accuracy (%)
0db	0.101	0.304	0.152	30.4348
5db	0.101	0.304	0.152	30.4348
10db	0.122	0.307	0.153	30.7453
15db	0.107	0.295	0.155	29.5031

The MMSE outperform among the enhancement methods and this system also used the MMSE to remove real world sport noise. The following tables from 6 to 9 show the result for the enhancement files.

Table 6. Emotion recognition results for enhanced speech signal with MMSE using Deep Learning

SNR	Precision	Recall	FMeasure	Accuracy (%)
0db	0.139	0.143	0.104	14.2857
5db	0.154	0.149	0.113	14.9068
10db	0.234	0.152	0.120	15.2174
15db	0.414	0.186	0.172	18.6335

Table 7. Emotion recognition results for enhanced speech signal with MMSE using Neural Network

SNR	Precision	Recall	FMeasure	Accuracy (%)
0db	0.525	0.096	0.032	9.6273
5db	0.445	0.118	0.077	11.8012
10db	0.359	0.236	0.185	23.6025
15db	0.307	0.314	0.230	31.3665

Table 8. Emotion recognition results for enhanced speech signal with MMSE using SVM

SNR	Precision	Recall	FMeasure	Accuracy (%)
0db	0.242	0.373	0.272	37.2671
5db	0.263	0.404	0.310	40.3727
10db	0.314	0.429	0.328	42.8571
15db	0.472	0.413	0.317	41.3043

Table 9. Emotion recognition results for enhanced speech signal with MMSE using Decision Tree

SNR	Precision	Recall	FMeasure	Accuracy (%)
0db	0.245	0.307	0.190	30.7453
5db	0.252	0.326	0.272	32.6087
10db	0.260	0.339	0.293	33.8509
15db	0.288	0.363	0.305	36.3354

To be clearly analyses the results, the average accuracy are calculated and the results are compared for various data types; clean signal, noisy signal and enhanced signals as shown in the following Fig.5.

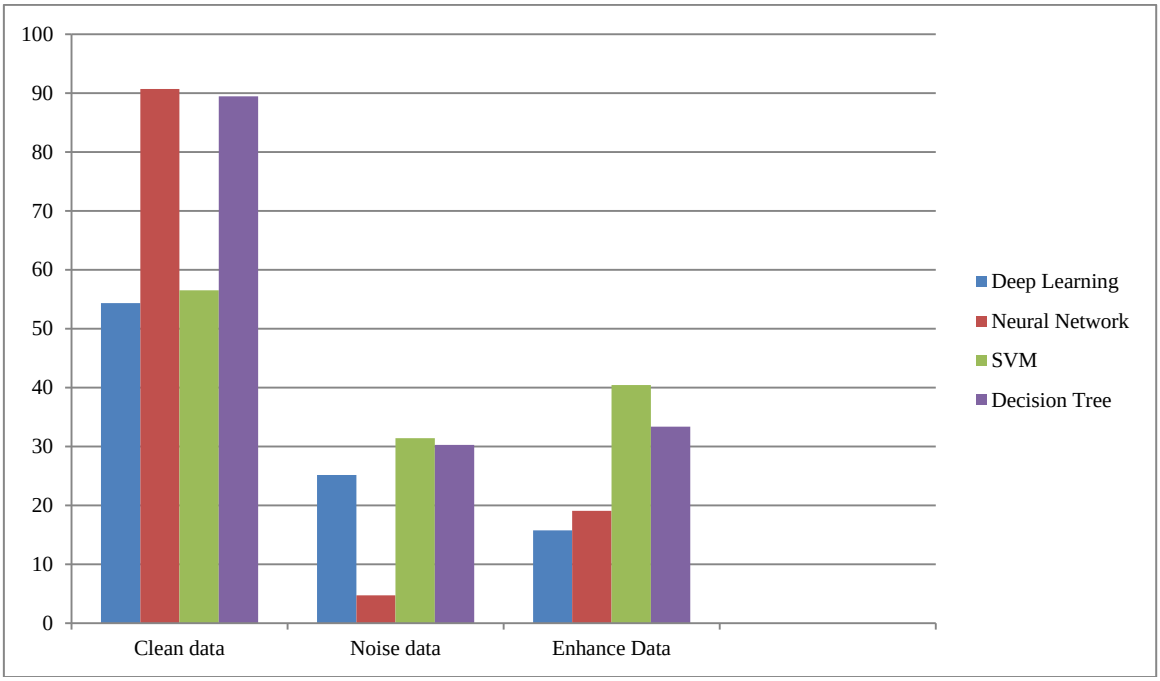

Fig.5. Comparison of average Accuracy results for various data signals

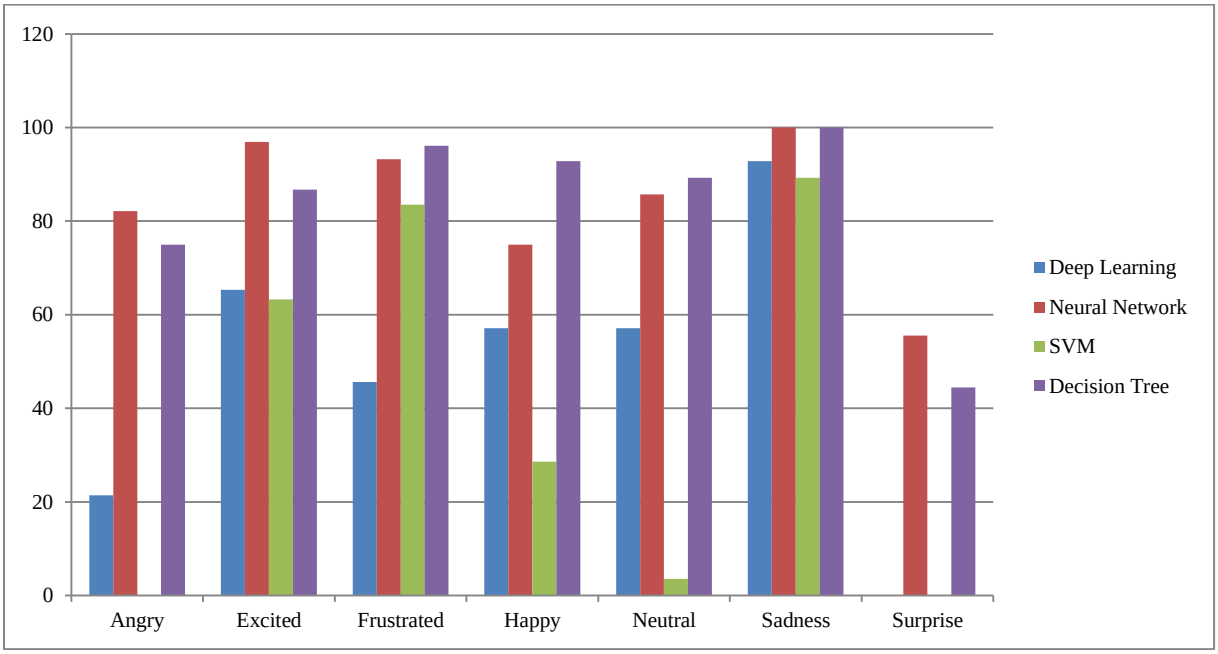

Fig.6. Comparison results for each emotion of the clean signals

As can be seen from the results of Fig.5, Neural Network and Decision Tree provide the best results for clean data. But, they provide degrade performance for the noisy signal and SVM gives the best performance results among them. SVM also hold the best classifier for those enhanced signals with the MMSE technique. Deep Learning gradually decreases the performance from the clean signals to noisy signals and enhanced signals. The results clearly show that the performance of the classifier depends on the signals data types and should be notice which classification model is suitable for which types of application.

The experiments are continued to be carried out to analyses which classifier can do the best for each emotion. The performance results for the best recognition of emotions are shown in the following figures. Fig.6 shows the accuracy result for each emotion of the clean signals for various classifiers, Fig. 7 illustrates for average noisy signals and Fig. 8 proves the enhanced signals results.

From the results of the Fig.6, it can be conclude that Neural Network and Decision Tree provide the full marks recognition results for sadness emotion. It can be the useful information for some system, such as health care system, because the tone of the sadness is very low and the human beings cannot recognize easily. Generally Neural Network outperform for all types of emotion for clean signals. But SVM and Deep Learning perform the worst recognition for surprise emotion.

From the noisy signal results of Fig.7, the classifier can perform only for both angry and excited emotions because the adding noise signal is high and the other emotion signals may not enough to be recognized. Although Neural Network can perform well for all clean signals, it can only do for angry emotion here. For excitement emotion, Decision Tree provides well results.

The results for the enhanced signals for each emotion are shown in Fig.8. SVM can't perform in Angry, happy and surprise emotions and Deep Learning and Decision Tree can perform well for Angry signal types.

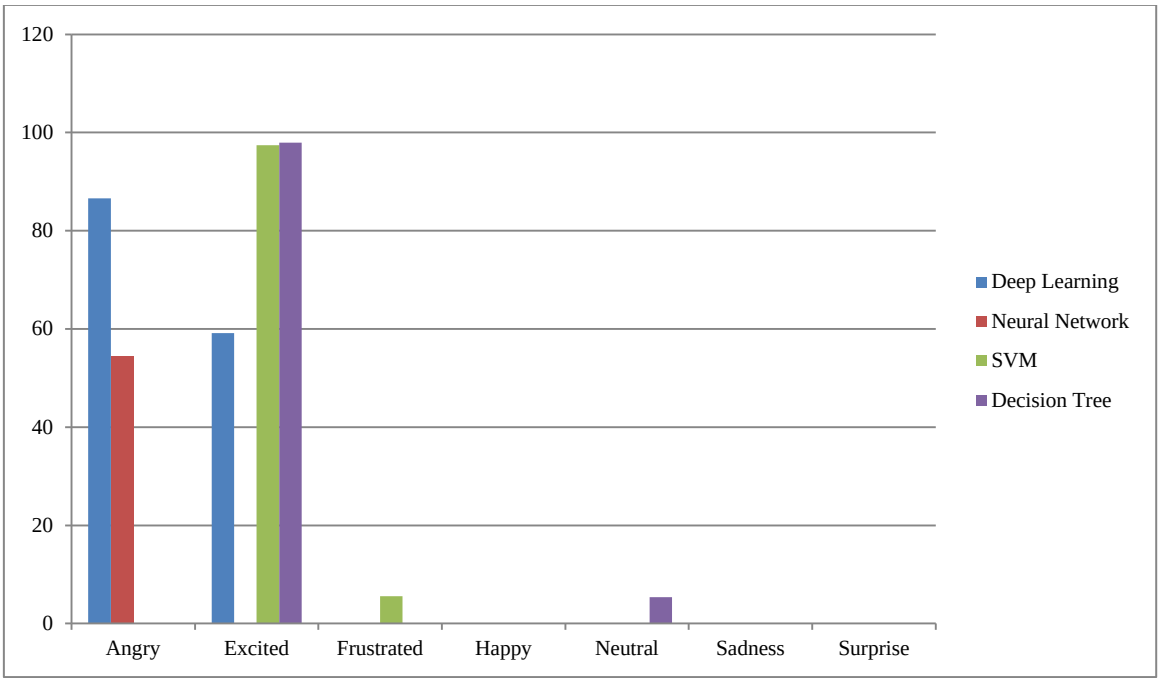

Fig.7. Comparison results for each emotion of the average noisy signals from 0db to 15db

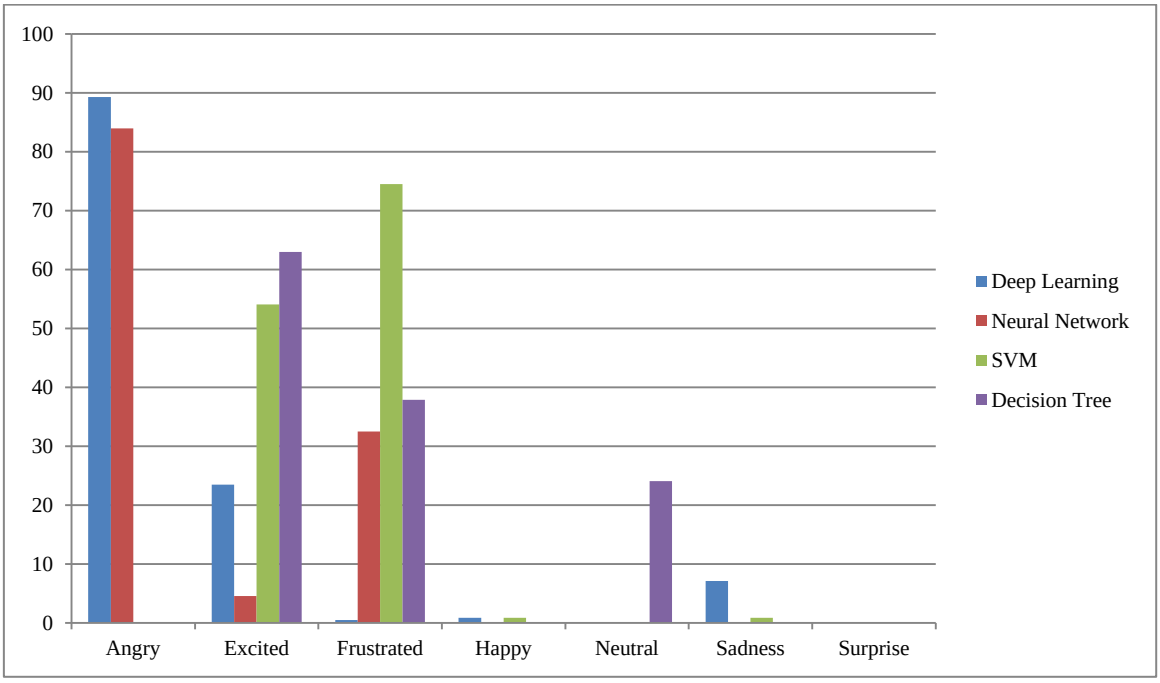

Fig.8. Comparison results for each emotion of the average enhanced signals from 0db to 15db

Deep Learning and Neural Network can also provide for enhanced angry signals. Decision Tree is still the best for excitement emotion and SVM outperform the best result for Frustration. Surprise emotion cannot recognize by none of the classifier except for clean signals.

V. Conclusion

This paper focuses on the development of speed emotion recognition system for noisy environment with the advantage of the MMSE enhancement technique. The performances are tested by using the IEMOCAP standard dataset. MFCC features can deal with various types of the signal including noise and enhancements techniques. Different types of popular machine learning techniques are used to measure the accuracy performance for each signal data type. The system is also tested the performance of the classifiers for each emotion to analyses which classifier is the best for which emotion to facilitate in real world environment. Neural Network outperforms the best result for clean data and SVM is the best among the classifier for both addition of the original signal. Deep Learning gives the superlatives performance for angry emotion of noisy signal and SVM provide the most recognizable classifier for frustration. Therefore, this research experiments show which types of classifier model is appropriate for which application or devices including the determination of signal data types whether it is clean, noise or enhanced signal. However, the system needs to extend to give the better results for all real world noises with the best enhancement techniques. Finding the better speed enhancement techniques still remain the ongoing works. This may be the future direction for this research work.

References

[1] Causse, M., Dehais, F., P éran, P., Sabatini, U., and 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.

[2] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Cleder (March 25th 2019). Automatic Speech Emotion Recognition Using Machine Learning [Online First], IntechOpen, DOI: 10.5772/intechopen.84856.

[3] Panikos Heracleous et.al, "Speech Emotion Recognition in Noisy and Reverberant Environments", 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), 23-26 Oct. 2017, DOI: 10.1109/ACII.2017.8273610

[4] Tayseer M. F. Taha, Ahsan Adeel and Amir Hussain, "A Survey on Techniques for Enhancing Speech", International Journal of Computer Applications ·February 2018 DOI: 10.5120/ijca2018916290

[5] Sun et al. , "Decision tree SVM model with Fisher feature selection for speech emotion recognition", EURASIP Journal on Audio, Speech, and Music Processing (2019) 2019:2 https://doi.org/10.1186/s13636-018-0145-5

[6] Xixin Wu, et.al, "Speech Emotion Recognition Using Capsule Networks", 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 12-17 May 2019, DOI: 10.1109/ICASSP.2019.8683163

[7] Michael Neumann, Ngoc Thang Vu, "Improving Speech Emotion Recognition With Unsupervised Representation Learning On Unlabeled Speech", 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 12-17 May 2019, DOI: 10.1109/ICASSP.2019.8682541

[8] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Julien Epps, "Direct Modelling of Speech Emotion from Raw Speech", DOI: https://arxiv.org/abs/1904.03833

[9] Evaggelos Spyrou 1,2,3,_, Rozalia Nikopoulou 4, Ioannis Vernikos 2 and Phivos Mylonas, "Emotion Recognition from Speech Using the Bag-of-VisualWords on Audio Segment Spectrograms", Technologies 2019, 7, 20; doi:10.3390/technologies7010020

[10] Htwe Pa Pa Win and Phyo Thu Thu Khine, "Speech Enhancement Techniques for Noisy Speech in Real World Environments", Proceedings of the 17th International Conference on Computer Applications, pp.238-244, 27th – 28th February, 2019.

[11] Farah Chenchah and Zied Lachiri, "Speech emotion recognition in noisy environment", 2nd International Conference on Advanced Technologies for Signal and Image Processing - ATSIP'2016, March 21-24, 2016, Monastir, Tunisia

[12] Bashirpour and Geravanchizadeh, "Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments", EURASIP Journal on Audio, Speech, and Music Processing (2018) 2018:9, https://doi.org/10.1186/s13636-018-0133-9

[13] 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.

[14] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Volume 2, Issue3, March 2010, ISSN 2151-9617.

[15] 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.

Authors' Profiles

Htwe Pa Pa Win received her Ph.D (IT) from University of Computer Studies, Yangon, Myanmar in 2012. She is currently working as a Lecturer at the University of Computer Studies, Hpa-an, Myanmar. Her research interests include Image Processing, Speech processing and Digital Signal processing.

Phyo Thu Thu Khine received her Ph.D (IT) from University of Computer Studies, Yangon, Myanmar in 2012. She is currently working as a Lecturer at the University of Computer Studies, Hpa-an, Myanmar. Her research interests include Image Processing, Speech

processing, Digital Signal processing, Database Management System and Big Data.

How to cite this paper: Htwe Pa Pa Win, Phyo Thu Thu Khine, " Emotion Recognition System of Noisy Speech in Real World Environment", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.12, No.2, pp. 1-8, 2020.DOI: 10.5815/ijigsp.2020.02.01

