A COMPARATIVE STUDY OF DIFFERENT CLASSIFIERS FOR EMOTION RECOGNITION-AN OVERVIEW

Piyu Sarcar
Assistant Professor
Electronics and Communication Engineering Dept.
Narula Institute of Technology, Kolkata, India

Abstract : With the advent of modern technology, emotion and sentiment recognition plays an important role in human machine interface. Various kinds of human characteristics such as facial expression, speech analysis, body movement, eye movement, body temperature, blood pressure are considered to extract emotion, which has an important contribution in medical field. This paper overviews the extracted features from speech followed by different classifiers that makes a comparative study. This paper also highlights the functions of classifiers RNN, LSTM, SVM, ANN, Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) which is very much crucial to recognize speech and facial emotion.

***IndexTerms* - speech analysis, Classifiers, RNN, LSTM, SVM, Hidden Markov Model, GMM**

I. INTRODUCTION

Speech and facial expression are the key features to recognize emotion in human which plays An important aspect of human-computer- interaction (HCI). With the rapid enhancement in the number of internet the multimedia data would help to improve more naturalness of human voice. Emotion is also varied by facial expression. To extract features from facial expression various audio-visual aid are used [1][2]. Different features of speech i.e pitch, formant, loudness, mel-frequency cepstral coefficients are considered.

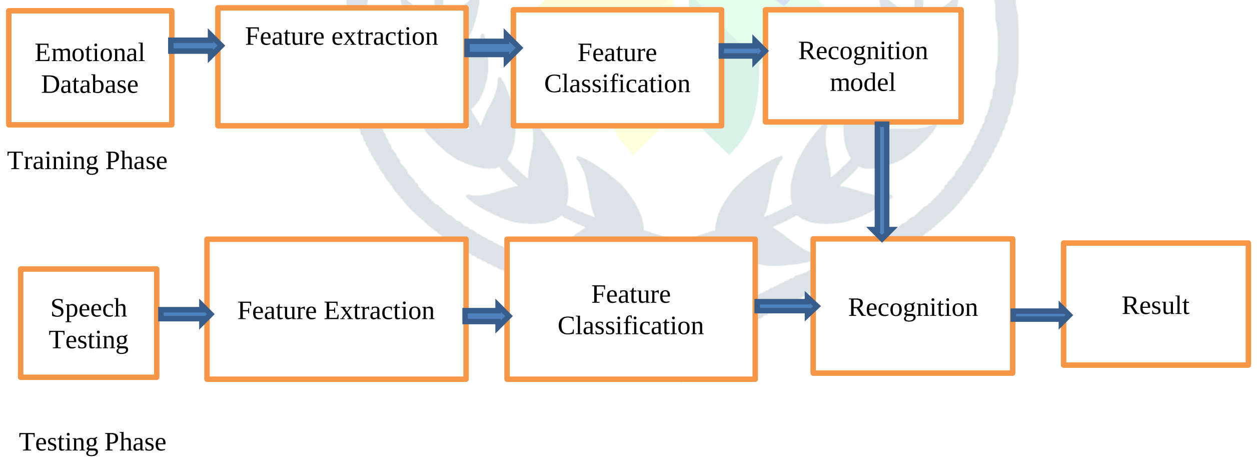

Fig.1 Training and testing phase of Speech Emotion Recognition

Types of Speech Features :

a) Pitch: Pitch indicates the degree of highness or lowness of a sound.
b) Formant: Formant is a spectral analysis that results from the vocal resonance of human voice. The maximum peak amplitude in the spectral part is considered as formants.
c) Loudness: Loudness is defined as the attribute of the sound which determines the auditory sensation which depends on maximum amplitude of the sound wave.
d) Mel Frequency Cepstral Coefficient (MFCC): Mel frequency is extensively applied in Automatic Speech Emotion Recognition system (ASR). To calculate MFCC, a signal is divided into short frames. Then for each frame periodogram of power spectrum is calculated. Then mel filter bank is applied in each power spectrum and add the energy of each filter. Then use Discrete Cosine

Transform (DCT) after taking the logarithm of all filter bank energies and keep DCT co-efficient 2-13 and neglect the other co-efficient.

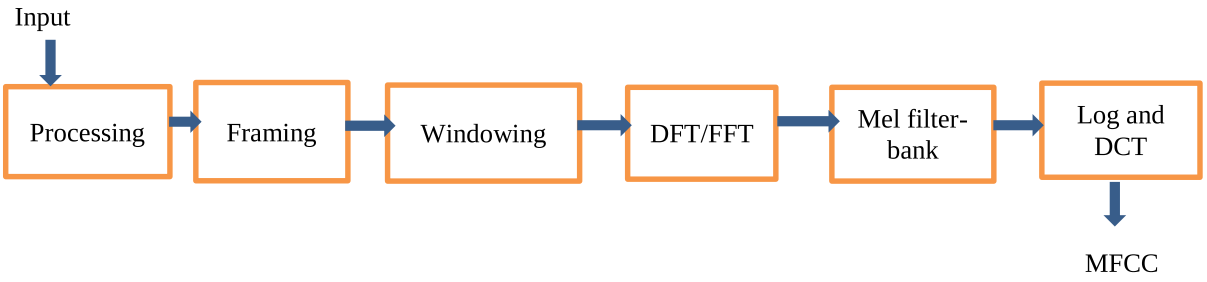

Fig. 2 Block Diagram of MFCC

After feature extraction, features would be classified by using various types of classification techniques. In this paper a comparative study of Recurrent Neural network (RNN), Long Short Term Memory (LSTM), Support Vector Machine (SVM), Hidden Markov Model (HMM) classifiers are discussed.

II. RELATED WORK:

Now a days many researches are conducted and many algorithms are developed in speech and facial emotion recognition. In [6] author presents how facial and vocal modalities are combined to improve the recognition accuracy. Xianxin Ke et.al [7] explain CASIA emotional data set based on which statistical values of emotional features are extracted speech using SVM and ANN classifier. [5]Presents accuracy and performance of Gaussian Mixture Model and parametric modelling technique to detect emotion in speech signal. Lili Guo et al. demonstrates feature fusion method [10] which combines Convolution Neural Network (CNN) based feature with heuristic based features using Deep Neural Network. Deep architecture using moderately deep temporal architecture which use gate-based skip connections and memory [11]. Sparse Auto Encoder (SAE) is used to learn local invariant features using CNN and this features are then used to salient discriminative feature analysis that provides prominence and orthogonality which constitutes a robust platform for SER [12]. Affective multimedia interfaces plays an important role in automatic speech recognition [13] which focus on how facial movement changes with emotions and the result shows combination of emotion recognition provides accuracy rate higher than individual emotion recognition. Affective saliency uses minimum classification error (MCE) criteria [14] which represents global and local prosodic features of speech and their combined effects shows higher classification accuracy. On the contrary of Automatic Personality Recognition, Automatic Personality Perception (APP) which attributes personality of a person [16]. LSTM gives temporal information of speech when computing attention vector using 5-fold cross validation method applied to state-of-the-art SER deep learning system [17]. Olga Krestinskaya Proposes a facial emotion recognition algorithm [18] which applies min-max metric for nearest neighbour classifier to eliminate pixel mismatch. For proper extraction of features from thousands of speech data harmony search (HS) algorithm plays a significant role which allows classifiers for accurate classification [19].Emily Mower [20] et al mention Emotion Profile-based (EP) emotion classification that describes multiple probabilistic class levels to clarify natural emotional state of human.

III. CLASSIFIERS

3.1 Recurrent Neural Network (RNN):

RNN is a powerful classification method to recognise emotion. RNN is a type of neural network which captures informations from a sequences or time series. It works on recursive function in which

$S_t = F_w (S_{t-1,}X_t)$

$S_{t-1=}$ tanh $(W_sS_{t-1+}W_xX_t)$

$Y= W_yX_t$

Where, $X_{t =}$ input at time step t

$S_{t =}$ state at time step t.

$Y_{t =}$ output at time step t.

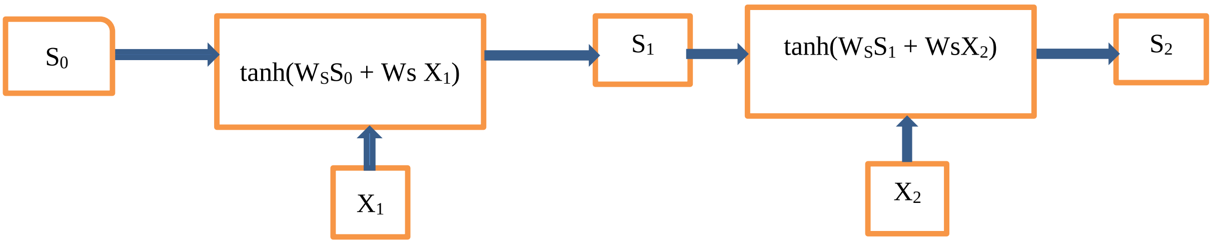

Fig.3 Recurrent Neural Network (Single layer)

The main drawback of RNN network is vanishing gradient problem. To solve this problem Long Short Term Memory Module (LSTM) classifier may be deployed which uses forget gate (f_t), input gate (i_t) and output gate (o_t).

$f_t = \sigma(W_f S_{t-1} + W_x X_t)$
$i_t = \sigma(W_i S_{t-1} + W_i X_t)$
$o_t = \sigma(W_o S_{t-1} + W_o X_t)$
$C_t = \tanh(W_c S_{t-1} + W_c X_t)$
$c_t = (i_t \times C_t) + (f_t \times C_{t-1})$
$h_t = o_t \times \tanh(c_t)$

c_t and h_t is known as cell state and new state respectively.

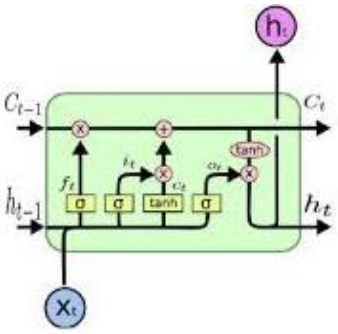

Fig.4 Long Short Term Memory [24]

3.2 Support Vector machine (SVM)

SVM is a discriminative classifier which separates two hyper-plane. For multi-classification problem three algorithms one to many, one to one and hierarchical SVM are deployed.

One to many algorithm: In this category sample is classified as positive and negative set. The number of negative set is more than positive set. This algorithm does not give satisfactory result.

One to one algorithm: In this algorithm k (k-1)/2 classifiers are used for k number of problem. The result shows highest number of occurrence.

Hierarchical SVM: Here samples are classified as sub-categories. This sub-category is further divided into sub-categories. This continues until an independent category is achieved. Cai and Hofmann proposes multilevel case in hierarchical SVM [4] [21].

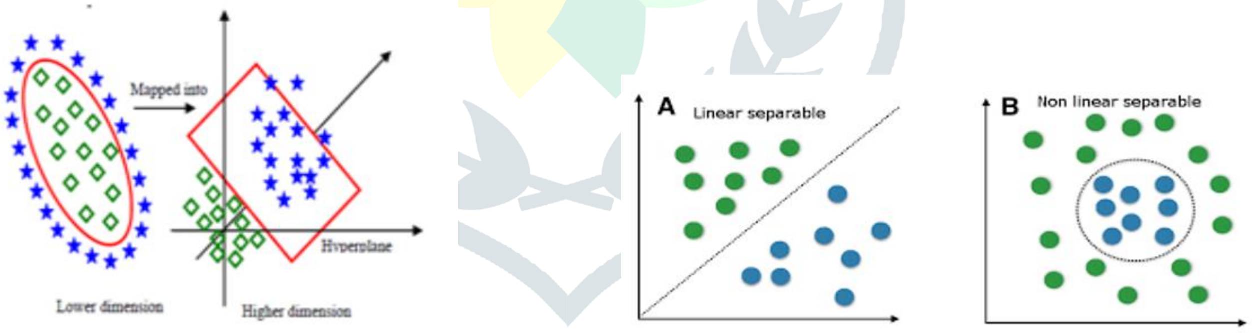

Fig.5 Hierarchical SVM [22]

3.3 Artificial Neural Network (ANN):

ANN is constructed from artificial neurons which performs similar as human brain. The input layer is multiplied by weight, these products are added and fed to the fuction to generate proper output. Now a days ANN plays an important role in speech emotion recognition [5].

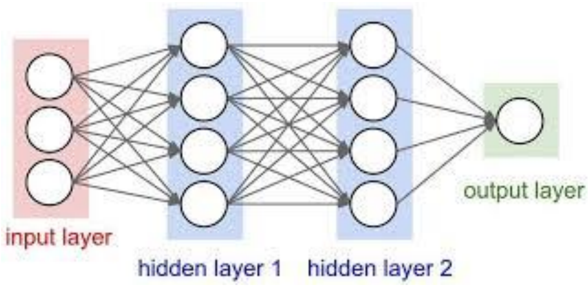

Fig. 6 Artificial Neural Network

3.4 Hidden Markov Model:

Hidden Markov Model is a very effective model which represents spectral analysis of speech sequences A hidden Markov Model represents probability distribution over sequence of observations at time t. Observation at time t has S_t states have hidden properties and these hidden states satisfies Markov property. The main pros of HMM classifier is that it can train the sequence of data and also it can predict each words from the trained data.

If the states T= $t_1,t_2,t_3.....t_N$. and observations W= $w_{1,}\ w_{2,}w_{3....}w_N$

$$P(t_1,t_2.......t_N) = \prod_{i=1}^{n} P\left(ti \mid t(i-1)\right)$$

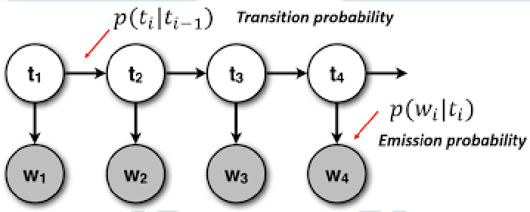

Fig. 7 Hidden Markov Model [23]

3.5 Gaussian Mixture Model (GMM)

GMM represents class conditional distribution of emotional features and their parameters are determined by Expectation Maximization (EM) algorithm or Maximum A Posteriori (MAP) which is based on training data. EM algorithm is an iterative optimization technique that is locally operated. This algorithm has two states: Estimation state and Maximization sate. In estimation state expected values of latent variables are estimated, whereas in the maximization state parameters are updated which depend on latent variable. GMM is a statistical method which represents the spectral components of a speech spectrum which gives approximately 70 % accuracy rate. GMM is similar to kernel density estimates but number of component is less. The spectral GMM has more flexibility than Gravity Centroid (GC) parameters. The Gaussian Mixture Model is parameterized by mean vectors, covariance matrices, and weight of the mixtures from component densities.

For a sequence of T training vectors

1) the total likelihoods of Gaussian pdf is expressed as

   $T_j = \sum_{I=1}^{N} Pi\ g(xi, \mu i, \sum i)$ j=1,2…N

2) The normalized likelihood is represented by the following equation

   $nij = Pi\,\frac{g(xi,\mu i,\sum i)}{Tj}$ i=1 to j-1.

IV. CONCLUSION

Emotion recognition in speech is a multimodal process which is the part of human computer interaction. Each classifier can recognize particular emotion correctly which enhances accuracy rate. Automatic speech emotion Recognition (ASR) is always trained in minimum atmospheric noise. So if the experiment is done in noisy ambient, the overall efficiency will be degraded which is the main drawbacks of the system. In this paper comparison among different classifiers and their performance in speech recognition are investigated. In the future work emotion recognition based on speech and movement using combination of different classifier will be included.

REFERENCES

[1] Pathak,Sandeep.,Kolhe,Vaishali. 2016. A Survey on Emotion Recognition from Speech Signal.IJARCCE.5(7).

[2] Iliou, Theodoros. and Anagnostopoulos ,Christos-Nikolaos.2009.Statistical Evaluation of Speech Features for Emotion Recognition. Fourth International Conference on Digital Telecommunications.121-126.

[3] Kerkeni, Leila., Serrestou ,Youssef. Mbarki, Mohamed., Raoof, Kosai. and Mahjoub ,Mohamed Ali.2018 Speech Emotion Recognition: Methods and Cases Study, SCITEPRESS. 176-182.

[4] Choi,Heejin., Sasaki ,Yutaka. and Srebro, Nathan. 2015.Normalized Hierarchical SVM. arXiv:1508.02479

[5] Vyas, Manan. 2013. A Gaussian Mixture Model Based Speech Recognition System Using MATLAB, An International Journal (SIPIJ) 4(4): 109-118.

[6] Metallinou, Angeliki., Lee, Sungbok. and Narayanan, Shrikanth. 2008. Audio-Visual Emotion Recognition using Gaussian Mixture Models for Face and Voice. IEEE International symposium. 250-257.

[7] Ke ,Xianxin., Zhu ,Yujiao., Wen, Lei. and Zhang, Wenzhen. 2018. Speech Emotion Recognition Based on SVM and ANN. International Journal of Machine Learning and Computing. 8(3): 198-202.

[8] Sarma , Mousmita., Ghahremani, Pegah., Povey,Daniel., Goel, Nagendra Kumar., Sarma Kandarpa Kumar. and Dehak ,Najim. 2018. Emotion Identification from raw speech signals using DNNs. Interspeech 2018. 3097-3101

[9] Albanie, Samuel., Nagrani, Arsha., Vedaldi ,Andrea. and Zisserman, Andrew. 2018. Emotion Recognition in Speech using Cross-Modal Transfer in the Wild, Creative Commons Attribution 4.0 Visual Geometry Group, Department of Engineering Science, University of Oxford.22–26.

[10] Guo ,Lili., Wang, Longbiao., Dang, Jianwu., Zhang ,Linjuan . and Guan ,Haotian. 2018. A feature fusion method based on extreme learning machine for speech emotion recognition. ICASSP 2018, 2666-2670.

[11] Kim, Jaebok. Englebienne , Gwenn., Truong , Khiet P. and Evers ,Vanessa.2017.Deep Temporal Models using Identity Skip-Connections for Speech Emotion Recognition. ACM.

[12] Mao, Qirong., Dong,Ming., Huang , Zhengwei. and Zhan, Yongzhao. 2014. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia 16(8): 2203–2213.

[13] Davood, G., Mansour, S., Alireza, N. and Sahar, G., 2012.Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network.Neural Comput. Appl. 21(8): 2115–2126.

[14] Rao,K. S., Koolagudi ,S. G. and Vempada, R. R.2012. Emotion recognition from speech using global and local prosodic features. Int. J. Speech Technol., 16(2): 143–160.

[15] Chorianopoulou, Arodami., Koutsakis ,Polychronis. and Potamianos, Alexandros.2016. Speech Emotion Recognition using Affective Saliency. ISCA Speech.

[16] Mohammadi, G., and Vincianelli, A. 2012. Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features", IEEE Transactions on Affective Computing, 3(3), 273-284.

[17] Ramet, Gaetan., Garner, Philip N., Baeriswyl, Michael. And Lazaridis, Alexandros. 2018.Context-Aware Attention Mechanism For Speech Emotion Recognition.

[18] Krestinskaya,Olga, James, Alex Pappachen. 2017. Facial Emotion Recognition using Min-Max Similarity Classifier. International Conference on Advances in Computing, Communications and Informatics (ICACCI).

[19] Tao,Yongsen., Wang, Kunxia., Yang, Jing., An, Ning and Li,Lian.2015. Harmony search for feature selection in speech emotion recognition. International Conference on Affective Computing and Intelligent Interaction (ACII). 362-367.

[20] Mower, Emily., Mataric, Maja J. and Narayanan, Shrikanth. 2011. A Framework for Automatic Human Emotion Classification Using Emotion Profiles. IEEE Transactions On Audio, Speech, And Language Processing. 19(5): 1057-1070.

[21] Vijayavani, E., Lavanya, S., Suganya,P. and Elakiya ,E. 2014.Emotion Recognition Based on MFCC Features using SVM. International Journal of Advance Research in Computer Science and Management Studies. 2(4): 31-36.

[22] Support Vector Machine images. medium.com .

[23] Hidden Markov Model images. davidsbatista.net.

[24] Long Short Term Memory image.colah.github.io