A study on Speech Emotion Recognition

Nitesh Kumar[1], Reema[2], Shabnam Kumari[3]
[1]*M.Tech scholar, Department of CSE, SKITM, Bahadurgarh, Haryana*
[2]*A.P., Department of CSE, SKITM, Bahadurgarh, Haryana*

***Abstract*- Emotion recognition is the task of conceding a person's emotional state such as sad, happy, anger, disguise, love, hate etc. Emotion recognition plays very key role in present-days to improve both openness and efficacy of human-computer interaction. An emotion is an amalgamated intellectual state that involves different elements such as an individual experience, a functional reaction, and a behavioral or expressive response. This paper introduce necessity and applications of speech emotion recognition. A big research has been addressed to enhance.**

***Index Terms*- ASR, SVM, KNN, pitch.**

1. INTRODUCTION

Emotions play a crucial role in modulating how humans experience and interact with the outside world and have a huge effect on the human decision making process. They are an essential part of human social relations and take role in important life decisions. Therefore detection of emotions is crucial in high level interactions. Each emotion has unique properties that make us recognize them. Acoustic signal generated for the same utterance or sentence changes primarily due to biophysical changes triggered by emotions. This relation between acoustic cues and emotions made speech emotion recognition one of the trending topics of the affective computing domain. The main reason of a speech emotion recognition algorithm is to identify the emotional condition of a speaker from recorded speech signals.

1.1. Speech emotion recognition system:

Speech emotion recognition is nothing but an application of the pattern recognition system in which patterns of derived speech features such as Pitch, Energy, MFCC are mapped using classifier like ANN, SVM, HMM etc.

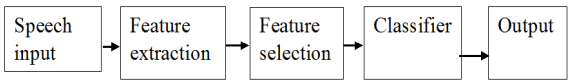

Fig.1.1 Block diagram of speech emotion recognition system

The system contains five major modules: speech input database, feature extraction, feature selection, classifier & recognized output as illustrated in figure 1.1 above. Overall, the system is based on deep analysis of the generation mechanism of speech signal, extracting some of features which contain information about speaker's emotion & taking appropriate pattern recognition model to identify states of emotion. Typically, a set of emotion having 300 emotional states. Whenever, signal is passed to the feature extraction & selection process, the extracted speech features are selected in terms of emotion relevance. All over procedure revolves around the speech signal for extraction to the selection of speech features corresponding to emotions. Forward step is generation of database for training as well as testing of extracted speech features. At the end, detection of emotions has been done using classifier with the usage of pattern recognition algorithm. The Speech emotion recognition is similar to the speaker recognition system but different types of approach to detect emotions make it secure & intelligent. The evaluation of the system is depending on naturalness of the input database.

2. LITERATURE REVIEW

During present scenario, for human emotion recognition an extensive research is made by using different speech information and signal [1]. Many researchers used different classifiers for human emotion recognition from speech such as Hidden Markov Model (HMM) [2], Neural Network (NN), Maximum likelihood bayes classifier (MLBC), Gaussian Mixture Model (GMM), Kernel deterioration and K-nearest Neighbours approach (KNN), support vector machine (SVM)[2] [3], Naive Bayes classifier. Commonly used classifiers for human emotion recognition from speech such as Hidden Markov Model (HMM), Neural Network

(NN), Maximum likelihood bayes classifier (MLBC), Kernel deterioration and K-nearest Neighbours approach (KNN) [7], support vector machine (SVM), Naive Bayes classifier[6], Gaussian Mixture Model (GMM)[8].

Wootaek Lim et al., (2017) [9] : investigated that with rapid developments in the design of deep architecture models and learning algorithms, methods referred to as deep learning have come to be widely used in a variety of research areas such as pattern recognition, classification, and signal processing. Convolutional Neural Networks (CNNs) especially show remarkable recognition performance for computer vision tasks. In addition, Recurrent Neural Networks (RNNs) show significant success in various sequential data processing tasks. In this study, we explore the result of the Speech Emotion Recognition (SER) algorithm based on RNNs and CNNs trained using an emotional speech database. The main goal of our work is to suggest a SER method based on concatenated CNNs and RNNs without using any fixed hand-crafted features.

Pavitra Patel et al., (2017) [10]: investigated that speech has several characteristic features such as naturalness and efficient, which makes it as attractive interface medium. It is possible to express emotions and attitudes through speech. In human machine interface application emotion recognition from the speech signal has been current topic of research. Speech emotion recognition is an important issue which affects the human machine interaction.

Prajakta P. Dahake et al., (2016) [11]: investigated that in human computer interaction, speech emotion recognition is playing a pivotal part in the field of research. Human emotions consist of being angry, happy, sad, disgust, neutral. In this paper the features are extracted with hybrid of pitch, formants, zero crossing, MFCC and its statistical parameters. The pitch detection is done by cepstral algorithm after comparing it with autocorrelation and AMDF.

Ritu D.Shah and Dr. Anil. C.Suthar (2016) [12]: In this paper methodology for emotion recognition from speech signal is presented. Some of sound features are removed from speech signal to analyze the features and behaviour of speech.

Kunxia Wang et al., (2015) [13]: Recently, studies have been performed on harmony characteristics for speech emotion recognition. It is found in our study that the first- and second-order differences of harmony features also play an important role in speech emotion recognition. Investigational results show that the offered Fourier parameter (FP) features are effective in recognising various emotional states in speech signals.

Rahul B.Lanjewar et al., (2015) [14]: investigated that the kinship between man and machines has become a new trend of technology such that machines now have to respond by considering the human emotional levels. The signal processing and machine learning technologies have boosted the machine intelligence that it gained the capability to understand human emotions.

3. METHODOLOGY USED FOR SPEECH EMOTION RECOGNITION

3.1. speech

A record of emotional speech data collections is undoubtedly useful for researchers interested in emotional speech recognition. It is evident that research into emotional speech recognition is limited to certain emotions, because the majority of emotional speech data collections encompass 5 or 6 emotions, although there are many more emotion categories in real life. Two speech corpora were used in this investigation: the first Database of Emotional Speech and contains semantic units made up of sentences: the second, in English, is called Speech Under Simulated and Actual Stress and comprises semantic units made up of single words.

3.2. Emotion recognition

Emotion detection from the speech signal is a relatively new field of research, that have much potential application in the current scenario. In human we can easily identify the emotion of a person by interacting with each other. But in human computer interaction systems, emotion recognition is not an easy task. Emotion system could provide users with improved services by being versatile to their emotions. In virtual world, emotion recognition could help in simulating more realistic interactions. Emotions in speech are quite limited, we are not hundred percent sure that always we can find the emotion of a person. Currently, researchers are still debating regarding the recognition of emotion in speech for accuracy. There is also considerable apprehension as to the best algorithm for classifying emotion and which emotions to class together or

what can be their feature. In this project we use K-Means and Support Vector Machine (SVM) algorithm to classify opposing emotions. We separate the speech by speaker gender to investigate the relationship between gender and emotional content of speech. Here we extract a variety of emotions as compare to other algorithms. This process also allows us to develop criteria to class emotions together. In this algorithm, our system accuracy is much high.

3.3. Emotion classification

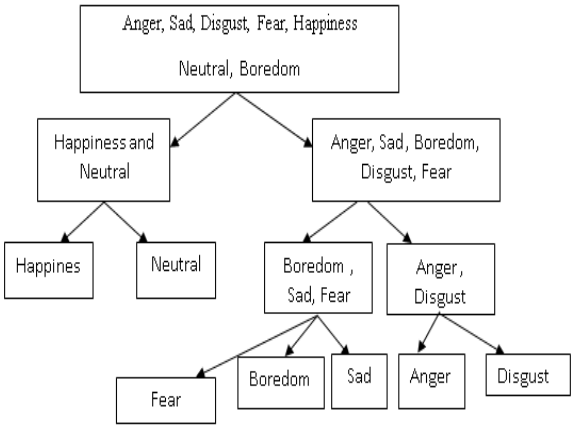

Fig3.1 Classification schema for speech emotion recognition

4. FEATURES WIDELY USED IN EMOTION RECOGNITION

As mentioned previously, extraction of speech features is a very important process in speech emotion recognition. Speech features can be separated into numerous categories. In (E. Ayadi et 13 al., 2011) [15], speech features are divided into 4 categories; spectral continuous, qualitative and TEO-based. Continuous features are pitch, energy and formants. Quantitative features are described as voice quality features which are harsh, tense and breathy voices. Most popular acoustic features used in emotion recognition process are outlined below.

4.1. Pitch features

Pitch is the fundamental frequency of the glottal excitation. Pitch depends on the tension of the vocal folds and subglottal air pressure. Pitch frequency is one of the widely used features in emotion from speech applications. Pitch frequency is also known as the fundamental frequency. The time elapsed between successive vocal fond openings determine the fundamental frequency (Ververidis & Koropoulos, 2006) [16]. From pitch features given features could be extracted which are min, max, mean, standard deviation, range at the turn level, slope (mean and max) in the voiced segments, regression coefficient and its mean square error and maximum cross-variation of F0 between two adjoining voiced segments (inter-segment) and with each voiced segment(intra-segment) (Vidrascu & Devillers, 2005) [17].

4.2. Teager energy operator

Produced number of harmonics due to the non-linear air flow in the vocal tract is another useful acoustic feature. In case of anger, the fast air flow causes nonlinear stress (Teager, 1990) [18]. In (E. Ayadi et al., 2011) [15], it was stated that TEO-based features can be used to detect stress in speech.

4.3. Vocal tract features

Formants are a vocal tract feature. Each formant has its own bandwidth and center frequency (Ververidis & Koropoulos, 2006) [16]. Slackened speech can be distinguished from an articulated speech using formant features. Other widely used feature is the energy of a certain frequency which corresponds to the critical bands of the human ear (Ververidis & Koropoulos, 2006)[16].

4.4. Spectral features

Mel-frequency cepstrum coefficients, linear predictive coding and log frequency power coefficients are the most popular spectral features. Mean and standard deviation of 13 Mel frequency cepstral coefficients (MFCC) are set as discriminating features in many studies. (D. Wu, Parsons, & Narayanan, 2010) [19].

4.5. Duration features

Mean and standard deviation of the duration of voiced and unvoiced segments, ratio between the duration of unvoiced and voiced segments are (D. Wu et al., 2010) duration features.

4.6. Energy features

Energy mean, standard deviation, maximum, 25% and 75% quantiles, and the inter quantile distance are the popular energy based features used in speech emotion recognition task (D. Wu et al., 2010) [19].

5. SPEECH EMOTION RECOGNITION

Speech emotion recognition is basically performed through pure sound processing without linguistic information. In terms of acoustics, speech processing techniques offer extremely valuable information derived mainly from prosodic and spectral features. Sometimes the process is assisted by Automatic Speech Recognition (ASR) systems, which contribute to classification using linguistic information. However, the use of ASR is limited due to fact that most of the experiments in the field have been assessed using databases of non-spontaneous and predefined speech and thus, there is no need for speech recognition. After sound processing and feature acquisition, it is quite common to follow a feature selection in search for the "golden set" of sound features. Finally, such a plethora of classification algorithms has been evaluated for speech emotion recognition that attempting their comparison in this paper is, unfortunately, an impractical task. This is also due to the fact that there is a lack of uniformity in the way these methods are evaluated (different test sets, feature vectors and evaluation frameworks) and, therefore, it is inappropriate to make direct comparisons or explicitly declare which methods demonstrate the highest performance. In the next sections, a brief classification of papers that follow the basic processing pipeline (as highlighted in Fig. 1) are surveyed and categorized according to their major methodology for feature processing (with or without linguistic information), as well as their classification schema for emotion recognition.

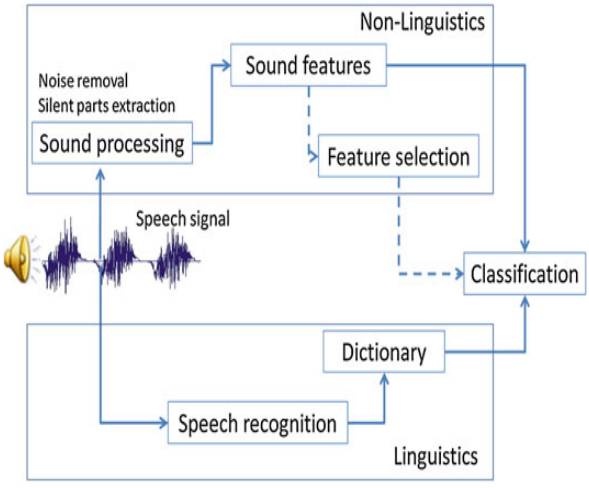

Fig 5.1 Speech emotion recognition pipeline

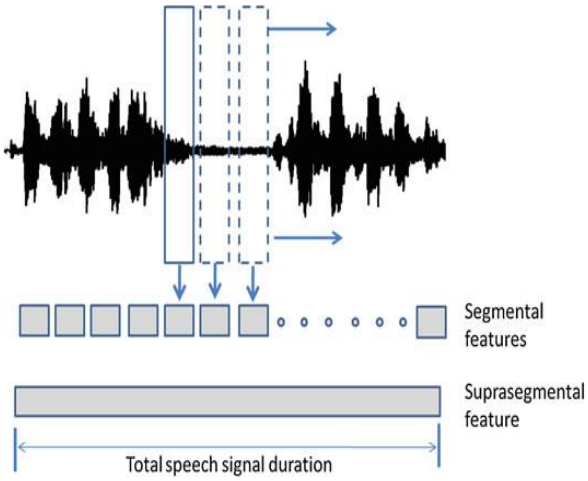

Fig 5.2 Segmental and Super-segmental features in a speech signal

6. CHALLENGES

This section describes some of the expected challenges in implementing a real time speech emotion detector. Firstly, discovering which features are indicative of emotion classes is a difficult task. The key challenge, in emotion detection and in pattern recognition in general, is to maximise the between-class variability so that classes are well separated. However, features indicating different emotional states may be overlapping, and there may be multiple ways of expressing the same emotional state. One strategy is to compute as many features as possible. Optimisation algorithms can then be applied to select the features contributing most to the discrimination while ignoring others, creating a compact emotion code that can be used for classification. This avoids making difficult a priori assumptions about which features may be relevant. Secondly, previous studies indicate that several emotions can occur simultaneously. For example, co-occurring emotions could include being happy at the same time as being tired, or feeling excited touched and surprised, when enquiry good news. This requires a classifier that can infer multiple temporally co-occurring emotions. Thirdly, real-time classification will require choosing and implementing efficient algorithms and data structures. Despite there existing some working systems, implementations are still seen as challenging and are generally expected to be imperfect and imprecise.

7. CONCLUSION

Substantial efforts have been made over the time to perceive more vigorous methods of assessing truthfulness, integrity, delusiveness and reliability during human interactions. Efforts have been made to catch speech expressions of anyone. Emotions are due to any liveliness in brain and it is expressed through face, as face has maximum sense organs. The objective of this research paper is to give brief introduction towards techniques, application and challenges of speech emotion recognition system. We studied emotion recognition by using various techniques. It will help to understand these techniques in short and to choose from them which can be well suited for future evaluation depending on the individual's work. It also summarizes the research work that has been done in this field.

REFERENCES

[1] Jeet Kumar, Om Prakash Prabhakar, Navneet Kumar Sahu, "Comparative Analysis of Different Feature Extraction and Classifier Techniques for Speaker Identification Systems: A Review", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE),Vol. 2, Issue 1, pg-2760-2769, January 2014.

[2] Liqin Fu, Xia Mao, Lijiang Chen "Speaker Independent Emotion Recognition Based on SVM/HMMs Fusion System" IEEE International Conference on Audio, Language and Image Processing(ICALIP), pages 61-65, 7-9 July 2008

[3] Peipei Shen, Zhou Changjun, Xiong Chen," Automatic Speech Emotion Recognition Using Support Vector Machine" IEEE International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT) volume2 , Page(s) : 621 - 625 , 12-14 Aug. 2011.

[4] Akalpita Das, Purnendu Acharjee , Laba Kr. Thakuria , " A brief study on speech emotion recognition" , International Journal of Scientific & Engineering Research(IJSER), Volume 5, Issue 1,pg-339-343, January-2014.

[5] Vinay, Shilpi Gupta, Anu Mehra,"Gender Specific Emotion Recognition Through Speech Signals", IEEE International Conference on Signal Processing and Integrated Networks (SPIN), 2014 , Page(s):727 – 733, 20-21 Feb. 2014.

[6] Norhaslinda Kamaruddin, Abdul wahab Rahman,Nor Sakinah Abdullah,"Speech emotion identification analysis based on different spectral feature extraction methods", IEEE Information and Communication Technology for The Muslim World, 2014 The 5th International Conference, Pages:1-5, 2014.

[7] A. D. Dileep, C. Chandra Sekhar, "GMM Based Intermediate Matching Kernel for Classification of Varying Length Patterns of Long Duration Speech Using Support Vector Machines", IEEE Transactions on Neural Networks and Learning Systems, Volume: 25, Issue: 8, Pages: 1421 - 1432, 2014.

[8] Wootaek Lim, Daeyoung Jang and Taejin Lee, Speech emotion recognition using convolutional and Recurrent Neural Networks, Published in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific

[9] Pavitra Patel et al, "Speech Emotion Recognition System Using Gaussian Mixture Model and Improvement proposed via Boosted GMM", IRA-International Journal of Technology & Engineering ISSN 2455-4480.

[10] Dahake, Prajakta P., Kailash Shaw, and P. Malathi. "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine." Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on. IEEE, 2016.

[11] Ritu D.Shah, Dr. Anil.C.Suthar, "Speech Emotion Recognition Based on SVM Using MATLAB" International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Vol. 4, Issue 3, pg-2916-2921, March 2016.

[12] Kunxia Wang, Jing Yang, Ning An, Lian Li ,"Harmony search for feature selection in speech emotion recognition", 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 978-1-4799-9953-8/15©2015 IEEE

[13] Rahul. B. Lanjewar, D. S. Chaudhari. "Speech Emotion Recognition: A Review", International

Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-4, March 2013

[14] E. Ayadi, E. et al; (2011), "Survey on speech emotion recognition features, classification schemes and databases" Pattern Recognition, 44, 572–587.

[15] Ververidis, D., & Koropoulos, C. (2006), "Emotional speech recognition: Resources, features, and methods" Speech Communication, 48, 1162–1181

[16] Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection, Journal of Neural Networks 2005. special issue: Emotion and Brain 18(4), 407–422 (2005)

[17] H. Teager, S. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract", in Speech Production and Speech Modeling, NATO Advanced Study Institute, vol. 55, Bonas, France, (Boston: Kluwer Academic Pub.), pp. 241-261, 1990.

[18] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In Interspeech 2014, 2014