Study and analysis of speech emotion recognition system: A review

[1]Maske Vijay, [2]V V Yerigeri
[1]M.Tech Student,[2] Professor,
Department of PG

MBESs College of Engineering, Ambajogai, MS, India

Abstract: The importance of automatically recognizing emotions from human speech has grown with the increasing role of spoken language interfaces in human-computer interaction applications. Research in understanding and modeling human emotions, a topic that has been predominantly dealt with in psychology and linguistics is increasingly attracting attention within the engineering community. A major motivation comes from the desire to develop human machine interfaces that are more adaptive and responsive to a user's behavior. Capturing the emotion from only speech is a difficult challenge. The proper selection of features in relation with both the time and frequency domains together is necessary to produce optimized results. In this review paper, we concentrate on recognizing four emotional states: Happy, Sad, Anger and Neutral from speech. For this we explore features like Energy, Pitch, Zero-crossing rate and Mel-Frequency Cepstrum Coefficients (MFCC). The performance of classification algorithms, such as Support vector machine (SVMs), Decision Tree, Linear Discriminate Analysis (LDA) and etc are compared for recognizing emotions from speech in different applications.

Index Terms **- Human-Computer Interaction (HCI), Support vector machine (SVMs) Classifier, Emotional states**

I. INTRODUCTION

Affective interaction is the high-level phase of human computer interaction (HCI). Rosalind Picard coined the term Affective Computing to describe a newly established field dealing with the automatic sensing, recognition and synthesis of human emotions from any biological modality such as speech or facial expressions. It is an interdisciplinary field widely involving computer science, psychology, and cognitive science [1-4]. Emotion recognition is a common instinct for human beings, which has been studied by researchers from different disciplines for more than 70 years.

The ability to recognize, interpret and express emotions commonly referred to as emotional intelligence and is the back bone in human communication. Computers with affect recognition and expression skills will allow a more natural and thus improved human-computer interaction. An affect recognizing computer can learn during an interaction by associating emotional expressions with its own behavior [5, 6]. There is a continuous interaction among emotions, behavior and thoughts, in such a way that they constantly influence each other. Research by psychologists and neuroscientists has shown that emotions are closely related to decision making and thus emotion plays a significant role in the rational actions of human beings. Researching emotions, however, is extremely challenging in several respects [8].

One of the main difficulties results from the fact that it is difficult to define what emotion means in a precise way [9]. Understanding emotions is essential in human social interactions. Studies suggest that only a 10% of human life is completely unemotional and the rest is effectively colored with emotions. Although having been studied since the 1950's, the investigation of emotional cues has made considerable advances in the recent years. This is mainly due to the new application developments with respect to human machine, human-robot interfaces and multimedia retrieval applications [10-12]. Interface design between humans and computers is attaining promising interest because of the goal of recognizing and understanding emotions automatically. Automatic emotion recognition will provide more natural interaction between humans and computers. As computers and computer based applications are deeply involved in our everyday life, it is very much important to made computers and personal robots with affect recognition capabilities.

The interactive process among human beings is ensuring convergence and optimization because of our ability to infer the emotional states of others based on their emotional states. This allows human beings to adjust our responses and behavioural patterns accordingly, thus ensuring the communication meaningful and effective [14]. The performance by a computer, and the emotional categories it can cover, are far limited compared with those capabilities by humans. One main difficulty comes from the fact that there is a lack of complete understanding of emotions in human minds, including a lack of agreement among psychological researchers [13-18].

The paper is organized as follows. Section II describes the Background of HCI, and its applications with different technologies used, Section III discusses about the Speech Emotion Recognition with different emotions classifiers. Section IV concludes the paper and gives detail on current Research.

II. SPEECH EMOTION RECOGNITION

The motivation of Speech Emotion Recognition (SER) is to provide more flexibility in human computer interaction, and make more intelligent machines. The interest behind this study is also beneficial to better speech, speaker recognition systems and more responsible human robot interactions. Our SER system should be able to recognize the different emotional patterns in an efficient and faster way. Since the emotion recognition is a hard problem for humans, the task would not be easy for machines [1-2].

Enormous advances have been made in speech emotion recognition research. Speaker independent speech emotion recognition is a sub domain in speech recognition and is a subject of intensive research since 1980. In the case of speaker independent and language independent databases, the problem becomes more complicated. Performance of SER depends on the number of classes used for recognition. Many research results are reported for languages like Hindi, Telugu, Marathi, Kannada, Bangle, Assamese, Urdu and Punjabi [2-8]. In Indian languages studies are active in Assamese, Bengali, and Kannada. However in Malayalam not many works were reported. So in depth research works are highly essential in Malayalam language towards speech emotion recognition. Figure 1 illustrates our emotion classification system. In the learning phase, for each utterance, the extracted speech features and the emotion labels are used to train each individual [10, 13]

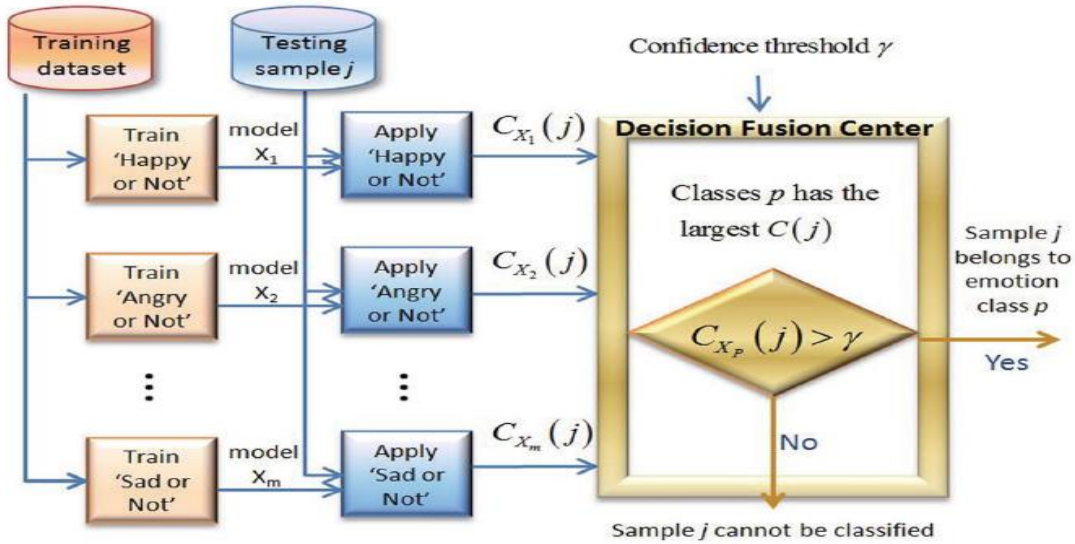

Figure 1: Speech Emotion Recognition

A. Challenges in Speech Emotion Recognition

Emotion does not have a commonly agreed theoretical definition. The task of speech emotion recognition is very challenging for the following reasons [1-3]. First, it is not clear which speech features are most effective in distinguishing between emotions. The acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates adds obstacles because these properties directly affect most of the common extracted speech features. Another challenging issue is that how a certain emotion is expressed generally depends on the speaker, his or her culture and environment. Most works have focused on monolingual emotion classification, making an assumption that there is no cultural difference among speakers. Another problem is that one may undergo a certain emotional state such as sadness for days, weeks, or even months. In such a case, other emotions will be transient and will not last for more than a few minutes. As a consequence, it is not clear which emotion the automatic emotion recognizer will detect: the long-term emotion or the transient one. However, people know emotions when they feel them. For this reason, researchers were able to study and define different aspects of emotions [3, 6].

B. Speech Emotion Classifiers

Classifiers or Learning classifier systems, or LCS, are a paradigm of rule-based machine learning methods that combine a discovery component (e.g. typically a genetic algorithm) with a learning component (performing either supervised learning, reinforcement learning, or unsupervised learning).For this system, the classifier algorithms based on Supervised learning are used. The following describes about the classifiers used:

Support Vector Machine (SVM): A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples [16,18].

To define an optimal hyperplane let us consider the problem: For a linearly separable set of 2D-points which belong to one of two classes, find a separating straight line. In the above picture we can see that there exist multiple lines that offer a solution to the problem. But, we need to know which line is the best line among them to separate. A line is bad if it passes too close to the points because it will be noise sensitive and it will not generalize correctly.

Therefore, we should find the line passing as far as possible from all points. Therefore, the operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. Twice, this distance receives the important name of margin within SVMs theory. Hence, the optimal separating hyperplane is the line that maximizes the margin of the training data [1].

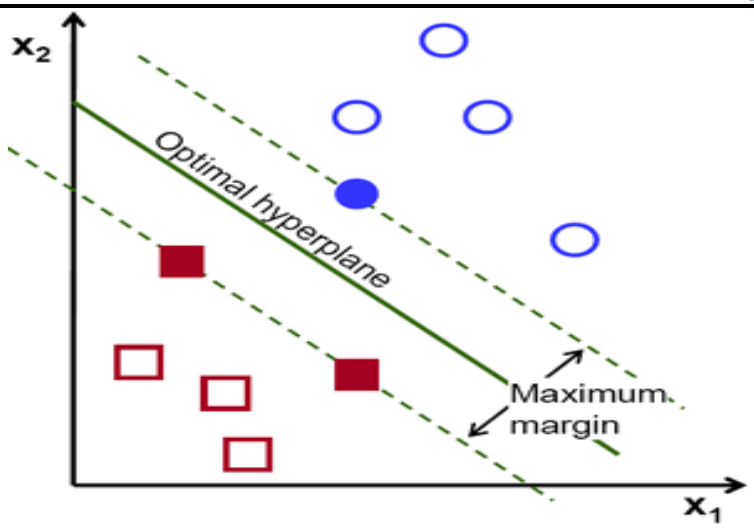

Figure 2: SVM Hyperplane.

Decision Tree: Decision trees are classification trees that divide the feature space into several regions, and in each region, if a category of samples is dominant; they are marked with the category labels. A decision tree is a tree whose internal nodes are tests and whose leaf nodes are categories. Each internal node is responsible for testing one attribute and each branch from the node selects one value for the attribute. The leaf node predicts a specific class. The decision trees are not limited to boolean functions, but multiple categories.

Linear Discriminate Analysis: Linear discriminate analysis (LDA) is a generalization of Fisher's linear discriminate, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. LDA takes multi-dimensional data, makes use of prior class information (Supervised Learning) and represents the data in a form which maximizes the distance between different classes.

Multi Class Classification: Multi classification is the problem of classifying into three or more classes unlike two classes in binary classification. In one v/s one strategy the samples from a pair of classes from the original training set are taken and are distinguished. At prediction time voting is applied and the one that gets the highest number of votes will be predicted by the combined classifier. Authors have used one v/s one strategy in all our classifiers as all of them deal with multi class classification.

C. Case study: Avoidance of school violence by speech emotion detection system

What kind of emotions pupils have during the violence events is an important issue for emotion detection research. A research focused on emotional reaction of school violence victims has carried on 6282 Maltese school children between 9 and 14 years of age. The results showed that most pupil victims felt angry, vengeful, helpless, and self-pity, and about 24% of the victims felt indifferent [7, 8]. Another research presented a survey on violence victims in three countries (England, Italy, and Spain) revealed that the victims are angry (42.7%), upset (34.8%), stressed (22.4%), worried (24.3%), afraid (18.1%), alone (14.3%), defenseless (14.3%), and depressed (18.9%) [9]. both the researches showed that the victims have negative emotions during the violence events.

Generally, the term emotion describes the subjective feelings in short periods of time which are related to events, persons, or objects [10, 11]. Since the emotional state of human is a highly subjective experience, it is hard to find objective and universal definitions. This is the reason there are different approaches to model emotions in the psychological literature. One approach is the definition of discrete emotion classes, the so-called basic emotions. Ekman defined seven emotions which humans are very familiar with: happiness, sadness, anger, anxiety, boredom, disgust, and neutral [12]. These seven emotions are considered as the basic emotions, and more emotions can be defined by mixtures of the basic emotions [13]. According to this theory, the emotions, which the school violence victims have in above two researches, can also be defined by mixtures of the basic emotions. Analyzing these emotions generated by school violence victims, the negative emotions consist of three basic emotions: anger, sadness, and fear. So the detection of these three basic emotions may indicate that violence events happened. Combined with movement and other bio-signal detection, the accuracy of violence detection might be improved [17].

III. CONCLUSION

In this review, we discussed the Speech emotion recognition system and challenges to implement for HCI based applications. Many researches are still in a developing stage, because of the complexity in defining and modeling emotions. Different groups of researchers are working hard with different approaches to find out a complete speech emotion recognition system. The ultimate aim of speech emotion recognition research is to create machines and user interfaces that can interact with the user with an emotional effectiveness. Therefore we can conclude that the performance of decision tree classifier was best with the chosen polish dataset for speech emotion recognition. The performance of SVM was better than the LDA classifier. But however the SVM classifier could give better results if the size of the dataset would increase.

REFERENCES

[1] D. Olweus, School bullying: development and some important challenges. Annu. Rev. Clin. Psychol. 9, 751–780 (2013).

[2] E. Menesini, C. Salmivalli, Bullying in schools: the state of knowledge and effective interventions. Psychol. Health Med. 22, 240–253 (2017).

[3] J. Wang, R.J. Iannotti, T.R. Nansel, School bullying among adolescents in the United States: physical, verbal, relational, and cyber. J. Adolesc. Health 45(4), 368–375 (2009).

[4] T. Vaillancourt, R. Faris, F. Mishna, Cyberbullying in children and youth: implications for health and clinical practice. Can. J. Psychiatry 62(6), 368–373 (2017).

[5] A. Foteini, H. Dimitris, K. Anderson Adam, ECG pattern analysis for emotion detection. IEEE Trans. Affect. Comput. 3(1), 102–115 (2012).

[6] L. Ye, H. Ferdinando, T. Seppanen, An Instance-Based Physical Violence Detection Algorithm for School Bullying Prevention (11th IEEE International Wireless Communications and Mobile Computing Conference, Dubrovnik, 2015), pp. 24–25

[7] M.G. Borg, The extent and nature of bullying among primary and secondary schoolchildren. Educ. Res. 41(2), 137–153 (1999).

[8] L.R. Barhight, J.A. Hubbard, C.T. Hyde, Children's physiological and emotional reactions to witnessing bullying predict bystander intervention. Child Dev. 84(1), 375–380 (2013).

[9] M. Giménez Gualdo Ana, C. Hunter Simon, D. Kevin, The emotional impact of cyberbullying: differences in perceptions and experiences as a function of role. Comput. Educ. 82, 228–235 (2015).

[10] M.A. Quiros-Ramirez, T. Onisawa, Considering cross-cultural context in the automatic recognition of emotions. Int. J. Mach. Learn. Cybern. 6(1), 119– 127 (2015).

[11] M. Waseem, M. Ryan, C.B. Foster, Assessment and management of bullied children in the emergency department. Pediatr. Emerg. Care 29(3), 389–398 (2013).

[12] H. Saarimaki, A. Gotsopoulos, I.P. Jaaskelainen, Discrete neural signatures of basic emotions. Cereb. Cortex 26(6), 2563–2573 (2016).

[13] H. Schlosberg, Three dimensions of emotion. Psychol. Rev. 61(2), 81–88 (195.

[14] X.D. Wu, V. Kumar, J.R. Quinlan, Top 10 algorithms in data mining. Knowl. Inf. Syst. 14(1), 1–37 (2008).

[15] A. Keskinarkaus, S. Huttunen, A. Siipo, MORE - a multimodal observation and analysis system for social interaction research. Multimed. Tools Appl. 75(1), 6321–6345 (2016).

[16] Hong-feng Ma, Jian-wu Dang, Xin Liu "Research of the Optimal Wavelet Selection on Entropy Function" Lecture Notes in Electrical Engineering Volume 173, PP. 35-42, 2012

[17] Humberto Perez-Espinosa, Carlos A. Reyes-Garcia, Luis Villase nor-Pineda "Acoustic Feature Selection and Classification of Emotions in Speech using a 3D Continuous Emotion Model" Biomedical Signal Processing and Control 7 PP. 79– 87, 2012

[18] Krishna Kishore K.V and Krishna Satish P. "Emotion Recognition in Speech using MFCC and Wavelet Features "Advance Computing Conference (IACC), IEEE 3rd International 2013, PP. 842 – 847, 2013