Speech emotion Recognition using Neural Networks

CH. Deepika, P.Swetha

Abstract: Emotion recognition is a procedure to identify the human emotion, the identification criteria may be facial expression analysis or may be verbal expression. Emotion plays a vital role in all aspects of cognitive learning processes. Identification of emotion from human expressions is a trending research topic in human computer interaction (HCI). Speech emotion recognition is one the area which can be used to identify the emotions from verbal expression of human. Speech Emotion recognition also become a main research topic in human computer interaction studies. In recent times, the attention of researchers was increased to study the emotional content of speech and verbal expressions. Implementation of Speech Emotion Recognition may involve several learning models, classification methods, feature extraction and pattern recognition. We reviewed many numbers of research articles, major challengesand applications of speech emotion recognition. At present many emotional speech databases and recognition applications are developed for research and development purpose. The results, limitations and performance of current speech emotion recognition system is based on different classifiers are discussed.

Keywords: Speech emotion recognition, Emotional speech databases, emotion classification, feature extraction

I. INTRODUCTION

SER is the most widely growing research topic in the field of deep learning, pattern recognition and HCI.SER identifies human discrete categories such as happy, sad, fear and disgust from speech signals. Speech signal is the most natural way to communicate among human beings. Speech signal combines linguistics(explicit) information like intention of speaker, along with paralinguistic(implicit) information like emotional aspect of speech. Linguistic information identifies qualitative patterns that the speaker has uttered, while paralinguistic information is usually measured by quantitative features describing variations in the way that the linguistic patterns (i.e. words or phrases) are pronounced [2].Most of the recent research field concentrate on three critical aspects of SER, namely (1) databases, (2) speech features, and (3) classification methods to improve the performance accuracy of SER systems. The different applications where speech emotion recognition isusedin education, entertainment, medical diagnosis, callcentresmultimedia retrieval etc.

In this paper, we present a survey of speech emotion recognition systems.Survey consists of threemain points in speech emotion recognition: (1) design criteria of emotional speech dataset, (2) the effect of speech features on the classification performance of speech emotion recognition, and (3) classification methods used in speech emotion recognition.Survey have different types of features and considered the benefits of combining the available acoustic information with other sources of information such as linguistic, paralinguistic and audio information. In this study wecovered, classificationmethods used in speech emotion recognition. However, the person who reads should understand the recognition rates of those systems carefully since different emotional speech corpora and experimental setups were used with each of them [1].

II. SPEECH EMOTIONAL DATABASES

A key point must be considered in the estimation of an emotional speech recognizer is the degree of naturalness of the database used to assess its performance. Moreover, the criticaldatabase design ismore important to the classification task being considered. For example, the emotions being classified may be infant-directed; e.g. soothing and prohibition [3,6], or adult-directed; e.g. joy and anger [4,5].The classification task is also defined by the number and type of emotions included in the database.

A suitable choice of speech databaseplays a main role inside the discipline of affect detection. A context rich emotional speech database is desired for a great emotion reputation system. Specifically, three forms of databases are used for growing a speech system [11], [12] they're

1)Natural speech database: This form of speech database is a part of real-world records. This type of data is very helpful for real international emotion status and completelynatural.

2)Elicited emotional speech database: This kind of corpora is gathered from speaker by using and creating artificial emotional situation. The advantage of these kind of database is that it is very close to the natural database, butstill there are a few issues additionally. All emotions might not be available and if the speaker is aware of that they're being recorded, then the emotion expressed by him can be artificial.

3)Actor based speech database: This form of speech corpora is collected from professional and trained artists. By collectingthese kinds of datasetsare very smooth, and a huge kind of emotion are available. The main difficulty of this type of database are v irregular in nature and it is very much artificial in nature.

In literature survey some speech corpora have been used by different authors for detecting the emotions from different databases is listed in TableI.From table, it may be observed that, there is a huge difference among the corpora,

Revised Version Manuscript Received on 16 September, 2019.

***** Correspondence Author
CH.Deepika *, Assistant Professor ,Vidya Jyothi Institute of Technology Aziz Nagar,Hyderabad
P.Swetha, Assistant Professor,VJIT,Aziz Nagar, Hyd

in terms of language, number of emotions, number of subjects, purpose and methods of corporacollection [17].The corpora include real and natural emotional speech spoken popularity and for emotion popularityin Indian context [11].Most of the speech corpora contains different emotion like anger, sad, neutral and happy. Some of the speech with the aid of a large quantity of maleand female persons in different languages.There aredifferent speech corpora for speaker corpora is used to capture the motions and based on thatemotions are detected

Table I
speech corpora used for emotion recognition

Corpus	Language	Size	Emotions
FAU AIBO(AEC)	German, English and French	21 male, 30 females	Anger, Emphatic, Neutral, Positive, fear, happiness, and neutral.
ABC	German	4 male,4 female	Neutral, tired, aggressive, cheerful, intoxicated, and nervous.
EMO	German	5 male,5 female	Anger, disgust, fear, happiness, neutral, and sadness.
GeWEC	French	2 female,2 male	Angry, fear, happiness, and neutral.
Remote COLlaborative and Affective (RECOLA)	French, Italian, German and Portuguese	10 female,6 male	Arousal and Valence
IEMOCAP	Motion Capture	5 males, 5 females	Anger,Happiness,Neutral,Sadness
EmoDB	German	10(5 female and 5male)	Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise
eNTERFACE	English	43 subjects	Anger, Disgust, Fear, Happy, Sadness, and Surprise
XiaoIce	Mandarin	17048 sentences	Happy, angry, sad, and neutral.

III.Features and classifications

The literature survey presents the features and classification techniques used by the authors in different papers. Features are used to construct the speech emotion recognition system through speech signals.The useful information regarding speech is carried out by the features. The collection and designing the features is the main challenging problemin speech emotion recognition.The testing and training vector are needed for the classification for speech verification development and identification.In this study most of the authors used Mel frequency cepstral coefficient for extracting the feature. The properties of MFCCCs are robust, dynamic method for speech feature extraction and power spectrum is computed by performing Fourier analysis. The spectral and cepstral features are Mel frequency cepstral coefficient (MFCC),linear prediction cepstral coefficient (LPCC) and acoustic features like pitch,energy, formants are also used in speech emotion recognition. Various types of classifiers have been proposed for the task of speech emotion recognition.

The classification methods, feature extraction and challenges of speech emotion recognition system are given in below table.

In survey majority of authors have used deep neutral networks as classifier to improve the performance of speech emotion recognition as shown in Table2.In machine learning the deep neural network is an emerging technologyin recent years. Main characteristic of DNNs is that they can learn high-level invariant features from raw data [18, 19], which is helpful for emotion recognition. A very recent studies utilized DNNs for speech emotion recognition.The types of DNN are convolutional neural networks and recurrent neural networks were applied in majority of the literature survey.The applications of CNN and DNN are facial expression and hand movements during scripted and spontaneous spoken communication scenarios, motion capture, emotion recognition using speech signals etc.Convolutional neural network model operates on the raw signal, to perform an emotion prediction task from speech dataset. The main purpose of CNN is to extract the features from speech and image. The RNN that often achieves the state-of-the-art performance in speech emotion recognition [8].This study presented different CNN and RNN architecture to extract the emotions from different speech databases

Table 2
Speech emotion recognition features and classifiers

S.no	Ref.	Features	Classifier	Description
1	Yue Xie, Ruiyu Liang, Fangmei Zhu, Li Zhao, Jie Wang, Guichen Tang	Zero Crossing Rate, MFCCs,	Recurrent neural networks	The performance of this algorithm is morestable than previous features and achieves 82.5% of accuracy
2	Panagiotis Tzirakis, Jiehao Zhang, Bjorn W. Schuller	MFCCs	Recurrent neural networks	Theconcordance correlation coefficient attained state-of-the-art resultsfor both arousal and valence.The performance has improved compared to the previously used state-of-the-art results . Arousal=.787 Valence=.4
3	Vladimir Chernykh, Pavel Prikhodko	Energy-based features, Mel-Frequency Cepstral Coefficients (MFCC)	Recurrent neural networks	The main benefits of this novel algorithmaudio based onConnectionist Temporal classification approach is that even the emotional utterance might contain parts where there are no emotions and it can guess the sequence of emotions for one utterance
4	Michael Neumann, Ngoc Thang Vu	MFCCS,Log Melfilter-bank, prosody features	Convolutional Neural Networks	Extensive experiments have conducted using an attentive convolutional neural network with multi-view learning objective function.
5.	Srinivas Parthasarathy, Ivan Tashev	Mel-spectrum, Energy	Deep neural networks, convolutional neural networks	With imbalanced datasets the training and evaluationhave used UWA and WA as a metric, which brought good outcome.The result ofUWA and WA is 116.5
6.	Zhichao Peng1, 2, Zhi Zhu1, Masashi Unoki1, Jianwu Dang1, 2, Masato Akagi1	Sampling frequency, Modulation filter bank, sampling frequency	three-dimensional convolution recurrent neural networks, spectral-temporal representation	Based on the experimental results and the method used in this paper is themost effective way to design an emotion recognitionsystem by representing the human auditory system.The results of this method areWA as 61.98 % and UA as 60.93%
7.	Po-Wei Hsiao and Chia-Ping Chen	MFCC, root-mean-square energy, zero-crossing rate,	Deep recurrent neural network	Within the dynamicmodelling framework, the best UA recallrate has ever achieved on the dataset used in this paper and the UA recall rate is46.3%.
8.	Jing Han, Zixing Zhang, Fabien Ringeval, Björn Schuller	LLD	bidirectional long short-termmemory, Reconstruction error based	Based on the correlation betweenthe Reconstruction Error and the performance improvement, itindicates that the RE informationhas a positive impact on the model and significantly overcomes other state-of-the-artmethods.
9.	JunDeng, Xinzhou Xu, Zixing Zhang, Sascha Fr¨uhholz, andBjörn Schuller	LLD, ZCR, RMS, HNR, MFCC	Supervised learning methods	The proposed method focuses on semi supervised autoencoders and it has been evaluated on five databases which has improved the recognition performance of the unweighted Average Recall (UAR).

III. RESULTS

In literature survey the authors have used different neural network algorithm, based on that the performance of different methods have been plottedin the graph. In [10] the author applied recurrent neural networks and achieved 82.5% of accuracy. In [15] the author achieved the recall rate as 46.5% by applying the classifier deep recurrentneural network. In [14] the author used three-dimensional convolution recurrent neural networks and achieved the 60.93% and 61.98%.

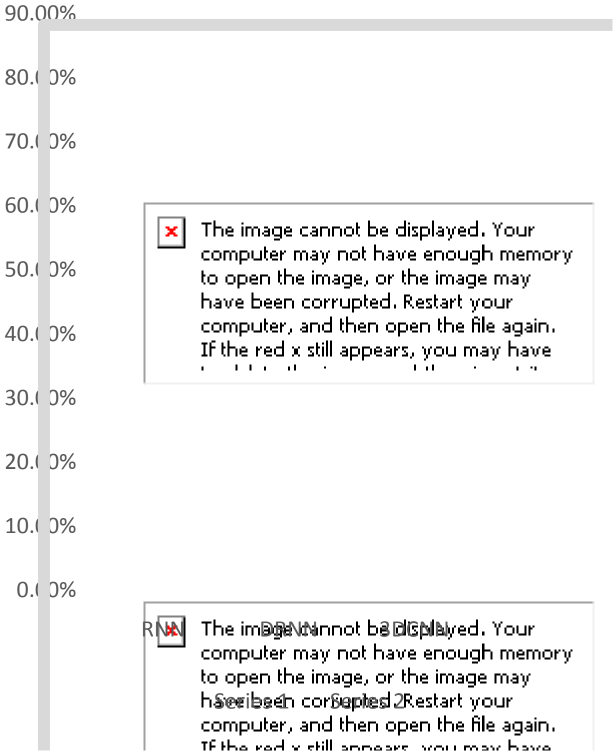

Fig.1. Results obtained for different algorithms

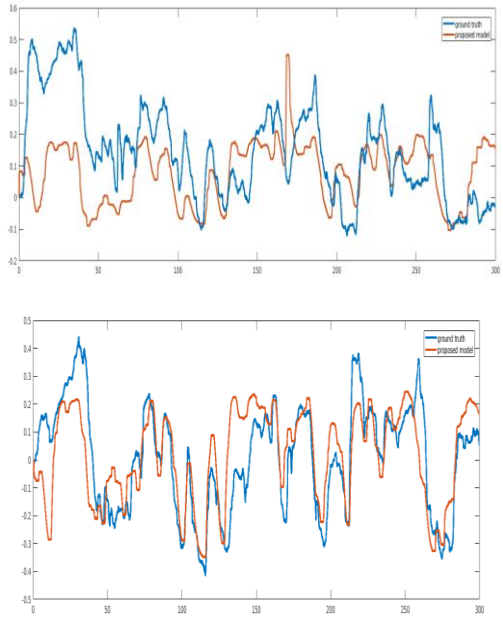

Fig. .2. Results obtained for a test subject for the arousal (Top)and valence (Bottom) dimensions

IV. CONCLUSION

In this paper, experiments done on speech emotion recognition. To improve the accuracy of speech emotion recognition process, the differentclassification methodshave been used. Also, by extracting more effective features of speech, accuracy of the speech emotion recognition system can be improved. The different speech corpora used in this survey is illustrated. The important issues in speech emotion recognition system are the signal processing unit in which appropriate features are extracted from available speech signal and another is a classifier which recognizes emotions from the speech signal..

REFERENCES

1. Moataz ElAyadi a, , MohamedS. Kamel b, FakhriKarray: Survey on speech emotion recognition: Features, classification schemes, and databases ,Volume 44, Issue 3, 2011, Pages 572-587
2. Christos-Nikolaos ,Anagnostopoulos , Theodoros Iliou , Ioannis Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, 2015, Volume 43,Issue 2, pp 155–177
3. C. Breazeal, L. Aryananda, Recognition of affective communicative intent in robot-directed speech, Autonomous Robots , 2002, Volume 12, Issue 1, pp 83–104
4. W. Campbell, Databases of emotional speech, in: Proceedings of the ISCA (International Speech Communication and Association) ITRW on Speech and Emotion, 2000, pp. 34–38
5.I.Engberg,A.Hansen,DocumentationoftheDanishemotionalspeech database des /http://cpk.auc.dk/tb/speech/Emotions/S, 1996
6. M. Slaney, G. McRoberts, Babyears: a recognition system for affective vocalizations, Speech Commun. 39 (2003) pp 367–384
7. Jun Deng, Xinzhou Xu, Zixing Zhang, Member, IEEE, Sascha Fr¨uhholz, and Bj¨orn Schuller, Senior Member, IEEE:Semi-Supervised Autoencoders for Speech Emotion Recognition,2018,Volume 26, Issue 1
8. Panagiotis Tzirakis, Jiehao Zhang, Bjorn W. Schuller,End-to End speech Emotion Recognition Using Deep Neural Network ,2018,
9.Vladimir Chernykh, Pavel Prikhodko, Emotion Recognition from Speech with Recurrent Neural Networks,2017
10.Yue Xie, Ruiyu Liang, Fangmei Zhu, Li Zhao, Jie Wang, Guichen Tang,Long-short term memory for emotional recognition with variable length speech ,2018
11.S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: A review. *International Journal of Speech Technology*, 2012.
12. M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011

AUTHORS PROFILE

CH.Deepika
Assistant Professor
Vidya Jyothi Institute of Technology
Aziz Nagar,Hyd.

P.Swetha
Assistant Professor,VJIT
Aziz Nagar, Hyd.