ISSN 2278-3091

Volume 9 No.2, March -April 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse54922020.pdf

https://doi.org/10.30534/ijatcse/2020/54922020

Speech Emotion recognition feature Extraction and Classification

Chandupatla Deepika[1], Swarna Kuchibhotla[2]

[1]Koneru Lakshmaiah Education Foundation,India,chandu.deepu9@gmail.com

[2]Koneru Lakshmaiah Education Foundation,India, drkswarna@kluniversity.in

ABSTRACT

Emotion recognition from speech is one of the most popular topics in human computer interaction (HCI). Emotions can be identified by facial expression analysis or by verbal expressions. Several researchers build systems to understand various emotions in human expression. Human emotions automatically recognized by machines for improving human machine interaction. In this article, we have identified three basics of speech emotion recognition system 1) databases 2) feature extraction and 3) various methods of classification. Discussed about the performance of speech emotion recognition system. Features are classified as Essential, Prosodic and Spectral characteristics. Different classifying techniques are used to classify different emotions from human speech like Hidden Markov Model (HMM), K-Nearest Neighbour (KNN),Gaussian Mixtures Model (GMM),Support Vector Machine (SVM) and deep learning.

Key words:Speech emotion reognition, features, classifiers

1.INTRODUCTION

Speaking emotion is one of the most common research fields. Several researchers worldwide work on different technology for language processing. Too many problems for researchers are added by speech processing, such as continuous recognition of speech, emotional intelligence, etc. Emotions are understood not only by facial expressions but also by words. Every individual's discourse is linked to an emotion. Emotions are very important because they allow a person to comprehend feelings. Speech reveals the individuals feelings to be happy,neutral,angry,sad, etc. Speech is a complex signal which contains message information, speaker, speech, emotion, etc. The emotional presence makes it more natural to speak. In a conversation significant information like the purpose of the speaker is given during non-verbal communication. The way the words are spoken also provides the essential unrelated information in addition to the message that is conveyed by text. The same text message would be conveyed by combining acceptable emotions with different semantism (meaning). Spoken text could be interpreted in several ways. In English, for example, the word ' Ok ' expresses admiration, disbelief, consent, disinterest or a claim. To interpret the semantines of a spoken speech is therefore not sufficient to understand the text alone. It is important however that language systems can process unrelated data like emotions. People can understand the information related to phonetic by using multimodal indicators and by perceiving the underlying emotions. Non language data may be detected by (1) video facial expressions; (2) voice emotional expression; and (3) Written text punctuation; The discussion in this paper is confined to speech-related expressions or emotions. The main goal of processing emotions are (a) comprehension of emotions in speech (b) desired emotions can be synthesized in speech according to the message intended. From a computer point of view, speech emotions can be seen as a category or psychological bias.

2.LITERATUE REVIEW

Björn Schuller et al. Present the new benchmark achieved through the fusion of predictions from participants and conclude by discussing 10 recent and emerging trends in speech and language analysis. Humberto Perez-Espinosa et al. In this paper we present the IESCChild, which is a guide that can contribute to research the interaction of affective reactions without speech when children communicate with their computers and to the creation of models for the perception of acoustic data in paralinguistic terms. Ocquaye Elias Nii Noi et al. To assess our approach, we use FAU AEC as the target dataset of the Emotion Challenge 2009 INTERSPEECH and as a source datasets ABC and Emo-DB. The proposed method is superior to other state-of - the-art approaches based on experimental results.Nan Song et al. We propose a sign language approach, which incorporates facial expression recognition, in this article. emotional Mandarin and Tibetan speech conversion. First, the recognized sign languages are converted into context-dependent labeling and corresponding emotional tags for sign language and facial expression information. In addition, an HMM-based emotional language-synthesis is being equipped by a positive and emotional corpus. Eventually, emotional speech synthesis are rendered according to the emotional tags and context-related labels of the sign language text in order to turn the sign language into emotional Mandarin-Tibetan language. Po-Wei Hsiao et al. On the FAU-Aibo tasks as described in Inter talk 2009 Emotion Challenge, the proposed recognition model is being assessed. Our deep-recurring, base line version of the neural network hits average (UA) recall of 37.0 per cent unweighted, equivalent to the official HMB modeling framework. The proposed implementation of the attention system in addition to the base deep RNN model results in a recall rate of 46.3% for UA. To the best of our knowledge, the AU recall level for complex modeling tasks FAU-Aibo has ever achieved. Vishnu Vidyadhara Raju V et al. This study is conducted using the EMO-DB server in Berlin. Compared to existing state of the art emotional recognition systems, the proposed approach resulted in a better solution. This paper shows an accuracy of 80.83%. Jun Deng et al. Evaluate thoroughly the proposed Emotion Challenge database model and four other databases in various scenarios in the INTERSPEECH 2009 model. Experimental results show the model proposed achieves state-of- the art performance with a very small number of labelled data on the challenge of the task and other tasks. Yue Xie, et al. The loss of frame-level functions from the perspective of information theory is less than fixed-length, and more appropriate to the input of deep learning with the ability to learn itself. Applied as a classifyer, the Bidirectional long-term memory (BiLS) system the variable feature length. Experimental results show that the proposed method exceeds substantially the CASIA repository features of the INTERSPEECH 2010. Ying Chen et al. The DSTL system proposed is developed by integrating the common domain and species data. The main contribution of our paper compared with the state of the art is that not only do our commonalities but the specific information often takes advantage of the DSTL system. Experiments on three corporations for emotional speech check whether our approach is elective. Common domain and species perform better results than those that use only common domain features. Stefano Rovetta et al. Based on fuzzy clustering, including probabilistic, possibility possibilistic c-means ,a novel based approach for emotion recognition from speech signal is prensented in this paper.Using K-means, fuzzy c-means look matted for this issue compared to crisp clustering, to analyze emotions conveyed by speech using membership degrees potentially o_er an innovative way.Dara Pir, et al. We analyze CGI-FS ' performance using four different classifiers and assess the pertinence of group characteristics to the work. The Licability SubChallenge baseline on heart-EAT design information is all four CGI-FS device results that deliver the best quality, achieving a relative unwounded recall improvement of 9.8 percent compared with it. Nattapong Kurpukdee et al. We have proposed a technique for speaking emotional recognition with convolutionary long-term recurrent neural network memory in this article. In the proposed technique, phonemic emotional probabilities at the frame level are extracted from ConvLSTM-RNN crude input speech signals and converted into numerical input characteristics. We then used SVM's or LDA classifiers to classify the emotional state on the basis of the pronunciation. Experimental results in the IEMOCAP database showed that the technique proposed could improve speech emotion recognition performance. Zhichao Peng et al. The experimental results indicate that speech emotional characteristics can effectively be used to use the proposed deep education approach, which is based on the auditive filterbank Gammatone. Shashidhar G et al. This study proposes the classification model that best fits a particular feature set. Unlike existing methods or meta-learning such numerical transactions on features are carried out and the classification system is recommended based on the results obtained. The research concludes that the reliability of the classification also depends on the chosen data set. GMM gives better precision in this respect if the data falls in the normal area of distribution. Statistically, emotional information are found to be mostly in the same area. Dr. Yogesh Kumar,et al. Adding a deep neural network shows the improvement in emotion recognition system by making automatic emotion recognition system. The analysis has also been performed using different ML techniques for Speech emotion recognition accuracy in different languages.Mumtaz Begum Mustafa et al. This paper discusses and analyzes the latest SER research and found that it has been an important field of research in the last 12 years and that it will certainly remain a subject of future research. The current research focused on a number of key concerns, including databases, classification, real time identification, and cross-language SER that have not yet been fully resolved and are most likely to be a priority in future for SER research.. Tian Han et al. The results were also compared with the adult database. Results showed that the voices of children and adults in speech emotional recognition research should be treated differently. The accuracy of the child database is lower than that of the adult database; other signals in the system will enhance the accuracy of violence detection. Wei Jiang et al. In this article, we suggested an architecture for speech emotional recognition to solve heterogeneous problems in acoustics that generally deteriorate the quality of classification. The proposed deep hybrid neuronal network consists mainly of an extraction module and a heterogeneous unification unit

3.FEATURE EXTRACTION

In recent research on language emotional recognition, the combination of the various features has been emphasized to improve the performance of recognition. In the previous paragraphs the sources, systems and prosodic features addressed represent largely mutually exclusive data on speech signal. Such features are therefore mutually compatible. The planned performance of the system will be increased by an insightful combination of complementary features. Various studies on feature combinations have demonstrated better classifications of emotions compared to systems that use individual features. Human speech is made up of many parameters that demonstrate the emotions it contains. These criteria are also changed as feelings are modified. Therefore, to identify emotions, you need to select the proper feature vector. Features, spectral features, and prosodic properties are categorized as a source of excitation. The source of excitation is achieved by deleting vocal tract (VT) characteristics. Spectral properties used to classify emotions are linear correlation coefficients (LPCs) coefficients (PLPCs), mel-frequency curve coefficients (MFCCs), linear cepstrum coefficients (LPCCs), and linear perceptual prediction (PLP). MFCC, LFPC, LPC, PLP, and RASTA-PLP[9, 10] will help to achieve the exactness of differentiating different emotions. Pitch, strength, and frequency are prosodic features used for emotion recognition. Statistical measurement is also applied to distinguish between emotions such as minimum, maximum, standard, range, average, median, variance, skewness, kurtosis, etc.

Table 1: Speech emotion recognition features

S. No	Features	Approach and Purpose	Ref.
1	Speech prosody,Combination of spectral energy, articulator activities.	Approximately 75 percent of typical emotional perception is archived on 4 emotions as an anger graded, positive and sad in English, and anger-glad and good have identical acoustic properties.	Yildirim et al.
2	Pitch-related characteristics of LPCC combination	Eight feelings category. A maximum of 50 male and 50 female native speakers record 100 phonetically balanced words. The artificial neural network is reporter with around 50% of the typical speaker-independent emotional category.	Nakatsu et al.
3	Pitch,energy and speech rate characteristics	The rage of identification fears positive, negative, and unhappy feelings represented by thirty non-professionals. Average 70 percent emotion score is obtained by artificial neural networks.	Petrushin
4	Functions with spectral, prosodium, and HMM	Emotional speech corpus description of five INTERSPEECH-2009 emotions. The average recorded quality of the emotion category is about 63%.	Bozkurt et al.
5	39 prosodic and spectral properties	Shorter expressions have better emotions ,characterization of 15 discreet emotions. More emotional information is provided in specific words in longer speeches	Tisecher et al

4.CLASSIFIER

After extracting speech features, it is important to select a correct classifier for the recognition of speech emotions. Emotions are classified by classifiers. Diverse classifications are used, including the Hidden Markov Model (HMM), Gussian

Mischofs (GMM), Support Vector Machine (SVM), Artificial Neural network (ANN). Identification instruments can also be combined and used for the development of a hybrid model. Generally, pattern classifier can be divided into two specific categories for the identification of speech emotion.

1.Non- Linear classifiers
2. Linear classifiers

4.1SVM: It is a classification algorithm used in problems with pattern detection. This classification is used to separate characteristics from other characteristics. Data is classified by constructing a hyperplane that is N-dimensional and gives greater operational range, i.e. a large gap between the nearest data points.

4.2 **HMM:** Also called as first-order chain is the Hidden Markov Model. The internal work is in this respect concealed from the observer. The temporal data structure is collected using HMM. By training with the extracted features, HMM differentiates between different emotions. This model is the result of the input given to the SVM.

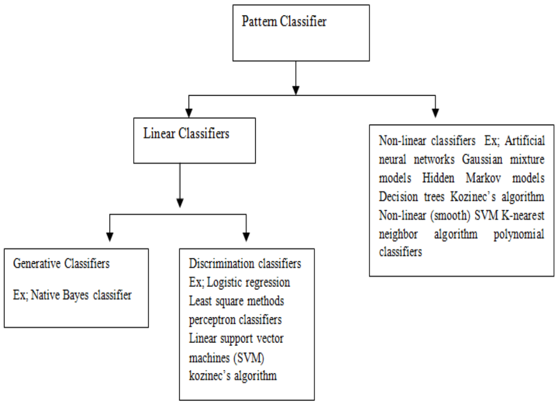
Pattern Classifier
Linear Classifiers
Non-linear classifiers Ex; Artificial neural networks Gaussian mixture models Hidden Markov models Decision trees Kozinec's algorithm Non-linear (smooth) SVM K-nearest neighbor algorithm polynomial classifiers
Generative Classifiers
Ex; Native Bayes classifier
Discrimination classifiers
Ex; Logistic regression
Least square methods
perceptron classifiers
Linear support vector machines (SVM)
kozinec's algorithm

Figure 1: Different types of classifiers

The use of classifiers depends mainly on the data's existence. If the essence of the data is previously known, it would be easier to decide the kind of classifier. If they are straightforward, linear classifying better and faster classifies features. Supervised learning would be beneficial if the training dataset was appropriately marked. Feature vectors which cannot be linearly separated will require non-linear grading. The essence of the data is seldom understood in most real-world situations. Scientists, therefore, use non-linear classification methods on an often complex and time-consuming basis. Nonetheless, if you choose the emotional recognition pattern classifiers, based on the nature of speech functions a systematic approach is needed. The different natures of the characteristics (excitation source, vocal and prosodic system) would contribute to deciding the correct classification. The systemic study is useful and appreciable in this respect as it saves a large number of computer resources.

5.RESULTS

Authors have used different classification algorithm, based on the accuracy graph is plotted. Vishnu Vidyadhara Raju V et al improved the accuracy of 80.83% by applying Support vector machine on Berlin emotion speech (EMO-DB) database.Po-Wei Hsiao and Chia-Ping Chen applied the LSTM-attention model on data set FAU-Aibo in German and achieved the recall rate as 46.5% .Zhichao Peng1 et al achieved the 60.93% and 61.98% by using 3DCNN on IEMOCAP dataset (figure 2).

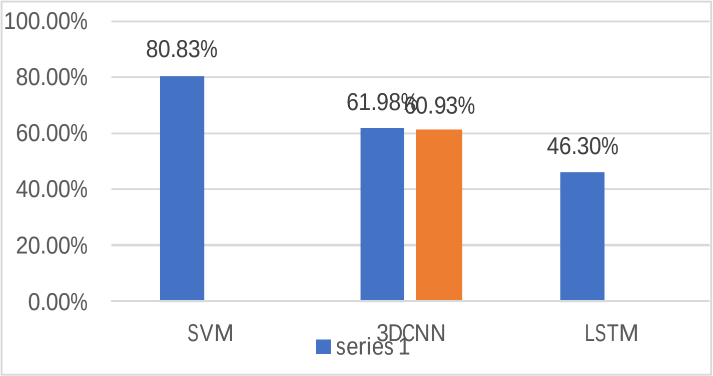

Figure 2: Accuracy for different classifiers

6.CONCLUSION

We also discussed the fundamentals of the speech recognition process in this article. We have also given a short overview of the production of speech emotions to ensure that current speech systems function naturally. In the recent past, substantial work has been done in this area. The absence of information and standardization is a common phenomenon, and many works overlap. In fact, in the Indian sense there are not many comprehensive papers on language emotion identification, has not been published since today. We, therefore, thought it possible to spark the research community by reviewing recent work on language emotion recognition in order to fill some primary research gaps. This paper explores recent work in understanding speech emotions from the perspective of emotional repositories, vocabulary, and classification models. The paper also discusses several major research issues regarding the recognition of language emotion.

REFERENCES

1. Björn Schuller , Stefan Steidl, Paralinguistics in speech and language—State-of-the-art and the challenge, Computer Speech and Language 27 (2013) 4–39.
https://doi.org/10.1016/j.csl.2012.02.005

2. HumbertoPerez-Espinosa, Juan Martınez-Miranda, ESC Child: An Interactive Emotional Children's Speech Corpus, Computer Speech&Language59(2020)55 _74.
https://doi.org/10.1016/j.csl.2019.06.006

3. Ocquaye Elias Nii Noi, Mao Qirong, COUPLED UNSUPERVISED DEEP CONVOLUTIONAL DOMAIN ADAPTATION FOR SPEECH EMOTION RECOGNITION, 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM).

4. Panagiotis Tzirakis1, Jiehao Zhang, END-TO-END SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORKS, 2018 IEEE.

5. Nan Song, Hongwu Yang, A Gesture-to-Emotional Speech Conversion by Combining Gesture Recognition and Facial Expression Recognition, 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia).

6. Po-Wei Hsiao and Chia-Ping Chen, EFFECTIVE ATTENTION MECHANISM IN DYNAMIC MODELS FOR SPEECH EMOTION RECOGNITION, ICASSP 2018.
https://doi.org/10.1109/ICASSP.2018.8461431

7. Vishnu Vidyadhara Raju , Krishna Gurugubelli, Differenced Prosody Features from Normal and Stressed Regions for Emotion Recognition, 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN).

8. Jun Deng, Xinzhou Xu, Semi-Supervised Autoencoders for Speech Emotion Recognition, 29-9290 (c) 2017 IEEE.

9. Yue Xie, Ruiyu Liang, Long-short term memory for emotional recognition with variable length speech, 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia).

10.Ying Chen, Zhongzhe Xiao, DSTL: Solution to Limitation of Small Corpus in Speech Emotion Recognition, Journal of Arti_cial Intelligence Research 66 (2019) 381-410.

11. Stefano Rovetta and Zied Mnasri, Emotion recognition from speech signal using fuzzy clustering, 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019).
https://doi.org/10.2991/eusflat-19.2019.19

12. Dara Pir, Cascaded Acoustic Group and Individual Feature Selection for Recognition of Food Likability, (ICPRAM 8th International Conference on Pattern Recognition Applications and Methods), pages 881-886.

13. Nattapong Kurpukdee, Tomoki Koriyama, Speech Emotion Recognition using Convolutional Long Short-Term Memory Neural Network and Support Vector Machines, Proceedings of APSIPA Annual Summit and Conference 2017.

14. Zhichao Peng, Zhi Zhu, Speech Emotion Recognition Using Multichannel Parallel Convolutional Recurrent Neural Networks based on Gammatone Auditory Filterbank, Proceedings of APSIPA Annual Summit and Conference 2017.

15. Shashidhar G. Koolagudi, Choice of a classifier, based on properties of a dataset: case study speech emotion recognition, International Journal of Speech Technology (2018) 21:167–183.
https://doi.org/10.1007/s10772-018-9495-8

16. Dr. Yogesh Kumar, Dr. Manish Mahajan, Machine Learning Based Speech Emotions Recognition System, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 07, JULY 2019.

17. Mumtaz Begum Mustafa· Mansoor A. M. Speech emotion recognition research: an analysis of research focus, International Journal of Speech Technology (2018) 21:137–156.
https://doi.org/10.1007/s10772-018-9493-x

18. Tian Han, Jincheng Zhang, Emotion recognition and school violence detection from children speech, EURASIP Journal on Wireless Communications and Networking (2018) 2018:235.

19. Wei Jiang , Zheng Wang, Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network, Sensors 2019, 19, 2730.

20. Mohit Shah, Ming Tu, Articulation constrained learning with application to speech emotion recognition, *EURASIP Journal on Audio, Speech, andMusic Processing* (2019) 2019:14.

21. Linhui Sun · Jia Chen, Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition, International Journal of Speech Technology (2018) 21:931–940.
https://doi.org/10.1007/s10772-018-9551-4

22. Zhichao Peng1, 2, Zhi Zhu1, Masashi Unoki1, Jianwu Dang1, 2, Masato Akagi1 Auditory-inspired end-to-end speech emotion recognition using 3D convolutional recurrent neural network based on spectral temporal representation.2018