A study on different approaches of machine learning on Emotion Recognition from human speech

Priyanka Joshi[1], Dr. Suman Pant[2]
[1]Research Scholar, Department of Computer Application and Information Technology
[2]Professor, Department of Computer Application and Information Technology
[1,2]Shri Guru Ram Rai University, Patel Nagar, Dehradun- 248001, Uttarakhand (India)

ABSTRACT:
This paper present the basic study of different approaches of machine learning towards emotion recognition from human speech. It presents the various approaches used for feature extraction, a set of different features, challenges towards speech recognition. The classifiers used for pattern matching with all its advantages and disadvantages. It is a survey on the existing approaches for recognizing emotions from human speech and accuracy rates achieved using different models and datasets by different authors. It also focuses on the different applications of emotion recognition in different fields like diagnosing psychological disorders, improved human machine interaction and other real life scenarios etc.

KEYWORDS: Speech Emotion recognition, feature extraction, pattern matching, classifiers.

Introduction

Speech is one of the most fundamental and natural way to communicate between humans. Spoken communication between human and machine is also increasing with the development of technology. A speech signal is a complex signal which contains the information about the speaker such as age, gender and also the emotional state of speaker.[1] Different researches are going on to make human – computer interaction more efficient. Therefore Speech Emotion Recognition by machine has been a goal of research since last decades. Speech Emotion Recognition has been used in many areas such as computer games , medical diagnosis, call centers ,automotive engineering, smart phones, household robots etc[2].Emotions play very important role in humans life and also responsible for one's decision making[3]. Though various methods are used for SER, there are several challenges in identifying the correct emotional state[4]. Therefore several researches are working in this area.

In SER various feature extraction methods are used to extract features sets by which find the relevant feature to predict emotions[5] and then pattern matching techniques are used ,also various classifiers are used for recognizing emotions from speech signal.

Speech is more convenient for humans to interact with computer or any kind of machines like robots etc. The idea which generated for making speech recognition system is because it is convenient for humans to interact with a computer, robot or any machine through speech or vocalization rather than difficult instructions. Human beings have long been inspired to create computer that can understand and talk like human. Various feature extraction methods and pattern matching techniques are used to make better quality speech recognition systems. Feature extraction technique and pattern matching techniques plays important role in speech recognition system to maximize the rate of speech emotion recognition[6].

ISSN: 2173-0695

After introduction part, the remaining part of paper is organized as follows: In Section II, we define the Emotion recognition from speech and machine learning concept. In Section III, we describe the methodology for Speech Emotion Recognition with different steps. In Section IV, we define a review of existing models with their results. Finally in the last Section V, we conclude our work with discussion by proposing some direction towards the future action for SER with improved performance.

Emotion Recognition And Machine Learning

A.Speech Emotion Recogniotion

Speech: Speech is one by which human can communicate with each other. The speech is comprised of vowels and consonants. It is also comprised with vocabularies, syntax and the set of speech sounds differ in creating different types of human languages. There is a strong relationship between the quality of voice in speech and the perceived emotions[7].

To recognize emotions on the bases of speech signal, is the current research topic in the field of human-computer interaction and several research has been done during the last decades[8].It has many potential benefits that result from correct identification of subject's emotional state. If human emotions can be correctly recognized by machine, it can improve performance and make human-machine interfaces much friendliness. This also makes monitoring of psychological state of individuals in different environments. It can also be used to automate medical or forensic data analysis systems in medical fields.

The studies observed that the information regarding emotions contained by speech signal is spread over different features. This is because the differences in speaking style and speaking rates of speaker[9].Therefore it is a challenge to identify exact emotion relevant features from speech signal. The main objective of emotional speech processing is, to clearly understand emotions present in speech and to synthesize emotions from the speech which are desired .

If we talk about machine, it can view or understand emotions from speech by classification or differentiation process of emotions. While emotion synthesis, it include the knowledge of emotion synthesis at the time of speech synthesis. The goal of Emotional speech recognition is to identifying the emotional or physical state of a human being from human voice.Emotional aspect of human speech includes some kind of para-linguistic aspects. It is one of the factor in human communication, however the emotion state of human speech does not change the linguistic content[10].

II. Methodology Used for Emotion Recognition from speech

(i)Pre-Processing: The pre-processing of speech signal is important where silence or ambient noise is completely undesirable which comes during the signal capturing[11]. Pre-processing can be done by many techniques such as noise cancelling, pre-emphasis techniques, and dimensionality reduction of speech and voice activity detection. Pre-processing adjust or modifies the speech signal so that it will be more acceptable for feature extraction.

ii) Feature Extraction: Feature extraction means extracting the desirable features to extract emotions of human from their speech. Feature extraction is the process of reducing data while retaining the information that can be used to discriminate the speakers[7]. The features can be of three types, syllable features, prosodic features and spectral features, Syllable Features: Syllable features are mostly associated with the quality of voice[12].

Prosodic Features: The prosodic features deals with the loudness, pitch, energy,the stress, and the rhythm for describing emotional state[13].

Spectral Features: The spectral features are frequency based features. The spectral features of voice signal are obtained by converting the time based signal into the frequency domain using different techniques like Fourier Transform. There are several spectral features in literature such as Mel Frequency Cepstral Coefficient (MFCC) Shifted Delta Cepstral Coefficient (SDCC), spectral centroid,

ISSN: 2173-0695

spectral roll off, spectral flatness, spectral contrast, Linear Prediction Cepstral Coefficients (LPCC), spectral subband centroid, etc[14]. Feature extraction is the process of converting a raw speech signal into a sequence of acoustic feature vectors which carries speaker specific information.MFCC: MFCC is one of the widely used spectral characteristics in emotional Speech recognition that are collection of coefficients that provide information on the shape of the speech signal spectrum[15].

iii) Classifiers: The classifiers are used to map the extracted features to corresponding emotions. A classifier detects the emotion from speech utterances of different speaker's. To perform emotion recognition from speech, various types of machine learning techniques or classifiers can be used to classify emotional states. Some of them are Gaussian Mixture Model(GMM), Hidden Markov model(HMM)[16], K-Nearest Neighbour(KNN)[17],Artificial Neural Network(ANN)[18],Support Vector machine(SVM) and fourier transform[19]etc.

iv) Emotional Speech Databases: In Emotional speech recognition, the databases are the necessary prerequisite for predicting emotions. We need a training set of samples. The different types of databases are designed for the prediction of emotion recognition from speech. These databases are Simulated, Induced and Natural database.

Simulated databases are pre-recorded by well trained and well performed artists, majority of databases are simulates that they use actors to reproduce emotions[20].

Induced databases are another kind of databases which are more naturalistic than simulated databases. An artificial emotional situation is created so that speaker speaks in a natural way.

Natural databases are real time database, contains on the spot recorded speech from real world experience or general public conversation. Databases used by early research for speech emotion used limited number of samples but now the modern databases have large number of samples of wide range of speakers.

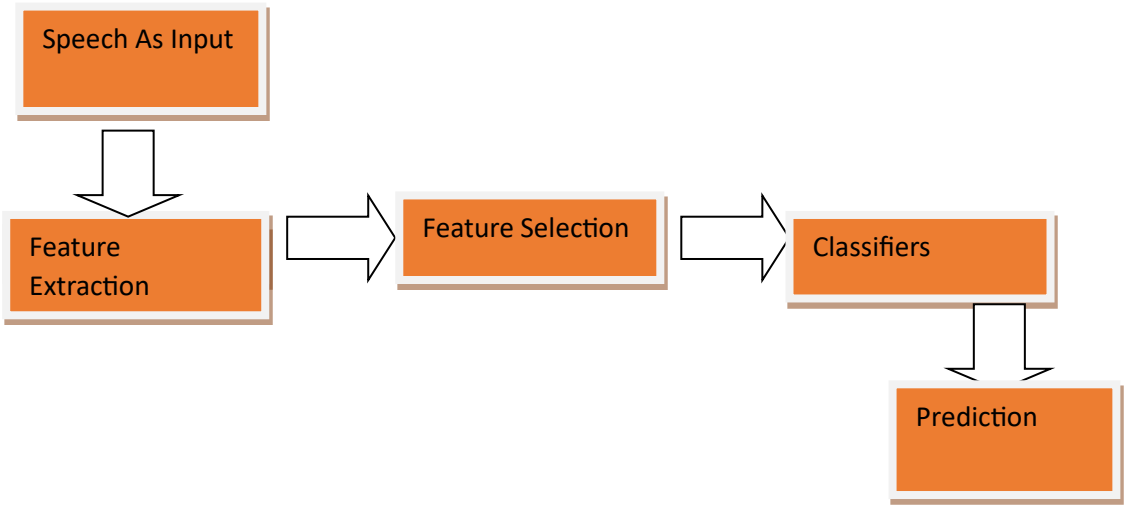

Figure1: Outline work for speech emotion recognition using machine learning

Literature Review

This paper, present a review of the recent advancements in speech emotion recognition research. Feature extraction and classification methodologies using deep learning have provided improved performance for audio signal processing[21]. Several studies are carried out in speech signal processing and pattern recognition algorithms for predicting the emotional state of speaker for different languages[22]. The SER is modelled by different classifiers. This review aims to access and upgrade real-time emotion detection systems to know the latest progress in this technology.

ISSN: 2173-0695

Ref.	Year	Authors	Work	Databases	Results
[23]	2020	M. Farooq , F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. B. Zikria, ‘‘Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network,’’ 2020	Optimized deep learning model consist of CNN is proposed. LSTM layer is used for learning long term correlations in Log-Mel Spectogram.	RAVDESS IEMOCAP SAVEE EMO-DB	Accuracy RAVDESS-81.30% IEMOCAP-83.80% SAVEE-83.80% EMO-DB-82.10%
[24]	2020	Shadi Langari,Hossein, Marvi,Morteza, Zahedi. ”Efficient Speech emotion Recognition using modified feature extraction”(2020)	Proposed a novel feature extraction method based on adaptive time frequency coefficient to improve SER	Berlin Emotional Database - EMO-DB, SAVEE and PDREC	Accuracy EMO-DB 79.57% SAVEE-80% PDREC 91%
[25]	2020	S.Ramesh. S.Gomathi,S.Sashikala,T.R Saravanam ”Automatic Speech detection using hybrid if grey wolf optimizer and naive bayes”	Proposed a new machine learning algorithim ,a hybrid of gray wolf optimizer and naive Bayes for classification. For Feature extraction , MFCC is used	They merged the two databases SAVEE and TESS to create a new database	It provide 15.76% higher accuracy 20.69% increased specification 15.59 % greater sensitivity in comparison of traditional.
[26]	2020	Mustaqeem, M. Sajjad, and S. Kwon, ‘‘Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM,’’	Introduced Redial based function network(RBNF). STFT algorithm is used to generate spectrogram features. CNN model is used Deep-BiLSTM is used for recognition of emotions.	IEMOCAP Emo-DB RAVDESS	Accuracy IEMOCAP-72.25% Emo-DB-85.57% RAVDESS-77.02%
[27]	2021	Stavros Ntalampiras “Speech emotion recognitioin via learning analogies”	Introduce the few-shot learning paradigm. Speech segments are characterized through analogies, designed the Siamese Neural Network (SNN) modeling to find out the differences	EMO-DB	The model did not trained for classification it specify and Learn similar and disslimilar relationship between input

ISSN: 2173-0695

			and similarities between novel and known recorded audio signals.		pairs.
[28]	2022	L. Li, K. Xie, X. Guo, C. Wen, and J. He, ‘‘Emotion recognition from speech with StarGAN and dense-DCNN,’’	Dense-DCNN in combination of StarGAN IS used. StarGAN generate Log-Mel Spectra features and DCNNis used to achieve high precision classiufication.	RAVDESS EMO-DB SAVEE	Accuracy RAVDESS-97.36% EMO-DB-91.06% SAVEE-92.97%
[29]	2022	Biswajit Nayak, Mitali Madhusmita Debendra Ku Sahu . “Speech Emotion Recognition using Different Centred GMM	The classifications were carried out using Gaussian Mixture Model. Mel Frequency Cepstral Coefficients (MFCCs) features are used for identifying the emotions. It can be observed that, when increasing the number of GMM centres then recognition performance increases.	IITKGP-SEHSC	Accuracy - 32 centered GMM-75.01% 16 centered GMM - 71.46% 8 centered GMM-66.81% (training and testing data were 80% and20% respectively). GMM model when increase the number of centres then recognition performance increases.
[30]		A Hyeon Jo and Keun-Chang Kwak. “Speech Emotion Recognition based on Two stream Deep Learning Model Using Korean audio Information”	Two stream based model. Bi-directional-LSTM and YAMNet(CNN based model) are used.	Korean speech emotion recognition database	Bi-LSTM 90.38% YAM-Net 94.91%
[31]	2023	Meera Mohan,P.Dhanlakshmi,R.satheesh kumar”Speech emotion Classification using ensemble model with MFCC”(2023)	Proposed an ensemble model for Speech emotion recognition,Spectra l features havebeen extracted using MFCC.Emotion Classification based on 2D CNN and extreme	RAVDESS	Accuracy-Machine learning Random Forest- 61% XG boost classifiers-68%

ISSN: 2173-0695

			Gradient boosting (XG-Boost)		
[32]	2023	Kishor Bhangale and Mohanaprasad Kothandaraman "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network"	Work presents acoustic feature set based on MFCC,LPCC,WPT . DCNN model is used.	EMO-DB RAVDESS	EMO-DB-93.31% RAVDESS-94.18%

Conclusion:

Emotion recognition performance can be improved by extracting high number of features from speech and by using appropriate dimension reduction techniques. It is also necessary to use a suitable classifier along with dimension reduction techniques for building a most suitable SER model. Meta-feature subsets fusion gives different representation of emotion clusters, and can contribute in improving emotion recognition performance. Identifying human emotions from speech by machine with good performance is still challenging task.Different speech emotional databases are contributing towards the improvement of emotion recognition performance but a difference is observed among different databases which creates difficulty and also ambiguity in emotion classification. The speech signals contain a spectrum of emotions, therefore there is a need to give more attention towards feature extraction methods.

References

1. B. Nayak, M. Madhusmita, D. K. Sahu, R. K. Behera, and K. Shaw, "Speaker Dependent Emotion Recognition from Speech," *Int. J. Innov. Technol. Explor. Eng.*, vol. 3, no. 6, pp. 40–42, 2013.
2. X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," *Proc. - Int. Conf. Artif. Intell. Comput. Intell. AICI 2010*, vol. 1, pp. 537–541, 2010, doi: 10.1109/AICI.2010.118.
3. J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Annu. Rev. Psychol.*, vol. 66, no. September 2014, pp. 799–823, 2015, doi: 10.1146/annurev-psych-010213-115043.
4. A. Al-Talabani, H. Sellahewa, and S. A. Jassim, "Emotion recognition from speech: tools and challenges," *Mob. Multimedia/Image Process. Secur. Appl. 2015*, vol. 9497, no. April, p. 94970N, 2015, doi: 10.1117/12.2191623.
5. T. Vogt and E. Andr, "COMPARING FEATURE SETS FOR ACTED AND SPONTANEOUS SPEECH IN VIEW OF Augsburg University , Germany Multimedia concepts and applications Applied Computer Science," 2005.
6. S. Ding, H. Zhu, W. Jia, and C. Su, "A survey on feature extraction for pattern recognition," *Artif. Intell. Rev.*, vol. 37, no. 3, pp. 169–180, 2012, doi: 10.1007/s10462-011-9225-y.
7. C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, no. 1–2, pp. 189–212, 2003, doi: 10.1016/S0167-6393(02)00082-1.
8. B. W. Schuller, "Speech Emotion Recognition two decades in a Nutshell," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018, [Online]. Available: http://delivery.acm.org/10.1145/3130000/3129340/p90-schuller.pdf?ip=128.125.20.233&id=3129340&acc=OPEN&key=B63ACEF81C6334F5.C52804B674E616B8.4D4702B0C3E38B35.6D218144511F3437&__acm__=1526412054_950badaff5c5f1011d85dc05734135c4
9. M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011, doi: 10.1016/j.patcog.2010.09.020.
10. T. Liu and X. Yuan, "Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques," *Eurasip J. Audio, Speech, Music Process.*,

ISSN: 2173-0695

vol. 2023, no. 1, 2023, doi: 10.1186/s13636-023-00290-x.
11. J. Han, Z. Zhang, and F. Ringeval, "RECONSTRUCTION-ERROR-BASED LEARNING FOR CONTINUOUS EMOTION RECOGNITION IN SPEECH Chair of Complex & Intelligent System , University of Passau , Passau , Germany Laboratoire d ' Informatique de Grenoble , Universit ´ e Grenoble Alpes , France Department of," pp. 2367–2371, 2017.
12. E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Commun.*, vol. 46, no. 3–4, pp. 455–472, 2005, doi: 10.1016/j.specom.2005.02.018.
13. M. Pervaiz and T. Ahmed, "Emotion Recognition from Speech using Prosodic and Linguistic Features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 8, 2016, doi: 10.14569/ijacsa.2016.070813.
14. S. M. Tsai, "A robust zero-watermarking algorithm for audio based on LPCC," *ICOT 2013 - 1st Int. Conf. Orange Technol.*, no. 1, pp. 63–66, 2013, doi: 10.1109/ICOT.2013.6521158.
15. M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," *Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017*, vol. 2018-Janua, pp. 2257–2260, 2018, doi: 10.1109/WiSPNET.2017.8300161.
16. B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 1, no. August 2003, pp. I401–I404, 2003, doi: 10.1109/ICME.2003.1220939.
17. M. J. Al Dujaili, A. Ebrahimi-Moghadam, and A. Fatlawi, "Speech emotion recognition based on SVM and KNN classifications fusion," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 2, pp. 1259–1264, 2021, doi: 10.11591/ijece.v11i2.pp1259-1264.
18. D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Syst. Appl.*, vol. 173, no. September 2019, p. 114683, 2021, doi: 10.1016/j.eswa.2021.114683.
19. K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 69–75, 2015, doi: 10.1109/TAFFC.2015.2392101.
20. N. Campbell, "Databases of Emotional Speech," *Proc. ISCA Work. Speech Emot. North. Ireland, Sept. 5-7, 2000*, pp. 114–21, 2000, [Online]. Available: http://www.isd.atr.co.jp/esp
21. H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Y. Chang, and T. Sainath, "Deep Learning for Audio Signal Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 206–219, 2019, doi: 10.1109/JSTSP.2019.2908700.
22. S. T. Alam Monisha and S. Sultana, "A Review of the Advancement in Speech Emotion Recognition for Indo-Aryan and Dravidian Languages," *Adv. Human-Computer Interact.*, vol. 2022, 2022, doi: 10.1155/2022/9602429.
23. M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. Bin Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors (Switzerland)*, vol. 20, no. 21, pp. 1–18, 2020, doi: 10.3390/s20216008.
24. S. Langari, H. Marvi, and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," *Informatics Med. Unlocked*, vol. 20, Jan. 2020, doi: 10.1016/j.imu.2020.100424.
25. S. Ramesh, S. Gomathi, S. Sasikala, and T. R. Saravanan, "Automatic speech emotion detection using hybrid of gray wolf optimizer and naïve Bayes," *Int. J. Speech Technol.*, 2021, doi: 10.1007/s10772-021-09870-8.
26. Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.
27. S. Ntalampiras, "Speech emotion recognition via learning analogies," *Pattern Recognit. Lett.*, vol. 144, pp. 21–26, Apr. 2021, doi: 10.1016/j.patrec.2021.01.018.
28. L. Q. Li, K. Xie, X. L. Guo, C. Wen, and J. B. He, "Emotion recognition from speech with StarGAN and Dense-DCNN," *IET Signal Process.*, vol. 16, no. 1, pp. 62–79, 2022, doi: 10.1049/sil2.12078.
29. N. Kaur, "Speech Emotion Recognition using Different Centred GMM," *Int. J. Adv. Res.*

ISSN: 2173-0695

Comput. Sci. Softw. Eng., vol. 3, no. 10, pp. 646–649, 2013.
30. A. H. Jo and K. C. Kwak, "Speech Emotion Recognition Based on Two-Stream Deep Learning Model Using Korean Audio Information," *Appl. Sci.*, vol. 13, no. 4, 2023, doi: 10.3390/app13042167.
31. M. Mohan, P. Dhanalakshmi, and R. S. Kumar, "Speech Emotion Classification using Ensemble Models with MFCC," *Procedia Comput. Sci.*, vol. 218, pp. 1857–1868, 2023, doi: 10.1016/j.procs.2023.01.163.
32. K. Bhangale and M. Kothandaraman, "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network," *Electron.*, vol. 12, no. 4, 2023, doi: 10.3390/electronics12040839.

ISSN: 2173-0695