Impact Factor (SJIF): 6.092 E-ISSN: 2583-1615

International Journal of Research in Academic World

Received: 15/April/2023 IJRAW: 2023; 2(5):94-97 Accepted: 10/May/2023

Speech Emotion Detection Using Deep Learning

[*1]Dr. Sharayu Deote, [2]Shreya Telang, [3]Sidra Sheikh, [4]Samruddhi Virkhare and [5]Janhvi Meshram

[*1]Professor, Computer Engineering, Cummins College of Engineering for Women, RTMNU, Maharashtra, India.

[2, 3, 4, 5]Student, Computer Engineering, Cummins College of Engineering for Women, RTMNU, Maharashtra, India.

Abstract

The speech emotion recognition is a very exigent assignment of human computer interaction (HCI). This subject has gained so much attention in recent time and will soon achieve a high position for the requirement in coming years. In this strenuous field of speech emotion recognition many techniques have been utilized to extract emotions from signals, including many experimented speech analysis and classification techniques. In the classical way of speech emotion recognition features are extracted from the signals, pitches and frequencies of speech and then the features are selected which is known as selection module and then the emotions are recognized. This is a time consuming process so this paper gives an overview of the modern technique which is based on a simple algorithm based on feature extraction and model creation which recognizes the emotion.

These methods of signal processing and machine learning are widely used to recognize human emotions based on features extracted from facial images, video files or speech signals. Various Experiments were performed to test the accuracy of the classified features extracted from audio files. Results show that random decision forest learning of this hybrid acoustic features is highly effective for speech emotion recognition.

The objective of this research paper is to develop a system which can analyze and predict the expression of the human being. The study proves that this procedure is workable and produces valid results of around 80%.

Keywords: Speech emotion recognition, SER, speech emotion recognition using deep learning

Introduction

Speech Emotion Recognition (SER) is the task of recognizing the emotional aspects of speech irrespective of the linguistic contents. While humans can perform this task effectively as a basic part of speech communication, the ability to conduct it techniquely automated using various devices is an ongoing subject of research. Studies of automatic emotion recognition systems aim to create efficient, real-time methods of detecting the emotions of mobile phone users, call center operators and customers, car drivers, pilots, child behaviour monitoring and many other human-machine communications. Machines have to understand emotions expressed by speech. Only with this approach, an entirely meaningful aspect based on human-machine intervention and understanding can be achieved.

Customary, machine learning (ML) involves the classification of feature framework from the raw data such as speech, images, video, ECG, EEG. The features are used to train a model that learns to produce the desired labeled output In general, it is not known which features can lead to the most efficient clustering of data into different categories. Some probability can be achieved by testing a large number of different features, combining different features into a common feature by applying various feature selection techniques.

An inventive solution in the way is the problem of a desirable feature selection has been given by the use of deep neural networks (DNN) classifiers. The purpose is to use an end-to-end network that takes raw data as an input and generates a class label as an output. There is no need to compute non computational or manual features, as It is all done by the network itself. This came to be very implied solution at the cost of much vast stipulation or demand.

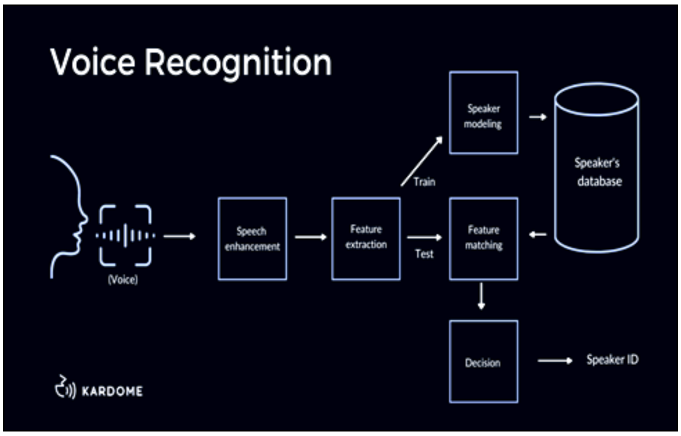

Fig 1: Voice recognition

Databases Used for SER

*Corresponding Author: Dr. Sharayu Deote

Speech emotional databases are used by many researchers in a variety of research activities. The quality of the databases used and the accuracy achieved are the most important factors in the evaluation the model. The databases used may vary depending upon the requirement of the particular model. Some of the databases available can also be described as:

Simulated Database: In these databases, the speech data has been recorded by experienced performers. Among all, this is considered the simplest way to gain the speech-based dataset of various emotions. It is considered nearly about 60% of speech databases are collected by this technique.

Induced Database: In this is type of database the set is collected by creating an artificial emotional situation. This is obtained without the knowledge of the performer. As compared to other databases, this is more naturalistic database. However, an issue arises, because the speaker should be aware that they have been recorded for the purpose of research-based activities.

Natural Database: Most realistic, these databases are difficult to collect due to the difficulty in recognition. Natural emotional speech databases are usually recorded from the general conversation, call center conversations, and so on.

Traditional Techniques of SER

An emotion recognition system based on automated speech is comprised of three fundamental components signal preprocessing feature extraction and classification. Acoustic preprocessing such as noising cancellation as well as dividation is carried out to determine effective classes of this acoustic signal. Feature extraction is required to identify the rare event feature available in the signal.

In this section, a detailed discussion of speech signal processing, feature extraction, and classification is provided. Also, the differences between spontaneous and acted speech are discussed due to their similarity to the topic.

In the first stage of signal processing, speech enhancement is carried out where the noisy components are cancelled. The second stage involves two parts, feature extraction, and feature selection. The required features are extracted from the preprocessed speech signal and the selection is carried out from the extracted features. Feature extraction and selection are usually based on the analysis of speech signals in the time and frequency domains. During the third stage, various classifiers such as GMM and HMM, etc. are carried out for the classification of these features. Lastly, based on feature classification different emotions are recognized and analyzed.

Speech Emotion Recognition System

Speech emotion recognition is similar to the pattern recognition system. This shows that the stages that are present in the recognition system are also present in the Speech emotion recognition system. The speech emotion recognition system contains five main modules emotional speech input, feature extraction, feature selection, classification, and recognized emotional output.

The analysis of the speech emotion recognition system is based on the level of originality of the database which is used as an input to the speech emotion recognition system. If the inferior database is used as an input to the system then incorrect conclusion may result. The database as an input to the speech emotion recognition system may contain the real emotions. It is more practical to use database that is collected from the real life situations and not acted on.

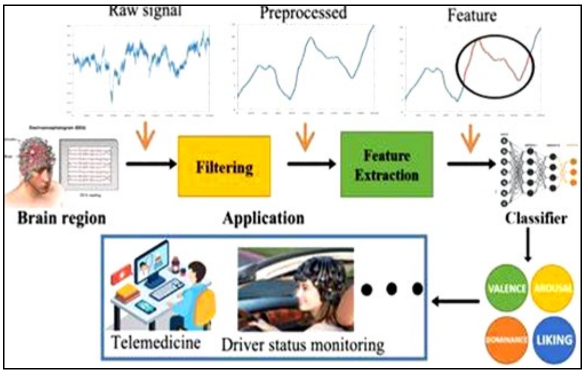

Fig 2: Speech emotion recognition using ECG signals

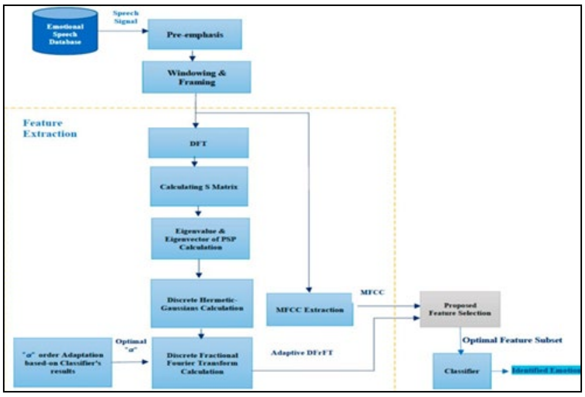
Emotional Speech Database
Speech Signal
Pre-emphasis
Windowing & Framing
Feature Extraction
DFT
Calculating S Matrix
Eigenvalue & Eigenvector of PSP Calculation
Discrete Hermetic-Gaussians Calculation
MFCC Extraction
MFCC
Proposed Feature Selection
Optimal Feature Subset
"a" order Adaptation based-on Classifier's results
Optimal "a"
Discrete Fractional Fourier Transform Calculation
Adaptive DFrFT
Classifier
Identified Emotion

Fig 3: Speech emotion recognition using MFCC

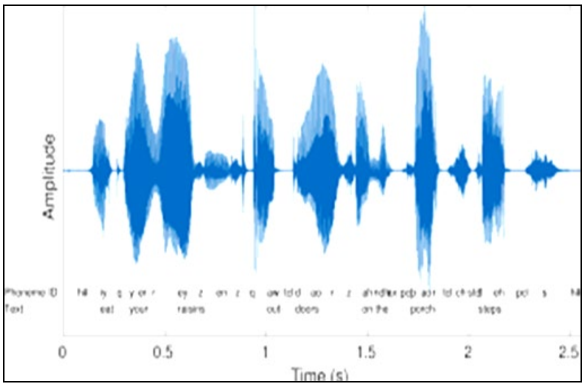

Fig 4: Audio Signal Processing

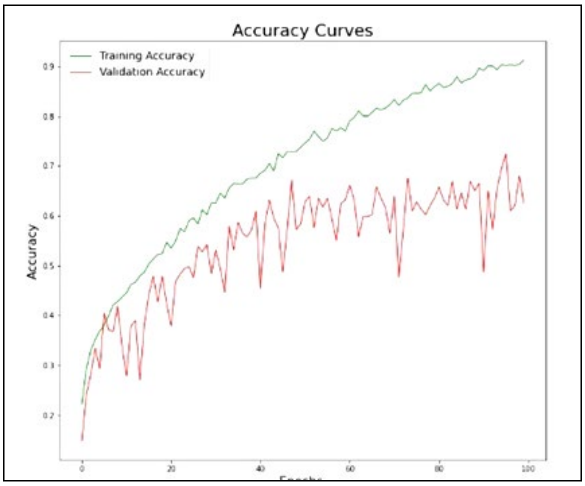

Fig 5: Accuracy Curve of SER

Conclusion and Result

Since a wide range of parameters are very drawn to human emotions, the automatic recognition of emotion is still a subject of active research. The objective of this work is to evaluate and validate a variety of acoustic features based on prosodic and spectral variables for enhancing speech emotion identification. The auditory characteristics have made it very simple to distinguish between the many types of human emotions. This study used a group of measures and insignificant features to achieve useful categorization outcomes. All eight of the emotions seen in this study—including the elusive feeling of fear—can be accurately identified using the suggested collection of auditory cues. Because integrating specific metric and spectral features raises the total viewpoint power of the features and improves classification accuracy, the current methodology appears to be effective. The idea that the same learning algorithms were applied to various feature sets in order to examine how well each collection of features performed against one another draws attention to this priority. Furthermore, we have observed that ensemble learning techniques worked well. Results from the experiments reveal that it was difficult to achieve high accuracy levels when identifying various emotions, whether using pure MFCC features or a mixture of MFCC, ZCR, energy, and fundamental frequency characteristics that were successful in identifying the surprise emotion. However, we demonstrated through extensive experimental that the suggested hybrid acoustic characteristics learned using random choice forest ensemble learning were effective for recognizing speech emotions.

References

1. Ayadi ME, Kamel MS, Karray F, “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases”, Pattern Recognition. 2011; 44:572-587.
2. Chiriacescu I. “Automatic Emotion Analysis Based On Speech”, M.Sc. Thesis Delft University of Technology, 2009.
3. Vogt T, Andre E, Wagner J. “Automatic Recognition of Emotions from Speech: A review of the literature and recommendations for practical realization”, LNCS. 2008; 4868:75-91.
4. Emerich S, Lupu E, Apatean A. “Emotions Recognitions by Speech and Facial Expressions Analysis”, 17th European Signal Processing Conference, 2009.
5. Nogueiras A, Moreno A, Bonafonte A, Jose B. Marino, “Speech Emotion Recognition Using Hidden Markov Model”, Eurospeech, 2001.
6. Shen P, Changjun Z, Chen X. “Automatic Speech Emotion Recognition Using Support Vector Machine”, International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.
7. Ververidis D, Kotropoulos C. "Emotional Speech Recognition: Resources, Features and Methods", Elsevier Speech communication. 2006; 48(9):1162-1181.
8. Ciota Z. “Feature Extraction of Spoken Dialogs for Emotion Detection”, ICSP, 2006.
9. Bozkurt E, Erzin E, Erdem CE, Tanju Erdem A. “Formant Position Based Weighted Spectral Features for Emotion Recognition”, Science Direct Speech Communication, 2011.
10. Jiang H, Hu B, Liu Z, Yan L, Wang T, Liu F, Kang H, Li X. Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Commun,* 2017.
11. Hess U, Thibault P. Darwin and emotion expression. *Am. Psychol,* 2009.
12. Palo HK, Chandra M, Mohanty MN. Emotion recognition using MLP and GMM for Oriya language. *Int. J. Comput*. Vis. Robot, 2017.
13. Stuhlsatz A, Meyer C, Eyben F, Zielke T, Meier G, Schuller B. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011, 22-27.
14. Prasada Rao K, Chandra Sekhara Rao, M.V.P.; Hemanth Chowdary, N. An integrated approach to emotion recognition and gender classification. *J. Vis. Commun. Image Represent,* 2019.
15. Bhavan A, Chauhan P, Shah RR. Bagged support vector machines for emotion recognition from speech. *Knowl. Based Syst,* 2019.
16. Ibrahim NJ, Idris MYI, Yakub M, Yusoff ZM, Rahman NNA, Dien MI. Robust feature extraction based on spectral and prosodic features for classical Arabic accents recognition. *Malaysian J Comput. Sci.* 2019, 46-72.
17. Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. *J Pers. Soc. Psychol.* 1996; 70:614-636.
18. McEnnis D, McKay C, Fujinaga I, Depalle P, J Audio. A feature extraction library. In Proceedings of the International Conference on Music Information Retrieval, London, UK, 2005, 600-603.
19. Hellbernd N, Sammler D. Prosody conveys speaker’s intentions: Acoustic cues for speech act perception. *J Mem. Lang.* 2016; 88:70-86.
20. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* 2011; 44:572-587
21. Guidi A, Gentili C, Scilingo EP, Vanello N. Analysis of speech features and personality traits. *Biomed. Signal Process. Control.* 2019; 51:1-7.
22. Pervaiz M, Ahmed T. Emotion recognition from speech using prosodic and linguistic features. *Int. J Adv. Comput. Sci. Appl.* 2016; 7:84-90.
23. Chen L, Mao X, Xue Y, Cheng LL. Speech emotion recognition: Features and classification models. *Digit. Signal Process.* 2012; 22:1154-1160.
24. Ernawan F, Abu NA, Suryana N. Spectrum analysis of speech recognition via discrete Tchebichef transform. In Proceedings of the International Conference on Graphic and Image Processing (ICGIP 2011), Cairo, Egypt, 1-3 October 2011, 8285-82856L.
25. James AP. Heart rate monitoring using human speech spectral features. *Hum. Cent. Comput. Inf. Sci.* 2015, 5:1-12.
26. Kajarekar S, Malayath N, Hermansky H. Analysis of sources of variability in speech. In Proceedings of the Sixth European Conference on Speech Communication and Technology, Budapest, Hungary, 1999, 5-9.
27. Pachet F, Roy P. Analytical features: A knowledge-based approach to audio feature generation. *EURASIP J. Audio Speech Music. Process.* 2009, 153017.
28. Turgut, Ö. The acoustic cues of fear: Investigation of acoustic parameters of speech containing fear. *Arch. Acoust.* 2018; 43:245-251.
29. Thakur S, Adetiba E, Olusgbara OO, Millham R. Experimentation using short-term spectral features for secure mobile internet voting authentication. *Math. Probl. Eng.* 2015, 564904.

30. Sagi O, Rokach L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 2018, 8:e1249.
31. Kotsiantis SB, Pintelas PE. Combining bagging and boosting. *Int. J Comput. Intell.* 2004; 1:324-333.
32. De Almeida R, Goh YM, Monfared R, Steiner MTA West A. An ensemble based on neural networks with random weights for online data stream regression. *Soft Comput.* 2019, 1-21.
33. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM ensembles in breast cancer prediction. *PLoS ONE*. 2017; 12:e0161501.
34. Xing HJ, Liu WT. Robust Ada Boost based ensemble of one-class support vector machines. *Inf. Fusion* 2020; 55:45-58.
35. Navarro CF, Perez C. A Color–texture pattern classification using global-local feature extraction, an SVM classifier with bagging ensemble post-processing. *Appl. Sci.* 2019; 9:3130.
36. Wu Y, Ke Y, Chen Z, Liang S, Zhao H, Hong H. Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. *Catena*. 2020; 187:104-396.
37. Zvarevashe K, Olugbara OO. Gender voice recognition using random forest recursive feature elimination with gradient boosting machines. In Proceedings of the 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 6-7, 2018, 1-6.
38. Yaman E, Subasi A. Comparison of bagging and boosting ensemble machine learning methods for automated EMG signal classification. *BioMed Res. Int.* 2019, 9152506.
39. Friedman J. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 2001; 29:1189.
40. Dong X, Yu Z, Cao W, Shi Y, Ma Q. A survey on ensemble learning. *Front. Comput. Sci.* 2020; 14:241-258.