WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech Emotion Recognition: An Updated Survey

Mrs Sophia[1], Komal Kumari[2], Nancy Choudhary[3], Nimesh Dash[4]

[1]Assistant Professor, Department of Computer Science and Engineering , MVJ College of Engineering, Bangalore, Karnataka, India.
[2,3,4]Undergraduate Scholar, Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore, Karnataka, India.

A B S T R A C T

Emotions are an essential element of human communication and relationships, as they can greatly impact a person's overall happiness and well-being. With the integration of speech emotion recognition (SER) modules into human-computer interaction (HCI) applications, speech signals can now be used as a means of electronic communication between humans and machines, thanks to various technological advancements.

Keywords:speech recognition, emotion recognition, deep learning, recurrent neural network (RNN), residual neural network (ResNet), mel-frequency cepstral coefficients (MFCC), feature classification, zero crossing rate (ZCR), and support vector machine (SVM).

1. Introduction

Words are the most effective means of expressing our thoughts. Depending on our state of mind, we may convey a range of emotions, such as anger, sadness, happiness, fear, and excitement. By precisely identifying and analysing emotional cues in speech signals, Speech Emotion Recognition (SER) technology has been created to help computers understand human feelings and enhance human-machine involvement.By employing SER, natural interaction between people and computers is made possible. Moreover, SER has broad applications in various fields, such as text-to-speech synthesis, medical diagnosis, multimedia content administration, the automotive industry, and education.

For instance, in real-time emergency call centers, SER is utilized to determine if a caller reporting an abnormal situation is stressed or afraid. This helps to increase the likelihood that the caller is telling the truth, enabling call centers to take the most appropriate action.

People can communicate and express emotions through various means, such as posture, gestures, speech, and facial expressions. Among these, speech signals are the most efficient and natural form of communication. Speech is a sophisticated signal that carries information about the speaker's gender, language, feelings, and message. The vocal tract system is stimulated by a time-varying source to produce speech. However, identifying emotions from speech signals can be challenging due to variations in speaking patterns and speed across individuals and locations.

2. Literature Review

2.1. Excitation Features of Speech for Speaker-Specific Emotion Detection

As This paper aims to detect emotion from speech in context to speaker-specific. In this author analyse deviations between emotional speech and neutral speech by using excitation features of voiced speech as parameters to generate an automated emotion detection system.

The excitation features consist of instantaneous fundamental frequency, the strength of excitation and the energy of excitation. For extracting these above features, two signal processing methods are used. The zero-frequency filtering (ZFF) method is used to extract instantaneous fundamental frequency and the strength of excitation and linear prediction analysis (LPA) is used to extract the energy of excitation. These features are all calculated around glottal closure instants (GCIs) that the ZFF method obtained. The proposed model uses two databases of speech. First one is the IIIT-H Telugu emotional speech database and the second one is the Berlin emotional speech database[1].

The algorithm used in proposed system is Kullback-Leibler (KL) where distance is calculated to evaluate the similarity between feature distributions of

two types of speech. Three 2-dimensional feature spaces are created from the three extracted excitation features by merging two of the features at once.Despite being speaker-specific, the emotion detection system can be made speaker-independent by initialising the system with a large number of neutral sounds and by selecting the proper emotion detection strategy. The primary benefit of the proposed system is that emotional speech is not required to train it.

2.2. Speech Emotion Recognition for Performance Interaction.

In this paper, it explores the potential of machine-driven Speech Emotion Recognition (SER) to enhance theatrical interactions and performances (such as adjusting stage lighting and colour, encouraging audience participation, actors' interactive training, etc.). The proposed system primary objective is to improve user interaction through the use of speech emotion analysis. The microphones fitted on the stage to capture the voice of a performer.The emotion analysis system, located on a computing environment, is used to analyze audio signals. From the audio signals, features are extracted and used in a decision model to identify the dominant emotion in each audio frame. The output of the classification model can be used as a suggestion for unique lighting design or as a complement to the original lighting scheme during different parts of a play. The audio data used for the analysis is collected during two long-term processes: 1) rehearsing projects that utilize the augmentation framework, and 2) acting training[2].
Experiments were conducted using the Surrey Audio-Visual Expressed Emotion (SAVEE) database. However, due to its limitations, the Acted Emotional Speech Dynamic Database (AESDD) was created. This database includes recordings from five professional actors, aged 25 to 30, with two male and three female actors in the group. The proposed model was trained using both the SAVEE and AESDD databases, and audio features were extracted from them. For the training process, k-fold validation was used, whereby the input data was divided into k-subsets, with k-1 subsets used to train the classifier and the last subset used to evaluate and calculate the accuracy of discrimination after k iterations. The AESDD library has proven to be a reliable source of data for SER classification issues.

2.3. New Trends in Speech Emotion Recognition

The proposed system is developed based on the energy and characteristics of sound, which are investigated to extract relevant features. The extracted features include frequency, wavelength, period, speed of travel, intensity, loudness, echo, pressure, amplitude, and resonance, among others[3].
The process of converting analog sound signals to digital signals and extracting features from them using signal processing techniques such as Fast Fourier Transform, Discrete-Time Fourier Transform, Discrete Fourier Transform, and Discrete Cosine Transform, is an important aspect of speech emotion recognition technology. In particular, Mel Frequency Cepstrum Coefficients (MFCC) is a popular method used for feature extraction in the frequency domain. By analyzing these features, speech emotion recognition systems can identify and classify emotions in speech signals. Process of Speech Emotion Recognition consists of 4 steps:
1. Pre-Processing 2. Feature Extraction 3. Emotion Classification 4. Emotion Categories
So, after data is acquired from the dataset, we start working on creating the model which follows the above steps/

2.4. Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation.

In this paper, the authors convey thatthe difficult job of automatic speech emotion recognition (SER) is essential for realistic human-computer communication. Data scarcity, or the lack of enough well labelled data to create and fully explore complicated deep learning models for emotion classification, is one of the primary problems in SER.The study proposes a combination of transfer learning and spectrogram augmentation techniques to enhance SER performance. Specifically, a transfer learning approach is utilized along with a Residual Network (ResNet) model, pre-trained with a statistics pooling layer from speaker identification using a substantial amount of labeled data. Additionally, spectrogram augmentation is employed to generate more training data samples by applying random time-frequency masks to log-melspectrograms.In particular, we suggest a transfer learning strategy that makes use of a residual network (ResNet) model that has already been trained and includes a statistics pooling layer from speaker identification trained on a sizable quantity of speaker-labeled data.The model can process variable-length input well because to the statistics pooling layer, which also eliminates the requirement for sequence truncation, which is frequently employed in SER systems. In order to reduce overfitting and boost the generalisation of emotion identification models, we also use a spectrogram augmentation strategy to provide more training data samples by applying random time-frequency masks to log-mel spectrograms. Using the interactive emotional dyadic motion capture (IEMOCAP) dataset, we assess the efficacy of our suggested methodology. According to experimental findings, the transfer learning and spectrogram augmentation procedures enhance SER performance and, when coupled, produce cutting-edge outcomes[4].

2.5. Lexical Dependent Emotion Detection Using Synthetic Speech Reference.

TheIn order to contrast the emotional content of a speech signal, this study attempts to produce neutral reference models using synthetic speech. Due to the diversity in how people perceive and describe their emotions, modelling emotional actions is a difficult endeavour. Relative judgements are more trustworthy than absolute assessments, according to earlier research.According to this research, using a reference signal with established emotional content (such as neutral feeling) to compare a target sentence to may result in more accurate metrics to detect emotional segments. The ideal sentence would have the same lexical content as the target sentence, be emotionally neutral, and have their contents be temporally matched. By comparing the acoustic characteristics of the target and reference utterances frame by frame, we would be able to recogniselocalised emotional cues in this hypothetical case.Using feature analysis and perceptual evaluation, this research examines whether the synthesised signals offer reliable template references to represent neutral speech. Last but not least, we show how this approach may be used to recognise emotions, outperforming classifiers trained using cutting-edge features in identifying low vs high degrees of arousal and valence[5].

2.6. Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest.

Theobjective of thissystem called speech emotion recognition may automatically identify different emotion types from provided attribute segments. The creation of high-accuracy speech emotion identification systems has emerged as a popular study area in the field of speech due to the rising need for emotion recognition in business, education, and other industries.In order for the computer to evaluate the speaker's individual emotional condition through speech and better humanise human-computer contact, speech emotion recognition uses speech as the carrier of emotion to research the genesis and evolution of various emotions in speech. A voice emotion recognition model based on feature representation of convolutional neural network CNN (Convolution Neural Network) is suggested in order to increase the precision of intelligent speech emotion detection system. The most popular technique for extracting speech characteristics, mel-frequency cepstral coefficients (MFCC), is used for the experiment[6].

2.7. Effective Attention Mechanism in Dynamic Models for speech Emotion Recognition.

The biological visual attention mechanism discovered in nature served as inspiration for the author's attention mechanism, which is presented in this paper. Each element in the output sequence in this mechanism is dependent on specific elements in the input sequence, or put another way, it relies on the elements that were chosen to make up the input sequence. The model becomes more accurate and performs better as a consequence, but the computational workload is increased. The majority of methods realise attention as a weight vector with a dimension that is roughly equal to the length of the input sequence[7].

In addition to the standard attention mechanism, the author has proposed a unique type of attention mechanism termed the structural attention mechanism. It differs from the typical attention process in that the contextual structural information is stored in a memory matrix.
In this study, a recurrent neural network (RNN) model is used as the foundation for an attention process. An artificial neural network called an RNN is used to represent sequential data. RNN frequently includes feedback circuits that let data move through time steps. Today, LSTM cells or gated recurrent units are frequently used in RNNs (GRUs). Recently, the use of bi-directional RNNs (BiRNN), which enable two-way information exchange, has become widespread.

The FAU-Aibo tasks are used to assess the suggested recognition model. A German speech collection is called FAU-Aibo. Emotion labels are annotated on statements in this instance. It is made up of 9 hours of German speech from 51 kids as they converse with Sony's companion robot Aibo. There are 8257 speech chunks in the test set and 9959 in the training group. The FAU-Aibo competition dataset is used in the Inter-Speech 2009 Emotion Challenge. The 5-class classification task with the emotion categories of Anger, Emphatic, Neutral, Positive, and Rest is the main focus of this study.

2.8. Adding Dimensional Features for Emotion Recognition on Speech.

The primary addition of the author to this paper is the automatic insertion of the Valence-Arousal-Dominance (VAD) model parameters in the characteristic's vector to enhance the performance of the emotion recognition system. Three stages make up the procedure[8]:
• Regression: Arousal and valence regression models are created using the labelled train data.
• Estimation: The automatic classifier estimates the regression variables (i.e., arousal and valence) for the test data by using the training data or prior models.
• Classification: To conduct emotion identification, the estimated values are concatenated or combined with the acoustic features.

The corpus comprises of 134 sessions, each lasting 10 minutes, in which elderly Spanish individuals spontaneously converse with a virtual coach. 80 participants interacted with a visual agent in two distinct situations, the first of which developed a nutrition coaching session and was more general in nature and concerned leisure preferences. Audio and video recordings from each encounter were kept. Here, only the talks' audio portion—which lasts for about 7 hours—is taken into account. After that, the audio files were appropriately labelled with respect to feelings. Three native collaborators worked on a manual annotation from start to achieve this. Both the categorical and the three-dimensional (arousal, valence, and dominance) models were used to label the observed emotion. Numerical labels between -1, 0, and 1 are given to the three-dimensional model parameters. Thus, four parameters can be used to characterise each emotion:
• its classification: tense, tense, sad, joyful, puzzled.
• its degree of arousal: neutral (0), slightly excited (1), and excited (1). (-1).
• its orientation, which is either positive (1), neither positive nor negative (0), or negative (-1).
• its degree of dominance: somewhat dominant (1), neither dominant nor intimidated (0), and somewhat intimidated (-1).
The annotators also established the time frames that denote shifts in emotional condition.

2.9. Speech Signal-Based Modelling of Basic Emotions to Analyse Compound Emotion: Anxiety.

The author's goal in this article is to model the fundamental emotions through the extraction of speech signal discriminating features and to analyse anxiety as a combination of the fundamental emotions of anger, fear, and sadness. The workflow of the suggested model is illustrated and elaborated in this article, which concentrates on analysing the stated compound emotion—anxiety—as a combination of the fundamental emotions of anger, fear, and sadness[9].
The Min-Max normalisation technique was used to normalise the voice samples of the compound emotion, anxiety, as well as the three basic emotions,

anger, sadness, and fear. The Voice Activity Detector was used to identify voiced sounds in the speech stream.
To recognise the emotions in voice, various segmental and suprasegmental acoustic features corresponding to each fundamental emotion have been used. The suggested model in this study enables the extraction of the appropriate voice sample characteristics for each of the three basic emotions under consideration. This job is accomplished by autoencoders.
In order to recognize the fundamental feelings in the provided voice sample, a powerful and effective classifier is required. The effectiveness and astounding accuracy of neural networks for classifying fundamental feelings in speech signals have been demonstrated in studies. For the job, a single hidden layer artificial neural network is used. The feedforward network picks up the characteristics of the three fundamental emotions, analyses them, and uses them to detect and pinpoint the fundamental emotion in a compound emotion.

2.10. Speech Based Emotion Recognition Using Spectral Feature Extraction and an Ensemble of KNN Classifiers.

Using a pattern recognition paradigm, spectral feature extraction methods, and an ensemble of k Nearest Neighbor (kNN) classifiers, the author has primarily concentrated on the study of speech-based emotion detection in this article. In contrast to physiological signals (such as the electrocardiogram (ECG), electroencephalogram (EEG), skin temperature and resistance, blood pressure, and respiration), which are highly intrusive and cannot be measured remotely, speech and image signals can be acquired with the help of remote audio and video surveillance. The main goal of this study is to determine whether the reliable spectral features used for speaker identification and determination of the blind signal to noise ratio can also be used to identify emotions[10].
The assumption of speaker and gender freedom underlies recognition. Linear predictive Cepstrum (CEP), Mel Frequency Cepstrum (MFCC), Line Spectral Frequencies (LSF), Adaptive Component Weighted Cepstrum (ACW), and Postfilter Cepstrum are the five spectral characteristics (PFL). The ACW, PFL, and MFCC features for speaker identification are reasonably resistant to channel, noise, and speech coding distortion or disturbance. Emotion detection has been used with the CEP and MFCC features. We still need to look into and ultimately learn more about the ACW, PFL, and LSF features' efficacy and applicability in recognising emotions. Although they have been used in the past to identify emotions, acoustic features like energy, speaking rate, zero crossing rate, pitch, and formants are not taken into account in this study.It only employs the k closest neighbour classifier (kNN). Both a solitary kNN and an ensemble system made up of several kNNs are taken into consideration.

3. Conclusion

Speech emotion recognition is an important research area with many potential applications, including mental health, education, marketing, and human-computer interaction.There are various approaches to speech emotion recognition, including rule-based methods, machine learning-based methods, and hybrid methods that combine both approaches. The accuracy of speech emotion recognition systems varies depending on factors such as the quality of the recording, the language used, and the cultural context. By reviewing several literature survey articles, we come to the conclusion that the Random Forest has been proven to be more perfect in its operation. In the future, we plan to use the Random Forest as one of the core algorithms to lay the foundation for our project.

REFERENCES

[1] S. R. Kadiri and P. Alku, "Excitation Features of Speech for Speaker-Specific Emotion Detection," in *IEEE Access*, vol. 8, pp. 60382- 60391, 2020, doi: 10.1109/ACCESS.2020.2982954.
[2] Vryzas, Nikolaos &Kotsakis, Rigas&Liatsou, Aikaterini&Dimoulas, Charalampos&Kalliris, George. (2018). Speech Emotion Recognition for Performance Interaction. Journal of the Audio Engineering Society. Audio Engineering Society. 66. 457-467. 10.17743/jaes.2018.0036.
[3] Y. Ü. Sonmez and A. Varol, "New Trends in Speech Emotion Recognition," 2019 7th International Symposium on Digital Forensics and Security (ISDFS), Barcelos, Portugal, 2019, pp. 1-7, doi: 10.1109/ISDFS.2019.8757528
[4] Sarala Padi, Seyed Omid Sadjadi, Ram D.Sriram and Dinesh Manocha(2021). Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation. Isbn: 9781450384810, doi1: 10.1145/3462244.
[5] R. Lotfian and C. Busso, "Lexical Dependent Emotion Detection Using Synthetic Speech Reference," in *IEEE Access*, vol. 7, pp. 22071-22085, 2019, doi: 10.1109/ACCESS.2019.2898353.
[6] L. Zheng, Q. Li, H. Ban and S. Liu, "Speech emotion recognition based on convolution neural network combined with random forest," 2018 Chinese Control And Decision Conference (CCDC), Shenyang, China, 2018, pp. 4143-4147, doi: 10.1109/CCDC.2018.8407844.
[7] P. -W. Hsiao and C. -P. Chen, "Effective Attention Mechanism in Dynamic Models for Speech Emotion Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 2526-2530, doi: 10.1109/ICASSP.2018.8461431.
[8] L. B. Letaifa, M. I. Torres and R. Justo, "Adding dimensional features for emotion recognition on speech," 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 2020, pp. 1-6, doi: 10.1109/ATSIP49331.2020.9231766
[9] R. A. Rammohan, J. Medikonda and D. I. Pothiyil, "Speech Signal-Based Modelling of Basic Emotions to Analyse Compound Emotion: Anxiety," 2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), Udupi, India, 2020, pp. 218-223, doi: 10.1109/DISCOVER50404.2020.9278094.
[10] S. A. Rieger, R. Muraleedharan and R. P. Ramachandran, "Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers," The 9th International Symposium on Chinese Spoken Language Processing, Singapore, 2014, pp. 589-593, doi: 10.1109/ISCSLP.2014.6936711.