IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 5, April 2024

Speech Emotion Recognition with Deep Neural Networks

Mahesh Anap[1], Late Sakshi[2], Sarode Kanchan[3], Varule Mayuri[4]

Department of Electronic and Telecommunication[1,2,3,4]

Pravara Rural Engineering College, Loni, Maharashtra, India

Abstract: *The recognition and categorization of emotions is becoming increasingly important in the field of Human-Computer Interaction (HCI). Body language, including tone of voice and facial expression, can be used to identify emotions.*

Speech Emotion Recognition (SER) is one of the most popular methods for identifying emotions, according to the current study. The EMO-DB dataset is utilized in this study, while the SER dataset comprises four distinct datasets. This technique is employed because it has a high temporal resolution at a low cost and no hazards. Many academics have used SER signals in sequence during the past few decades to deal with Human-Computer Interface (HCI) and detect emotions. It entails filtering out background noise from audio signals, obtaining temporal or spectral properties from them, analyzing the signals in the time or frequency domains, and finally creating a multi-class classification plan. The method of recognizing and categorizing human emotions using auditory cues is covered in the paper.

Artificial Neural Network (ANN), KNN Classifier, Support Vector Machine (SVM), Random Forest Classifier, Convolution Network (CNN), and Decision Tree (DT) Models were among the machine learning techniques employed in the method. The experimental result that was achieved is promising, exhibiting good accuracy in the classification of emotions.

Keywords: Emotions, Audio Signal, Random Forest (RF), Support Vector Machine (SVM) Convolution Network (CNN), Decision Tree (DT), EMO-DB Dataset

I. INTRODUCTION

Emotions are vital to human existence and the communication process between individuals. There are numerous ways to communicate emotions, such asthrough body language, facial expressions, and conversation. As a result, researchers are increasingly using audio signals to identify human emotions. Emotions are essential to human existence. It's one method to let people know how you feel. Emotion recognition has emerged as one of the most popular research topics these days. Emotion has facilitated simple and efficient communication between humans and machines. Many communication channels, including body language, facial expressions, voice recognition, etc., can be used to identify emotions. In certain analyzed through his/her facial expression and body language, whereas the conversation is made through the medium. It can be challenging to forecast the person's emotion when communication and connection through the channel is what people are expecting from one another. Speech Emotion Recognition (SER) is a technique that allows a person to communicate their emotional state through speech. Modulated vocal sounds are the primary way that humans are different from other living things. Human voices can be classified according to a number of characteristics, including vocal tone, pitch, loudness, and timbre. Different vocal characteristics allow us to easily analyze human emotions. Any system can be trained to recognize a few common emotions, such as fear, surprise, happiness, sorrow, rage, and neutrality.

Support Vector Machines, or SVMs for short, are supervised learning algorithms. For tasks involving regression and classification, it performs better. It is mostly applied to categorization issues. Finding the ideal boundary line that enables us to quickly categorize the n- dimensional space into the desired classes is the primary objective of support vector machines (SVMs). In order to make it simple to add additional data in the same category—which will be helpful in the future—thisboundary line is referred to as a hyper-plane. The feed- forward method of a synthetic neural network

DOI: 10.48175/568

called Multilayer Perceptron, or MLP for short, produces a large number of outputs from a small number of inputs. There are three levels in total: input, concealed, and output.

It is most effective when applied to regression and classification issues, namely those involving classification. Because the approach is tree-structured, internal nodes reflect the features of the dataset, branches indicate the decision rules, and the leaf node displays the result. The term "convolution neural network" (NNN) An sophisticated neural network is used to categorize emotions. Without human assistance, CNN automatically identifies the salient characteristics. Its computational performance comes from the use of pooling and convolution processes.

II. LITERATURE SURVEY

The author suggests two-stage methods for emotion recognition from voice signals: feature extraction and classification engine. The author of this work employs 42 features as feature vectors, including the harmonic to noise ratio, Teager energy operator (TEO), and 39MFCC (12MFCC+energy, 12deltaMFCC+energy). Using an RML database and SVM (support vector machine) as a classifier, accuracy was attained rate of 74.07% for speech emotion detection which is better than that of random forest(88 features, accuracy 65.28%),SVM using RML database(MFCC,FBE ,ZCR, pitch energy feature with accuracy 69.3) and SVM(MFFC and modulation spectral features with accuracy 69.6%)

Author proposed Fourier parameter-based harmonic features. First extracted frame-level harmonic features and then computed the statistics of these frame-level features along with their derivatives. The extracted features were normalized for each speaker using z normalization. The accuracy of SER was improved by 16.2 percentage, 6.8%, and 16.6%, respectively on the German ,CASIA, and EESDB databases when compared with the MFCCfeatures and using the Fourier parameters. After combiningthe Fourier parameters with MFCC, the accuracy was improved up to17.5%, 10%, and 10.5% on the German, CASIA, and EESDB databases, respectively.

Author uses a leveraging parallel combination of BLSTM(bidirectional long sort term memory) neural network and attention based fully convolutional network(FCN) as a classifier .The experiments are conducted using two types of databases IEMOCAP and FAU aibo emotion corpus(FAU-AEC) to show the effectiveness of the approach. The proposed model well suited for the SER, with achieving a weighted accuracy of 68.1% and unweighted accuracy of 67.0% on IEMOCAP database and 45.4% for unweighted accuracy on FAU- AEC dataset. these accuracy shows better when compare with DNN+ELM[[20],[26]](WA 57.9%,UA52.1%) and RNN+ELM[26](WA62.9% and UA 63.9%).

Author proposed spectral features based on the local Hu moments of Gabor-spectrogram. feature are computed using the following steps:(1)finding log energy of the spectram2computation of Gabor spectrogram by convolving the logarithmic energy spectrum with wavelet(3)computation of Gabor local Hu (GSLHu) moments spectrogram then applied DCT to decorrelate the features and finally use the PCA to eliminate the redundancy in the features .The proposed features called GSLHu-PC, compared with the MFCC compared with theMFCC, Hu-weighted spectral, and PLP features. The SER accuracy using GSLHu-PCA was improved by 7.47%, 4.97%, and 1.35% over the MFCC, PLP ,and Hu-weightedspectral features, respectively, on the ABC database.

Author proposed modulation spectral features (MSFs) for the automatic recognition of human affective information from speech. The features are extracted from an auditory- inspired long-term spectro temporal representation. Using an auditory filter bank and a modulation filter bank for speech analysis, the representation captures acousticfrequency and temporal modulation frequency components, thereby conveying information important for human speech perception but missing from conventional short-term spectral features. On an experiment assessing the classification of discrete emotion 30 categories, the MSFs show promising performance compared with features based on mel-frequency cepstral coefficients and perceptual linear prediction coefficients, two commonly used short-term spectral representations. The MSFs furthershows a substantial improvement in recognition performance when used to augment prosodic features, which have been extensively used for emotion recognition.Using both types of features, an overall recognition rate of 91.6% is obtained for classifying seven emotion categories. 13 Department of ETC, PREC, Loni The accuracy of SER was improved by 6.8% and 6.6% using the MSF (modulation spectral features) when compared with the MFCC and PLP features on Verlin emotional speech database and VAM database respectively.

Author proposes RSPA(residual sinusoidal peak amplitude) features for the speech emotion recognition. SVM used as a classifier and got better accuracy as compare with linear combination of LPC(linear prediction coefficient) and TEO-CB-Auto-Env features. Then Combination of the RSPA and mel frequency cepstral coefficients (MFCC) further improves the recognition performance of the emotion classification. EMODB database used for the experiment point of to test the features. Author found that the RSPA feature has accuracy of 67% which is better as compare with LPC(63.4%) and TEO-CB-Auto-Env(64.8%) features

Author proposed features based on wavelet packets (WP). They used a tree pruning algorithm to optimize the wavelet packet tree for the emotion classification task. The optimal sub bands were selected and then the discriminative band WP power coefficient (dB-WPPC) features were derived. The accuracy of SER was improved by 5.1%, 8.16%, and 5.61% when compared with the MFCC, PLP, and Mel wavelt packet features, respectively.

proposed features based on wavelet packets (WP). They used a tree pruning algorithm to optimize the wavelet packet tree for the emotion classification task. The optimal sub-bands were selected and then the discriminative band WP power coefficient (dB-WPPC) features were derived. The accuracy of SER was improved by 5.1%, 8.16%,and 5.61% when compared with the MFCC, PLP ,and Mel wavelt packet features, respectively. In[9] Author used 88 features for speech emotion recognition. Features include pitch, intensity, percentile, Formants ,formant bandwidth, MFCC, delta MFCC, filter bank energy(FBE).Random forest classified used to classify the emotion. RML database used in the experiment to test the accuracy of SER. The Accuracy rate achieved for SER is 65.28%.

III. DATASETS AND METHODOLOGY

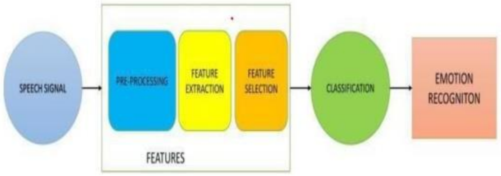


Fig. Speech Emotion Recognition Model

The goal of this project is to use deep neural networks (DNNs) and multi-feature extraction to identify emotions from EMO-DB audio data. Several audio extraction techniques, including Mel Frequency Cepstral Coefficients (MFCC), Chromogram, Mel-Spectrogram, Tonnetz, and Contrast, will be used to extract the audio characteristics from the EMO-DB dataset. Next, the collected features are subjected to PCA (Principal (DNN). The next stage is to test using the testing set as the test data once the training results model has been established. The trained DNN model will use the emotion type prediction results from the testing procedure to createa confusion matrix table, which will be used as a guide to determine the performance parameters of the suggested model. Accuracy was one of the factors used to assess the suggested model's performance.

Dataset

The EMODB database is the freely available German emotional database. The database is created by the Institute of Communication Science, Technical University, Berlin, Germany. Ten professional speakers (five males and five females) participated in data recording. The database contains a total of 535 utterances. The EMODB database comprises of seven emotions: 1) anger; 2) boredom; 3) anxiety; 4) happiness; 5) sadness; 6) disgust; and 7) neutral. The data was recorded at a 48- kHz sampling rate and then down-sampled to 16-kHz.

Feature Extraction

One feature extraction method that is frequently applied to audio data is Mel Frequency Cepstral Coefficients (MFCC). It is commonly advised to utilize MFCC to identify monosyllables in audio without revealing the speaker's identity. The Pre-emphasis stage of the MFCC feature extraction process in audio involves boosting the audio stream at high frequencies. The next step involves applying the Fast Fourier Transform, Mel Filter Bank, and Discrete Cosine. These are followed by the windowing and framing stages, where the windowing technique is used to limit the occurrence of disturbances at the beginning and end of the audio and the framing stage aims to divide the length of the audio into several time intervals between 20 ms to 30 ms. Transform as a transformational process the windowing results into MFCC. The advantage of using MFCC (Mel-Frequency Cepstral Coefficients) in speech emotion recognition is that it accurately captures the acoustic characteristics of human speech. The MFCC employs the mel scale, which is comparable to how frequency is perceived by humans through hearing. In order to create MFCC features, the power spectrum's logarithm is transformed into cepstrum. This helps to lower the dimensionality and processingcomplexity of the features. Using short-duration frame splitting techniques, MFCC can capture fluctuations in speech signals linked with temporal emotional changes and express temporal information in those signals.

Chroma is an audio feature extraction method that focuses on tones related to music. With the help of this capability, audio tonal fluctuations can be distributed as a basic feature. The output of the Chroma feature is a chromagram constructed using 12 (twelve) tone levels. The goal of using chroma in audio is to identify the high and low pitches of an actor's speech, as the tone of the speech can convey a particular kind of emotion.

The Mel-Spectrogram is an audio feature extraction technique designed to address the issue of humans' limitedcapacity to differentiate high-frequency values through hearing. In this study, the Mel-Spectrogram is used toextract information on variations in frequency values, specifically in recognizing the types of emotions that performers are expressing.

Tonnetz is an audio harmony and tone class- focused feature extraction method that comes fromChroma.

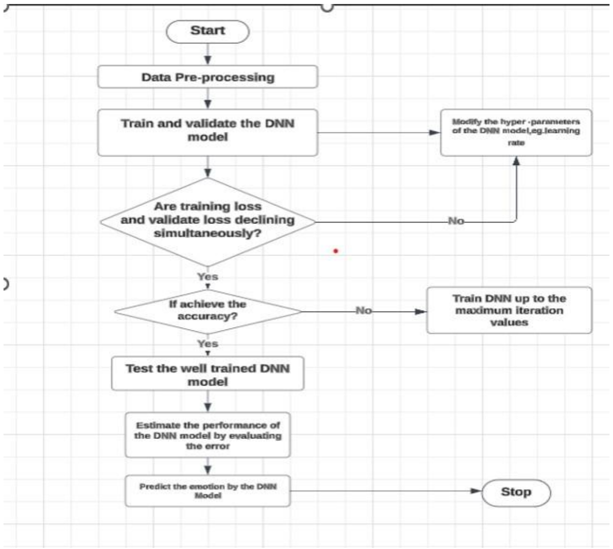
An audio feature extraction called contrast can be used to determine the average sound energy based on the peak and valley spectral values of each sub-band.

Min-Max Normalization

Normalization is the process of bringing features with notable value discrepancies into balance so that, when used as a guide for classifier model training, the weights and effects on each feature are the same. The value of each feature in this study will be dispersed throughout a range of values between 0 and 1, using the Min-MaxNormalization approach. An overview of the effects of applying normalization in the creation of classifier modelswill be given by the application of this technique.

This method works by first determining the maximum($xmax$) and minimum ($xmin$) values of each variable or feature.

IV. FLOWCHART


Start
Data Pre-processing
Train and validate the DNN model
Modify the hyper -parameters of the DNN model,eg.learning rate
Are training loss and validate loss declining simultaneously?
No
Yes
If achieve the accuracy?
No
Train DNN up to the maximum iteration values
Yes
Test the well trained DNN model
Estimate the performance of the DNN model by evaluating the error
Predict the emotion by the DNN Model
Stop

V. CONCLUSION

The two main challenges in speech emotion identification are feature extraction and classification, which make it a challenging undertaking. We suggest a novel approach to voice emotion identification that leverages a one-dimensional deep neural network (DLNN) and five distinct audio feature combinations as input. To increase the performance for EMO-DB, we gradually present a collection of models based on our initial framework. With the exception of one method, our top-performing model, attains greater accuracy than any earlier research. On the other hand, our method compares well with that one in terms of simplicity, universality, and application. Unlike some other

DOI: 10.48175/568

methods, all of their suggested models operate straight from audio input without the need to convert it to visual representations

REFFRENCES

[1]. Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access, 8, 79861-79875.

[2]. Mitra, Ayushi. ‖ Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset). ‖ Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 03 (2020): 145-152.

[3]. M. Li, X. Qiu, S. Peng et al., ―Multimodal Emotion Recognition Model Based on a Deep Neural Network with Multiobjective Optimization, ‖ Wireless Communications and Mobile Computing, vol. 2021, Article ID 6971100, 10 pages, 2021.

[4]. M. Abdel-Basset, R. Mohamed, and S. Mirjalili, A novel whale optimization algorithm integrated with Nelder-Mead simplex for multi- objective optimization problems, ‖ Knowledge- Based Systems, vol. 212, p. 106619, 2021.

[5]. W. Rahman, M. K. Hasan, and A. Zadeh, ―M- BERT: injecting multimodal information in the BERT structure, ‖ in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2359 – 2369, 2020.

[6]. M. D. Ding and L. Li, ―CNN and HOG dual-path feature fusion for face expression recognition, ‖ Information and Control, vol. 49, no. 1, pp. 47 – 54,2020.

[7]. Ullah, R.; Asif, M.; Shah, W.A.; Anjam, F.; Ullah, I.; Khurshaid, T.; Wuttisittikulkij, L.; Shah, S.; Ali, S.M.; Alibakhshikenari, M. Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer. Sensors 2023

[8]. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W. Survey of deep representation learning for speech emotion recognition. IEEE Trans. Affect. Comput. 2021, 14, 1634–1654.

[9]. Singh, P.; Srivastava, R.; Rana, K.P.S.; Kumar, A multimodal hierarchical approach to speech emotion recognition from audio and text. Knowl.-Based Syst. 2021.

[10]. Huddar, M.G.; Sannakki, S.S.; Rajpurohit, V.S. Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN. Int. J. Interact. Multimed. Artif. Intell. 2021.

[11]. Akçay, M.B.; O˘guz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun.2020

[12]. Tuncer, T.; Dogan, S.; Acharya, U.R. Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. Knowl.-Based Syst. 2021, 211, 106547.

[13]. Saleem, N.; Gao, J.; Khattak, M.I.; Rauf, H.T.; Kadry, S.; Shafi, M. Deepresgru: Residual gated recurrent neural network-augmentedkalman filtering for speech enhancement and recognition.

[14]. Knowl.-Based Syst. 2022

[15]. Shilandari, A.; Marvi, H.; Khosravi, H.; Wang,

[16]. W. Speech emotion recognition using data augmentation method by cyclegenerative adversarial networks. Signal Image VideoProcess. 2022, 16, 1955–1962.

[17]. Chen, Q.; Huang, G. A novel dual attention- based BLSTM with hybrid features in speech emotion recognition. Eng. Appl. Artif.Intell.2021, 102, 104277

[18]. M. Abdel-Basset, R. Mohamed, and S. Mirjalili, A novel whale optimization algorithm integrated with Nelder-Mead simplex for multi- objective optimization problems, ‖ Knowledge- Based Systems, vol. 212, p. 106619, 2021.

DOI: 10.48175/568

[19]. W. Rahman, M. K. Hasan, and A. Zadeh, M- BERT: injecting multimodal information in the BERT structure, ‖ in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2359 – 2369, 2020.

[20]. M. D. Ding and L. Li, —CNN and HOG dual-path feature fusion for face expression recognition,‖ Information and Control, vol. 49, no. 1, pp. 47 – 54,2020.

[21]. Ullah, R.; Asif, M.; Shah, W.A.; Anjam, F.; Ullah, I.; Khurshaid, T.; Wuttisittikulkij, L.; Shah, S.; Ali, S.M.; Alibakhshikenari, M. Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer. Sensors 2023

[22]. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W. Survey of deep representation learning for speech emotion recognition. IEEE Trans. Affect. Comput. 2021, 14, 1634–1654.

[23]. Singh, P.; Srivastava, R.; Rana, K.P.S.; Kumar, A multimodal hierarchical approach to speech emotion recognition from audio and text. Knowl.- Based Syst. 2021.

[24]. Huddar, M.G.; Sannakki, S.S.; Rajpurohit, V.S. Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN. Int. J. Interact. Multimed. Artif. Intell. 2021.

[25]. Akçay, M.B.; O˘guz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun.2020

[26]. Tuncer, T.; Dogan, S.; Acharya, U.R. Automated accurate speech emotion recognition systemusing twine shuffle pattern and iterative neighborhood component analysis techniques. Knowl.-Based Syst. 2021, 211, 106547.

[27]. Saleem, N.; Gao, J.; Khattak, M.I.; Rauf, H.T.; Kadry, S.; Shafi, M. Deepresgru: Residual gated recurrent neural network-augmentedkalman filtering for speech enhancement and recognition.

[28]. Knowl.-Based Syst. 2022

[29]. Adversarial networks. Signal Image VideoProcess. 2022, 16, 1955–1962.