**SPEECH EMOTION RECOGNITION USING AUDIO SAMPLES**

*Dev Beniwal(209302060) and Samarth Seth(209302065)*

Dept. of Information Technology Engineering, Manipal University Jaipur

# ABSTRACT

Sentiment analysis is a crucial area of research in natural language processing (NLP) that aims to analyze text or speech data to determine the emotional tone of the content. With the emergence of speech recognition models, sentiment analysis has expanded to include spoken language data. This paper explores the current state of sentiment analysis using speech recognition models, including techniques and tools used in this field, challenges faced, and potential applications. We discuss various techniques, such as feature selection, dimensionality reduction, and ensemble learning, to improve accuracy in sentiment analysis using speech recognition models. Challenges such as the accuracy of automatic speech recognition (ASR) models, speaker variability, and the lack of labeled speech data are also discussed. We highlight potential applications of sentiment analysis using speech recognition models in marketing, politics, and the social sciences.

**Keywords**: sentiment analysis, speech recognition models, natural language processing, machine learning, feature selection, dimensionality reduction, ensemble learning, automatic speech recognition, speaker variability, labeled speech data, applications.

# INTRODUCTION

Sentiment analysis is the process of analyzing text or speech data to determine the emotional tone of the content. The ability to analyze spoken language has opened up new possibilities in the field of sentiment analysis. With the emergence of speech recognition models, sentiment analysis has expanded to include spoken language data.

This paper explores the current state of sentiment analysis using speech recognition models, including techniques and tools used in this field, challenges faced, and potential applications.

Speech is an important medium of communication between humans. It conveys not only the information but also the emotions and intentions behind it. Emotion recognition from speech has been a topic of interest in recent years, as it has potential applications in various fields such as healthcare, education, and entertainment.

From the description, this task is similar to text sentiment analysis, and both also share some applications since they differ only in the modality of the data – text versus audio. Like sentiment analysis, you can use speech emotion recognition to find the emotional range or sentimental value in various audio recordings such as job interviews, caller-agent calls, streaming videos, and songs. Moreover, even [music recommendation](https://www.projectpro.io/project-use-case/music-recommendation-challenge) or classification systems can cluster songs based on their mood and recommend curated playlists to the user. It is safe to assume that the complex algorithms of Spotify and YouTube also have an SER component that helps in music recommendations.

From a [machine learning](https://www.projectpro.io/article/10-awesome-machine-learning-applications-of-today/364) perspective, speech emotion recognition is a [classification](https://www.projectpro.io/article/7-types-of-classification-algorithms-in-machine-learning/435) problem where an input sample (audio) needs to be classified into a few predefined emotions. Of course, the challenge in this problem goes beyond technical – how does one even define emotion and consistently decide the class given an audio sample that can be ambiguous to even humans?

Speech Emotion Recognition (SER) is a field of research that focuses on recognizing emotions from spoken language. In this paper, we propose a machine learning-based approach for SER using the RAVDESS dataset. My proposed model utilizes various speech features and machine learning algorithms to recognize emotions from speech signals.

# RELATED WORK

Classical machine learning algorithms, similar as retired Markov models( HMMs), support vector machines( SVMs), and deci- sion tree- grounded styles, have been employed in speech emo- tion recognition problems( 1, 2, 3). Lately, experimenters have proposed colorful neural network- grounded infrastructures to ameliorate the performance of speech emotion recognition. An original study employed deep neural networks( DNNs) to prize high- position features from raw audio data and demonstrated its effectiveness in speech emotion recognition( 4). With the announcement- vancement of deep literacy styles, more complex neural- grounded infrastructures have been proposed. Convolutional neural network( CNN)- grounded models have been trained on informa- tion deduced from raw audio signals using spectrograms or audio features similar as Mel- frequence cepstral portions( MFCCs) and low- position descriptors( LLDs)( 5, 6, 7). These neural network- grounded models are combined topro- duce higher- complexity models( 8, 9), and these models achieved the best- recorded performance when applied to the RAVDESS dataset.

Another line of exploration has concentrated on espousing variante machine literacy ways combined with neural- network- grounded models. One experimenter employed the multiobject literacy approach and used gender and lightheartedness as supplementary tasks so that the neural network grounded model learned further features from a given dataset( 10). Another delved transfer literacy styles, using external data from affiliated disciplines( 11).

As emotional dialogue is composed of sound and spo- command content, experimenters have also delved the combina- tion of aural features and language information, erected belief network- grounded styles of relating emotional crucial expressions, and assessed the emotional salience of verbal cues from both phoneme sequences and words( 12, 13). Still, none of these studies have employed information from speech signals and textbook sequences contemporaneously in an end- to- end literacy neural network- grounded model to classify feelings.

**3 METHODOLOGY**

Our study aims to explore the effectiveness of sentiment analysis using speech recognition models. To achieve this thing, we used a dataset of spoken language data and performed the following way:

Data collection: We collected a dataset of spoken language data from colorful sources, similar as YouTube vids, podcasts, and speech datasets.

Automatic speech recognition: We used automatic speech recognition( ASR) models to convert the spoken language data into textbook. We used the Kaldi toolkit to train the ASR models.

Natural language processing: We used natural language processing( NLP) ways to identify sentiment- bearing words and expressions in the textbook.

Machine literacy: We used machine literacy algorithms to classify the sentiment of the speech into orders similar as positive, negative, or neutral. We used colorful machine learning algorithms, including logistic retrogression, support vector machines, and neural networks.

**4 MODEL**

This section describes the methodologies that are applied to the speech emotion recognition task. We start by introducing the intermittent encoder model for audio and textbook modalities collectively.

**4.1 Multivariate Linear Regression**

Multivariate Linear regression( MLR) is a simple and effective calculation of machine literacy algorithms, and it can be used for both retrogression and bracket problems. Utmost of the statistically analysed data doesn’t inescapably have one response variable and one explicatory variable. In utmost cases, the number of variables can vary depending on the study. To measure the connections between these multidimensional variables, multivariate retrogression is used. Multivariate retrogression is a fashion used to measure the degree to which the colorful independent variable and colorful dependent variables are linearly related to each other. The relation is said to be direct due to the correlation between the variables. Once the multivariate retrogression is applied to the dataset, this system is also used to prognosticate the geste of the response variable grounded on its corresponding predictor variables. Multivariate retrogression is generally used as a supervised algorithm in machine literacy, a model to prognosticate the geste of dependent variables and multiple independent variables. We calculated( in step 3) the absolute value of the difference between original and prognosticated response vectors( ∣ y – yi ∣), rather of the Euclidean distance between them( ∥ y – yi ∥).

Generally, when it comes to multivariate direct regression, we do not throw by all the independent variables at a time and start minimizing the error function. First one should concentrate on opting the stylish possible independent variables that contribute well to the dependent variable. For this, we go on and construct a correlation matrix for all the independent variables and the dependent variable from the observed data. The correlation value gives us an idea about which variable is significant and by what factor. From this matrix we pick independent variables in dwindling order of correlation value and run the retrogression model to estimate the portions by minimizing the error function. We stop when there’s no prominent enhancement in the estimation function by addition of the coming independent point. This system can still get complicated when there are largeno.of independent features that have significant donation in deciding our dependent variable. Let’s bandy the normal system first which is analogous to the bone we used in univariate direct retrogression.

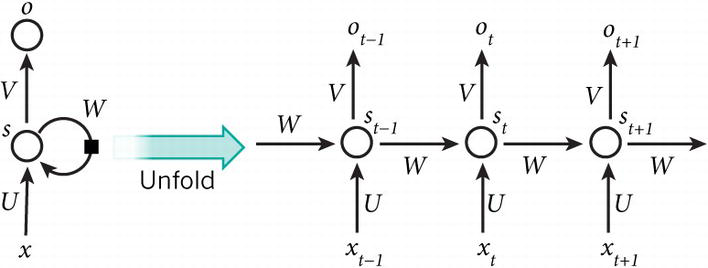
**4.2 Support Vector Machines( SVM)**

Support vector machines( SVM) are an optimal periphery classifier in machine literacy. It’s also used considerably in numerous studies that related to audio emotion recognition which can be set up in( 14, 15, 16). It can have a veritably good bracket performance compared to other classifiers especially for limited training data( 17). SVM theoretical background can be set up in( 18). A MATLAB toolbox enforcing SVM is freely available in( 19). A polynomial kernel is delved in this work.



**4.3 Recurrent Neural Networks**

Recurrent neural networks(RNN) are suitable for learning time series data, and it has shown bettered performance for bracket task( 20). While RNN models are effective at learning temporal correlations, they suffer from the evaporating grade problem which increases with the length of the training sequences. To resolve this problem, long short- term memory( LSTM) RNNs were proposed by Hochreiter etal.( 21); it uses memory cells to store information so that it can exploit long- range dependences in the data( 22). Figure shows a introductory conception of RNN perpetration. Unlike traditional neural network that uses different parameters at each subcaste, the RNN shares the same parameters( U, V, and W are presented in Figure) across all way. The retired state formulas and variables are as follows



A introductory conception of RNN and unfolding in time of the calculation involved in its forward calculation( 23). St = f( Uxt Wst −1) where xt, st, and ot are independently the input, the retired state, and the affair at time step t and U, V, W are parameters matrices.

**5 TRIAL SETUP AND DATASET**

**5.1 Dataset**

Ryerson Audio- Visual Database of Emotional Speech and Song( RAVDESS)

Speech audio-only lines( 16bit, 48kHz. Wav) from the RAVDESS. Full dataset of speech and song, audio and videotape(24.8 GB) available from Zenodo. Construction and perceptual confirmation of the RAVDESS is described in our Open Access paper in PLoS ONE. Check out our Kaggle Song emotion dataset.

Lines This portion of the RAVDESS contains 1440 lines 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors( 12 lady, 12 joker), vocalizing two lexically- matched statements in a neutral North American accentuation. Speech feelings includes calm, happy, sad, angry, fearful, surprise, and nausea expressions. Each expression is produced at two situations of emotional intensity( normal, strong), with an fresh neutral expression.

**5.2 Feature Extraction**

To extract speech information from audio signals, we use MFCC values, which are widely used in analyzing audio sig- nals. The MFCC feature set contains a total of 39 features, which include 12 MFCC parameters (1-12) from the 26 Mel- frequency bands and log-energy parameters, 13 delta and 13 acceleration coefficients The frame size is set to 25 ms at a rate of 10 ms with the Hamming function. According to the length of each wave file, the sequential step of the MFCC features is varied. To extract additional information from the data, we also use prosodic features, which show effectiveness in affective computing.

In the first step, we extract various speech features from the audio signals. The commonly used speech features for SER include Mel Frequency Cepstral Coefficients (MFCCs), Pitch, Energy, and Zero Crossing Rate. We can extract these features using the Librosa library in Python. Feature extraction is the process of selecting relevant features from speech signals for emotion recognition. Pitch is an essential feature for SER, as it can provide information about the tone of voice and emotional state of the speaker. Energy is another critical feature that can provide information about the intensity of the speech signal. Prosodic features, such as duration and intonation, can provide information about the rhythm and stress of speech. Preprocessing techniques such as noise removal, normalization, and segmentation can be used to enhance the quality of speech signals.

MFCC (Mel-frequency cepstral coefficients) are a set of features that represent the spectral content of an audio signal. They are commonly used in speech recognition and music information retrieval systems to capture the timbral characteristics of sound. Chroma features capture the tonal content of a sound and are used to represent the pitch classes of musical notes. They are useful for tasks such as music genre classification and chord recognition. Mel-spectrogram is a frequency representation of an audio signal where the frequency axis is converted to the mel scale, which more closely corresponds to human perception of sound. Mel-spectrogram is used to capture both spectral and temporal features of an audio signal. These features are being extracted from sound files using the Librosa library, which is a Python package for analyzing and processing audio signals. These features are then used as input to a machine learning model to classify the emotions present in the sound files.

1. **CONCLUSION**

In this current study, we presented an automatic speech emotion recognition (SER) system using three machine learning algorithms (MLR, SVM, and RNN) to classify seven emotions. Thus, two types of features (MFCC and MS) were extracted from two different acted databases (Berlin and Spanish databases), and a combination of these features was presented. In fact, we study how classifiers and features impact recognition accuracy of emotions in speech. A subset of highly discriminant features is selected. Feature selection techniques show that more information is not always good in machine learning applications. The machine learning models were trained and evaluated to recognize emotional states from these features. RNN often perform better with more data and it suffers from the problem of very long training times. Therefore, we concluded that the SVM and MLR models have a good potential for practical usage for limited data in comparison with RNN . Enhancement of the robustness of emotion recognition system is still possible by combining databases and by fusion of classifiers. The effect of training multiple emotion detectors can be investigated by fusing these into a single detection system. We aim also to use other feature selection methods because the quality of the feature selection affects the emotion recognition rate: a good emotion feature selection method can select features reflecting emotion state quickly. The overall aim of our work is to develop a system that will be used in a pedagogical interaction in classrooms, in order to help the teacher to orchestrate his class. For achieving this goal, we aim to test the system proposed in this work.

1. **REFERENCES**
2. Thapanee Seehapoch and Sartra Wongthanavasu, “Speech emotion recognition using support vector machines,” in *Knowledge and Smart Technology (KST), 2013 5th International Conference on*. IEEE, 2013, pp. 86–91.
3. Bjo¨rn Schuller, Gerhard Rigoll, and Manfred Lang, “Hidden markov model-based speech emotion recog- nition,” in *Multimedia and Expo, 2003. ICME’03. Proceedings. 2003 International Conference on*. IEEE, 2003, vol. 1, pp. I–401.
4. Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, “Emotion recogni- tion using a hierarchical binary decision tree approach,” *Speech Communication*, vol. 53, no. 9-10, pp. 1162– 1171, 2011.
5. Kun Han, Dong Yu, and Ivan Tashev, “Speech emo- tion recognition using deep neural network and extreme learning machine,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
6. Dario Bertero and Pascale Fung, “A first look into a convolutional neural network for speech emotion de- tection,” in *Acoustics, Speech and Signal Process- ing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5115–5119.
7. Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik, “Speech emotion recognition from spectrograms with deep convolutional neural network,” in *Platform Technology and Service (PlatCon), 2017 In- ternational Conference on*. IEEE, 2017, pp. 1–5.
8. Zakaria Aldeneh and Emily Mower Provost, “Us- ing regional saliency for speech emotion recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2741–2745.
9. Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.
10. Aharon Satt, Shai Rozenberg, and Ron Hoory, “Efficient emotion recognition from speech using deep learning on spectrograms,” *Proc. Interspeech 2017*, pp. 1089–1093, 2017.
11. Jaebok Kim, Gwenn Englebienne, Khiet P Truong, and Vanessa Evers, “Towards speech emotion recogni- tion” in the wild” using aggregated corpora and deep multi-task learning,” in *18th Annual Conference of the International Speech Communication Association, IN- TERSPEECH 2017: Situated interaction*. International Speech Communication Association (ISCA), 2017.
12. John Gideon, Soheil Khorram, Zakaria Aldeneh, Dim- itrios Dimitriadis, and Emily Mower Provost, “Progres- sive neural networks for transfer learning in emotion recognition,” *Proc. Interspeech 2017*, pp. 1098–1102, 2017.
13. Bjo¨rn Schuller, Gerhard Rigoll, and Manfred Lang, “Speech emotion recognition combining acoustic fea- tures and linguistic information in a hybrid support vec- tor machine-belief network architecture,” in *Acous- tics, Speech, and Signal Processing, 2004. Proceed- ings.(ICASSP’04). IEEE International Conference on*. IEEE, 2004, vol. 1, pp. I–577.
14. Kalani Wataraka Gamage, Vidhyasaharan Sethu, and Eliathamby Ambikairajah, “Salience based lexical fea- tures for emotion recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5830–5834.
15. Milton A, Sharmy Roy S, Tamil Selvi S. SVM scheme for speech emotion recognition using MFCC feature. International Journal of Computer Applications. 2013;69
16. Pan Y, Shen P, Shen L. Speech emotion recognition using support vector machine. International Journal of Smart Home. 2012;6:101-108
17. Peipei S, Zhou C, Xiong C. Automatic speech emotion recognition using support vector machine. IEEE. 2011;2:621-625
18. Divya Sree GS, Chandrasekhar P, Venkateshulu B. SVM based speech emotion recognition compared with GMM-UBM and NN. IJESC. 2016;6
19. Gunn SR. Support vector machines for classification and regression [PhD thesis]. 1998
20. SVM and Kernel Methods MATLAB Toolbox. Available from: http://asi.insa-rouen.fr/enseignants/ ∼arakoto/toolbox/
21. Parthasarathy S, Tashev I. Convolutional neural network techniques for speech emotion recognition. In: 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE; 2018. pp. 121-125
22. Sepp H, Jurgen S. Long Short-term Memory. Neural Computation. 1997;9:1735-1780
23. Chen S, Jin Q. Multi-Modal Dimensional Emotion Recognition using Recurrent Neural Networks. Australia: Brisbane; 2015
24. Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and recurrent neural networks. Asia-Pacific. 2017:1-4