



SPEECH EMOTION RECOGNITION USING DEEPLARNING

Stella J, Srushti Rathod, Alisha Joy, Aman Khakhi

DEPT. OF INFORMATION TECHNOLOGY,
XAVIER INSTITUTE OF ENGINEERING

Abstract— Communication is an integral part of the human social life. We communicate using the different semantics known to us and we express our thoughts, ideas and opinions via tone and linguistics. It has been observed that humans convey information using various emotions such as anger, fear, sadness, calm, disgust, happiness, etc. Humans have been able to perceive such emotions because of years of social interaction among them, but such gifted abilities are restricted and limited to humans only. Machines, which us humans are dependent on nowadays do not have this luxury. Machines are not as capable as us humans to process such delicate and intricate information on their own and hence, there is a need to train them accordingly resulting in easier and better communication. In the field of human computer interaction (HCI), emotion recognition from the computer is still a challenging issue, especially when the recognition is based solely on voice, which is the basic mean and the most integral part of human communication. In human computer interaction systems, emotion recognition could provide the users with improved personalization services by being adaptive to their emotions. Therefore, emotion detection from speech could have many potential applications in order to make the machine more flexible to the user and thus enriching and bettering user's needs and experience. In the proposed system we will be using various Deep Learning (ML) techniques to build and train a model which is capable of detecting and recognizing the various emotions known to man. By this project, one of our aims is to enrich customer experiences in call centers by analyzing various call recordings and recognize the emotional aspects of speech irrespective of the semantic contents.

Keywords: Speech, Emotion, Noise.

I. INTRODUCTION

In order for any conversation to be fruitful, the use of emotions is very necessary. Emotions play a

of speech emotion recognition is one of the most important problems in the field of para linguistics. This field has recently expanded its applications

very important and crucial role in communication. It is a medium of expression which humans rely on heavily. The importance of emotion recognition in human speech has increased in recent times to improve both the naturalness and efficiency of human-machine interactions. Recognizing human emotions is itself a very complex task due to the ambiguity of classifying emotions as occurring and natural. A number of studies have been conducted with the aim of extracting spectral and prosodic features that would lead to the correct identification of emotions. Speech Emotion Recognition (SER) can be stated as the extraction of the emotional state of a user from their speech signal. There exists few universal emotions like Anger, Happiness, Sadness. These emotions with the help of any intelligent system and finite computational resources can be trained to identify as required. In this work, spectral and prosodic features are used for speech emotion recognition because both of these features contain emotional and semantic content. Fundamental frequency, loudness, pitch, intensity of speech and glottal parameters are prosodic features that are used to model the different emotions. Potential features are extracted from each utterance for computational mapping between the emotions and speech patterns. Pitch can be detected from the selected elements. Emotions can be classified as Natural and Artificial emotions and can be further divided into an emotion set like anger, sadness, joy, fear, etc. Various machine learning techniques, including k-nearest neighbor, radial basis function, and neural network back propagation have been used to create recognition agents. In particular, the role because it is a decisive factor in optimal human-computer interaction, including dialog systems. The goal of speech emotion recognition is to predict the emotional content of speech and classify speech

according to one of several labels (i.e., happy, sad, neutral, and angry). Different types of deep learning methods have been used to improve the performance of emotion classifiers; however, this task is still considered challenging for several reasons. First, due to the cost associated with involving humans, there is not enough data available to train complex neural network-based models. Second, emotion characteristics must be learned from low-level speech signals. Feature-based models show limited skill when applied to this problem.

To overcome this problem, we propose a model that uses a high level text transcription to utilize the information contained within the information contained within low resource datasets to a greater degree. With this project we plan to build a system given the problem definition stated above along with gathering relevant data regarding speech expressing different emotions. Along with this, we aim to improve man and machine interface. We are building a Multi-Layer Perceptron and a classifier to recognize emotion given a speech signal. The models are evaluated on the IEMOCAP dataset under different settings, namely, Audio-only, Text-only and Audio + Text .Our project involves building a model using Natural Language Processing (NLP) and deep learning techniques to improve the learning performance, lowering computational complexity, and building better generalizing models and decreasing the required storage. With this project we generate a model to recognize the emotions available and recognize the emotional aspects of speech irrespective of the semantic contents.

II. LITERATURE SURVEY

In this section, we explore and cover some of the research on speech emotion recognition that has been done. The SER's mission is not new, and has been a topic of literature study for a long time.

1. The difficulty of recognizing emotions from speech signals as a component of Human-Computer Interaction. Many techniques, including traditional and Deep Learning methods, have been used for speech emotion recognition. The paper presents an overview of Deep Learning techniques in speech-based emotion recognition, covering the databases used, emotions extracted, contributions made, and limitations. Understanding emotions from voice signals is a crucial yet difficult aspect of human-computer interaction (HCI).

2. Many methods, including many well-known speech analysis and classification methods, have been used to extract emotions from signals in the literature on speech emotion recognition (SER). Recently, deep learning approaches have been put out as an alternative to conventional SER techniques. In order to recognise speech-based emotions, this paper provides an overview of deep learning approaches and reviews some current work that uses these techniques. The review discusses the databases used, the emotions that were retrieved, the advancements made in speech emotion identification, and any associated constraints.

3. Although a few good surveys exist for SER, they either don't cover all aspects of SER in natural environments or don't discuss the specifics in detail. This survey focuses on SER in a natural environment, discussing SER techniques for natural environment along with their advantages and disadvantages in terms of speaker, text, language, and recording environments. In the recent past, the deep learning techniques have become very popular due to minimal speech processing and enhanced accuracy. Special attention has been given to deep-learning techniques and the related issues in this survey. Recent databases, features, and feature selection algorithms for SER, which have not been discussed in the existing surveys and can be promising for SER in a natural environment, have also been discussed in this paper.

III. OBJECTIVES

- Gather relevant data regarding speech expressing different emotions
- To Improve man-machine interface.
- Build a model using Natural language processing (NLP) and deep learning techniques to improve the learning performance, lowering computational complexity, building better generalizable models and decreasing required storage.
- To identify and analyse the emotional aspects of speech irrespective of the semantic contents.
- To evaluate the performance of the model in recognizing different emotions in speech signals and to compare the performance of the proposed model with existing methods for speech emotion recognition.

IV. PROPOSED METHODOLOGY AND APPROACH

A. DATASET

In our paper we make use of the IEMOCAP, which researchers at the University of Southern California published in 2008(USC). There are five recorded sessions in it, ten speakers, totaling almost 12 hours of audio and video with transcriptions of the conversations. It is labeled with eight classified emotion categories, including surprised, fear, rage, happiness, sadness, neutral, and thrilled. Additionally, it has dimensional labels with activation and valence values ranging from 1 to 5, but they are not employed in this work. We further divided each utterance file to produce wav files for each sentence because the dataset has already been divided into numerous utterances for each session.

B. IMPLEMENTATION

This section shows the process of our data pre-processing and feature extraction. Four models are ensemble to generate better results.

1) *Data Pre-Processing*: The data pre-processing steps are as follows:

a) For Audio: A first frequency analysis of the dataset showed that it is unbalanced. The under representation of the emotions "fear" and "surprise"

was addressed using upsampling methods. Once we recognised that "happy" was underrepresented and that the two feelings were very similar to one another, we combined instances from the "happy" and "excited" classes. Additionally, we eliminate examples labelled as "others" since they matched situations that even a human would have labelled as unclear. The aforementioned operations were used to generate a total of 7837 samples.

b) For Text: The available transcriptions were first normalized to lowercase and any special symbols were removed.

2) *Feature Extraction*: We now describe the handcrafted features used to train both, the ML- and the DL-based models.

i) Audio Features:

a) Standard Deviation: Standard deviation is a statistical measure that describes how much variation or dispersion there is in a set of data. In speech emotion recognition, standard deviation can be used to analyze the variability of features extracted from speech signals, such as pitch, intensity, and formant frequencies.

b) RMS: RMS also known as root mean square shows the average loudness of the audio. RMS loudness measures the audio signal taking into account the energy of the wave. This measurement represents the perceived loudness more accurately than peak levels.

c) Silence: Silence in audio refers to the absence or near-absence of sound in an audio signal. In other words, it is a period of time during which there is little or no audible sound present in the audio signal. Silence can occur for various reasons, such as during pauses in speech or music, in between tracks on an album, or due to technical issues during audio recording or processing. In audio processing, detecting silence can be useful for tasks such as audio segmentation, where we want to divide a long audio signal into smaller

chunks based on the presence or absence of sound. Detecting silence can also be useful for removing or reducing noise in an audio signal, as we can filter out or attenuate parts of the signal that are below a certain threshold.

d)Harmonics: In audio, harmonics refer to the additional frequencies that are produced by a sound source in addition to its fundamental frequency. These additional frequencies are integer multiples of the fundamental frequency and are also known as overtones. The presence and strength of harmonics in an audio signal can have a significant impact on its perceived timbre or tone color. The unique pattern of harmonics produced by different instruments or voices is what allows us to distinguish between them even when they are playing the same note at the same volume.

ii) Text Features:

a) Term Frequency-Inverse Document Frequency (TFIDF): TFIDF is a numeric value that represents the relationship between a word and a document in a collection or corpus. This value has two components:

- Term Frequency: Term frequency refers to the number of times a specific word or term appears in a given document or text. It is a simple measure that quantifies the importance or relevance of a term within a document.

- Inverse Document Frequency: Inverse Document Frequency (IDF) is a statistical measure that quantifies how rare or unique a term is in a collection or corpus of documents. It is calculated by taking the logarithm of the ratio of the total number of documents in the corpus to the number of documents that contain the term. TFIDF value for a term is calculated by taking the product of TF and IDF values.

V. PROPOSED SYSTEM

To achieve results for our proposed system, we performed audio classification and text

classification separately. But our classification did not yield the accuracy we were aiming for, hence to improve our accuracy rate we combined both, the audio and text classification. After performing these combined classification problems, we observed that we obtained a better accuracy rate when both the classification problems are combined.

Our combined data of audio and text yielded us better accuracy results but not how much we expected it to be. Following are the results of the various machine learning models we used to get results.

1. Random Forest:

Random Forest consists of a large number of Decision Trees that operate as an ensemble. Each individual decision tree has a class prediction and the one class with the most votes becomes the prediction of the model. After applying random forest model on our dataset we achieved an accuracy rate of **66%**

2. XGBoost:

XGBoost is an optimized distributed gradient boosting library for efficient training of machine learning models. This method combines the results of multiple weak models to achieve a stronger and better result. After applying XGBoost model on our dataset we achieved an accuracy rate of **61.8%**

3. Support Vector Classifier:

Support Vector Classifier is a supervised machine learning algorithm typically used for classification tasks. SVC works by mapping data points to a high-dimensional space and then finding the optimal hyperplane that divides the data into two classes. After applying support vector classifier model on our dataset we achieved an accuracy rate of **63.8%**

4. Multinomial Naive Bayes:

Multinomial Naive Bayes (MNB) is a popular machine learning algorithm for text classification problems in Natural Language Processing (NLP). It is particularly useful for problems that involve

text data with discrete features such as word frequency counts. MNB works on the principle of Bayes theorem and assumes that the features are conditionally independent given the class variable. After applying support vector classifier model on our dataset we achieved an accuracy rate of **59.8%**

5. Multilayer Perceptron: A fully connected multi-layer neural network is called a Multilayer Perceptron (MLP). It has 3 layers including one hidden layer. If it has more than 1 hidden layer, it is called a deep ANN. An MLP is a typical example of a feedforward artificial neural network. After applying support vector classifier model on our dataset we achieved an accuracy rate of **66.3%**

6. Logistic Regression: Logistic regression predicts the output of a categorical dependent variable. It is a Machine Learning method that is used to solve classification issues. It is a predictive analytic technique that is based on the probability idea. After applying support vector classifier model on our dataset we achieved an accuracy rate of **63.3%**

The above mentioned accuracy rates are a better result than what we had achieved earlier from individually training on text and speech. to generate even better results we ensembled the following models, i.e, MultiLayer Perceptron, Random Forest, Logistic Regression , XGBoost and ensembled using the soft voting ensembling technique which yielded us an accuracy rate of **70%**

A. SYSTEM FLOWCHART

The below figure demonstrates our system flowchart for our proposed system. We begin our process by importing libraries.

Importing Libraries: Importing libraries in speech emotion recognition involves the process of importing pre-existing code libraries that contain useful functions and tools for developing speech

emotion recognition models. These libraries can significantly simplify the development process and save time and effort by providing pre-built functionality for common tasks in speech emotion

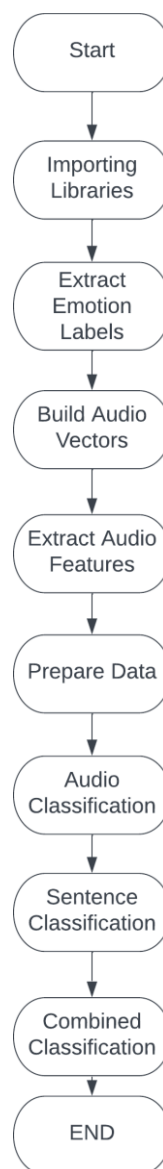


Fig. 1: Flowchart diagram

recognition. There are various libraries that can be used for speech emotion recognition, like Librosa, TensorFlow, Keras, Sklearn etc. When developing speech emotion recognition models, it is important to carefully choose the appropriate libraries based on the requirements of the project and the expertise of the developers. Additionally, it is important to keep the libraries up-to-date to take advantage of the latest advancements in speech emotion recognition research and technology.

Extracting Emotion Labels: We extracted the emotion labels that have been assigned to our dataset. Extracting emotion labels in speech emotion recognition involves the process of identifying and labeling emotional states or categories within a given audio signal or text data. The goal of extracting emotion labels is to accurately identify and label the emotional content of the input data with us. Accurate and consistent extraction of emotion labels is crucial for the development of effective speech emotion recognition models. It is important to ensure that the labeled data accurately reflects the emotional content of the input data and that the emotion categories are well-defined and consistent across different datasets.

Build Audio Vectors: In audio classification, we build the audio vectors which involves representing the audio signals in a way that can be used as input to a machine learning model for the purpose of classification. The process of building audio vectors include pre-processing, feature extraction and vectorization.

Extract Audio Features: Extracting the audio features involves extracting the relevant information or characteristics from the audio signal that can be used to classify the emotional state of the speaker. In other words, audio features are mathematical representations of specific aspects of the audio signal that are informative for distinguishing different emotions. Deep learning models can be used to learn feature representations directly from the raw audio signal. These models

typically use convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to extract features from the audio signal.

Prepare Data: We prepare the data which we could gather after our audio and text classification. It is done so as to ensure that the data is in a format that can be used as input to the deep learning models. The steps involved for preparing the data are to collect a representative sample of audio or text data that covers the range of emotions or classes that the model is intended to classify. It is important to ensure that the data is of high quality and accurately labeled. The collected data may contain noise, errors, or missing values that can affect the performance of the model. Data cleaning involves removing or correcting any errors or inconsistencies in the data.

Audio Classification: In audio classification we make use of machine learning algorithms to classify an audio signal into different emotional states or categories. The goal of audio classification in speech emotion recognition is to accurately identify the emotional state of the speaker based on the audio signal. Audio classification can be performed using a wide range of machine learning algorithms like SVM, Random Forest and CNN. Accurate audio classification is essential for effective speech emotion recognition, as it allows the model to make predictions based on meaningful and relevant information within the audio signals.

Sentence Classification: With the available text with us, we performed sentence classification which involves feature extraction and classification. The goal of sentence classification in speech emotion recognition is to accurately identify the emotional content of a given sentence or a sequence of sentences. Once the features have been extracted, they are fed into the machine learning algorithm, which then classifies the input data into different emotional states or categories. Accurate sentence classification is essential for effective speech emotion recognition, as it allows the model to

make predictions based on meaningful and relevant information within the text data.

Combined Classification: Combined classification in our project involves the integration of multiple modalities, such as audio and text, to improve the accuracy and robustness of emotion classification models. The idea behind combined classification is that different modalities can provide complementary information about the emotional content of a given audio signal or text data, which can be used to improve the performance of the emotion recognition model. Combined classification in speech emotion recognition has been shown to improve the accuracy and robustness of emotion classification models, particularly in scenarios where the audio or text data alone may not provide sufficient information for accurate emotion classification. However, the success of combined classification depends on the quality of the input data, the choice of classification models, and the design of the fusion approach.

VI. RESULTS AND TEST CASES

The mentioned accuracy rates in section V are a better result than what we had achieved earlier from individually training on text and speech. To generate even better results we ensemble the following models, i.e, MultiLayer Perceptron, Random Forest, Logistic Regression , XGBoost and ensemble using the soft voting ensembling technique which yielded us an accuracy rate of **70%**.

1) Multilayer perceptron: Following is the confusion matrix for the combined audio+text classification. We obtained an accuracy of **66.3%**

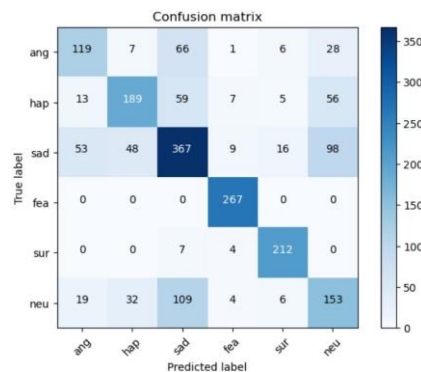


Fig. 2: Multilayer Perceptron

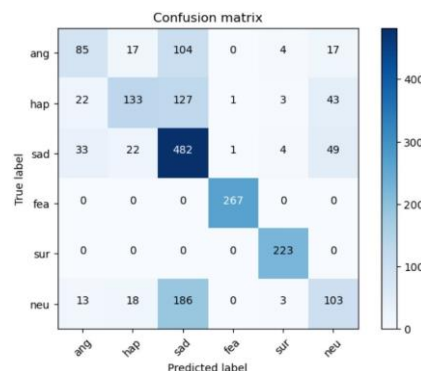


Fig. 3: Random Forest

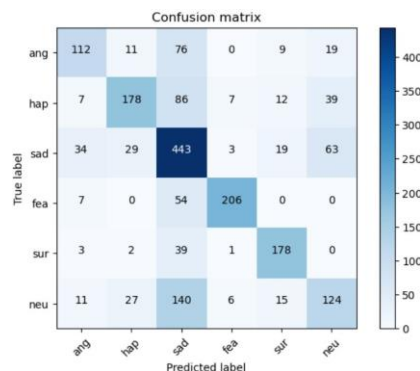


Fig. 4: Logistic Regression

2) Random Forest classifier: Following is the confusion matrix for the combined audio+text classification. We obtained an accuracy of **66%**

3) Logistic Regression: Following is the confusion matrix for the combined audio+text classification. We obtained an accuracy of **63.3%**

4) XGBoost: Following is the confusion matrix for the combined audio+text classification. We

obtained an accuracy of **61.8%**

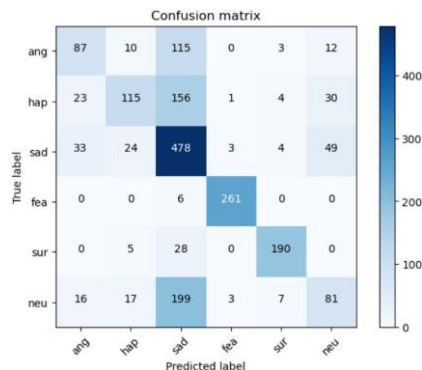


Fig. 5: XG boost

After ensembling the above models using soft voting we got an accuracy of **71.3%**

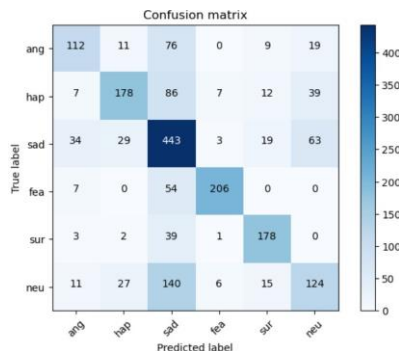


Fig. 6: Soft Voting Ensembling

VII. CONCLUSION

In this project, we have proposed a system that plans to make use of the deep neural network, to extract the emotional characteristic parameter from emotional speech signal automatically. We have combined deep learning networks and proposed a classifier model which is based on it. Through this project, we showed how we can make use of Deep Learning Techniques to obtain the underlying emotion from speech audio data as well as a text model and obtain some insights on the human expression of emotion through voice. This

system can be employed in a variety of setups like call centers for complaints or marketing, in voice based virtual assistants or chat bots, in linguistic research, etc. In future work we will continue to further study speech emotion recognition based on Deep Learning Techniques and further expand the training dataset. Our ultimate aim is to study how to improve the recognition rate of speech emotion and better the process of finding any trigger words if possible applicable in our dataset.

VIII. FUTURE WORK

A few possible steps that can be implanted to make the models more robust and accurate are the following:

1. An accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model.
2. Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition.

REFERENCES

- [1] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee and Shrikanth S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach", Speech Communication, 2011.
- [2] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," Signal Processing, vol. 90, no. 5, pp. 1415–1423, May 2010
- [3] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009
- [4] [1] <https://ieeexplore.ieee.org/abstract/document/8805181>
- [5] Emily Mower, Maja J. Mataric and Shrikanth S. Narayanan, "A Framework for Automatic Human Emotion Classification Using Emotional Profiles", IEEE Transactions on Audio, Speech and Language Processing, 19:5(1057-1070). May 2011
- [6] <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/IS140441.pdf>
- [7] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digit. Signal Process. vol. 22, no. 6, pp. 1154–1160, Dec. 2012.