Speech Emotion Classification

Arya Rajiv
Computer Science
PES University
Bangalore, India
aryarajivchaloli@gmail.com

Greeshma Karanth
Computer Science
PES University
Bangalore, India
greeshmakaranth.13@gmail.com

Shivangi Gupta
Computer Science
PES University
Bangalore, India
shivangig078@gmail.com

***Abstract*—Audio emotion analysis is trying to extract features from audio clips and classifying the emotion of the speaker. Many papers have attempted to solve this problem by using various techniques in machine learning and trying to achieve maximum accuracy. This report consists of the exploratory data analysis and the literature survey for various works in the related field.**

***Index Terms*—audio, emotion, learning, classification, features**

I. Introduction

Emotion has always strongly influenced a person's conversation. A message can be conveyed correctly only when the words along with the context of the message is understood. A truism: "It's not what you say, but how you say it". Expressions matter, as do the sentiment behind each encounter and the emotions raised. Emotion is entwined with the literal meaning of words used. Currently, machines are only capable of getting text from an audio clip. However, extending a machine's capabilities to make it understand the emotions with which the audio clip was produced can help interpret the intended message better and with more efficiency.

II. Literature Survey

Emotion is inferred from audiovisual inputs using decision based multimodal systems in one approach [1]. The inferences are classified into basic emotions such as happy, sad etc. using SVM and Radial Basis functions. While the algorithm yields a 98% classification rate, the accuracy is given mainly by the fusion of audio and video features. With just audio inputs - the classification rate is found to be 53%, which shows major scope for improvement. However, the decision based multimodal approach is proven to work on signals from environments which are highly susceptible to noise.

This approach focuses on emotion recognition from music - combining data from audio, MIDI and lyrics. The main aim of this paper is Music Emotion Recognition (MER) and to that extent, the success achieved by the SVM classifier used is 64% F-measure. This approach indicates that a multimodal approach works better than most models used for emotion recognition. However, the classification is done for music and not speech, which lack the supplementary features of songs such as instruments and lyrics, and therefore cannot be applied to speech signals [2].

Another method classifies seven emotions found on a Spanish and Berlin database using recurrent neural networks. The features used for classification are MFCC and MSFs and the combination has yielded high accuracy rates greater than 85% for both data-sets. However, the features used in this study are fewer than optimal and increasing the number of features for study can be looked into [3].

The relationship between emotions and learning was investigated using a device called AutoTutor which has an automated emotion classifier. By applying multiple regression on results on test subjects, conclusions were drawn for comparisons between predicted and actual values. This method is sufficient but it relies on the AutoTutor device for accurate results [4].

The emotion was detected based on speech and facial gestures with PCA and LDA applied to select features and Gaussian classifiers used for classification. The experiment was conducted with varying number of emotion classes with each yielding around 96% accuracy when both types of features were fused at the decision level. However, with only audio features, the recognition rate received was just 67%. Better results can be gained with other classifiers such as SVM [5].

Another approach talks about using partial least squares models on 2-5 predictors selected from the data-set to predict basic emotions from audio. The paper explains around 60%-85% of the variance seen between predicted and actual values. Data transformation has been done in multiple steps for optimal feature extraction and step-wise linear regression (MLR) has been applied to predict future values. The project however has been pursued in MATLab and can therefore provide different and improved results when implemented in R [6].

Another project attempted to use deep learning, instead of the commonly used machine learning approaches, as well as image classification in order to recognize emotion and classify them according to the speech signals. This model was able to achieve an overall accuracy of 35.6% using the Inception Net v3, which is a much lesser rate than other techniques. Transfer learning is used to train the model and although it reduces the computation cost and training time, it also results in reduced accuracy. Moreover, the data-sets used for each emotion were not enough. This approach does not work on purely audio based data-sets [7].

The main feature attribute considered in the prepared data-set was the peak-to-peak distance obtained from the graphical representation of the speech signals. For studying the different features of each emotion, the audio of every test subject was recorded and MATLAB functions were used for extracting features from these. The varying peak distance in the graphs, served as the distinguishing feature for the emotions. For classification, 30 different subjects were used and the prepared data-set had three attributes viz., feature distance, heart-rate and class. The data was classified using Weka software. The experiment was repeated for 30 times and the average value was chosen to estimate the standard deviation. However, the results show that the best accuracy is when the training size is small, thus, not proving useful for large data-sets [8].

III. Summary

The problem statement chosen is 'Speech Emotion Classification - identifying and classifying emotions in audio'.

A. *Issue to Address*

To classify emotions into four basic categories - sad, happy, angry and neutral using an optimal number of features and the best possible classification model such that accuracy gained is highest possible. The main challenge of this project will be to do appropriate pre-processing on the audio data and subsequently perform feature selection to choose suitable predictors for classification of emotions in speech.

B. *Innovation in Approach*

Many speech classifiers use visual features to aid audio clips while classifying emotions. The study here will be solely based on speech or audio clips with no visual aid. The data-set for this project is taken not only from one source but from multiple sources to enhance efficiency of learning. The aim is to reach an accuracy of 70% or higher with only audio clips.

References

[1] Ntombikayise Banda, Peter Robinson , 'Noise Analysis in Audio-Visual Emotion Recognition', unpublished

[2] R. Panda , R. Malheiro , B. Rocha , A. Oliveira and R. P. Paiva,'Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis', Centre for Informatics and Systems of the University of Coimbra, Portugal

[3] Leila Kerkeni, Youssef Serrestoul, Mohamed Mbarki, Kosai Raoof and Mohamed Ali Mahjoub, 'Speech Emotion Recognition: Methods and Cases Study', 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2, pages 175-182

[4] Arthur GRAESSER, Patrick CHIPMAN, Brandon KING, Bethany McDANIEL, Sidney D'MELLO, 'Emotions and Learning with AutoTutor', unpublished

[5] Sanaul Haq and Philip J.B. Jackson, 'Speaker-Dependent Audio-Visual Emotion Recognition',AVSP 2009 – International Conference on Audio-Visual Speech Processing, University of East Anglia, Norwich, UK September 10–13, 2009

[6] Tuomas Eerola, Olivier Lartillot, Petri Toiviainen, 'Prediction Of Multidimensional Emotional Ratings In Music From Audio Using Multivariate Regression Models', 10th International Society for Music Information Retrieval Conference (ISMIR 2009)

[7] Nithya Roopa S., Prabhakaran M, Betty.P, 'Speech Emotion Recognition using Deep Learning', International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4S, November 2018

[8] Assel Davletcharova, Sherin Sugathan, Bibia Abraham, Alex Pappachen James 'Detection and Analysis of Emotion From Speech Signals' Procedia Computer Science , 2015