ISSN No:-2456-2165

Feature Extraction Techniques Comparison for Emotion Recognition Using Acoustic Features

[1] Abhijeet Kalluray [2] Ankita Deshmukh [3] Rutuja Pacharne [4] Akash Sambhudas [5] Mr. Nikhil Dhavase
[1][2][3][4] Student, Information Technology Department, Marathwada Mitra Mandal's College of Engineering, Pune, Maharashtra, India
[5] Assitant Professor, Information Technology Department, Marathwada Mitra Mandal's College of Engineering, Pune, Maharashtra, India

Abstract:- As technology has developed so has our methods of interacting with them. Voice Interaction has become a major field in Human Computer Interaction. Today, we can successfully use our voice to give commands to different software. This is achieved using Natural Language Processing (NLP), i.e., the linguistic aspect. In this paper, we try to understand if the acoustic approach can be used for emotion recognition. Here, we use the sound waves from human voices to analyze the underlying tone of the speaker and classify it according to their emotion. This paper compares between different features extraction techniques of the audio and compares how different feature extraction techniques perform individually and together on audio data.

I. INTRODUCTION

Graphical User Interfaces (GUIs) have been the industry standard for interacting with computers for decades now. While GUI has virtual keys and buttons, a voice-user interface (VUI) makes spoken human interaction with computers possible, using speech recognition to understand spoken commands and answer questions, and typically respond with a text to speech module as a reply. This field has been mainly driven based upon the linguistic part of the speech. The new advancements in Voice Interaction have been made mainly in the fields of Natural Language Processing, whereas in the acoustic part of the speech, the underlying emotion behind the speech hasn’t been explored much. Here in this paper, we look at the acoustic part of the speech. For this purpose, there are many techniques to extract features from the audio clips, here in this paper we compare a few of them individually. A comparison with different permutations of these techniques is also done which shows how combined features inputs perform in comparison to their individual counterparts.

II. LITERATURE REVIEW

This paper focuses on using acoustic features for SER, and not the content (directly) in the speech. Hence different papers are studied which provided used techniques to extract the features from an audio signal and have used them for some application. The papers studied where Mel Frequency Cepstral Coefficients for Music Modelling [1], this paper investigates the applicability of Mel Frequency Cepstral Coefficients (MFCCs) to modelling music by examining two assumptions. The use of the MEL frequency scale to model the spectra, and the use of the Discrete Cosine Transform (DCT) to decorrelate the Mel-spectral vector. Music type classification by spectral contrast feature [2], is another paper where the authors use octave based spectral contrast feature for classifying music into different genres. Speech Emotion Recognition from 3DLog-Mel Spectrograms [3] uses 3D Log Mel spectrums paired with deep learning for speech emotion recognition. It uses a novel ADRNN architecture for that purpose. Chroma Feature Extraction [4]; this paper presents the details of chroma feature extraction from any audio files and the different types of extraction methods of the chroma feature are also explained. Detecting Harmonic Change In Musical Audio [5] explains a method for finding the changes in the harmonic content of musical audio signals by using a 12 bin chroma vector.

A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns [6] this paper uses multiple feature extractions for the classification purpose along with binary and ternary binning to take as inputs for the model.

III. METHODOLOGY

The methodological procedure is presented further below.

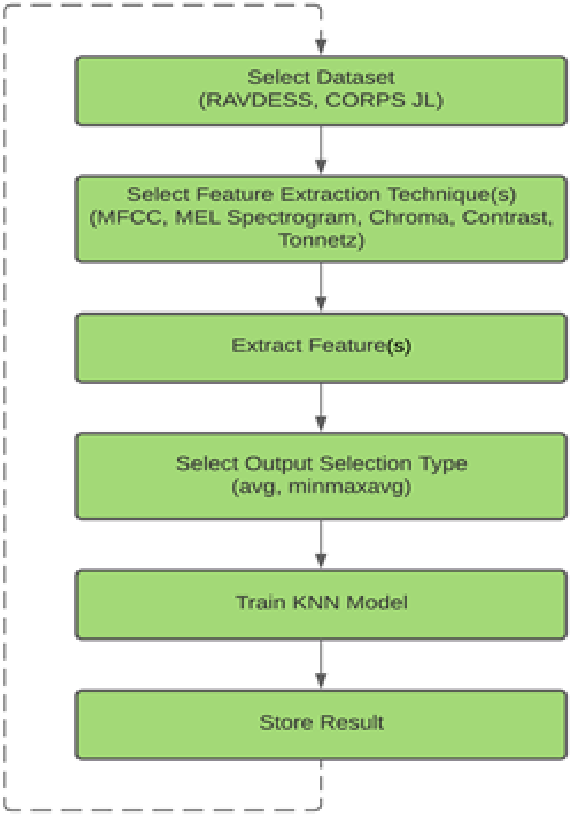
Select Dataset
(RAVDESS, CORPS JL)
Select Feature Extraction Technique(s)
(MFCC, MEL Spectrogram, Chroma, Contrast, Tonnetz)
Extract Feature(s)
Select Output Selection Type
(avg, minmaxavg)
Train KNN Model
Store Result

Fig. 1. Process flowchart

Firstly, a dataset is selected between the two datasets at hand, then a combination of feature extraction technique(s) is selected. After selecting the technique(s) the feature(s) are extracted. The sound clips that are loaded have different duration length, hence the number of features we would get would be different for different clips. This would cause problem while feeding inputs to the machine learning model. Hence to tackle two selection criteria are used. Selection criteria specifies how the features are taken in the end, either they are ‘avg’ meaning mean is taken of the features and that is used to train the model, or they are ‘minmaxavg’ meaning the minimum and maximum value along with the mean. NumPy library is utilized for this function. A KNN algorithm is used for classification purpose. After training the model the results are stored and the same procedure is repeated for a new combination of the dataset, feature extraction technique and the output selection type.

The combined audio dataset contains different categories of emotions, out of which 4 emotions are selected to train the model (Angry, Sad, Neutral, Happy). These 4 emotions are chosen as the dataset contained a high number of records for these particular emotions.

TABLE I. AUDIO CLIP DISTRIBUTION

Emotion	Total Audio Clips
Angry	324
Sad	324
Happy	324
Neutral	252

The previously mentioned feature extraction techniques are well written in the Librosa library. Hence rather than writing them from scratch. The Librosa package is used.

IV. RESULTS

The study focuses on 5 feature extractions in total, and the comparison between them. First individual features are compared to find the most important ones. Shown in the *TABLE 1*, are the accuracy comparisons for a single feature selection. The data is split as 80% training set, and 20% testing split, with stratified splitting applied.

TABLE II. INDIVIDUAL FEATURE COMPARISON TABLE

Dataset	Feature(s)	Selection Type	Test Score	Train Score	Average Score
CORPUS JL	chroma	avg	0.5469	0.6901	0.6185
		minmaxavg	0.5521	0.7201	0.6361
	contrast	avg	0.6667	0.7747	0.7207
		minmaxavg	0.6667	0.7747	0.7207
	mel	avg	0.8542	0.9362	0.8952
		minmaxavg	0.8385	0.9089	0.8737
	mfcc	avg	0.8698	0.9219	0.8958
		minmaxavg	0.8802	0.9232	0.9017
	tonnetz	avg	0.4323	0.6237	0.5280
		minmaxavg	0.3229	0.5951	0.4590
RAVDESS	chroma	avg	0.3556	0.6164	0.4860
		minmaxavg	0.3556	0.6462	0.5009
	contrast	avg	0.5259	0.6611	0.5935
		minmaxavg	0.5259	0.6611	0.5935
	mel	avg	0.5259	0.7430	0.6345
		minmaxavg	0.5556	0.6890	0.6223
	mfcc	avg	0.6741	0.8343	0.7542
		minmaxavg	0.6667	0.8287	0.7477
	tonnetz	avg	0.3556	0.5885	0.4720
		minmaxavg	0.2741	0.5512	0.4126

As it is evident from the table below, the top 3 performing feature extraction techniques are Mel Spectrogram Frequency, MFCC and Spectral Contrast. Further, more tests are carried out, to check how the model performs with different combinations of these 3 features.

Following are the results of these tests.

TABLE III. COMBINATON OF FEATURE COMPARISON TABLE

Dataset	Feature(s)	Selection Type	Test Score	Train Score	Average Score
CORPUS JL	contrast,mel	avg	0.547	0.690	0.8880
		minmaxavg	0.552	0.720	0.8770
	mffcc, contrast	avg	0.667	0.775	0.8926
		minmaxavg	0.667	0.775	0.9010
	mfcc, mel	avg	0.854	0.936	0.9010
		minmaxavg	0.839	0.909	0.9121
	mfcc, contrast, mel	avg	0.870	0.922	0.8971
		minmaxavg	0.880	0.923	0.9128
RAVDESS	contrast, mel	avg	0.356	0.616	0.6817
		minmaxavg	0.356	0.646	0.6316
	mfcc, contrast	avg	0.526	0.661	0.7644
		minmaxavg	0.526	0.661	0.7356
	mfcc, mel	avg	0.526	0.743	0.7644
		minmaxavg	0.556	0.689	0.7598
	mfcc, contrast, mel	avg	0.674	0.834	0.7644
		minmaxavg	0.667	0.829	0.7598

Here we can see that the model with CORPUS JL dataset, and all the features mel, mfcc, and contrast performs the best out of all, with a combined accuracy of 0.9128. Though it should be noted that the accuracy is not a lot from the combination individual features only. A similar thing can be seen with RAVDESS dataset. MFCC alone being the best feature with an accuracy of 0.9017 does almost as good a job as all the features combined. Without a significant upgrade in the accuracy, it cannot be said that the model with the combination of more than 1 feature performs better than their individual counterparts. It should also be noted that MFCC majorly works well with small duration data, and the audio clips used here are clips with duration of 1 to 2 seconds. Hence the other features or a combination of multiple features may show different results for audio clips with longer duration.

Following are more detailed results about the CORPUS JL, mel, mfcc, contrast with minmaxavg combination.

TABLE IV. CLASSIFICATION REPORT

The classification report on the test set is shown below.

	precision	recall	f1-score	support
angry	0.98	0.98	0.98	48
happy	0.92	0.92	0.92	48
neutral	0.81	0.88	0.84	48
sad	0.89	0.81	0.85	48
accuracy			0.90	192
macro avg	0.90	0.90	0.90	192
weighted avg	0.90	0.90	0.90	192

The following is a plot of the confusion matrix. As seen in the confusion matrix the model sometimes cannot predict correctly between sad and neutral. This can be a cause as both sad and neutral emotion have a lower tone resulting in weak features.

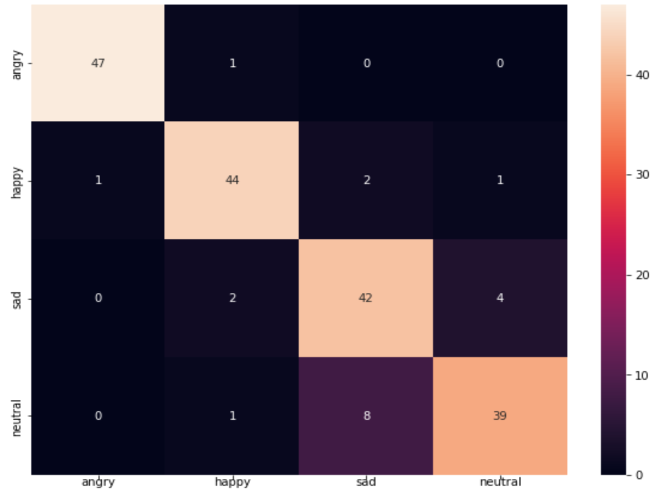

Fig. 2. Confusion Matrix

V. CONCLUSION

Using just the intensity and the variance of the sound waves in a human speech, the underlying emotions can be understood to some extent without linguistic processing. This means that acoustic features can act as an important factor in Speech Emotion Recognition (SER) with its big brother Natural Language Processing (NLP). Even without knowing the actual words, the sound waves can be analyzed to understand the emotion, this result can further be developed by classifying the words in the speech based on their sentiments, positive and negative. This combination would yield much better results.

MFCC still stands out to be the best feature extraction technique for speech sentiment analysis for short duration sound clips. A combination of the feature extraction techniques or the minimum and the maximum values of the features does not provide a substantial increase in the accuracy as compared to a single technique alone.

REFERENCES

[1]. Logan, Beth. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. Proc. 1st Int. Symposium Music Information Retrieval.

[2]. Jiang, Dan-Ning & Lu, Lie & Tao, Jian-Hua & Cai, Lian-Hong. (2002). Music type classification by spectral contrast feature. 113 - 116 vol.1. 10.1109/ICME.2002.1035731 K. Elissa, “Title of paper if known,” unpublished.

[3]. H. Meng, T. Yan, F. Yuan and H. Wei, "Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network," in IEEE Access, vol. 7, pp. 125868-125881, 2019, doi: 10.1109/ACCESS.2019.2938007.

[4]. Shah, Ayush & Kattel, Manasi & Nepal, Araju & Shrestha, D.. (2019). Chroma Feature Extraction.

[5]. Harte, Christopher & Sandler, Mark & Gasser, Martin. (2006). Detecting harmonic change in musical audio. Proceedings of the ACM International Multimedia Conference and Exhibition. 10.1145/1178723.1178727.

[6]. Y. Ü. Sönmez and A. Varol, "A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns," in IEEE Access, vol. 8, pp. 190784-190796, 2020, doi: 10.1109/ACCESS.2020.3031763.

[7]. B. T. Atmaja and M. Akagi, "On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers," 2020 IEEE REGION 10 CONFERENCE (TENCON), 2020, pp. 968-972, doi: 10.1109/TENCON50793.2020.9293852.

[8]. J. Mehta, D. Gandhi, G. Thakur and P. Kanani, "Music Genre Classification using Transfer Learning on log-based MEL Spectrogram," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1101-1107, doi:10.1109/ICCMC51019.2021.9418035.

[9]. N. Moradi, B. Nasersharif and A. Akbari, "Robust speech recognition using compression of Mel sub-band energies and temporal filtering," 2010 5th International Symposium on Telecommunications, 2010, pp. 760-764,doi:10.1109/ISTEL.2010.5734124.

[10]. S. Maghilnan and M. R. Kumar, "Sentiment analysis on speaker specific speech data," 2017 International Conference on Intelligent Computing and Control (I2C2), 2017, pp. 1-5, doi: 10.1109/I2C2.2017.8321795.