MACHINE LEARNING REPORT

SPEECH EMOTION RECOGNITION

PARTH RATHOD (A20458817)

CS584 SPRING 2021

TEAM - SOLO

ABSTRACT

Emotion is one of the most basic factors with respect to the communication among humans. It would be ideal to have human emotions automatically recognized by machines for improving human machine interaction. This is the motive behind the constantly increasing attention that this particular scientific field has been receiving lately. The first issue is the choice of suitable features for speech representation, the second issue is the proper preparation of an emotional speech database and the third issue is the design of an appropriate classification method.

INTRODUCTION

Speech is the quickest and most common way for humans to communicate with one another. This reality has led researchers to regard speech as a fast and efficient means of human-machine interaction. This, though, necessitates that the computer be intelligent enough to understand human voices. Since the late 1950s, there has been a great deal of study into speech recognition, which relates to the method of translating human speech into a series of phrases. Despite significant advances in speech recognition, we are still a long way from creating a normal relationship between man and machine because the machine does not recognize the emotional state of the speaker.

This has introduced a relatively recent research field, namely Speech Emotion Recognition (SER), which is defined as extracting the emotional state of a speaker from his or her speech. It is believed that speech emotion recognition can be used to extract useful semantics from speech, and hence, improves the performance of speech recognition systems.

For the reasons mentioned below, the process of speech emotion detection is extremely difficult. Firstly, it is unclear which facets of speech are most successful at discriminating between emotions. The acoustic heterogeneity introduced by the inclusion of various words, speakers, speaking modes, and speaking frequencies adds another difficulty since these properties have a significant effect on the majority of the typical extracted speech features such as pitch and energy contours. Moreover, there may be more than one perceived emotion in the same utterance; each emotion corresponds to a different portion of the spoken utterance. In addition, it is very difficult to determine the boundaries between these portions.

Secondly, the issue is that how a certain emotion is expressed generally depends on the speaker, his or her culture and environment. Most work has focused on monolingual emotion

classification, making an assumption there is no cultural difference among speakers.

The need to specify a selection of important emotions to be identified by an automated emotion recognizer is an important topic in speech emotion recognition. Linguists have developed inventories of the most typical emotional states we experience in our everyday lives. A typical set is given by Schubiger and O'Connor and Arnold, which contains 300 emotional states. However, classifying such a large number of emotions is very difficult. Primary emotions are Anger, Disgust, Fear, Joy, Sadness, and Surprise. These emotions are the most obvious and distinct emotions in our life. They are called the archetypal emotions.

PROBLEM DESCRIPTION

The task of speech emotion recognition is very challenging for following reasons. First it is not clear which speech features are powerful in distinguishing between emotions. Second issue is that how a certain emotion is expressed generally depends on the speaker, his or her culture and environment. Most of the work has focused on monolingual emotion classification making an assumption there is no cultural difference among speakers. This project aims to capture the correct prediction of speech emotions by using Machine Learning techniques.

DATA DESCRIPTION

An important issue to be considered in the evaluation of an emotional speech recognizer is the degree of naturalness of the database used to assess its performance. Incorrect conclusions may be established if a low-quality database is used. Moreover, the design of the database is critically important to the classification task being considered. There should be some criteria that can be used to judge how well a certain emotional database simulates a real-world environment. Using speech data obtained from real-life scenarios is more practical. Such recordings involve utterances that express feelings in a very natural way. Unfortunately, there might be ethical and moral considerations that prevent them from being used for scientific purposes. Emotional sentences can also be elicited in sound laboratories, as in the majority of current libraries. It has long been argued that acted feelings are not the same as genuine ones.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 3120 files. The database contains 10 professional actors (5 female, 5 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound).

<u>Audio-only files</u>

Audio-only files of all actors (01-10) are available:

- Speech file (Audio_Speech_Actors_01-10.zip) contains 600 files: 60 trials per actor x 10 actors = 600.
- Song file (Audio_Song_Actors_01-10.zip) contains 440 files: 44 trials per actor x 10 actors = 440.

<u>Audio-Visual and Video-only files</u>

Video files are provided as separate zip downloads for each actor (01-10), and are split into separate speech and song downloads:

- Speech files (Video_Speech_Actor_01.zip to Video_Speech_Actor_10.zip) collectively contains 1200 files: 60 trials per actor x 2 modalities (AV, VO) x 10 actors = 1200.
- Song files (Video_Song_Actor_01.zip to Video_Song_Actor_10.zip) collectively contains 880 files: 44 trials per actor x 2 modalities (AV, VO) x 10 actors = 880.

In total, the RAVDESS collection includes 3120 files (600+440+1200+880 files).

File Naming Conventions:
Each of the 3120 RAVDESS files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 02-01-06-01-02-01-10.mp4). These identifiers define the stimulus characteristics:

Filename identifiers

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 10. Odd numbered actors are male, even numbered actors are female).

DATABASE DESIGN CRITERIA

There should be some parameters for determining how accurately an emotional database simulates a real-world climate. According to some reports, the following are the most important things to consider: Of the real-world emotions or performed emotions? It is more realistic to use speech data that are collected from real life situations. Such recordings contain utterances with very natural conveyed emotions.

Who expresses the emotions? Many emotional expression libraries invite trained actors to articulate pre-determined sentences with the appropriate emotions. However, in some of them, semi-professional actors are used instead to prevent exaggeration in portraying feelings and to be more realistic.

How will the utterances be simulated? Many emotional speech datasets include utterances that were not generated in a conversational context. As a result, utterances can lack any naturalness because it is assumed that most feelings are the result of our reactions to various circumstances. Generally, there are two approaches for eliciting emotional utterances. In the first approach, experienced speakers act as if they were in a specific emotional state, e.g. being glad, angry, or sad. In many developed corporations such experienced actors were not available and semi-professional or amateur actors were invited to utter the emotional utterances.

Are utterances evenly spread across emotions? Such corpus authors, such as the Berlin corpus, prefer that the number of utterances for each emotion be almost the same in order to better test classification accuracy. Many other scholars, on the other hand, assume that the distribution of emotions in the database should reflect their occurrence in the real world. For eg, the neutral emotion is the most common in our everyday lives. As a consequence, the number of utterances containing neutral emotion should be the highest in the emotional expression corpus.

Is it the same sentence of different emotions? It

is popular in many libraries to document the same sentence with various emotions in order to research the explicit impact of emotions on the acoustic features of speech utterances. One drawback of such a database is that it ensures that human judgement on the perceived emotion is primarily dependent on the emotional nature of the sentence rather than its lexical content.

DATA VISUALIZATION

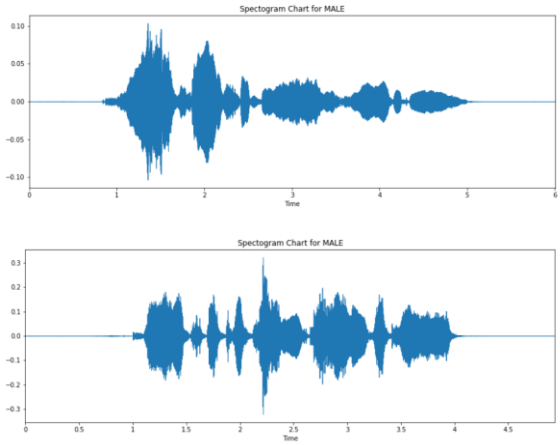


Fig:- MALE SPECTOGRAM

The first figure represents Spectrogram for Male when they are in Calm and Strong pitch. The Second figure represents a Males spectrogram for Angry and Strong pitch.

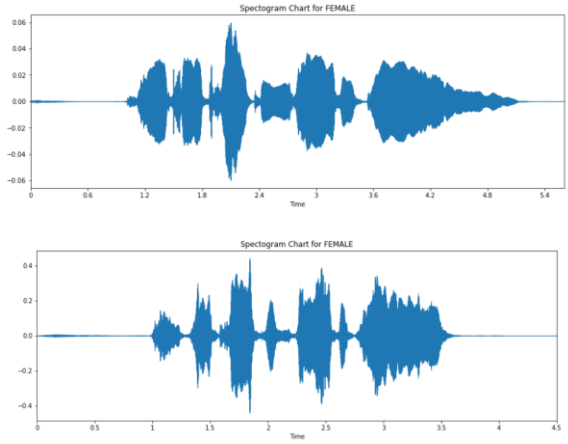


Fig:- FEMALE SPECTOGRAM

The first figure represents Spectrogram for Female when they are in Calm and Strong pitch. The Second figure represents a female spectrogram for Angry and Strong pitch.

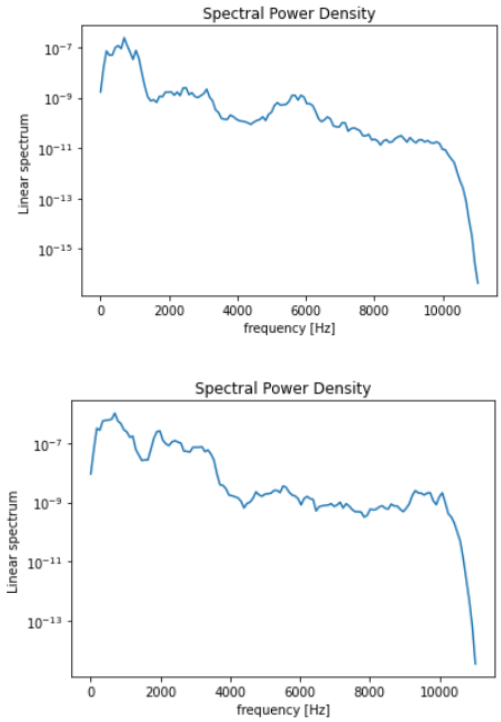
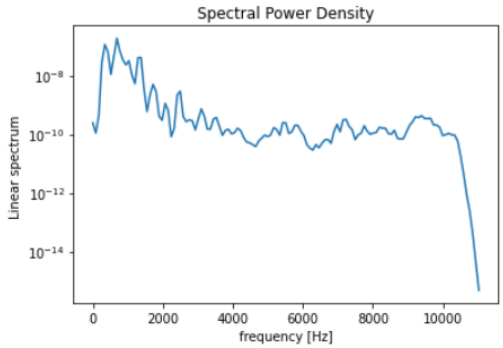


Fig:- MALE SPECTRAL POWER DENSITY

The first figure represents Spectral Power Density for Male when they are in Calm and Strong pitch. The Second figure represents a Males Spectral Power Density for Angry and Strong pitch.



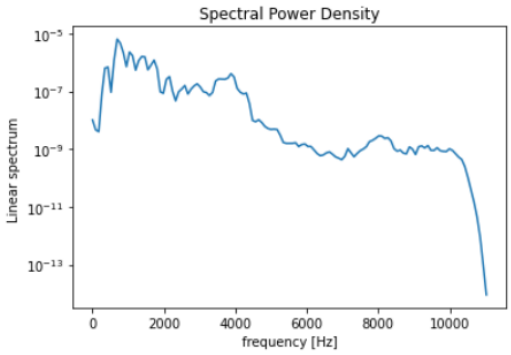


Fig:- FEMALE SPECTRAL POWER DENSITY

The first figure represents Spectral Power Density for Female when they are in Calm and Strong pitch. The Second figure represents a females Spectral Power Density for Angry and Strong pitch.

METHODOLOGY

Here we have used sklearn, librosa and soundfile libraries in python for predictions. We have used MLPClassifier to create a model. The MLP classifier optimizes the log-loss function using stochastic gradient descent. We have also used Adam optimizer here as we have relatively large data. Multilayer Perceptron is feedforward Artificial Neural Network. MLP uses supervised Learning for backpropagation for training. Learning occurs in perceptron by changing connection weights after data is processed, based on error. MLP has an internal neural network for purpose of classification. Using library soundfile we will extract,

- MFCC: Mel Frequency Cepstral Coefficient, represents short term power spectrum of a sound.
- Chroma: Pertains to 12 pitch class.
- Mel: Mel Spectrogram Frequency.

MFCC: This is the first step in automatic speech recognition system i.e to identify components of audio signal that are good for identifying linguistic content and discarding background noise.

Chroma: A chroma vector is a typically a 12-element feature vector indicating how much energy of each pitch class, is present in the signal.

Mel: The Mel Scale, mathematically speaking, is the result of some non-linear transformation of the frequency scale. This Mel Scale is constructed such that sounds of equal distance from each other on the Mel Scale, also "sound" to humans as they are equal in distance from one another.

In contrast to Hz scale, where the difference between 500 and 1000 Hz is obvious, whereas the difference between 7500 and 8000 Hz is barely noticeable.

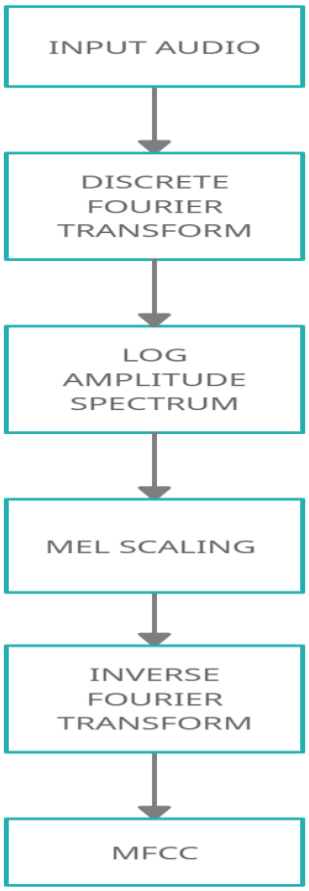


Fig:- Representing MFCC Flow Chart

THE CEPSTRUM:-

One way of evaluating periodic structures in a signal on different scales is to use the Fourier transform. Specifically, we can take the discrete Fourier transform (DFT) or the discrete cosine transform (DCT) of the log-spectrum, to obtain a representation known as the *cepstrum*. The name attempts to be an amusing reflection of the fact that this representation is a complicated rearrangement of time-frequency transforms. In technical terms, for a time signal *x(t)*, the cepstrum is defined as

$$F(x(t)) = F^{-1}\,[LOG\,(\,F\,[\,x(t)\,]\,)\,]$$

Here F represents Fourier Transform and F inverse represents Inverse Fourier Transform. A useful piece of information in the cepstrum is the harmonic structure of the log-spectrum.

MEL FREQUENCY CEPSTRAL COEFFICIENT: -

A classical approximation is to define the frequency-to-mel transform function for a frequency f as

$$M = 2595 * \log_{10}(1 + f/700)$$

The inverse transform can be readily derived as

$$F = 700\,(10^{m/2595} - 1)$$

SPEECH EMOTION RECOGNITION: -

A language-independent emotion recognition system based on Neural Network approach is implemented. Eight emotional classes were considered and ten subjects are selected. The speech utterances were recorded in English. 3120 speech utterances each delivered with one of eight particular emotions were used for training and testing. From these samples 936 utterances were selected for training the network and the rest were used for testing. A total of 180 prosodic features are extracted by analyzing the speech spectrogram. The Neural Network uses selected features achieved overall classification accuracy of 63.31% on the test set. I trained the model with MLP Classifier and achieved an accuracy of nearly 78% which is the best out of all the classifiers I used. I also used Random Forest Classifier to train the model and achieved an accuracy of 65%.

Refer below table for Classification Report

Classification Report Using Random Forest CLassifier:

	precision	recall	f1-score	support
angry	0.85	0.80	0.82	44
calm	0.71	0.91	0.80	44
disgust	0.58	0.42	0.49	26
fearful	0.55	0.51	0.53	47
happy	0.60	0.74	0.66	46
neutral	0.78	0.54	0.64	26
sad	0.67	0.61	0.64	54
surprised	0.43	0.48	0.45	25
accuracy			0.65	312
macro avg	0.65	0.63	0.63	312
weighted avg	0.66	0.65	0.65	312

Fig (a) :- Random Forest Classifier

Classification Report Using MLP CLassifier:

	precision	recall	f1-score	support
angry	0.91	0.87	0.89	45
calm	0.75	0.91	0.82	54
disgust	0.77	0.65	0.71	26
fearful	0.69	0.71	0.70	41
happy	0.90	0.80	0.85	45
neutral	0.80	0.57	0.67	21
sad	0.73	0.81	0.77	53
surprised	0.73	0.70	0.72	27
accuracy			0.78	312
macro avg	0.79	0.75	0.76	312
weighted avg	0.79	0.78	0.78	312

Fig (b) :- Multi Layer Perceptron Classifier

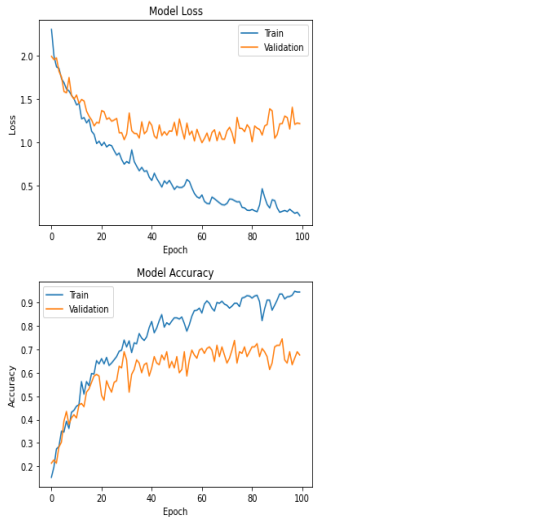


Fig (c) :- Neural Network

EVALUATION METRICS

Classifier	No. of Emotions	Speech Database	Average Accuracy
MLP	8	RAVDESS	78%
Neural Networks	8	RAVDESS	63.31%
Random Forest	8	RAVDESS	65%

APPLICATIONS

Speech emotion recognition is particularly useful for applications which require natural man–machine interaction such as web movies and computer tutorial applications where the response of those systems to the user depends on the detected emotion. It is also useful for in-car board system where information of the mental state of the driver may be provided to the system to initiate his/her safety. It can be also employed as a diagnostic tool for therapists. It may be also useful in automatic translation systems in which the emotional state of the speaker plays an important role in communication between parties. In aircraft cockpits, it has been found that speech recognition systems trained to stressed-speech achieve better performance than those trained by normal speech. The main objective of employing speech emotion recognition is to adapt the system response upon detecting frustration or annoyance in the speaker's voice.

CONCLUSION

In this paper, a survey of current research work in speech emotion recognition system has been given. Three important issues have been studied: the features used to characterize different emotions, the classification techniques used in research, and the important design criteria of emotional speech databases. There are several conclusions that can be drawn from this study.

The first one is that while high classification accuracies have been obtained for classification between high-arousal and low arousal emotions. Moreover, the performance of current stress detectors still needs significant improvement. The average classification accuracy of speaker-independent speech emotion recognition systems is less than 80% in most of the techniques.

Three classifiers have been tried for speech emotion recognition such as the Neural Networks, MLP and Random Forest. However, we see MLP performs better than rest.

FUTURE WORK

Most of the current body of research focuses on studying many speech features and their relations to the emotional content of the speech

utterance. New features have also been developed such as the TEO-based features. There are also attempts to employ different feature selection techniques in order to find the best features for this task. However, the conclusions obtained from different studies are not consistent. The main reason may be attributed to the fact that only one emotional speech database is investigated in each study. Most of the existing databases are not perfect for evaluating the performance of a speech emotion recognizer. There are some other problems for some databases such as the low quality of the recorded utterances, the small number of available utterances, and the unavailability of phonetic transcriptions. Therefore, it is likely that some of the conclusions established in some studies cannot be generalized to other databases. To address this problem, more cooperation across research institutes in developing bench mark emotional speech databases is necessary.

In order to improve the performance of current speech emotion recognition systems, the following possible extensions are proposed. The first extension relies on the fact that speaker-dependent classification is generally easier than speaker-independent classification. At the same time, there exist speaker identification techniques with high recognition performance such as the GMM-based text-independent speaker identification system proposed by Reynolds. Thus, a speaker-independent emotion recognition system may be implemented as a combination of a speaker identification system followed by a speaker-dependent emotion recognition system. It is also noted that the majority of the existing classification techniques do not model the temporal structure of the training data. The only exception may be the HMM in which time dependency may be modeled using its states. However, all the Baum–Welch re-estimation formulae are based on the assumption that all the feature vectors are statistically independent. This assumption is invalid in practice. It is sought that direct modeling of the dependency between feature vectors, e.g. through the use of autoregressive models, may provide an improvement in the classification performance.

REFERENCE LINKS

1. Multimodal Speech Emotion Recognition Using Audio and Text – Seunghyun Yoon, Seokhyun Byun, Kyomin Jung.
2. Emotion Recognition from Speech – Kannan Venkataramanan, Haresh Rengaraj Rajamohan.
3. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.
4. https://en.wikipedia.org/wiki/Multilayer_perceptron.
5. Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, 2009.
6. Rosenblatt, Frank. x. Principles of Neurodynamic: Perceptron's and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961
7. Davis, S. Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366
8. J. Nicholson, K. Takahashi, R. Nakatsu, "Emotion recognition in speech using neural networks", Neural Computing & Applications, 2000, Vol. 9, No. 4, pp. 290-296.
9. R. Banse, K. Scherer, "Acoustic profiles in vocal emotion expression", Journal of

Personality and Social Psychology, 1996, Vol.70, No. 3, pp.614-636.

10. M. Schubiger, “English intonation: its form and function”, Tübingen, M. Niemeyer Verlag, Germany, 1958.
11. J. O’Connor, G. Arnold, “Intonation of Colloquial English”, second ed., Longman, London, UK, 1973.
12. W. Campbell, “Databases of emotional speech”, in Proceedings of the International Speech Communication and Association (ISCA) ITRW on Speech and Emotion, 2000, pp. 34-38.
13. C. Lee, S. Narayanan, “Toward detecting emotions in spoken dialogs”, IEEE Transactions on Speech and Audio Processing,2005, Vol. 13, No. 2,pp. 293-303.
14. W. Campbell, “Databases of emotional speech”, in Proceedings of the International Speech Communication and Association (ISCA) ITRW on Speech and Emotion, 2000, pp. 34-38.
15. C. Lee, S. Narayanan, “Toward detecting emotions in spoken dialogs”, IEEE Transactions on Spee
16. Emily Mower, Maja J Matarić and Shrikanth Narayanan, “A Framework for Automatic Human Emotion Classification Using Emotion Profiles”,IEEE Transactions on Audio, Speech, and Language Processing, July 2011, vol. 19, No. 5, pp. 1057-1070.
17. Tomas Pfister and Peter Robinson,“Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis”,IEEE Transactions A
18. J. Hansen, D. Cairns, Icarus,“source generator based real-time recognition of speech in noisy stressful and Lombard effect environments”, Speech Commun., 1995, Vol. 16, No. 4, pp. 391- 422.
19. J. Ma, H. Jin, L. Yang, J. Tsai, “Ubiquitous Intelligence and Computing”, Third International Conference, UIC 2006, Vol. 4159.
20. T. Nwe, S. Foo, L. De Silva, “Speech emotion recognition using hidden Markov models”, Speech Commun., 2003, Vol. 41, pp. 603-623