Speech Emotion Recognition Using HMMs

*

1st Md. Sajeebul Islam Sk.
dept. of CSE
Brac University
Dhaka, Bangladesh
sajeebul.islam.sk@g.bracu.ac.bd

2nd Saikat Halder Tuhin
dept. of CSE
Brac University
Dhaka, Bangladesh
tuhin.halder@g.bracu.ac.bd

3rd Md. Golam Rabiul Alam
dept. of CSE
Brac University
Dhaka, Bangladesh
rabiul.alam@bracu.ac.bd

***Abstract*—This research used Ryerson University's Audio-Visual Database of Emotional Speech and Song (RAVDESS) audio recordings to analyze speech-based emotion recognition algorithms. Features including the Log-Mel Spectrogram, Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy were taken into account after the raw audio files had been pre-processed. Hidden Markov Models (HMMs) were employed to assess the importance of these features for the classification of emotions. Additionally, we note that the choice of audio features has a much greater influence on the outcomes than model complexity does on emotion recognition.**

***Index Terms*—Emotional Speech and Song audio, Log-Mel Spectrogram, MFCCs, HMMs**

I. Introduction

In the next five years, more individuals will engage with voice assistance machines than with their partner, according to a United Nations report. As Virtual Personal Assistants (VPAs) like Siri, Alexa, and Google Assistant, they serve a function of promptly and accurately responding to our inquiries and completing our requests. Although these helpers comprehend our orders, they are not skilled enough to gauge our emotions and respond appropriately. As a result, it is important to create an effective emotion detection system that can improve these assistants' abilities and transform the whole business [1].

Speech is a complex, dense mode of communication that is good at transferring information. It has two different kinds of information: linguistic and paralinguistic. While the latter relates to the implicit information such body language, gestures, facial expressions, tone, pitch, emotion, etc., the former refers to the verbal content, the underlying linguistic code. Paralinguistic traits can be used to understand a person's emotional state, gender, attitude, accent, and other factors [2]. Key characteristics of recorded speech can be used to systematically extract information, such as emotion. Since emotion colors ordinary human interactions, gathering this information would be crucial in enabling the virtual assistant and the user to have more natural dialogues.

Both continuous and discrete representations of emotion are often utilized. The emotion of a statement can be described as continuous values across a variety of psychological variables in the continuous representation. [3] states that "emotion can be characterized in two dimensions: activation and valence." Research has revealed that happiness, rage, and fear may be associated to high energy and pitch in speech, but sorrow can be linked to low energy and sluggish speech. Activation is the "amount of energy required to express a certain emotion." Since both enhanced activation and emotions like anger and happiness might be present, valence provides greater depth and aids in separating them. Emotions can be discretely stated in the discrete representation as distinct categories, such as anger, sadness, happiness, etc.

An emotion identification system's performance only depends on characteristics and representations taken from the audio. They may be roughly divided into aspects that are time-based and frequency-based. The advantages and disadvantages of these traits have been well researched. There isn't a single audio feature that excels at all jobs involving sound signal processing.

Several studies have been done in the past to determine speech emotions for various languages and accents.Using Hidden Markov Models (HMM) and Support Vector Machines (SVM), Chenchah and Lachiri [4] investigated the efficacy of Mel-Frequency Cepstral Coefficients and Linear Frequency Cepstral Coefficients (LPCC) in recognizing the emotions. The goal of this study is to select the most effective audio feature and HMM architecture implementation for speech emotion identification. The "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" dataset was used for the studies [5].

II. Datasets

A. Data Selection

Designing a SER (Speech Emotion Recognition) involves three basic steps: selecting an emotional speech database, selecting features from audio input, and using classifiers to identify emotions. A verified multi-modal database of emotional speech and music is the RAVDESS dataset. The 24 professional actors that make up this gender-balanced

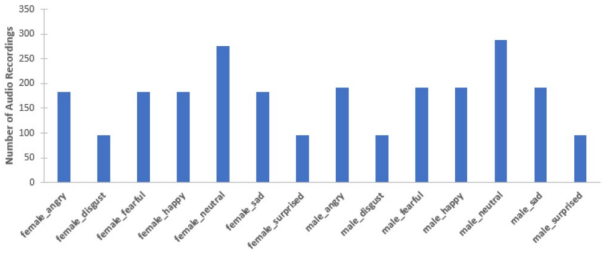

Fig. 1. Data Distribution across gender and emotion

database "perform 104 different vocalizations with emotions that include: happy, sad, angry, fearful, surprise, disgust, calm, and neutral." For each emotion, each actor performed two sentences, such as "Kids are talking by the door" and "Dogs are sitting by the door." For each emotion, with the exception of neutral (normal only), these remarks were also recorded in two distinct emotional intensities: normal and strong. The actors voiced each word twice. There are 1012 music utterances and 1440 speaking utterances combined [1].

Given that it is free of gender bias and includes a wide spectrum of emotions at various intensities, the RAVDESS dataset has a very rich nature. Other datasets, like Toronto Emotional Speech Set [6] (TESS) and Surrey Audio-Visual Expressed Emotion [7] (SAVEE), contained solely the audio of male and female performers, respectively. We also note that the RAVDESS dataset does not exhibit any class-imbalance issues because it is evenly divided across all emotion classes (15%).

Additionally, the RAVDESS dataset's developers conducted comprehensive validation and reliability testing. 247 naive individuals were asked to rate the "category of the emotion, strength of the emotion, and genuineness of the emotion" from a "pseudo-randomly chosen set of 298 stimuli, consisting of 174 speech and 124 song presentations" [5].

B. Data Preparation

Each audio file has a 7-part number identity that stands for the actor, voice channel, modality, emotion, emotional intensity, statement, and repetition. There was a pattern to the naming, with odd actors designating male and female sex, respectively. We turned all of this data by extracting it from the file names. The emotion that was assigned to the audio recording is the target variable.

C. Data Cleaning

Each recording lasted for around 3 seconds. Silences in audio recordings were cut off at the beginning and finish. Even though the audios were well captured, the data had very little noise patterns. To remove the noise, we experimented with a number of signal processing approaches, including filtering and voice-activity detection (VAD). A well-known noise reduction approach called spectral subtraction [8] seeks to eliminate background noise by deducting an estimate of the noise spectrum from the noisy speech spectrum. To remove the noise from the damaged signal and offer a clear representation of the underlying signal, we employ Wiener filtering [9]. After deployment, we saw a sizable boost in dataset quality without losing any crucial audio information.

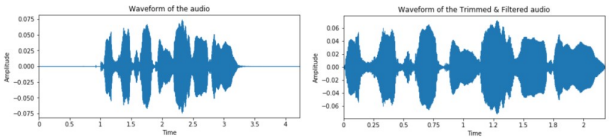

Fig. 2. Waveform of pure audio pre and post data cleaning

III. Feature Engineering and Modelling

A. Feature Selection

Time-domain features and frequency-domain characteristics are the two primary categories into which audio features may be divided. The short-term energy of the signal, zero crossing rate, maximum amplitude, lowest energy, and energy entropy are examples of time-domain properties. These characteristics are relatively simple to extract and offer a more straightforward method of audio signal analysis. In the presence of sparse data, frequency domain characteristics reveal more complex patterns in the audio signal, which may enable us to determine the emotion that underlies the signal. Spectograms, Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, spectral rolloff, spectral entropy, and chroma coefficients are examples of frequency-domain properties [10].

B. Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) are coefficients which capture the envelope of the short time power spectrum. MFCCs are computed as follows:

- The audio signal is cut into short frames to ensure the stationarity of the audio signal
- For each frame, a periodogram, which identifies the frequencies present in the frame, is estimated
- A Mel filter bank, which merges periodogram bins by summing up the energy, is performed to get an idea of how much energy exists in various frequency regions
- Since human perception of sound does not follow a linear scale, the spectral energies are transformed into log scale to transform the non-linear frequency pattern to a linear scale
- Generally these filterbanks overlap with each other resulting in energies across frequency band being correlated with each other. A discrete cosine transform (DCT) decorrelates these energies

C. Hidden Markov Models

For voice recognition tasks like emotion detection, HMMs were frequently utilized [11]. HMMs operate on the presumption that observations originate from a set of hidden states,

each of which has transitions that adhere to the Markov principle. Generative models include HMMs. The fact that HMM perform badly when modeling non-linear data is a severe drawback. We used Gaussian HMMs Model. The smallest discrete sound, or the hidden states in these HMMs, are called phones, and we categorize using the highest chance that a series of phones will belong to a certain class.
One HMM is trained for each class using data from that class's many hidden states (the number of HMMs is equal to the number of emotion classes). In testing, the likelihood probability p(D—c) of the utterance given class is computed for all the classes, and the label is equal to argmaxcp(D—c). D is the utterance, and c is the class.

IV. Experiments

A. Data

With a window size of 10 ms and a hop length of 5 ms, MFCCs were retrieved. With a window size and hop length of 0.014 sec and 0.0035 sec, respectively, 128 Log-Mel spectrogram features were recovered from the input audios.

B. Train, Validation and Test Data

Speaker-dependent and speaker-independent trials are the two primary categories under which emotion recognition studies fall. In the training, validation, and test datasets of speaker-dependent studies, there are audio examples of the same actor displaying various moods. This indicates that the dataset was divided at random and that the training set may have been biased by the over-fitting of one actor. Therefore, randomly splitting is not recommended for this activity since it might result in data leaking. Speaker-independent experiments, on the other hand, use audio examples from several actors as the training, validation, and test data. Training models without consideration of the speaker makes sure they are reliable and can recognize emotions from any performer. Therefore, audio examples of actors 1–20, 21–22, and 23–24 were used for training, validation, and testing, in that order. Since there are examples of both male (odd actors) and female (even actors) actors in both the validation and testing datasets, we have made sure that the models are not geared to a certain gender.

C. Methodology

Since the data are fairly evenly distributed, accuracy may be used to compare the performance of different models. Unweighted accuracy was therefore chosen as the model selection metric. All of the input audios were compressed into a 3 second length for training models with fixed size inputs.
The depiction of MFCCs and the Log-Mel spectrogram demonstrate how little information is there in MFCCs at higher frequencies. So, in this investigation, we anticipate that log-mel spectrogram features will produce superior outcomes. Utilizing MFCCs and Log-melspectrogram features, Gaussian HMM was implemented. Around accuracy was provided by HMMs using MFCCs and log-mel spectrogram models. This is most likely a result of HMMs' generally low representational power.

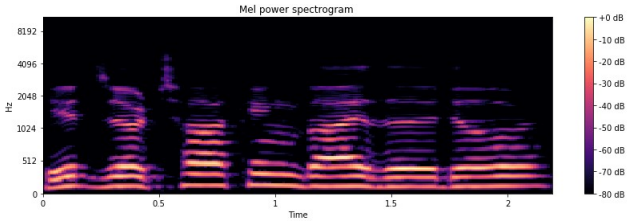

Fig. 3. Mel Power Spectrogram

V. Results and Conclusions

With designed characteristics like MFCCs and Log-mel spectrograms, we don't get significantly better results than the raw audio input, which is probably because there isn't enough data.

predictedvalues	actualvalues
angry	24
calm	25
disgust	18
fearful	8
happy	38
neutral	18
sad	14
surprised	63

Fig. 4. Actual Values of the Dataset

actualvalues	predictedvalues
angry	32
calm	32
disgust	16
fearful	32
happy	32
neutral	16
sad	32
surprised	16

Fig. 5. Predicted Values by Machine

Despite the fact that MFCCs are often employed features for speech-based emotion identification, we discover that the Log-Mel spectrogram features do this task unquestionably superior.

```
Normalized confusion matrix
[[0.34 0.   0.06 0.22 0.28 0.   0.03 0.06]
 [0.03 0.34 0.16 0.16 0.03 0.12 0.12 0.03]
 [0.31 0.   0.56 0.06 0.06 0.   0.   0.  ]
 [0.06 0.06 0.12 0.41 0.19 0.   0.06 0.09]
 [0.09 0.09 0.06 0.25 0.22 0.   0.   0.28]
 [0.   0.06 0.   0.25 0.25 0.12 0.19 0.12]
 [0.06 0.25 0.19 0.12 0.12 0.09 0.16 0.  ]
 [0.   0.   0.12 0.19 0.   0.19 0.06 0.44]]
```

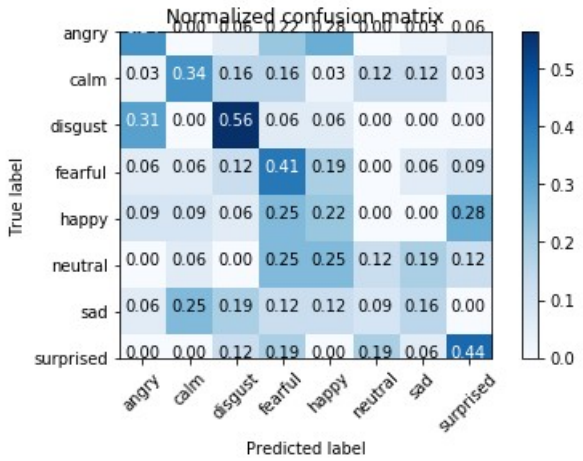

Fig. 6. Confusion matrix of the Model

Here we see that disgust to disgust captured almost 56%. Surprised to surprised almost close to 45%. Overall angry,calm,disgust,fearful,surprised capturing is pretty decent. But happy is closed to surprised, neutral close to fearful and sad close to calm, that are the mismatch. If the training data is high then accuracy will increase and also more care about for annotation to get high accuracy.

REFERENCES

[1] K. Venkataramanan and H. R. Rajamohan, "Emotion recognition from speech," *arXiv preprint arXiv:1912.10458*, 2019.

[2] Y. Yamashita, "A review of paralinguistic information processing for natural speech communication," *Acoustical Science and Technology*, vol. 34, no. 2, pp. 73–79, 2013.

[3] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern recognition*, vol. 44, no. 3, pp. 678–693, 2011.

[4] F. Chenchah and Z. Lachiri, "Acoustic emotion recognition using linear and nonlinear cepstral coefficients," *Int. J. Adv. Comput. Sci. Appl*, vol. 6, no. 11, pp. 1–4, 2015.

[5] S. Livingstone and F. Russo, "Ryerson audiovisual database of emotional speeches and songs (ravdess): a dynamic, multimodal set of north american english face and voice expressions," *Plos One*, vol. 13, no. 5, p. e0196391, 2018.

[6] K. Dupuis and M. K. Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set," *Canadian Acoustics*, vol. 39, no. 3, pp. 182–183, 2011.

[7] S. Haq, P. J. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition.," vol. 2009, pp. 53–58, 2009.

[8] T. Fux and D. Jouvet, "Evaluation of pncc and extended spectral subtraction methods for robust speech recognition," pp. 1416–1420, 2015.

[9] N. Wiener, N. Wiener, C. Mathematician, N. Wiener, N. Wiener, and C. Mathématicien, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, vol. 113. MIT press Cambridge, MA, 1949.

[10] G. S. Tomas, *Speech Emotion Recognition Using Convolutional Neural Networks*. PhD thesis, PhD thesis, Institute of Language and Communication, Technical University of . . . , 2019.

[11] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," vol. 2, pp. II–1, 2003.