Speech Emotion Recognition using Convolutional Neural Networks (CNN)

Gourav Verma

Data Science, Bellevue University

DSC-680 T302: Applied Data Science

Fadi Alsaleem

04/10/2021

Abstract

In this project, Speech Emotion Recognition (SER) is performed. To achieve this Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) data from ZENODO was used [1]. People express their emotions directly or indirectly through their speech, facial expressions, gestures or writings. Many different sources of information, such as speech, text and visual can be used to analyze emotions. The RAVDESS database contains 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent. Speeches include neutral, calm, happy, sad, angry, fearful, surprise, and disgusted expressions. My focus for the research is only on acoustic representation, with the assumption that audio speech signal contains sufficient human emotion information to recognize and retrieve. Several two-dimensional features such as Mel-frequency cepstral coefficients (MFCC), Short-Time Fourier Transform (STFT), Zero-Crossing Rate, Log-Mel Spectrogram, and Spectral Centroid are used to represent features of speech emotions. Convolutional Neural networks (CNN) was used to recognize human emotions from audio recordings. I achieve a recognition rate of approximately 60% when testing eight emotions for both male and female.

Keywords: Speech emotion recognition, Speech Analysis, Mel-frequency cepstral coefficient, CNN

Table of Contents

Introduction

Emotion recognition from speech signals has been a research topic for many years. Among all the ways of the communication speech signal is one of the fastest. Due to the subjective nature of emotions, recognition of emotion is natural for humans but very difficult for machines. The increase in the number of people interacting with voice assistants such as Alexa, Siri, and Google Assistant, has created challenges to understand human commands in different emotions [12]. Therefore, continuous development is happening to create an efficient human emotion recognition system that can enhance the performance of these voice assistance and revolutionize the whole industry. The speech of a human being is the most natural way of expression. Recognition of emotion was not felt so important till we were exposed to other communication forms like text messages and emails. Emojis of often used to express the emotion associated with messages. SER system is a collection of different methodologies that process and classify speech signals to identify their embedded emotions.

We can categorize features of speech into three classes namely, the lexical features (Vocabulary used), the visual features (expression of the speaker), and the acoustic features (properties of sound like pitch, tone, jitter, etc.). In this project, I will be performing emotion detection using acoustic features. This research will be capitalizing on the fact that voice often reflects underlying emotion through tone and pitch.

“Emotion can be characterized in two dimensions: activation and valence.” Activation is the “amount of energy required to express a certain emotion” [13]. The research has also shown that joy, anger, and fear can be linked to high energy and pitch in speech, whereas sadness can be linked to low energy and slow speech. Valence gives more nuance and helps distinguish between emotions like being angry and happy since increased activation can indicate both. In the discrete representation, emotions can be discretely expressed as specific categories, such as angry, sad, happy, etc.

SER (Speech Emotion Recognition) is used in a call center for classifying calls according to emotions and can be used as the performance parameter for conversational analysis thus identifying the unsatisfied customer, customer satisfaction, and so on. for helping companies improving their services. It can also be used in-car board system based on information of the mental state of the driver can be provided to the system to initiate his/her safety preventing accidents to happen. With the help of emotion recognition, computers can make better decisions to help users. With the increase in popularity of robotic research, emotion recognition will also help making human–robot interaction more natural.

Prior Studies

The efficiency of an SER completely relies on the information extracted from the audio. Voice signal features are categorized as time-based and frequency-based features. To solve the problem in hand suitable feature is used. Previous studies [2-9] show that the choice of feature is much more important than the model architecture.

Various studies have been performed in the past to recognize emotion from voice signals. Venkataraman and Rajamohan, 2019 [2] studied the performance of Log-Mel Spectrogram and Mel-Frequency Cepstral Coefficients (MFCCs). They compared emotion classifications by applying methods such as Long Short-Term Memory (LTSM), Convolutional Neural Networks (CNN), and Hidden Markov Models (HMMs). The CNN using Log-Mel Spectrogram features yielded 68% accuracy. Other recent approaches [3-9] were also used methods like artificial neural networks, convolutional networks, and LSTM. These are the new replacements for HMM and SVM architectures especially in speech emotion recognition and sentiment analysis.

Dataset

Data Selection

For this project, I will be using a portion of data available from ZENODO [1]. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and the 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). A portion of the RAVDESS from Kaggle [10] which contains 1440 files: 60 trials per actor x 24 actors = 1440, was used for this research.

Bias and Limitations

There are certain limitations of the RAVDESS data. Firstly, the data is rooted strongly in the language, cultural background, and accent. The model trained with this English dataset will not be able to correctly recognize other language speech emotions. We have selection bias because the dataset was created using 24 English-speaking actors from Toronto, Canada, who exhibit strong North American characteristics. Secondly, the speeches were created using trained actors, rather than natural instances of emotions.

Data Preparation

Each of the 1440 audio files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). 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 24. Odd numbered actors are male, even numbered actors are female).

Filename example: 03-01-06-01-02-01-12.wav

1. Audio-only (03)
2. Speech (01)
3. Fearful (06)
4. Normal intensity (01)
5. Statement "dogs" (02)
6. 1st Repetition (01)
7. 12th Actor (12)
   Female, as the actor ID number is even.

Data Cleaning

Preliminary visualization shows that voice records include silences both at the start and at the end. For this research I used the data as it is without trimming the silences. All the audio files were read into a dataframe. Based on the file naming gender, emotion and actor column were created. For classification gender and emotion column were merged to form 16 distinct categories, 8 emotions for each men and women [Table 1].

Feature Selection and Modelling

Feature Selection

Audio features are broadly classified into two domains: Time-domain [Fig.1] and frequency domain [Fig.2]. Time-domain features are Amplitude envelope, Root-mean-square energy, and Zero crossing rate. These are very easy to extract and provide an easier way to analyze audio signals. The frequency-domain features are Band energy ratio, spectral centroid, and spectral flux. Most sound is defined for its frequency and in time-based features, we will not have any frequency information. Similarly, with the frequency-domain representation, we will not be able to see changes in sound for the time. To overcome this issue, we have time-frequency features such as spectrogram, Mel-spectrogram, and Constant-Q transform. During EDA and extensive analysis of each feature was analyzed. However, for the model creation Mel Spectrogram (MFCC), Mel-Spectrograms were used.

MFCC

Mel-frequency cepstral coefficient (MFCC) is a well-known good feature which you can slice and dice in many ways [Fig.4]. Mel-frequency Cepstrum is a representation of the short-term power spectrum of a sound by transforming the audio signal through a series of steps to mimic the human cochlea. Pitch is one of the characteristics of a speech signal and is measured as the frequency of the signal. Mel scale is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency to match more closely what the human ear can hear (humans are better at identifying small changes in a speech at lower frequencies). This scale has been derived from sets of experiments on human subjects [14]. Mel-Frequency Cepstral Coefficients (MFCC) are coefficients that capture the envelope of the short-time power spectrum. MFCCs are commonly derived as follows: [wiki, 15]

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the Mel frequencies.
4. Take the discrete cosine transform of the list of Mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

For this research I have extracted Forty MFCCs for each audio file using librosa function *librosa.feature.mfcc.* All the MFCCs were added to the main dataframe as forty different columns.

Mel-Spectrograms

A spectrogram is a time vs frequency representation of an audio signal [Fig.3]. Different emotions exhibit different patterns in the energy spectrum. Mel-spectrogram is a representation of the audio signal on a Mel-scale. The logarithmic form of Mel-spectrogram helps understand emotions better because humans perceive sound on a logarithmic scale. Therefore, the log Mel-spectrogram corresponds to the time vs. log Mel-frequency representation, which was obtained in step 4 during MFCC computation. Refer to appendix (Section 1) for the plots of log Mel-spectrogram of different emotions in the dataset.

We observe that MFCC features, and Log-Mel Spectrograms can be represented as images, and these images can be fed to deep learning techniques such as CNN, RNN networks to classify the emotion of audio.

Models

Emotion recognition is the part of speech recognition which is gaining more popularity and need for it increases enormously. Although there are methods to recognize emotion using machine learning techniques, this project attempts to use deep learning to recognize the emotions from data. There are several algorithms and methods used for SER in prior research. Each of these models are focused on specific feature and had their advantages and shortcomings. The most effective method is to use neural networks which seems to provide higher accuracy than others. As I mentioned earlier in this paper, as audio signal can have many defining features and among those MFCC is highly capable of representing an audio signal completely. In this paper I have created CNN model with multiple layers using MFCC feature.

Librosa

Librosa is a Python library for analyzing audio and music. It has a flatter package layout, standardizes interfaces and names, backwards compatibility, modular functions, and readable code.

Convolutional Neural Networks (CNN)

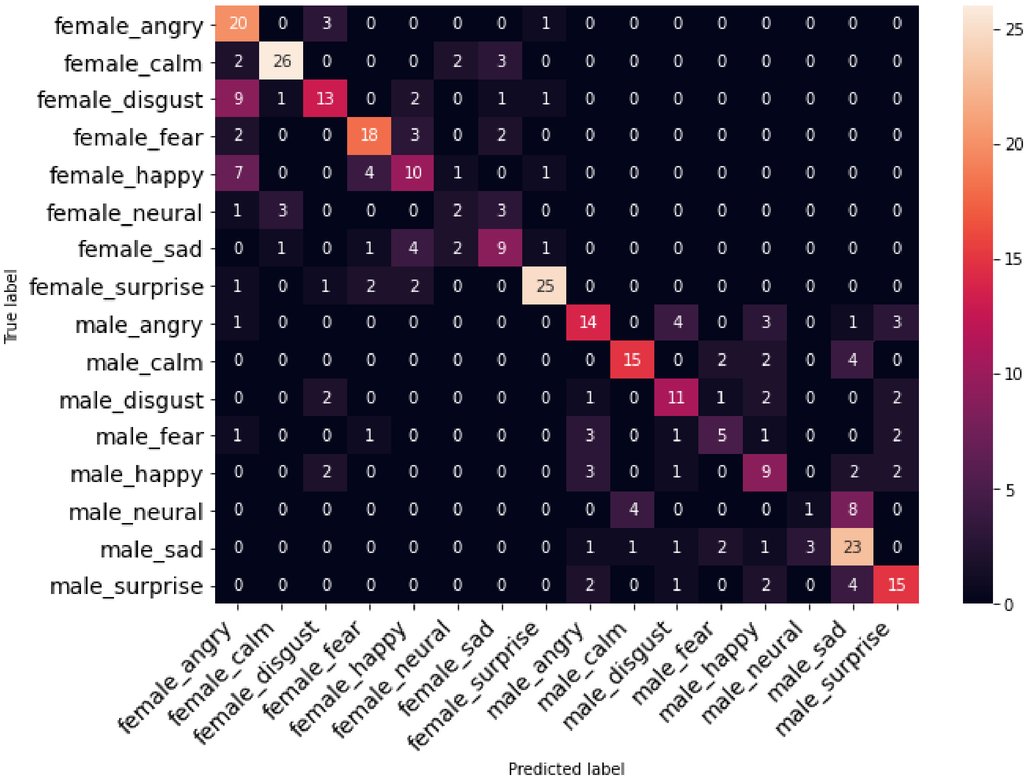
The tremendous strides made in the recent years in image recognition tasks is in large part due to the advent of Convolutional Neural Networks (CNNs). CNNs are good at automatically learning relevant features from high dimensional input images. CNNs use shared kernels (weights) to exploit the 2D correlated structure of image data. Max-pooling is added to CNNs to introduce invariance wherein only the relevant high dimensional features are learned for various tasks like classification, segmentation etc. Surprisingly, CNNs are good at audio recognition tasks. One positive perspective of CNNs is the ability to learn features from high-dimensional input data; however, it also learns features from small variations and distortion appearance that leads to the large storage requirement at the time of development. Hence, in CNNs, there usually exists a layer of convolution followed by a down sampling mechanism. CNN’s superior representation power helps the model learn the underlying patterns effectively from short timeframes resulting in state-of-the-art performance in Speech based Emotion recognition systems. Multi-layer 1D CNNs were trained on raw audio to identify patterns in the sound waveform.

Methodology

As the research data was almost equally distributed, I used accuracy as valid matric to evaluate model’s performance. Complete dataset was split into train and test ratio of 75-25%. The 8-layer 1D CNN model was trained for 100 epochs with batch size of 16. ADAM optimizer was used with default parameters. Model was saved by monitoring the accuracy on test set. The model was implemented on 16 gender + emotion class classification. The model was trained on MFCC feature and the resulting model was able to make classification with 60% accuracy.

Conclusion

I conducted feature focused analysis with only MFCC and 1D CNN. Achieved good results with widely used engineered feature MFCC for speech emotion recognition. High performance was seen with gender specific emotion classification [Table 2]. This is due to the difference in the pitch and energy in the average male and average female voice. This was seen in different plots. I tested the model on RAVDESS audio dataset and achieved 60% accuracy. Future studies an explore other audio features such as Log-mel spectrogram. Also, other neural networks like 2D CNN, 3D CNN, and HMM could be used to test-train and choose between best performing model.



	female_angry	female_calm	female_disgust	female_fear	female_happy	female_neural	female_sad	female_surprise	male_angry	male_calm	male_disgust	male_fear	male_happy	male_neural	male_sad	male_surprise
female_angry	20	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0
female_calm	2	26	0	0	0	2	3	0	0	0	0	0	0	0	0	0
female_disgust	9	1	13	0	2	0	1	1	0	0	0	0	0	0	0	0
female_fear	2	0	0	18	3	0	2	0	0	0	0	0	0	0	0	0
female_happy	7	0	0	4	10	1	0	1	0	0	0	0	0	0	0	0
female_neural	1	3	0	0	0	2	3	0	0	0	0	0	0	0	0	0
female_sad	0	1	0	1	4	2	9	1	0	0	0	0	0	0	0	0
female_surprise	1	0	1	2	2	0	0	25	0	0	0	0	0	0	0	0
male_angry	1	0	0	0	0	0	0	0	14	0	4	0	3	0	1	3
male_calm	0	0	0	0	0	0	0	0	0	15	0	2	2	0	4	0
male_disgust	0	0	2	0	0	0	0	0	1	0	11	1	2	0	0	2
male_fear	1	0	0	1	0	0	0	0	3	0	1	5	1	0	0	2
male_happy	0	0	2	0	0	0	0	0	3	0	1	0	9	0	2	2
male_neural	0	0	0	0	0	0	0	0	0	4	0	0	0	1	8	0
male_sad	0	0	0	0	0	0	0	0	1	1	1	2	1	3	23	0
male_surprise	0	0	0	0	0	0	0	0	2	0	1	0	2	0	4	15

Figure 1 Confusion Mertix of 1D CNN Model

References

1. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS One, 13*(5), e0196391.
2. Venkataramanan, K., & Rajamohan, H. R. (2019). Emotion recognition from speech. *arXiv preprint arXiv:1912.10458*.
3. Pareek, M., Sharma, K., Sharma, H., & Kumar, A. (2021). Emotion Recognition System using Speech.
4. Munot, R., & Nenkova, A. (2019, June). Emotion Impacts Speech Recognition Performance. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 16-21).
5. Mustaqeem, & Kwon, S. (2019). A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. *Sensors (Basel, Switzerland), 20*(1).
6. Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors (Basel, Switzerland), 21*(4).
7. Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(1), 235-238.
8. Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014, November). Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 801-804).
9. Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., & Vepa, J. (2018, September). Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In *Interspeech* (pp. 3688-3692).
10. Livingstone, S. (2019, January 19). RAVDESS emotional speech audio. Retrieved March 22, 2021, from https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio
11. Python mini project - Speech emotion recognition With librosa. (2021, March 14). Retrieved March 22, 2021, from https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/

12. United Nations Educational, Scientific, and Cultural Organization. (2019). I’d blush if I could: closing gender divides in digital skills through education (Programme Document GEN/2019/EQUALS/1 REV 2). Retrieved from http://unesdoc.unesco.org/images/0021/002170/217073e.pdf

13. Ayadi, M. E., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587. DOI: 10.1016/j.patcog.2010.09.020

14. https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

15. https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

Appendix

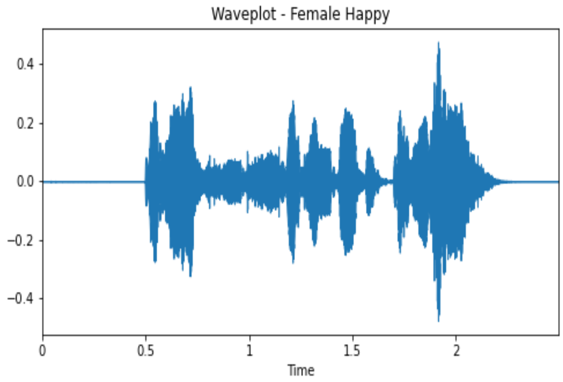


Figure 2: Waveplot

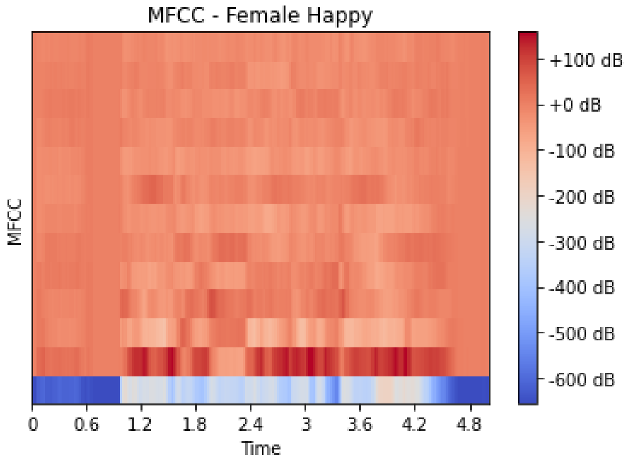


Figure 5: MFCC

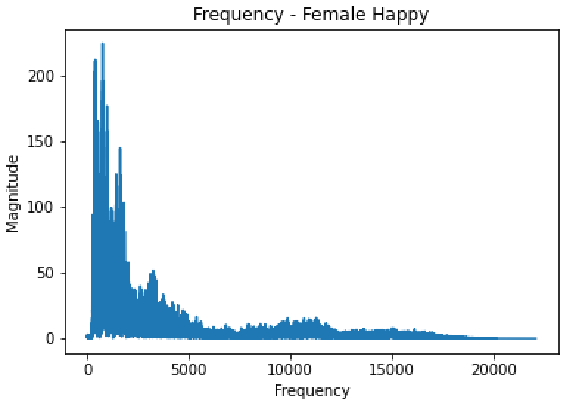


Figure 3: Frequency Spectrum

Table 1: Categories Count

Label	Count
female_calm	96
male_sad	96
male_disgust	96
female_fear	96
female_sad	96
female_happy	96
male_angry	96
female_disgust	96
male_fear	96
female_angry	96
male_happy	96
male_calm	96
female_surprise	96
male_surprise	96
male_neural	48
female_neural	48

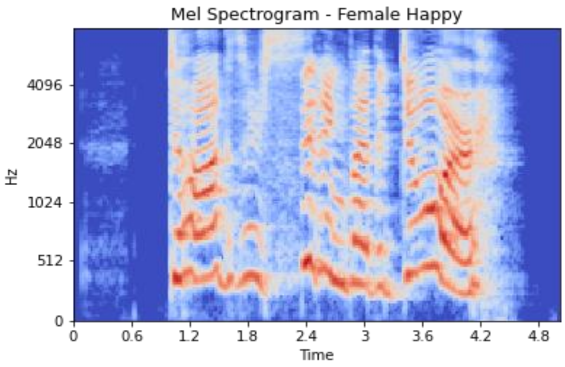
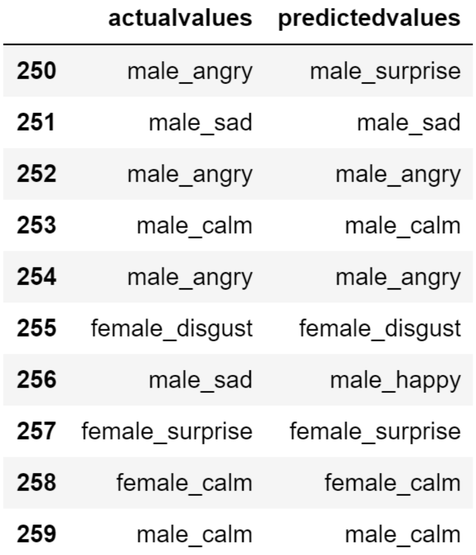


Figure 4: Mel Spectrum

Table 2 Predication



	actualvalues	predictedvalues
250	male_angry	male_surprise
251	male_sad	male_sad
252	male_angry	male_angry
253	male_calm	male_calm
254	male_angry	male_angry
255	female_disgust	female_disgust
256	male_sad	male_happy
257	female_surprise	female_surprise
258	female_calm	female_calm
259	male_calm	male_calm