Final Project
Speech Emotion Recognition

Shunit Avni

October 12, 2020

1 Introduction

In this project the problem we are dealing with is the classification of audio clips into emotions.

The idea for the project came from the world of sales. We wanted to help the seller by giving him an indication of what situation the customer is in and thereby allow him to increase sales.
This tool can also be useful in the world of criminal investigations or emergency support centers where from the audio section it will be possible to classify the emotional state of the person.

The goal of this project was to create an emotion recognition application based on models from the world of deep learning. The user enters as an input an audio segment in which he speaks and the machine knows how to classify the emotion expressed in the audio segment. We did in-depth research both around working with audio and around the topic of emotion.

So, lets start with a bit of background emotions are biological states associated with the nervous system brought on by neurophysiological changes variously associated with thoughts, feelings, behavioural responses, and a degree of pleasure or displeasure. There is currently no scientific consensus on a definition. Emotions are often intertwined with mood, temperament, personality, disposition, creativity and motivation.

In our work, we concentrated on identifying what is known in the professional literature as Expression.

Expression - facial and vocal expression almost always accompanies an emotional state to communicate reaction and intention of actions.
facial and vocal expression almost always accompanies an emotional state to communicate reaction and intention of actions.
`With focus on vocal expression of curse.`

So we can say that emotion is an abstract thing, and so we found it appropriate to classify emotion with the tone of a sentence, that is, to refer to the intonation in which things were said. In fact with the help of the tone reference we can make a better generalization of our model and classify emotion regardless of what language the things were said in.

2 Audio

Audio is defined as anything related to sound in terms of receiving, transmitting or reproducing or its specific frequency.

An audio signal is representation of sound of frequencies corresponding to sound waves that can normally be heard by the human ear. Audio signals include speech but also music and all types of sounds.
In this process by which we automatically assign an individual item to one of a number of categories or class, based on its characteristic.

In our case

1. The items are audio signals ('wav.' files).

2. Their characteristics are the features we extract from them. (MFCC,chroma ..)

The complexity lies in finding an appropriate relationship features that describes well each audio.

2.1 Features

The first step in any automatic speech recognition system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise etc.

2.2 Mel Spectrogram

Well first let's start with the mel. The term mel comes from the word 'melodic' and the mel scale is intended to regularize the intervals between notes. A mel is a number that corresponds to a pitch, similar to how a frequency describes a pitch.If we consider a note, A4 for example, its frequency is 440 hz. If we move up an octave to A5 its frequency doubles to 880 hz, and doubles again to 1760 at A6.The problem is that the human ear doesn't hear that way. The difference between two notes feels the same whether we jump from C to D or from F to G. But the logarithmic relationship gives different hz values for these different intervals. A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time.

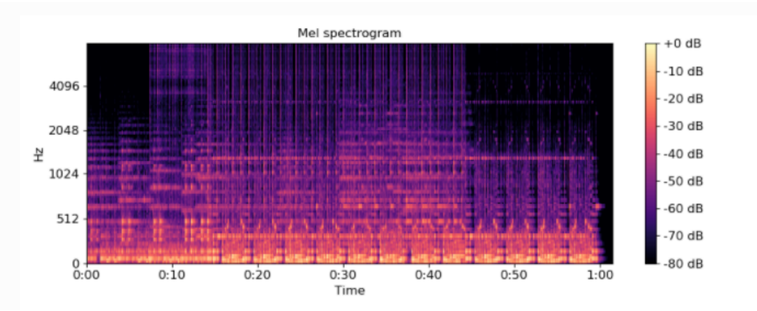

Figure 1: Example of a mel spec

2.3 MFCC

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since.

MFCCs are commonly derived as follows:[**1**]

1. Frame the signal into short frames.
2. For each frame calculate the periodogram[**2**] estimate of the power spectrum.
3. Apply the mel filterbank to the power spectra, sum the energy in each filter.
4. Take the logarithm of all filter-bank energies.
5. Take the DCT of the log filterbank energies.
6. Keep DCT coefficients 2-13, discard the rest.

Because the human ear is more sensitive to some frequencies than others, it's been traditional in speech recognition to do further processing to this representation to turn it into a set of Mel-Frequency Cepstral Coefficients, or MFCCs for short. This is also a two-dimensional, one-channel representation so it can be treated like an image too. If you're targeting general sounds rather than speech you may find you can skip this step and operate directly on the spectrograms.

2.4 Chroma

The chroma feature is a descriptor, which represents the tonal content of a musical audio signal in a condensed form. Therefore chroma features can be considered as important prerequisite for high-level semantic analysis, like chord recognition or harmonic similarity estimation. A better quality of the extracted chroma feature enables much better results in these high-level tasks. Short Time Fourier Transforms and Constant Q Transforms are used for chroma feature extraction.

This feature compute a chromagram from a waveform or power spectrogram. Which is a chroma vector (Wikipedia) (FMP, p. 123) that is a typically a 12-element feature vector indicating how much energy of each pitch class, C, C, D, D, E, ..., B, is present in the signal.

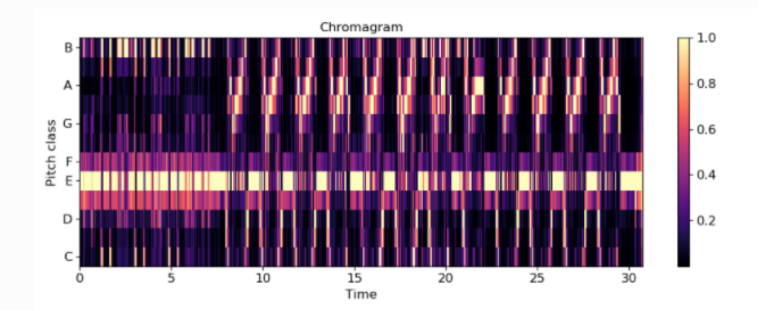

Figure 2: Example of a chromagram

2.5 Tonnetz

This feature gives us the pitch space (tonation) defined by a network of connections between the musical notes that are heard in the audio, close harmonic connections will be modeled as having a short distance between them and vice versa.

2.6 Spectral contrast

This feature gives us the difference between the frequencies within a certain time window.
When it refers within a defined time window to the peak point and the valley point (a type of local minimum) and the difference between them.

For further references see :
Jiang, Dan-Ning, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. “Music type classification by spectral contrast feature.” In Multimedia and Expo, 2002. ICME‘02. Proceedings. 2002 IEEE International Conference on, vol. 1, pp. 113-116. IEEE, 2002.

3 Architecture

When I started working on our model, I had no direction on what could give us good performance, In the research and information gathering stage we conducted many experiments that eventually led us to four key architectures that I will now present. Each of them uses different tools, we tried to test what architecture will bring us to the best performance.

3.1 VGG

1. Network Information
   This network is a modifed version of Visual Geometry Group's "VERY DEEP CONVOLUTIONAL NETWORKS" (VGG) as presented in their respective article

   - Very Deep Convolutional Networks for Large-Scale Image Recognition.
     - https://arxiv.org/pdf/1409.1556.pdf

   We have decided to use a variant of the architecture which can be trained in reasonable time yet produce state-of-the-art results.

   As input the network takes (1,161,101) which it transmits in 9 layers of convolution with 3X3 filter that eventually connects to a single fully connected layer. Differentiate from the original VGG11 network that ends in 3 layers of fully connected. Finally we run on the last layer softmax which aims to make the numbers Probabilities for each class.

   1.1 Hyper Parameters
   All hyper Parameters were chosen through validation tuning where consideration also included memory complexity and run-time complexity and not only test accuracy. our model runs on CUDA and on CPU and can start predicting after less than 25 minutes.

1. Epochs : 80 - Perfect time to stop before overfitting occurs and when time complexity is still accepted.
2. Optimizer : Adam -Worked better than all other options, maybe a custom one with Conjugate Directions may perform better.
3. Learning Rate : 0:001 - Standard learning rate, usually works well.
4. Batch Size : 5
5. Dropout rate = 0.5 - Standard drop out rate, works well.
6. Kernel Size = 3 - Chosen specifically for the VGG Architecture.
7. Stride = 1 - Chosen specifically for the VGG Architecture.
8. Padding = 1 - Chosen specifically for the VGG Architecture.
9. Activation: RELU - Chosen specifically for the VGG Architecture.

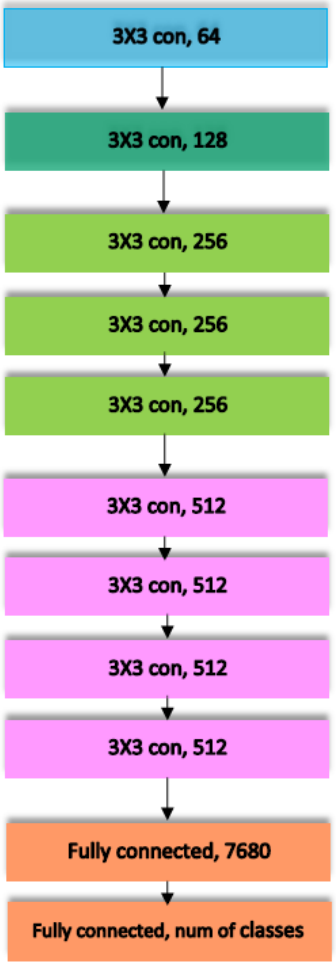

Figure 3: VGG Model Structure

3.2 Hierarchical architecture

The idea of this model came as a result of seeing that previous models tended to be more confused between specific classes. We first began to establish the distinction that our model makes between 2 classes that in the real world are considered as one extreme of the other - happy and angry.

In the initial stage we make a separation into 2 departments each of which unites 2 additional departments. At this initial stage we are actually advancing our model When every step in the tree is basically a new learning problem. We attacked the problem this time from the bottom up.

3.2.1 Hyper Parameters

1. Classifier : MLPClassifier
2. Optimizer : Adam -Worked better than all other options.
3. Learning Rate : 0:001 - Standard learning rate, usually works well.
4. Batch Size : 256
5. Momentum = 0.9 - Standard drop out rate, works well.
6. Activation: RELU - Chosen specifically for the VGG Architecture.
7. Maximum number of iterations: 500.

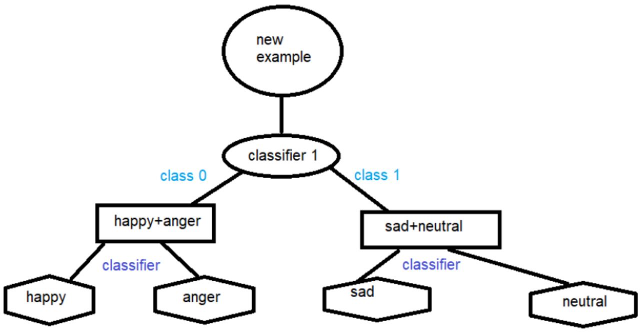

Figure 4: Hierarchical Model Structure