POLYTECHNIC SCHOOL OF THE UNIVERSITY OF SÃO PAULO

Emotion Recognition in Speech

written by
Daniel Soares Gieseler
Rodrigo Hideki Kido Narita

oriented by
Prof. Marcos Barretto

A project for the conclusion
of Engineering graduation

MECHATRONICS ENGINEERING DEPARTMENT

November 2019

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

1.1.1 Human-computer interaction

It is hard to think of modern life without any contact with technology. To avoid it, one would have to go extreme lengths and purposely seclude from civilization to a state of nature. Technology is ubiquitous. It starts off as a potential tool and, if proven useful, quickly makes its way into our culture and personal lives. And recently It has also been getting more sophisticated – computers are getting ever more intelligent and more capable at assisting our routines. Be it navigating the city, shopping on the internet or learning virtually, a human-computer interaction (HCI) naturally installs.

But, historically, HCI systems have displayed complete insensitivity towards affective states in interactions, which had a character of being an exchange of unilateral commands by the human for rule-based responses by the computer. If the task in question is highly procedural and devoid of implicit social communication like in the stock market or a nuclear plant, then such simplification would suffice. But technology has become profoundly integrated with our social networks and, in such emotionally rich environment, discussions are not only about facts, but part of a larger social interplay. The affective cognition is indeed recognized to be more impactful to success in social life than raw intellect. So, for machines to excel in our social arenas, a deeper computational understanding of emotions is required.

1.1.2 Affective computing

In response to that need of more natural HCI systems, an emerging field inside HCI has been taking notoriety: Affective Computing, as introduced by Picard in 1997 [1]. It is a highly multidisciplinary area ranging from computer science to psychology and neuroscience. It tries to tackle our problem by learning how to recognize, interpret

and simulate emotions with the aid of computational tools.

But one might ask: how can affective computing have an impact on improving HCI? There are at least a couple of ways: by making systems more flexible in face of different affective states and by empowering computers with complex decision-making abilities.

The first one is more intuitive and there is a great example, borrowed from [1], to illustrate it, which is the quintessential emotional experience - learning. A learning episode usually begins with curiosity and fascination. As difficulty increases, so does anxiety and frustration. At this point, the whole learning process may be abandoned, driven by avoidance of these negative emotions. In this scenario, the role a teacher (or a computer) can be seem as maximizing the first stage and minimizing the last. And a particularly good one would be able to detect these emotional states and act accordingly. Flexibility in face of different affective states.

The second way to improve HCI is inspired by the studies on frontal-lobe damaged patients conducted by the neurologist Antonio Damasio and reported in his book Decartes' Error. Damage on that area would impair the patient's ability to feel and, contrary to what classical thinking suggests, would turn decision-making into an even more laborious task. As Damasio puts himself:

> "Reason may not be as pure as most of us think it is or wish it were... Emotions and feelings may not be intruders in the bastion of reason at all: they may be enmeshed in its networks for worse and for better." [2]

His explanatory theory is that emotions and feelings are evolutionary mechanisms that regulate and facilitate the reasoning process. It basically serves as a bias to shortcut a virtually infinite logical search of paths in decision-making. This suggests that emotionally intelligent machines are not only better than their counterparts on how to interact with us, but also on being able to make decisions in the first place.

1.1.3 Speech and its relevance

Dating back to 100 thousand years, speech is the most ancient form of complex communication in our evolutionary history. It is ingrained in our psychology and tends to be the most practical and natural mode of communication. And, although, there exist other forms of communication, like facial expression, gesture and plain writing, the prevalent form has always been speech. So, it is no stretch of the imagination to see a natural preference for it. Indeed, it is reflected on the steady growth of speech recognition industry [3].

But the matter is that emotions are subjective and hard to define. And there is still no convergency by the affective community on the understanding of acoustic features emotional content nor on the modelling behind emotions. Many different approaches have been proposed throughout the years, but they all converge on a common objective: improving the computational understanding of emotions.

1.2 Objective

In this work, machine learning algorithms well-established in the affective community have been developed and optimized in order to tackle the problem of emotion classification from speech. By exploring each result and making comparisons, the expectation is to gain a better understanding of how different choices affect the performance of emotion recognition. These choices are the model implemented, settings of model's hyperparameters, format of inputted data points and format of outputted categories.

Other important contribution include exploring and documenting some different experiments in the area of emotion classification from speech, particularly testing methods developed for other use cases such as image classification and Natural Language Processing problems. By doing this, the group expects to provide new insights on these newly tested methods for coming projects to further explore and expand the knowledge on these matters.

1.3 Scope

The scope of this project is limited to three machine learning algorithms:

- Gaussian Mixture Model (GMM)
- Support Vector Machine (SVM)
- Convolutional Neural Network (CNN)

They are all implemented using the same database, namely Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [4]. However, as some emotions are expressed similarly through audio but can be differentiated mostly through facial expressions on video, this project has chosen to work with a restricted set of emotions from the database, namely: "neutral", "calm", "happy", "sad" and "angry". The other emotions of the dataset - "fearful", "disgust" and "surprised" - were then not considered in this project.

As methodology goes, a traditional machine learning framework (figure 1.1) will be used, which is composed of extraction and formatting of certain characteristics from the speech, usage of these as input to a machine learning classification model and the comparison of the models' classifications with the audio's real label.

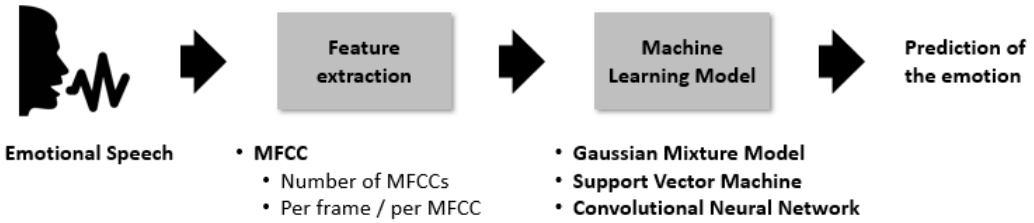


Figure 1.1: A general system architecture for absolute emotion recognition

The main feature used on this work is the Mel Frequency Cepstral Coefficient (MFCC). The many variations of how this feature can be inputted into the machine learning methods will be considered into the analysis. Be it frame by frame or the entire audio as a whole. It will also be analyzed the impact on the results of different set of categories in the output, considering an additional fragmentation by sex.

Chapter 2

State of the art

2.1 Theory and Notation

The whole process of emotion recognition consists of 3 main parts and this section will follow that segmentation: (1) Emotion modelling, in which emotions are attempted to be defined and represented to form a theoretical understanding of emotions; (2) Feature extraction and preprocessing, in which emotionally informative features are extracted from the datasets files; (3) Computational analysis (either classification or regression), in which machine learning techniques are used to process the emotion feature space into practical models.

2.1.1 Emotion Modelling

The modelling of emotions subject is rather fragmented over many disciplines. The ones most contributing to it are Psychology, Linguistics and, to a lesser extent, Biology. That coupled with the inherent fuzziness of the subject, which rely on a highly variable context of language, culture and symbolic meanings, makes any consensus in the community a hard ask.

The theory of component process model has taken progressively more support in the field. And its most concise definition of emotion is:

> "An episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism." [5]

The 5 components referred above are: cognitive component (evaluation of appraisal), neurophysiological component (bodily symptoms), motivational component (action tendencies), motor expression component (facial and vocal expression), subjective feeling component (emotional experience).

Considering the theory above, it is a natural conclusion to take a comprehensive measurement of emotion that captures these five components. But that has never been done before and limitations on the scope of the theory have to be set for practicality [5]. Following this trend, in this study, we will be focusing on the motor expression component, specifically vocal expression.

The advantage of using vocal expressions is the easy accessibility of vocal recording. Compare it, for example, with the physiological measurements that often require sensors physically attached to the individual for skin conductivity or electrocardiogram. On the other hand, this approach can lose the spontaneity value of its data, given that datasets available are usually artificially created by actors.

So far, we have covered the causalities behind the emergence of emotions. But it is also important to know how to represent them and there are two main choices on that regard.

Categorical or Dimensional

In affective science, emotions are either represented by a subset of basic emotions (referred by popular words) or by a dimensional space (emotions are localized based on their score on each dimension).

The first noticeable aspect of categorical approach is that it relies heavily on language to define emotions when compared to the dimensional approach. From it, two aggravations arise: higher redundancy of factors and ambiguity of emotions identify. The redundancy (or incompleteness, in the opposite extreme) is recognizable in the literature by the wildly differing number of emotions researchers have adopted.

Although the most widely used are the "Big Six", namely happiness, sadness, fear, disgust, anger and surprise, due to their popularity, dominance and occurrence in our daily life [6], it can range from 1 emotion to 18, depending on the author [7]. This lack of consensus on the number of emotions is matched by lack of consensus on the emotions identify. For example, different words can refer to different emotions on the perspective of different researchers or, even harsher, an emotion utilized by one researcher might not even be considered an emotion at all by another.

The former approach is analogous to the theory of color, where a few primary colors could be combined to form the entire visible spectrum. Although it can be very useful as a research strategy on a very limited scope of emotions, it has been criticized for its simplistic and somewhat erroneous portray of reality [7]. The counter analogy here would be trying to study animals with a subset of basic animals or to study language with a set of elementary languages, which is a counterproductive way of understanding it given the evolutionary nature of its subject. The bottom line being, a few emotions can't represent all possible emotions, but they do share universal building blocks innate to humans [8].

As a result, the affective community has been increasingly giving more atten-

tion to the dimensional approach and, in parallel, emotion researchers have been looking for the dimensional space that most economically accounts for the differences and similarities between emotions. In this space, categorical affect words are plotted according their dimensional coordinates and thus establishing a relationship between emotions.

The most fundamental dimensional format is the arousal-valence space and the earliest attempt at this is Russell's circumplex 2.1 [9]. Conceptually similar models came after, being the most preeminent: Whissel's model [10] with an astounding mapping of approximate 9000 affect words, and Plutchik's model [11], attempting to incorporate evolutionary principles.

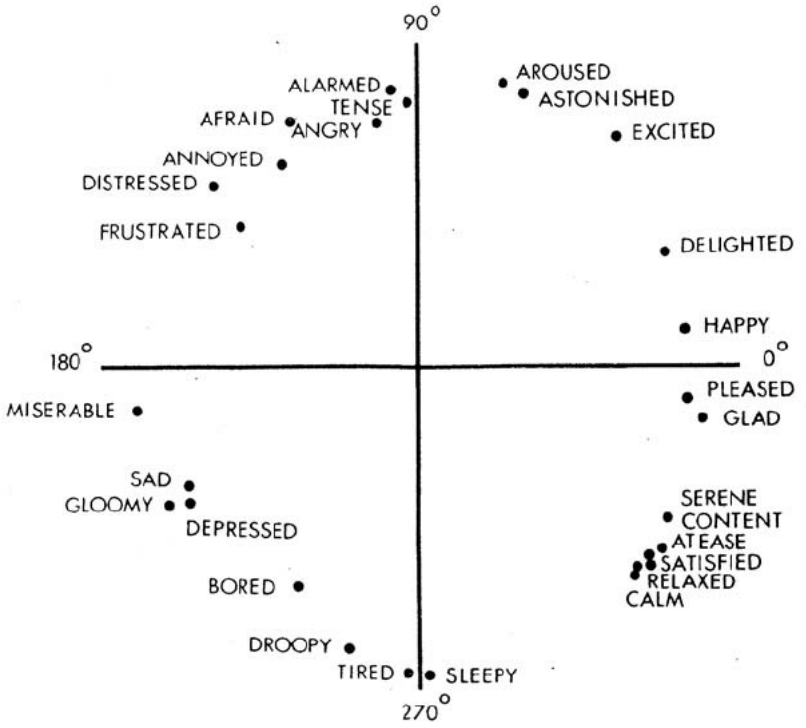


Figure 2.1: Russell's Circumplex Model

Simply put, valence measures a more subjective aspect of emotion, whether the individual finds it positive or negative. While arousal measures the degree of activation in the organism. And although they have proven to be the most impactful dimensions and certainly the preferred emotional space in recent emotion recognition studies [12], the world of emotions is not limited to two-dimensional. Other dimensions have been proposed to complement the emotion space: unpredictability and dominance [13], for example.

The advantage is that more dimensions decrease the ambiguity between emotions, but it naturally increases the complexity of the research. So, it is important to understand what aspects of emotion are the most important to the studied context. For example, anger and fear in the 2D space can be both categorized as negative for valence and high for arousal, but we intuitively know they are not the same. That is where the dimension of dominance can become handy. Anger is high dominance

and fear is low.

At the same time, unpredictability accomplishes a clearer distinction between stress and fear. Fontaine would argue that those 4 dimensions are the necessary bunch to define the emotional space at the state of the art [13], but as we know there is a catching up to do between the theoretical and practical front. Which is especially true for the creation of dataset that involves a considerable amount of effort and bureaucracy.

Static or Dynamic

Two important design features, not dealt with previously, from the component process theory are: rapidity of change and duration of emotion. These two are related to the nature of emotion regarding time. Emotions are understood to be an affect state of high intensity and highly adjustable to revaluations/change of circumstances [5]. So, emotions tend to have short duration (otherwise it would be unsustainably exhaustive) and adapt fast (to fit the organism immediate needs appropriately).

It is also possible to make a case, similar to the categorical one, where the static representation is a trade-off for simplicity in detriment of completeness of representation. A lot of information is lost when an entire recording of audio, for example, is labelled as a single emotion. And it is especially detrimental when the evolution or tendencies of emotions over time are more informative than the absolute labels.

2.1.2 Emotional Audio Database

As new studies and methods on emotion recognition emerge, a common starting point is needed in order to better compare the efficiency of each one of them. In order establish this common ground, different groups have created emotional speech databases in the past. However, as the language of speech greatly affects the training process of machine learning models and their efficiency, and noting there were not many databases containing audiovisual recordings of speakers in North American English, Steven R. Livingstone and Frank A. Russo created the The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [4].

The RAVDESS database is a multimodal, sex balanced database of emotional speech and song, which is composed of 7356 recordings of 24 professional actors expressing 8 different emotions with 2 different intensities. Some interesting characteristics of this database involve, for example, the fact that on all the recorded speeches, only 2 statements are said. Also, even though every file on the database was recorded by an actor, a very thorough validation step was done in order to maintain a high level of genuineness to the speeches. Bellow is a list of the multiple identifiers and respective factor levels of the recorded files:

- Modality: 01 = Audio-video, 02 = Video-only, 03 = Audio-only

- Channel: 01 = Speech, 02 = Song
- Emotion: 01 = Neutral, 02 = Calm, 03 = Happy, 04 = Sad, 05 = Angry, 06 = Fearful, 07 = Disgust, 08 = Surprised
- Intensity: 01 = Normal, 02 = Strong
- Statement: 01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"
- Repetition: 01 = First repetition, 02 = Second repetition
- Actor: 01 = First actor, ..., 24 = Twenty-fourth actor

As expressed on the first section, as the scope of the project is restricted to "Audio-only" "Speeches", the group has opted to not work with emotions such as "Fearful", "Disgust" and "Surprised", as these can be interpreted much more through visual cues than audio ones. However, all other identifiers will be included in the total database of the project.

2.1.3 Feature Extraction

The human voice is a very good indicator of someone's emotions at the time of speech. On a physiological level, humans produce speech by pressing air from the lungs through the vocal cords. The larynx, commonly called the voice box, holds the vocal folds, and is able to manipulate pitch and volume of the voice. The oral and nasal cavities, on the other hand, have the ability to amplify some frequencies while attenuating others.

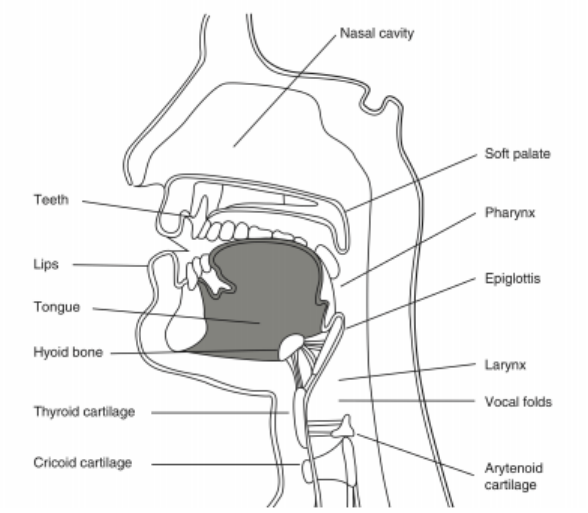


Figure 2.2: Phisiology of the human's vocal apparatus [14]

And because an emotional episode, altogether with some degree of intentional manipulation from the speaker, produces changes in the vocal apparatus, the perceived characteristics of the speech are changed and thus can be perceived. [14] By

extracting these most relevant characteristics from the person's speech the model can produce better results in emotion recognition.

In terms of the characteristics used to recognize emotion in speech, many different approaches have given result to novel processes to generate new and optimized feature sets. Most approaches fall into one of two categories: the first one generating optimized packs of features and the second one changing which segmentation to use when extracting them.

The Mel Frequency Cepstral Coefficient (MFCC)

Nowadays, one of the most used acoustic features is the Mel Frequency Cepstral Coefficient (MFCC). The MFCC is part of the category of spectral features, as it carries details of short-term speech in the frequency domain and carry information regarding the timbre of the sound. One of its main differential is that it treats audio in a human manner, as it incorporates a filtering technique that mimics the human audio perception - more sensitive to low frequencies and more insensitive to higher ones. Therefore, it works on a different scale – approximately linear before 1kHz and logarithmical after.

To extract the MFCCs, it is necessary to sample the audio, framing it afterwards. The rationale behind it is that frequencies change throughout the signal, so it doesn't make sense to apply Fourier Transform to the entire signal. Instead, we divide the signal into smaller frames of time in which frequencies are assumed to be stationary. And the frames are overlapping in the signal so to not lose information between frames. The typical values here are frames of 25ms with a stride of 10ms (15ms overlap).

Afterwards, a window function is applied to every frame in order to satisfy Fourier Transform requirements. With the application of the Discrete Fourier Transform to the sample, it results on a Periodogram estimate of the power Spectrum, in which we can evaluate the audio in terms of frequency. This is important as the MFCC tries to emulate the human cochlea (an organ in the ear), which vibrates at different spots depending on the frequency of the sound it's being exposed to, each frequency firing nerves that send different information to the brain.

On the next step, it is necessary to compute mel-spaced filterbanks. As much as humans can distinguish different frequencies through its cochlea, this organ has much more trouble distinguishing changes at low pitch than at the higher one. To emulate this phenomena, the Mel Scale was created to take it into account, becoming a much more closer match to the human ear. In this step, we multiply each of these filterbanks to the power spectrum, adding up each coefficient to result in how much energy was present at each filterbank.

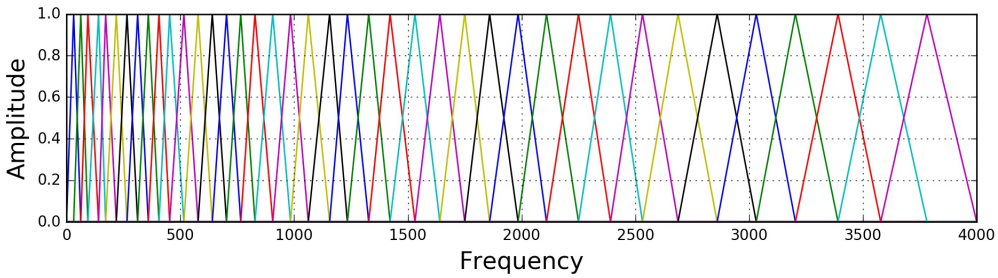


Figure 2.3: Mel filters, which emulate the human auditory system, used to calculate the MFCCs

The final step is to apply a Discrete Cosine Transform (DCT) of the log filterbank energies and consider only the first 12 to 13 of the cepstrums. While the DCT is important to decorrelate the overlapping filterbank energies, dismissing the higher DCT coefficients is necessary as these represent fast changes in the filterbank energies, and these fast changes actually harm the performance of the Automatic Speech Recognition model.

With these past steps implemented, we end up with the Mel-frequency cepstral coefficients, a state of the art feature which is the most used one in speech recognition nowadays.

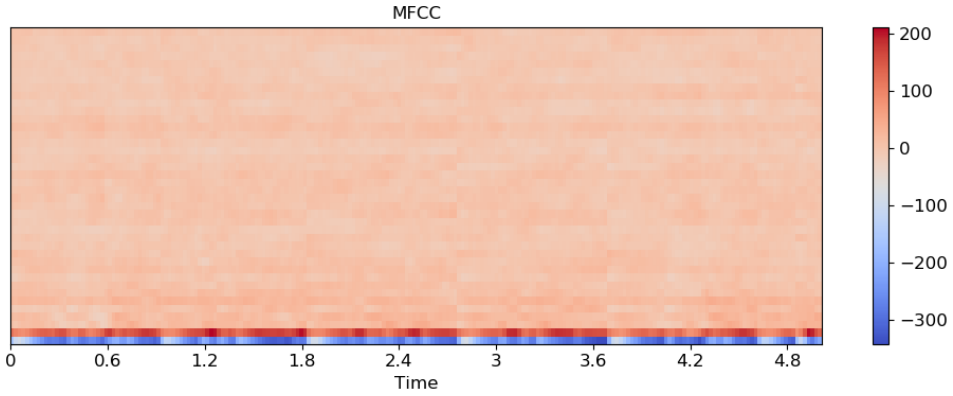


Figure 2.4: MFCCs

Other state-of-the-art features

In order to produce a state-of-the-art standard in acoustic features set, Floarian Eybein et al. [15] compared different sets of features and were able to produce a minimalist acoustic parameter set for various areas of automatic voice analysis, namely the GeMAPS (Geneva Minimalistic Acoustic Parameter Set). This work was very important in order to provide a common baseline for evaluation of future research and eliminate differences caused by varying parameter sets, as most works done in this field could not compare results, even when using the same database. [15]

Other works, such as Dr. Swarna Kuchibhotla's [16], explored a solution to what is called the curse of dimensionality. This problem occurs when the number of speech emotional samples available for training is smaller than the number of features extracted from the speech sample, and the sparsity in available data becomes a problem in any method that requires statistical significance. Regarding this problem, he used Sequential Forward Selection and Sequential Floating Forward Selection feature selection algorithms in order to extract more informative features and reach a more optimal feature set.

Ali Meftah et al. [17], for instance, made an analysis comparing emotion recognition in two different languages: English and Arabic. Thought his work, he used Analysis of Variance (ANOVA) to determine which acoustic features should be used in their emotion recognition system, and discovered that, specifically for some Arabic words, there are some acoustic features presented a benefit in terms of emotion recognition. This has shown that the speaker's native language has high impact on the choice of the acoustic feature set to work with.

Apart from the choice of which features to work with, some projects have been studying which segmentation of the speech is best to work with, as there are various possibilities to explore, such as syllables, phones, vowels and consonants. Suman Deb et al. [18] has proposed a classification method using vowel-like regions (VLRs) and non-vowel-like regions (non-VLRs) instead of processing the entire speech region, as these two segmentation types contain emotion-specific information which resulted in an out performance of state-of-the-art approaches using the same database.

2.1.4 Computational Analysis

For better a discussion of the findings from next chapters, it is necessary to align our notations and basic theorical background of the computational techniques explored. There won't be much mathematical deduction presented, because it is not the intent of this study, but rather build upon standardized open source implementations of these techniques. So, it will actually consist of narratives about the overall functioning and hyperparameters explored for the optimization.

Gaussian Mixture Model (GMM)

A good starting point to introduce GMM is the wildly popular clustering algorithm K-means. Both share the same clustering nature based on updating centroids, but GMM can also incorporate uncertainty. It models the data in terms of probabilities instead of hard categorizations. In fact, GMM can be considered a generalization of K-mean. Its constitution is a combination of ellipsoidal shaped gaussian distribution.

In practice, GMM starts as a Bayesian inference problem: we want to update the probability of an initial model (the prior, $p(\lambda)$) as more data about the phenomena becomes available (the likelihood, $p(X|\lambda)$), which yields the posterior, $p(\lambda|X)$. And we want to maximize it to have the model that most likely explains that data.

Put mathematically, it is the Bayes rule:

$$p(\lambda|X) = \frac{p(X|\lambda)p(\lambda)}{p(X)} \tag{2.1}$$

But, because the prior the normalizing factor remain constant, the problem reduces to maximizing the likelihood. And the most common way to find the maximum likelihood, in this context, is the expectation maximization algorithm. It is an iterative process of 2 steps, that is repeated until an arbitrary convergency measure:

i Initialization: it is not an iterative step per se. The model parameters can be initialized by two ways: randomly or, for better results, set by a previously run K-means algorithm, which gives a good starting point for the gaussian means. And the prior of each gaussian set to the same value – uniform distribution.

ii Expectation: first, for every pair of data point and gaussian, the likelihood of a point belonging to a gaussian is determined. Then, it is determined the posterior, which is how well a gaussian explains a data point. The higher it is, the higher the updating weight for the next step.

iii Maximization: the model parameters (mean and covariance matrix) are then updated to best fit the probabilities previously assigned.

Number of Components First of the two most impactful hyperparameters in the GMM optimization process. It is the number of gaussian components (each one having a mean a covariance matrix) to be arranged in the feature space during the clustering.

Covariance Matrix The covariance matrix is what gives a gaussian its shape in the feature space: whether it is circular or ellipsoidal and whether it is aligned or angled to the feature axis, see 2.5. Ellipsoidal means different features vary differently. Angled means different features can vary coupled.

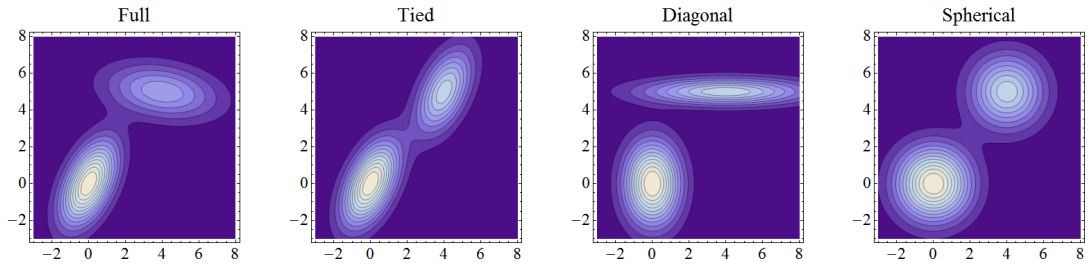


Figure 2.5: Visual example of different covariance types

Depending on the shape restriction imposed, different covariance types can be achieved for different objectives. In 2.6, the main 4 types are described and compared in the trade-off of complexity vs performance. The full covariance achieves the most detailed description, and consequently the best performance and highest

complexity. On the other hand, the spherical achieves the extreme opposite. And finally balancing in the middle, there are the diagonal and tied covariances.

Covariance type	Description	Number of model parameters	$n_{comp} = 56$ $n_{feat} = 13$
Full	each component has its own general covariance matrix	$n_{comp}\ x\ [n_{feat}, n_{feat}]$	= 9464
Diagonal	each component has its own diagonal covariance matrix	$n_{comp}\ x\ [n_{feat}, 1]$	= 728
Spherical	each component has its own single value variance	$n_{comp}\ x\ [1, 1]$	= 56
Tied	all components share a general covariance matrix	$1\ x\ [n_{feat}, n_{feat}]$	= 169

Figure 2.6: Table of Covariance Types. Descriptions are from [19]. n_{comp} is the number of components and n_{feat} is the number of features. The last column is exemplifying the different of complexity with a standard set of parameters.

Support Vector Machine (SVM)

Support Vector Machines (SVMs) are a set of supervised machine learning methods used in classification problems. The basis for the SVM method is the usage of hyperplanes (subspaces whose dimension is less than that of its ambient space) in N-dimensional spaces that segregate data points from different categories. These hyperplanes are defined as decision surfaces or decision boundaries.

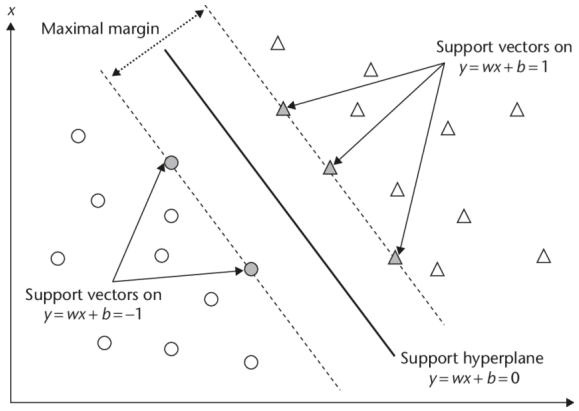


Figure 2.7: Example of chosen decision boundary given the support vectors

In order to separate these different types of data points, the method uses an element called support vector. Support vectors are vectors composed of the data points which are the most difficult to classify, meaning they are the closest ones to the ideal decision surface, and influence the position and orientation of the hyperplane. Considering this definition of support vectors, the SVM model aims to

choose the hyperplane which maximizes its margin, calculated by the maximization of the distance between support vectors from each category.

Kernel type The Support Vector Machine is a kernel method, as it uses kernel functions to operate and classify its data points. These include:

- Linear: uses a linear decision boundary to categorize the data points
- Polynomial: processor that generates new features by applying the polynomial combination of all the existing features
- Sigmoid: processor that generates new features using the sigmoid transform, a neural net activation function
- Radial Basis Function: processor that generates new features calculated by measuring the distance between all other dots to a specific dot

Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) is a supervised machine learning method, and a is considered to be a subclass of Artificial Neural Networks (ANN) which have at least one convolution layer. Also, the CNN is categorized as being a deep learning algorithm, as it uses multiple layers to progressively extract higher level features from the input.

The main component of Neural Networks is the neuron, which at first was designed to reflect and model biological neural systems, but afterwards has diverged to machine learning tasks as it has achieved very good results in this area. The way the neuron was modelled follows closely the biological way it functions on the human brain. Dendrites receive input signals from the axons of other neurons, the cell body considers all these inputs and generates an output signals to its axons. Computationally, the signals that run along the axon interacts multiplicatively with the strength of the synapse, and the body sums all these interactions and decide, based on an activation function, if it will fire a signal do its axons or not.

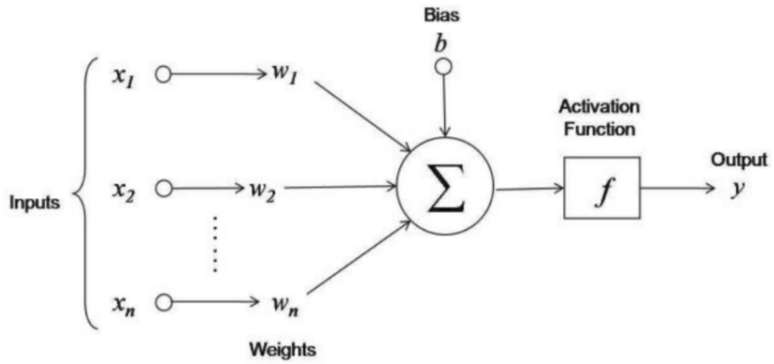


Figure 2.8: Single artificial neuron [20]

Neural networks are modeled as a collection of neurons, which are organized into distinctive layers. In this method, the most used type of layer is the fully-connected one, where every neuron from the next layer receives the output signal from each neuron from the last one. However, the CNN is considered a subclass as it presents at leas t one convolution layer. In this particular case, instead of the fully-connected layer previously seen, the neurons in a layer are only connected to a small number of neurons from the layer before. Also, the layers are positioned in a sequential manner, so the first one receives the features of the input and the last outputs the class scores.

The main layers used in CNNs are the following:

Convolutional layer The main building block of the CNN, and holds the most of its computational processing. It consists of a set of learnable filters, with defined length, height, depth and stride, the last one meaning the distance between a filter and the next one. As the filters slide over the input, a dot product is performed, resulting in a 2d activation map of that filter. If there is more than one filter, the activation maps are stacked, providing depth to the output volume of this layer.

Pooling layer Reduces the spatial size of the representation and the amount of parameters and computational processing in the network, also controlling overfitting as it simplifies the outputs. The pooling layer also has parameters such as lenght, height and stride, and performs functions to simplify the previous result, such as average and maximum pooling.

Dense layer A fully-connected layer where its output is the dot product of its input and its weights. It is often used at the end of CNNs as a way to consider all different outputs from the previous layer.

Dropout layer Reduces overfitting by randomly setting a fraction of the inputs to 0 at each update during the training of the model.

And even though CNNs are widely used in image classification problems, this method has also had very good results in audio and speech recognition problems.

2.2 Peer Developments

Gaussian Mixture Model (GMM)

In 2011, Je Hun Jeon [21] compared multiple decision combination methods, including the Gaussian Mixture Models (GMM), in order to generate emotion probabilities

for segments in a sentence. In their work, they compared the GMM with the majority vote and the average segment of probabilities ones, in which case the first one gave the best results.

In this experiment, a GMM model was trained for each emotion class using feature vectors containing their segments. Then they used this separate classifier to model the distribution of the posterior probabilities of all the segments in a sentence.

On a more conventional vein, MFCC-GMM systems on 4 emotions (i.e. neutral, sadness, happiness and anger) have been developed and achieved accuracies of around 53% [22] [12]. These performances can vary drastically from study to study when different databases are utilized. They can contain issues that worsen the classification process - overlapping speaker, some background noise or unrealistic acting of script.

Support Vector Machine (SVM)

As the SVM, compared to other machine learning methods, presents advantages in solving nonlinear, small samples, and high dimensional recognition problems, it has been used in speech emotion recognition.

Linhui Sun et al. have made use of these capabilities by creating a decision tree SVM model with Fisher feature selection for speech emotion recognition [23]. In their work, they used a Fisher criterion combined with a decision tree SVM in order to remove the redundant features and thus, improve emotion recognition performance.

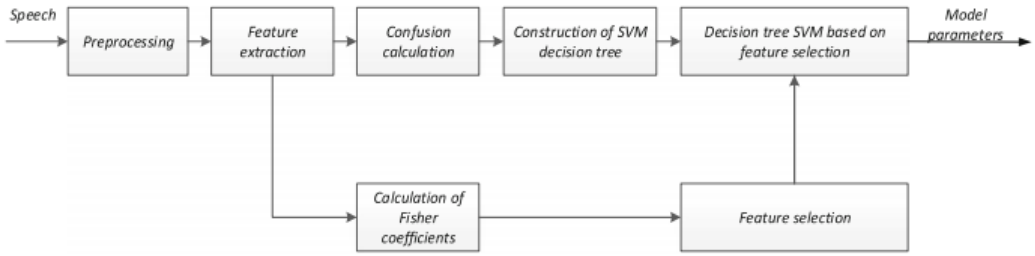


Figure 2.9: Construction flow chart of decision tree SVM with feature selection [23]

In this experiment, they reached a 83.75% accuracy rate on the CASIA Chinese speech emotion corpus, showing noticeable improvements over both the pure SVM model and the decision tree SVM without feature selection.

Convolutional Neural Network (CNN)

In 2018, Somayeh Shahsavarani used a deep learning algorithm on speech emotion recognition [24]. In his work, he used a baseline Convolutional Neural Network architecture and, after creating multiple models with different parameters on each

layer, compared the results on four different databases with different languages or accents, namely:

i EMODB: Berlin Database of Emotional Speech

ii SAVEE: Surrey Audio-Visual Expressed Emotion

iii EMOVO Corpus: Italian Emotional Speech Database

iv BTNRH: American English speech database

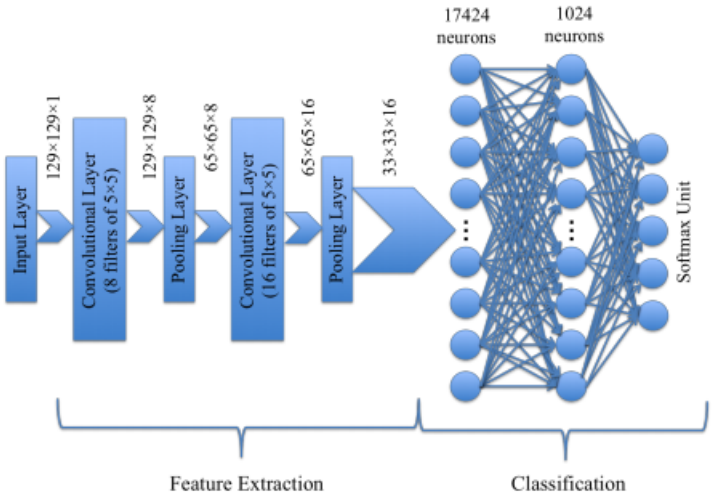


Figure 2.10: Baseline architecture of the CNN used in Somayeh Shahsavarani's study [24]

For instance, some of the changes made on each model involved changing the pooling method, the dropout percentage, the number of epochs and the kernel sizes of the convolutional layers. After testing each different model, the author compared his results both with benchmark comparisons on each database and with human listeners' performance on the audio classification.

f_1	f_2	pooling	$p_{dropout}$	augmentation	epoch	CV accuracy (%)
5×5	5×5	Max	0	10x	100	58.62
5×5	5×5	Max	0.5	10x	100	78.01
10×10	5×5	Max	0.5	10x	100	85.29
10×10	5×5	Average	0.5	10x	100	87.58
5×5	5×5	Max	0.5	20x	100	95.84
10×10	5×5	Max	0.5	20x	100	96.61
10×10	**5×5**	**Average**	**0.5**	**20x**	**100**	**96.78**
10×10	**5×5**	**Average**	**0.5**	**20x**	**800**	**99.5**
10×10	**5×5**	**Average**	**0.5**	**20x**	**4000**	**99.83**

Figure 2.11: The summary of the architectures and the results of Somayeh Shahsavarani's experiments on the EMODB database [24]

Another interesting development using CNNs in emotion recognition from speech is the work of Yawei Mu et al.[25]. Their work has used a new approach by combining CNNs with Bidirectional Recurrent Neural Networks (BRNN) and an Attention Mechanism.

By using the BRNN, the model was able to get information from past (backwards) and future (forward) states simultaneously, as both past and future audio information can be relevant. Also, the attention model served to allow the neural network to focus on the most pertinent pieces of information by training the attention vector together with the neural network through the backpropagation algorithm.

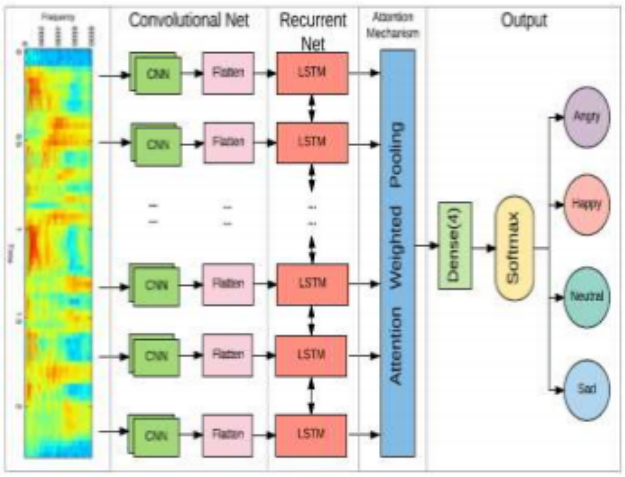


Figure 2.12: The proposed CRNN with attention model [25]

In their experiments with the Interactive Emotional Dyadic Motion Capture (IEMOCAP) and by using this innovative approach, the group reached state-of-the-art performance in this particular database, and noted that further steps include the prediction of arousal and valence instead of the four emotions they've worked with.

Also, as multiple layers, each one with its own parameters, are stacked to form a CNN model, some high performance patterns were found, developed and left for the public to use. These patterns involve a specific sequence of layers with pre-trained parameter, in which its weights are trained previously on a specific database. Nowadays, one of the most famous CNN structure is the Visual Geometry Group (VGG).

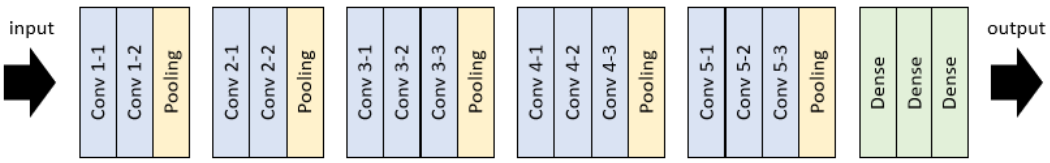


Figure 2.13: VGG16's layer structure

The VGG was a structure developed at the Large Scale Visual Recognition Challenge 2014 (ILSVRC2014) challenge, in which groups develop algorithms for object detection and image classification at large scale. In this competition, Karen Simonyan and Andrew Zisserman [26] developed an algorithm that reached the top classification in the competition by evaluating networks with increasing depth and very small convolution filters.

Chapter 3

Development

As presented before, three modelling techniques will be used: SVM, GMM and CNN. And it was initially desired to base our methodology on the well-established image recognition approach. That is, the audio samples would be treated as 2D images: frames on one dimension and features on the other. For the modelling techniques, it means inputting an entire audio at once. But due to reasons discussed further ahead the GMM approach had an alternative methodology.

3.1 Dataset

Before the modelling, a preparation of the raw RAVDESS dataset had to be done. After finished, the same dataframe would be used for all three modellings with the exception of some minor changes.

The audio samples were read at a 16000Hz sampling rate. They averaged around 3.7 seconds of duration, but varying from 2.5 seconds up to 4.7 seconds. So, for conformity of input, they were all brought to the same length of 3.7 seconds. If it was shorter, then a symmetrical (left and right) padding of zeroes would be applied. If it was longer, a symmetrical cutting would be applied.

After that, the feature extraction is performed. First, the framing of the audio is done with the typical frame size of 25ms and shift of 10ms. The windowing function of each frame is the well-known hamming function. The size of the frequency domain after the Fast Fourier Transform is $NFFT = 512$. Finally, 40 filter-banks in the Mel-space are calculated and then decorrelated to 13 MFCCs.

The target categories are processed, but just considering the main 5 emotions: Neutral, Calm, Happy, Sad, Angry. Two set of target categories are then constructed: 5 categories of the emotion and 10 categories of the emotions divided by sex. It is important to notice that, the emotion are unbalanced. Due to how the database was created, neutral emotions are half the number of data of other emotions, that is for having only one level of intensity.

At last, the group has used 80% of these audios to train each model and the 20% rest to test their accuracy and weighted F1 scores. In absolute number, that yields around 700 data points for training and 150, for testing.

3.2 CNN Approach

Regarding the usage of Convolutional Neural Networks, there is an additional layer of complexity in its usage if compared to the rest of the models, as it is necessary to develop a sequential set of layers to train the model. In order to construct the main model's network, the group used the Keras library in Python, which provides usage of the main CNN layers and facilitates their usage in the sequential model.

The CNN is a model that is widely used in the scope of image classification, and many frameworks have been created as they presented state of the art classification in a very large database. One of the most famous is the Visual Geometry Group (VGG), a network that has reached the top places in the ImageNet ILSVRC-2014 challenge regarding image classification. As a benchmark for our work, the group ran several different networks based on the macro structure of the VGG, but adapted it to our empirical trials and the limited processing available.

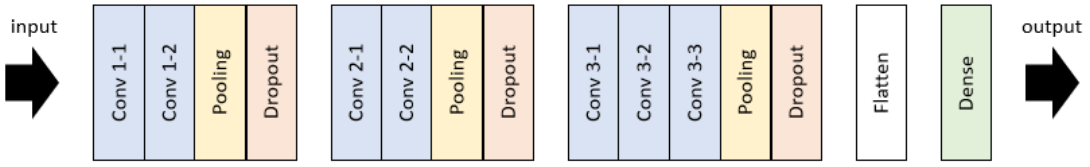


Figure 3.1: Final structure based on the VGG network

As seen in figure 3.1, the structure is composed of 3 main blocks, summing up to 7 convolutional layers total, which was determined as the most optimal empirically. In each block after the final convolutional layer, a maximum value pooling is done exclusively in the time dimension, halving the size of the audio length. Doing the same on the MFCCs dimension worsen the results and logically so, because the MFCCs points do not necessarily share the same similarity by proximity as time points do. Next, before leaving the block, a 10% dropout of connections is done for inciting further generalization of pattern recognition.

Finally, the loss function was categorical cross-entropy and the optimizer was RMSprop with learning rate of 10^{-5} and decaying rate of 10^{-6}. And it run for 100 epochs. The group has then formatted the 13 MFCCs per frame into a image-like input, with the MFCCs in the rows and each frame in the columns. Afterwards, a test was done using the first group of labels: the 5 categories of emotions, in which the model has achieved a final accuracy of 74.1% and a weighted F1 score of 0.77.

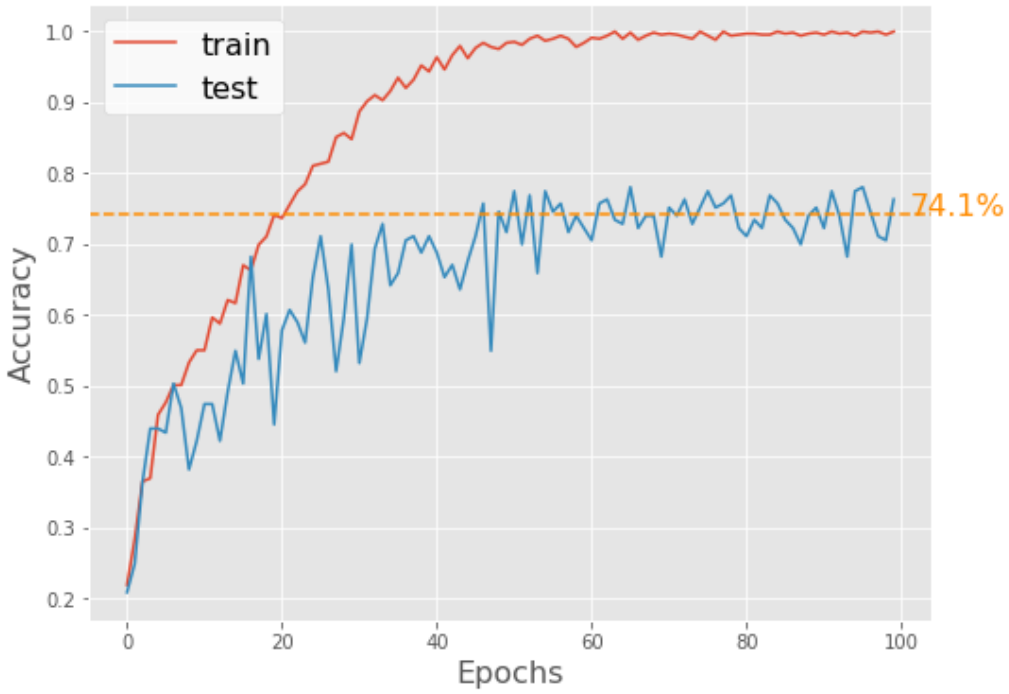


Figure 3.2: Traning history of CNN without sex fragmentation

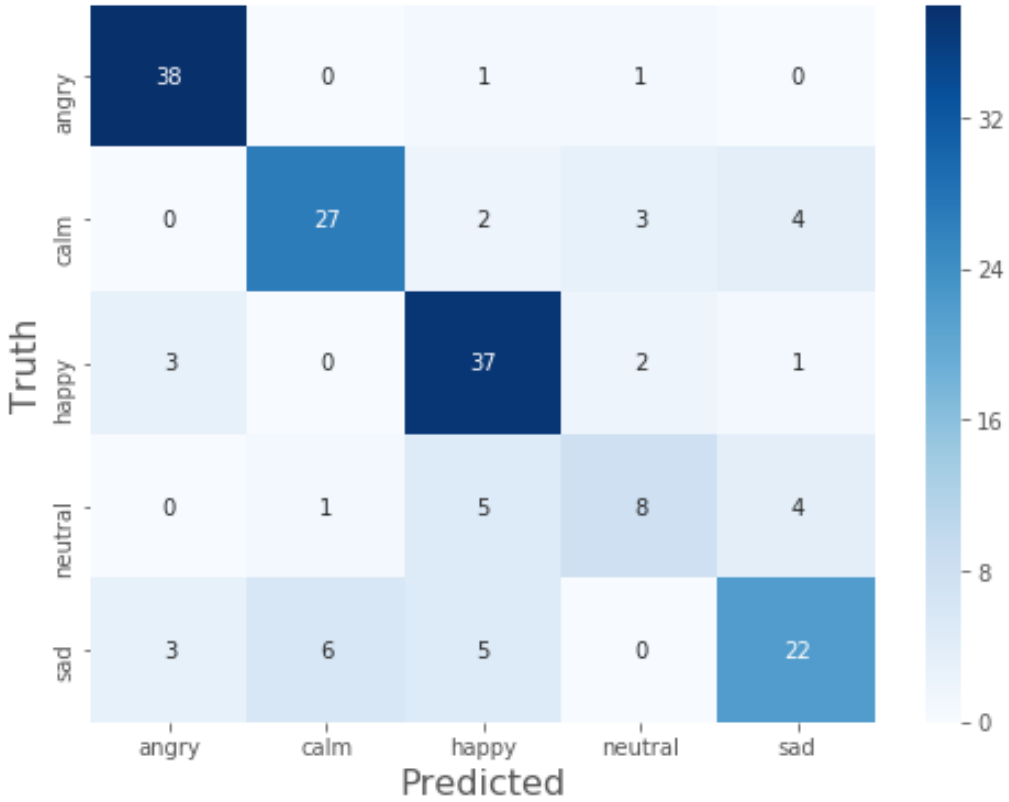


Figure 3.3: Confusion matrix for CNN without sex fragmentation

The second test using CNN was done using more specific labels. In this case, the group added the sex-factor to the labels from each emotion (e.g. male person

representing the emotion of anger), doubling the number of labels to 10. However, even though the number of labels increased, the model reached a higher accuracy at 80.7% and a higher weighted F1 score of 0.80.

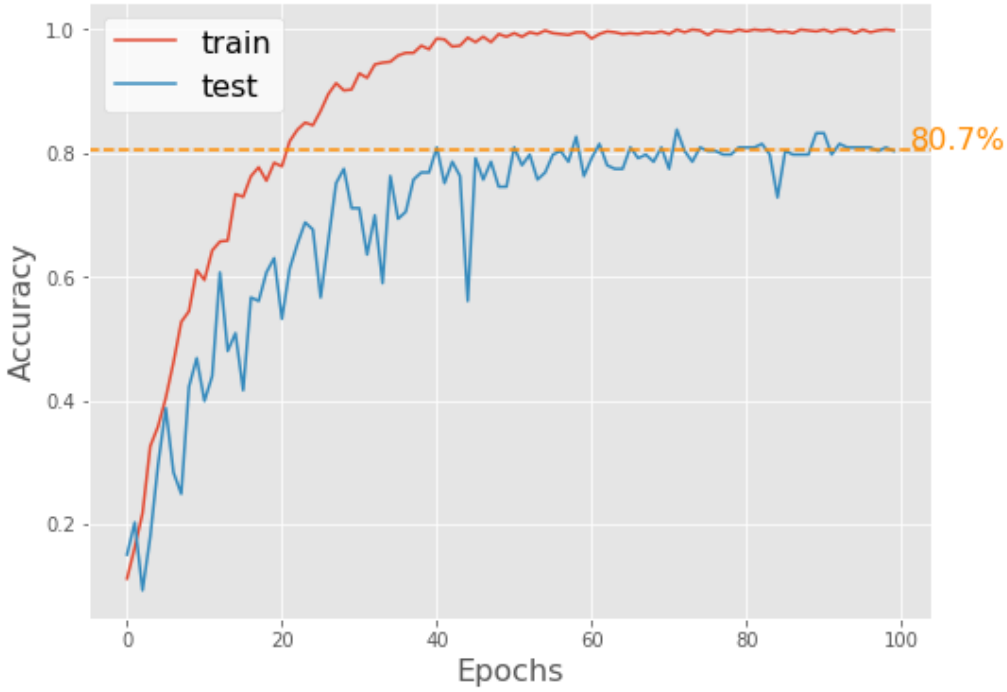


Figure 3.4: Traning history of CNN with sex fragmentation

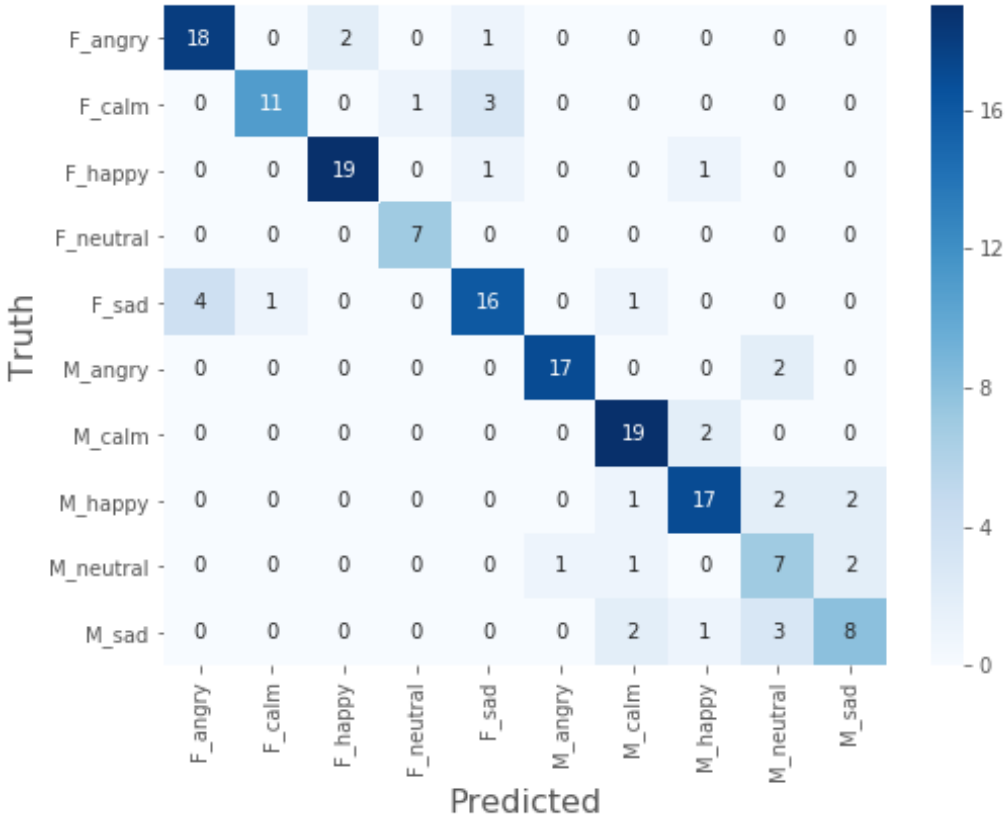


Figure 3.5: Confusion matrix for CNN witho sex fragmentation

3.3 GMM Approach

An initial attempt to use the image methodology yield a very poor accuracy of around 45%. The reason could be that the GMM is by far the model most sensitive to the number of data points. Its mathematical nature requires a large sample size to build well-formed gaussians and reach statistical significance. It is commonly referred as the curse of dimensionality [16]. Which is when the number of features exceeds the number of data points.

As an attempt to circumvent this speculative reasoning, an alternative was tried. One that loses discrimination of the time domain but produces more data points. And sure enough, the results improved drastically. The alternative was inputting frame by frame instead. Next, its detailed development is presented.

Differently from the other two modelling techniques, the GMM classification is done computing an individual model per classifiable category. It does not consider pattern from different emotions at once. Metaphorically, it computes a statistical fingerprint for each category. Then, for a given audio sample, the likelihood of it having each emotion fingerprint is calculated and the highest one is chosen as prediction. So, the first step was dividing the dataset into train/test subset at the proportion of 4:1 and its categories for each individual model.

Next, hyperparameters configuration for the GMM were explored to find the most optimal set:

- The initialization is set to be done once and by a simple K-means.
- The expectation maximization is set to iterate 1000 times.
- Type of covariance matrix is further optimized.
- The number of gaussian components is further optimized.

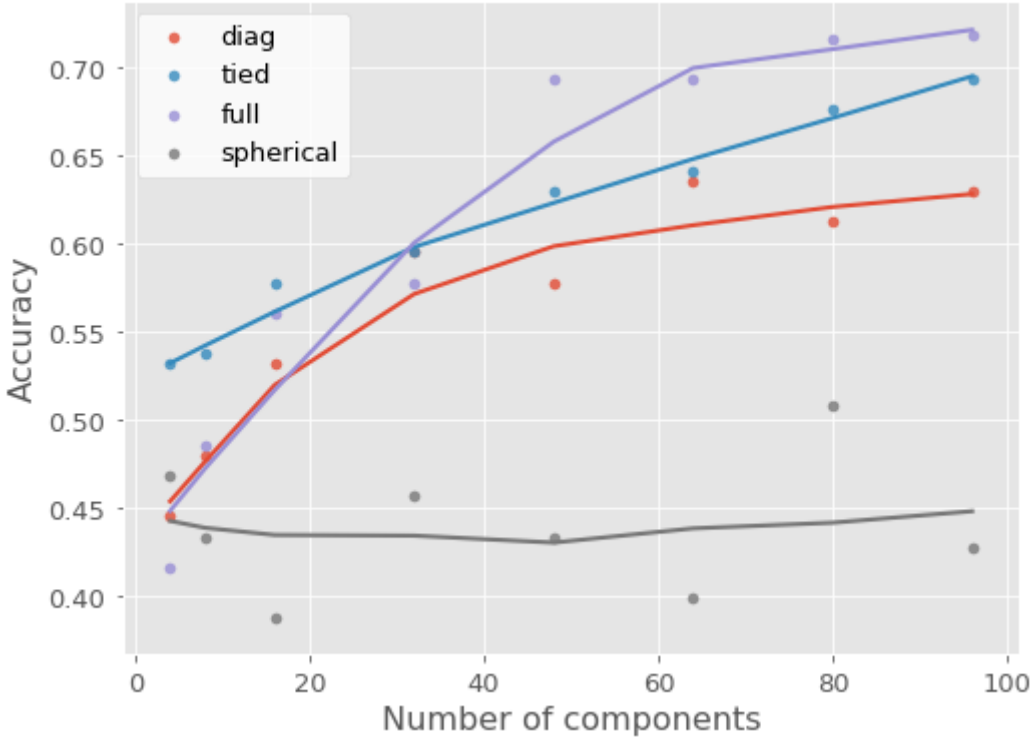


Figure 3.6: GMM frame by frame - optimization of covariance type

The spherical covariance had the worst result and the full covariance had the best 3.6, as expected since they are respectively the most simple and complex models available. What is surprising is that the tied covariance outperformed the diagonal covariance, although being a much simpler model. But that is not all, because the tied covariance also performed surprisingly close the full covariance. Which begs the question, is that small performance different worth the extraordinarily higher model complexity? So, for the rest of the experiments we will be using the tied covariance and number of components above 200.

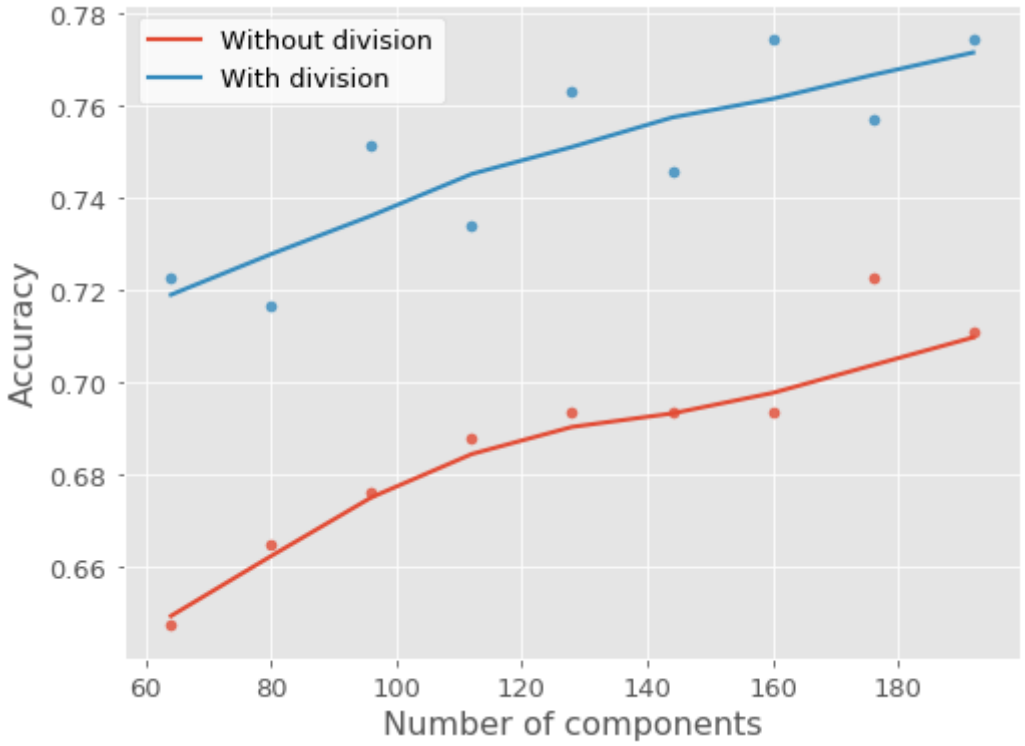
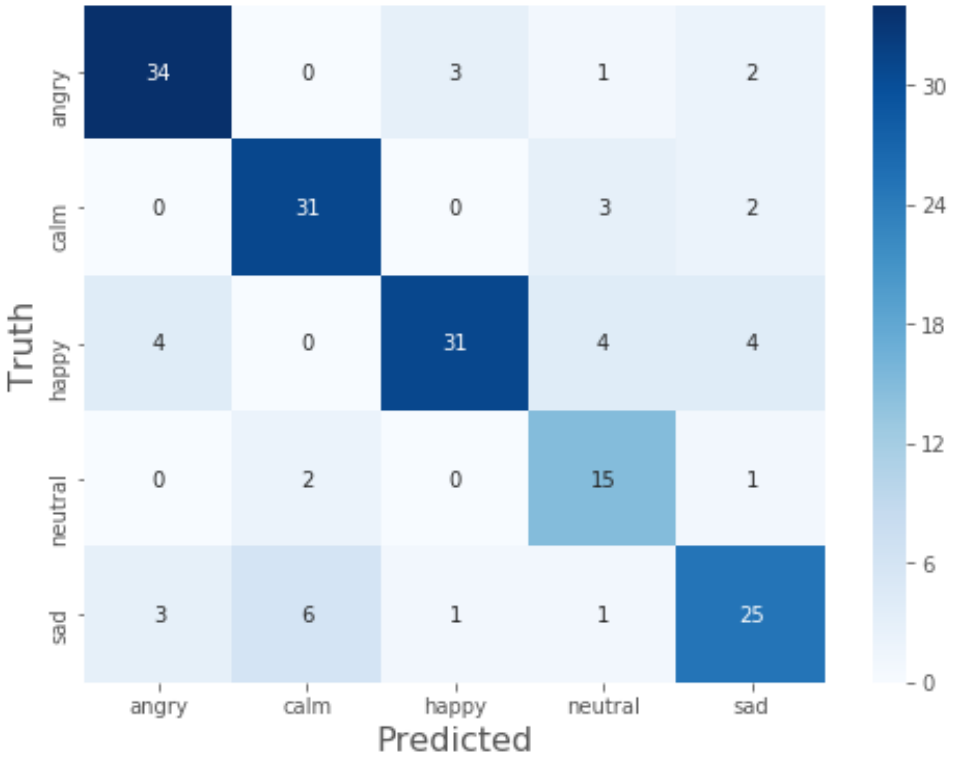


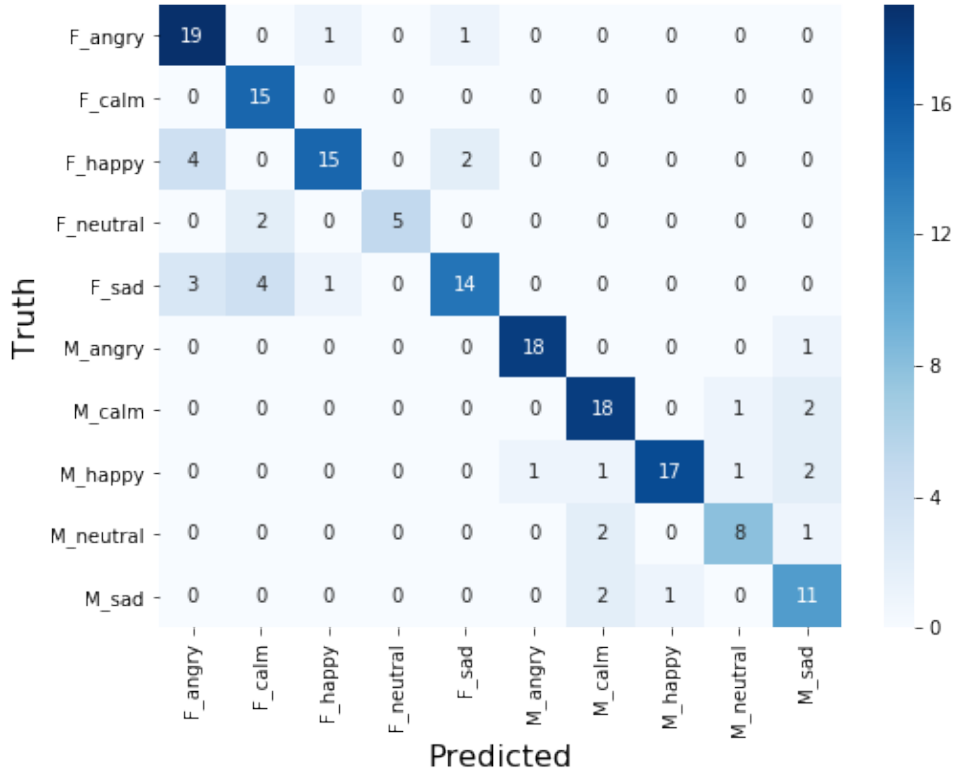
Figure 3.7: GMM frame by frame - comparison of fragmentation vs no fragmentation by sex

As seen in 3.7 , the range of number of components is increased, reaching up to 196. Throughout this range, the model with categories divided by sex consistently outperformed the other by around 6%. And, further out of the this range, its determined that no significant improvement is noticeable after 512 components. In fact, the gap between sex and no sex fragmentation diminishes, but do not disappear, which indicates that this trick does not only facilitates the learning with fewer components, but also ends up with a better representation for the emotion recognition task.

As shown in the confusion matrices, the final most optimal run yields an accuracy of 80.9% and 78.6% for fragmentation and no fragmentation by sex, respectively. It also results in weighted F1 scores of 0.81 and 0.79 in the same order. Furthermore, it is noticeable in (b) that the identification of sexes is perfect, without a single wrong prediction.



(a) without sex fragmentation



(b) with sex fragmentation

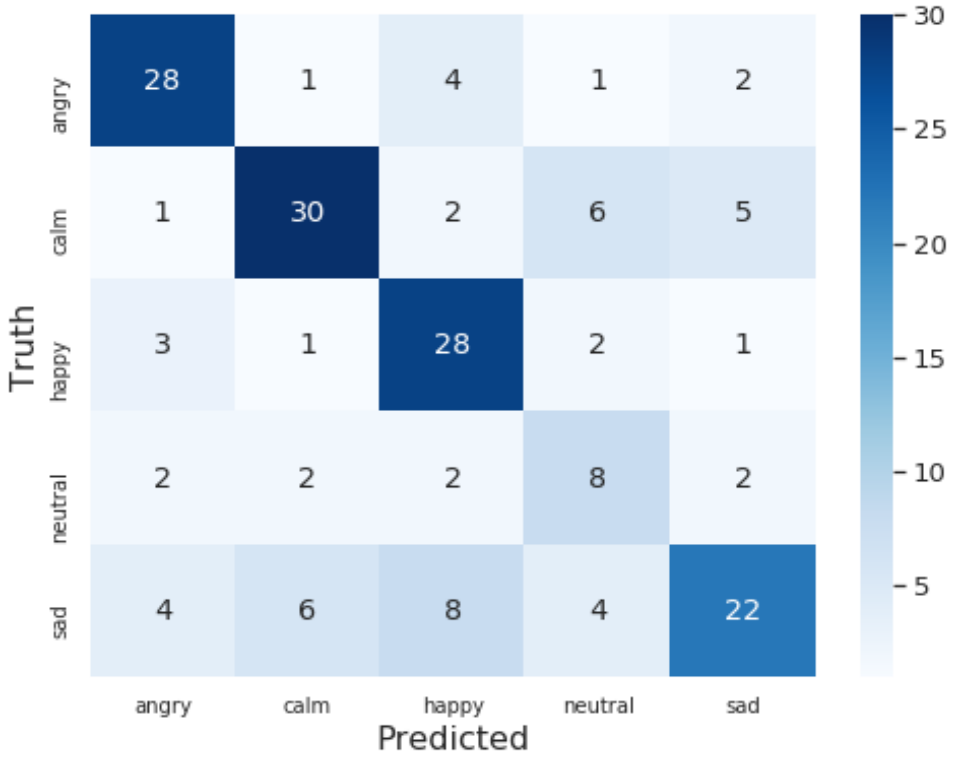
Figure 3.8: GMM Confusion Matrices

3.4 SVM Approach

For our tests with the Support Vector Machine, the same steps were taken in order to extract the features and prepare the model to receive the 2 types of inputs. However, there was a change in the way the group formatted the input, as the temporal factor of the input was taken into consideration and the group opted to line each MFCC in sequence in order to do so.

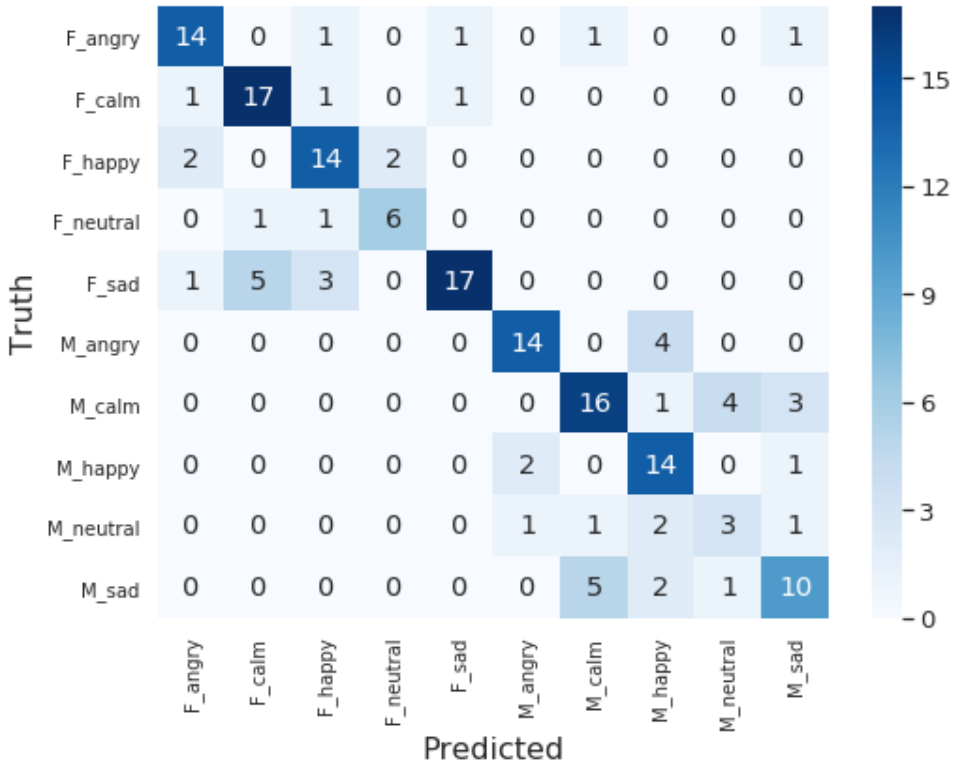
The scikit-learn library was used to implement the SVM, as it has a good implementation of it with the benefit of being able to control multiple parameters of the model. The main parameters used to build it were the following:

- Penalty parameter C: 1.0
- kernel: linear
- decision_function_shape : One vs. Rest



(a) without sex fragmentation

For the SVM experiment without taking sex into consideration, the group has reached 66.3% accuracy and a 0.66 weighted F1 score on predicting the correct label.



(a) with sex fragmentation

Figure 3.10: GMM Confusion Matrices

As expected, as sex was added to the labeled input, the results improved to a 71.4% accuracy and a 0.71 weighted F1 score.

Also, the group ended testing the SVM using different parameters on its set up, namely the type of kernel and the penalty parameter C of the error term. However, both of these changes resulted in marginal gains and losses, and were not high enough for the group to consider them relevant in this specific problem.

3.5 Another different approach - Transfer learning with pre-trained VGG16

Apart from the project main developed methods, the group has also tested another approach to emotion recognition from speech. As mentioned before, the VGG structure was created in 2014 at the ILSVRC2014 challenge, and how its pre-trained layers would fare in a different categorization problem was explored, in this case emotion recognition from speech using the RAVDESS database as input.

In order to better understand the effect of the pre-trained layers into speech recognition, the concept of transfer learning was used. This is a method where part of the already pre-trained model is used and combined with another set of layers, a technique which provides a head start given the vast compute, time and database

resources needed to train the model.

The group used the VGG16 as the base of the combined model, and took off the layers from the third group of convolutionals forward, remaining only the other layers pre-trained with the ImageNet database (and which weights would remain unchanged). Then, a simple trainable layer structure was put in the place of the taken layers, forming a new combined model.

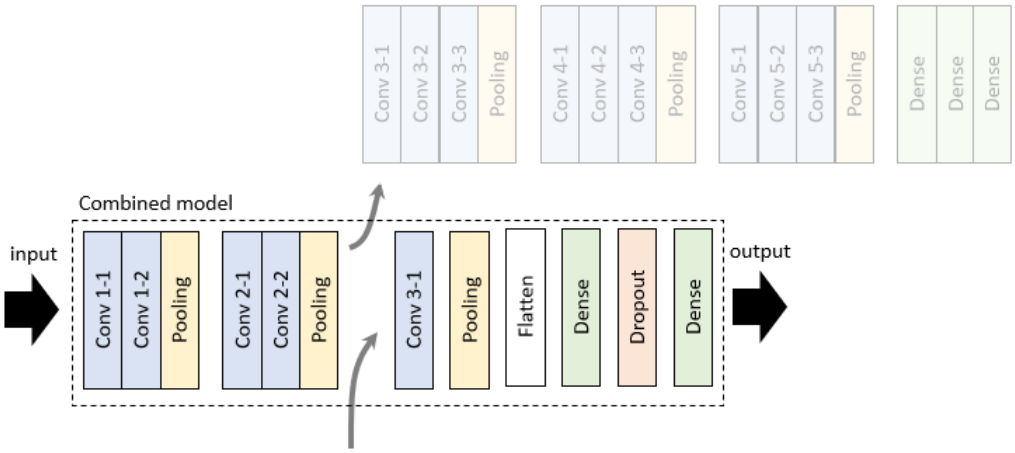


Figure 3.11: Transfer learning from VGG16 structure

Also, as the the VGG was mainly created to recognize images, the input to its model is a 3 parameter input (RGB) in a matrix of pixels. In order to reframe the input to use it with the VGG, the group extracted 33 MFCCs from the audios and used them as one of the colors, zeroing the others.

The test resulted in a very poor classification model, as the accuracy of the recognition remained very low through the epochs (around 10%). Another test was done without the fixed weights of the pre-trained layers, and the results were much better, reaching 65% accuracy.

Chapter 4

Conclusion

4.1 Final Results

Averaging 120Hz for males and 210Hz for females [27], the adult human voice is fairly easily separable by means of the fundamental frequency alone. So, although the duplication of category set size halves the expected accuracy by chance, the introduction of this sex division improved the modelling of emotion space without any or little propagation of error.

Some portion of these performances can certainly be due to the limitations of datasets. Because, they have both a small sample size and low variability of phonetic content or speakers, which causes them to be easily overfitted. And in some cases, as seen in CNN, the model completely specializes in the dataset, even learning internal patterns not related to emotion recognition. In turn the model reaches a higher score but can't be appropriately generalized to other datasets or real world situations.

Table 4.1: Results: F1 score

Model	Acc. w/o fragmentation	Acc. w/ fragmentation
Gaussian Mixture Model	0.79	0.81
Support Vector Machine	0.66	0.71
Convolutional Neural Network	0.77	0.80

As for the experiments of the three models including and excluding the sex factor, the ones with labels considering sex have reached in average 5% higher F1 score than the ones which didn't. Perhaps this phenomena can be explained by a confusion between an emotion from one sex and another emotion from the other (e.g. F_neutral being confused with M_sad).

4.2 Future Developments

Here we expose some of the questionings that arose during examination of results, and initial speculation to what might be the answer. But as they required further investigation that escaped the scope of the project, we present them as potential future development.

The superior performance of GMM inputting frame by frame may suggest that most of the emotional information is localized in the frequency domain inside a single frame. As opposed to being localized in the temporal domain across many frames. In other words, it may be the case that for emotion recognition, determining if something happened at all is vastly more important than when it happened or in what sequence.

Regarding the GMM covariance types, although they are mostly determined empirically, they can reflect interesting aspects about the limitation of our feature space. In our findings, the tied covariance had a superior performance to the diagonal covariance, even though its model is much simpler. It indicates that the emotional information is not so much localized along a feature axis, but rather angled across several feature axis. Therefore, it might be that the MFCC feature space is far from a latent principal component space for emotion recognition.

Also, even though there are multiple emotional databases nowadays, there is none that we've found in Brazilian portuguese. As seen in studies such as Somayeh Shahsavarani's one [24], each model's performance is adapted to some specific language in which the model is trained. Because of this specificity, the creation of a common ground for model testing in emotional speech classification restricted to Brazilian portuguese would facilitate and even develop the boundary of emotion classification through speech in this language.

Bibliography

[1] R.W. Picard. "Affective Computing". In: *The MIT Press* (1997).

[2] A. R. Damasio. "Descartes' Error: Emotion, Reason, and the Human Brain". In: *Gosset/Putnam Press* (1994).

[3] Grand View Research. "Voice and Speech Recognition Market Analysis Report". In: (2018).

[4] Livingstone SR and Russo FA. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". In: *PLoS ONE 13* (2018).

[5] K. Scherer. "What are emotions? And how can they be measured?" In: *Social Science Information* 44 (2005), pp. 695–729.

[6] P. Ekman, T. Dalgleish, and M. Power. "Handbook of Cognition and Emotion". In: (1999).

[7] A. Ortony and T. Turner. "What's basic about basic emotions?" In: *Psychological review* (1990).

[8] P. Ekman. "Cross-cultural studies of facial expression". In: *Darwin and facial expression: A century of research in review* (1973), pp. 169–222.

[9] J. Russell. "Affective space is bipolar". In: *Journal of Personality and Social Psychology* 37 (1979), pp. 345–356.

[10] C. Whissell. "The dictionary of affect in language". In: *Emotion: Theory, Research, and Experience* 4 (1989), pp. 113–131.

[11] R. Plutchik. "The nature of emotions". In: *American Scientist* 89 (2001), pp. 344–350.

[12] Z. Huang. "Speech-based Emotion and Emotion Change in Continuous Automatic Systems". In: (2018).

[13] J. Fontaine et al. "The world of emotions is not two-dimensional". In: *Psychological Science* 18 (2007), pp. 1050–1057.

[14] H. Nordström. "Emotional Communication in the Human Voice". Stockholm University, 2019.

[15] Florian Eyben et al. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". In: *IEEE Transactions on Affective Computing* 7 (2016), pp. 190–202.

[16] S. Kuchibhotla and M. Niranjan. "Emotional classification of Acoustic information with optimal feature subset selection methods". In: *International Journal of Engineering & Technology* 2 (2017), pp. 39–43.

[17] A. Meftah, Y. Alotaibi, and S. Selouani. "Emotional speech recognition: A multilingual perspective". In: *International Conference on Bio-engineering for Smart Technologies (BioSMART)* (2016), pp. 1–4.

[18] S. Deb and S. Dandapat. "Emotion Classification using Segmentation of Vowel-Like and Non-Vowel-Like Regions". In: *IEEE Transactions on Affective Computing* (2017), pp. 1–1.

[19] Scikit-learn. *GaussianMixture documentation*. URL: `https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html`.

[20] *CS231n: Convolutional Neural Networks for Visual Recognition*. `http://cs231n.github.io/`. Accessed: 2019-12-17.

[21] J. Jeon, R. Xia, and Y. Liu. "Sentence level emotion recognition based on decisions from subsentence segments". In: *ICASSP* (2011), pp. 4940–4943.

[22] M. Li et al. "Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling". In: *ICASSP* (2012).

[23] L. Sun, S. Fu, and F. Wang. "Decision tree SVM model with Fisher feature selection for speech emotion recognition". In: *EURASIP Journal on Audio, Speech and Music Processing* (2019).

[24] S. Shahsavarani. *Speech Emotion Recognition using Convolutional Neural Networks*. 2018.

[25] Y Mu et al. "Speech Emotion Recognition Using ConvolutionalRecurrent Neural Networks with Attention Model". In: *2nd International Conference on Computer Engineering, Information Science and Internet Technology (CII 2017)* (2017).

[26] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Visual Recognition". In: *ICLR15* (2014).

[27] Hartmut Traunmüller and Anders Eriksson. "The frequency range of the voice fundamental in the speech of male and female adults". In: (1991).

Catalogação-na-publicação

Gieseler, Daniel
Emotion Recognition in Speech / D. Gieseler, R. Narita -- São Paulo, 2019.
38 p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Mecatrônica e de Sistemas Mecânicos.

1.EMOÇÕES 2.APRENDIZADO DE MÁQUINA ver APRENDIZADO COMPUTACIONAL 3.REDES NEURAIS 4.INTERAÇÃO HOMEM-MÁQUINA I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Mecatrônica e de Sistemas Mecânicos II.t. III.Narita, Rodrigo