

*THE MEL- FREQUENCY CEPSTRUM
COEFFICIENT FOR MUSIC EMOTION
RECOGNITION IN MACHINE LEARNING*

Ai-Phee Chris Yong

Department of Computing

Faculty of Science and Engineering

Macquarie University

This thesis is submitted for the Master of Research degree

September 2019

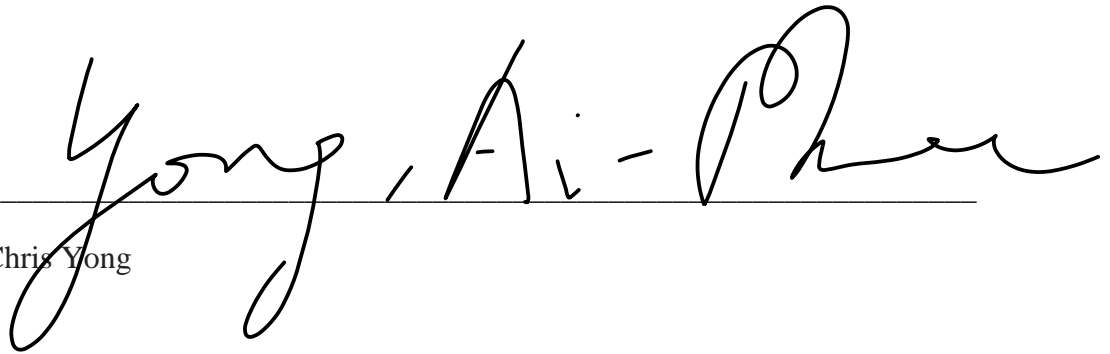
DECLARATION

This thesis is the result of my own work and includes nothing that is the outcome of work done in collaboration, except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

In accordance with the Faculty of Science and Engineering guidelines, this thesis does not exceed 50 pages.

Signed: _____

Ai-Phee Chris Yong

A handwritten signature in black ink, reading "Yong, Ai-Phee", written over a horizontal line. The signature is stylized with large, flowing letters.

Date: 18 October 2018

ABSTRACT

The Mel-frequency Cepstrum Coefficient (MFCC), a technique designed initially for speech analysis, has in recent years become very popular in music emotion recognition projects. MFCC uses the Mel scaling method to simulate human auditory properties, logarithmic noise reduction techniques, and the Discrete Cosine Transformation (DCT) to generalise all salient features, without losing critical information. These techniques, while applicable to speech analysis, may not always be suitable for music analysis. We suggest, in Music Emotion Recognition (MER) analysis, spectral and temporal (which have a deep historical foundation) should be the more relevant features to use.

We propose extracting three feature types, MFCC, Spectral, and Temporal, from the clips of songs in the ‘1000 songs’ dataset to train a simple Artificial neural network (ANN). The trained ANN model will subsequently be able to predict the emotion value of songs. The prediction error is calculated based on the predicted value and actual annotated value. The feature that produces the lowest prediction error is judged as the most suitable feature for MER. Our results show that spectral features produced the lowest error, whereas MFCC produced the highest prediction error; this suggests that MFCC may not be a suitable feature for MER.

ACKNOWLEDGEMENTS

I must first thank Prof. Mark Dras to introduce me to Dr Malcolm Ryan when I desperately needed a supervisor last year. Now under Malcolm supervision, I did not get drown in the vast ocean of papers. I am so grateful that Malcolm steered me back in focus of the project; while the aspect of the whole project is off course.

I thank my wife, Shirley, in patiently listen to my not so interesting technical talk during my studying.

I also want to thank the staff from Virgin Active gym for all the encouragement.

CONTENTS

1	INTRODUCTION.....	1
1.1	MOTIVATION.....	1
1.2	INTRODUCTION OF THE MUSIC FEATURES.....	1
1.3	REPRESENTING EMOTIONS.....	2
1.4	ABOUT MACHINE LEARNING.....	2
1.5	POPULARITY OF MFCC.....	3
1.6	THE FOLLOWING CHAPTERS OF THE THESIS.....	3
2	MUSIC FEATURES.....	5
2.1	MUSIC FEATURES.....	5
2.2	THE MYSTERIES OF THE MAJOR AND MINOR KEY.....	5
2.3	THE STYLE OF MUSIC.....	6
2.4	FEATURES BELIEVED TO TRIGGER EMOTION.....	6
3	EMOTION MODELS.....	8
3.1	THE PERCEPTION OF EMOTION FROM MUSIC.....	8
3.2	DEVELOPMENT OF THE EMOTIONS MODEL.....	9
3.3	DISCRETE MODEL.....	9
3.4	DIMENSIONAL MODEL.....	9
3.5	COMBINED MODEL.....	10
4	MEL-FREQUENCY CEPSTRAL COEFFICIENT.....	12
4.1	THE MEL-FREQUENCY CEPSTRAL COEFFICIENT MFCC.....	12
4.2	MFCC IS AN AMALGAMATION OF THESE CONCEPTS/TECHNIQUES.....	12
	<i>Power Spectrum.....</i>	<i>12</i>

	<i>Fourier Transformation</i>	13
	<i>Cepstrum</i>	13
	<i>Mel scaling</i>	14
	<i>Discrete cosine transformation (DCT)</i>	16
	<i>Windowing/Framing</i>	17
4.3	MFCC PROCESSES	18
4.4	WHY DO WE NEED MFCC?	19
	<i>Features extraction</i>	19
	<i>Segmentation</i>	20
	<i>Noise reduction and filtering</i>	20
	<i>The size of the spectrum vector</i>	20
5	TEMPORAL AND SPECTRAL PROPERTIES	21
5.1	TEMPORAL FEATURES	22
5.2	SPECTRAL FEATURES	22
5.3	SPECTRAL CENTROID	23
5.4	SPECTRAL BANDWIDTH	23
6	EXISTING WORK IN MUSIC EMOTION RECOGNITION	24
7	THE EXPERIMENT	26
7.1	INTRODUCTION	26
7.2	DATASET AND PROPERTIES	27
7.3	THE ANNOTATION OF EMOTION	27
7.4	INTRODUCTION OF THE ARTIFICIAL NEURAL NETWORK (ANN)	27
7.5	APPROACHES/METHODS	35
	<i>Features Extraction</i>	35
	<i>Program and library setup</i>	37

	<i>The hyper-parameter of the perceptron network</i>	<i>37</i>
8	RESULTS AND DISCUSSION	39
8.1	RESULTS	39
8.2	DISCUSSION OF THE STUDY	41
	<i>The logarithm suppressed spectral information</i>	<i>41</i>
	<i>Mel scale may not be suitable for music</i>	<i>41</i>
	<i>The dimensional reduction causes some loss.....</i>	<i>41</i>
	<i>Redundant with spectral and temporal</i>	<i>41</i>
	<i>The missing part of MFCC</i>	<i>41</i>
	<i>Human factors.....</i>	<i>42</i>
	<i>Problem with dimension reduction</i>	<i>42</i>
	<i>The relationship between human voice and music</i>	<i>42</i>
	<i>Machine learning aspect.....</i>	<i>43</i>
8.3	FUTURE WORKS	44
9	REFERENCES.....	45

LIST OF TABLES

TABLE 1 - MUSICAL FEATURES ASSOCIATED WITH EMOTIONS	7
TABLE 2 - RMSE VALUE OF THE PREDICTION ERROR	39
TABLE 3 - PREDICTION ERROR COMPARISON	40

LIST OF FIGURES

FIGURE 1 - CIRCUMPLEX MODEL (JAMES RUSSELL, 1980).	11
FIGURE 2 - C MAJ CHORD SOUND WAVE FIGURE 3- C MAJOR CHORD’S SPECTRUM	13
FIGURE 4 - CEPSTRUM OF THE C MAJOR CHORD	14
FIGURE 5 - MEL SCALING CURVE	15
FIGURE 6 - HAMMING WINDOWING	18
FIGURE 7 - MFCC OVERALL PROCESS BLOCK DIAGRAM.....	19
FIGURE 8 - SIMPLIFIED BRAIN NEURON	28
FIGURE 9 - A COMPUTATIONAL NEURON	28
FIGURE 10 - A MULTI-LAYER TRAINING PROCESS	30
FIGURE 11 - EXTRACTION DIFFERENT FEATURES FROM A SONG	35
FIGURE 12 – MACHINE LEARNING MODELS FOR TRAINING / PREDICTION OF EMOTION ...	36

LIST OF ABBREVIATIONS AND ACRONYMS

Artificial Intelligence	AI
Artificial Neural Network	ANN
Categorical Model	CM
Convolutional Network	CNN
Content-Based Music Information Retrieval	CBMIR
Discrete Cosine Transformation	DCT
Dimensional Model	DM
Discrete Fourier Transformation	DFT
Fourier Transformation	FT
Inverse Fourier Transformation	IFT
Long Short-Term Memory	LSTM
Music Emotion Recognition	MER
Mel-Frequency Cepstral Coefficient	MFCC
Machine Learning	ML
Principal Components Analysis	PCA
Recurrent Neural Network	RNN

1 INTRODUCTION

1.1 Motivation

Audio features extraction is a crucial technique in machine learning used by Content-Based Music Information Retrieval (CBMIR). CBMIR is popular, in recent years, due to the high growth in audio and music demand in the mass media entertainment via the internet. With the proliferation of Songs/Movies Recommendation System to introduce services to the consumer based on their preference and mood. To recommend movies and songs based on a user's emotion status is a new trend these days with the thrust of social media. At the current stage, it is almost impossible to create or understand human emotion by computer (algorithm). Music Emotion Recognition (MER) system provides a useful guideline and approaches towards the understanding of human emotion; subsequently, help in the development of humanised robots. Developers of these systems are desperate to find a suitable audio features extraction tool for the growing media market. With the widespread demand for audio features extraction, understanding the mechanism of the MFCC proved importance.

1.2 Introduction of the Music Features

Music can be fascinating, and not merely regarding the beautiful sounds of melodic arrangements. Music can make us feel emotional – sometimes delighted, other times, melancholic. How does music trigger our emotions? We will leave this question to music psychologists. In this paper, we attempt to discover the emotional quality of music via computing. It is not an easy task for a non-musician, as musicology and emotion in the context of psychology is an area of study vast in scope.

We can discover some clues from the use of features that describe the characteristics of music. Music has many features, e.g. **Pitch**, **Tempo**, and **Timbre**. These features are concerned with the frequency of notes (pitch), the timing for when specific sound notes should be played (tempo), and the quality of the sound of notes (timbre). The style of music is not a feature; different composers, when creating music, choose to use a specific set of note patterns or a combination of features that is familiar to them and apply it to express emotions or moods. Another common practice by composers is the use of the **Major** mode to conduct 'happy' music, whereas the **Minor** mode is used for 'sad' music. This practice has a long history that dates to ancient Greece when different music modes were employed to present a diverse range of expressions. Can we easily dissect music using these features to find evidence of the emotions therein?

1.3 Representing Emotions

Emotions can be observed as unusual human behaviour. One would ask, why do humans have emotions at all? However, answering this question is not the aim of this study. In this research, we only require understanding the ‘**representational form**’ (Rep. Form) of emotions for measurement and computation purposes. Among the numerous psychological emotion models, the categorical model (CM) and the dimensional model (DM) apply to our study. The discrete model employs the actual wording related to emotion, e.g. ‘delighted’, ‘joy’, and ‘laugh’ to describe how emotions feel. The dimensional model uses valence and the arousal components, with assigned values (from -1.0 to 1.0) to represent emotion; for example, a smile in DM will be valence = 0.5 and arousal = 0.5.

For a problem to be computable, variables (input and output data) for the said problem must be available in a quantised numerical value. All musical features and the emotional quantity were converted or extracted to a Rep. Form. Many musical features may not be available in a computable form, however, such as the style of music, and adaptation of the Major/Minor mode, the latter being a good indication of the emotions in music. Since this paper deals with sound as a media form of music, we can utilise acoustic physics processing techniques. In brief, the **frequency aspect** and the **time aspect** of sound, which we referred to as **spectral** and **temporal** properties, respectively, are provided in Rep. Form, thus rendering them computable. Pitch, tempo, and timbre belong to a spectral features group and are therefore included in the Rep. Form.

To recognise emotion in music, we need the latter’s most representable features, as a variety of features may reflect different types of emotions. The options in this context generally revolve around either pitch spectrum features, time series features, or both. Moreover, it seems like common sense, as both features are related to acoustic physics, and therefore, related to music. Surprisingly, in many recent music research projects, a features extraction technique known as MFCC (Mel-Frequency Cepstrum Coefficient) is gaining popularity.

1.4 About Machine Learning

In recent years, machine learning (ML) techniques have become a fast-growing area in the field of computing. It represents promising technology that can potentially resolve many severe problems, e.g. those related to image recognition. With advancements in the development of machine learning – inspired by human intelligence – this area of research has branched off into a unique field known as artificial intelligence (AI). An

artificial neural network (ANN), a concept based on the human neuron network, is the critical component of AI. A large body of research has discovered that AI/ML can help to resolve abstract and complex problems, such as facial recognition, speech recognition, and pattern recognition problems. An ANN is designed to learn from input data and deduce a predictive model from the input data. This model is subsequently used to predict new outcomes for a new set of input data.

1.5 Popularity of MFCC

MFCC is well-recognised for its timbre texture and spectrum features, which are the highest features used in speech recognition. Many different sound analysis projects included MFCC as a primary feature set. The results achieved by Li and Chan (T. L. H. Li & Chan, 2011) in their music genre classification project, speech recognition project (Dave, 2013), as well as many others, are laudable. It raised the question, “What is the effect of audio quality on the robustness of MFCC and chroma features”, particularly in the context of music (Urbano, Bogdanov, & Herrera, 2014)?

MFCC was initially developed for speech analysis (Davis & Mermelstein, 1980), and has since served as a state-of-the-art feature in speech research. The popularity of MFCC did not spread to music analysis until recently, likely due to its speech-oriented approach. Although music and speech share similarities, they are different in many ways. In human speech, the physiological structure of our trachea, throat, and tongue govern the quality of the voice, whereas, in most cases, instruments dominate the characteristics of a music piece. Simply stated, they are both spectra of a sound wave; however, music tends to have a broader spectrum range, while the speech spectrum is more clustered.

Considering these differences, the real reason for the good results derived from using MFCC for music is not yet fully understood. An early paper (Logan, 2000) questioned the suitability of MFCC for music by examining the structure of processes included in the MFCC. The purpose of this study is to extend the research on Music Emotion Recognition by investigating the relationship between the components of music and emotion with the MFCC features.

1.6 The following chapters of the thesis

Music Features - Discuss the characteristic features involved in music that give us hints of emotions. Moreover, how the traditional understanding of the musical features is related to emotion.

Emotion Models - Emotion is abstract and hard to understand. Be able to resolve with computing model, the basic structure and the psychological models, namely, the

Categorical and the Dimensional model are discussed. What the Valence and Arousal value in the models is?

Mel-Frequency Cepstrum Coefficient - Theoretically, MFCC is a complicated formula; we broke down the processes and discussed the general concept involved. In each step, we illustrated with the graphical result of input sound wave in the processes to understand the transformation and the mathematical functions involved.

Existing Work in MER - discusses some works using the more complicated Neural Network for MER, and the problems they are facing. Moreover, other projects using different options of MFCC to achieve a better result. Thus, raised the question if the MFCC is still suitable for musical analysis.

The experiment - describes the approach, model and the procedures we used to test the MFCC, spectral and temporal features with a simple perceptron.

Result and Discussion - summarised the observation of the result and discussion of the issues and the suitability in the use of the MFCC in MER. We concluded that it might not be the desirable features in comparison to the spectral and temporal features.

2 MUSIC FEATURES

Music comprises groups of sound put together in a pattern based on the composer's previous experiences. According to the active pattern, human listeners may or may not enjoy listening to a musical piece. The degree of enjoyment involved is different for many people, is influenced according to their cultural background, and their understanding of music. Music listeners are likely to experience music enjoyment if the sound is harmonised.

2.1 Music features

The tempo of the music is a good indicator of the excitement present in the music. Fast music can give people a feeling of significant excitement, perhaps that something is going to happen, whereas slower music tends to induce feelings of calmness, mystery, and sadness.

The beat is also another important feature in music to control the mood of the music. It is the pulse of rhythm in music. It is usually the beginning of a bar (section of notes in a repeating sequence stressed and unstressed in loudness ("strong" and "weak" pattern).

A music **genre** is the property of music that provides clues to the emotions that correspond to the music; for example, 'blue jazz' is a representation of a type of despondent music. The genre is not a music structure; instead, it is like a specific category of music. Musicians generally use a set of conventional structures in their music, which shapes the properties of the genre.

Slow tempo, low pitch, and low loudness level are associated with sad expressions (Juslin & Sloboda, 2010). However, none of these features can conclusively point to an emotion; instead, a blend of different features is needed.

Kate Hevner (an expert in music psychology) showed that individual features are not enough for understanding emotion in music. Extraction techniques are needed that can summarise all features rather than individual features (Hevner, 1937).

Other studies have shown that specific music structures arouse the same emotion in listeners (Meyer, 2017).

2.2 The mysteries of the major and minor key

The use of major and minor modes in Western music has a long historical evolution from its origins in Greek musical cultures. Musicians in ancient times used different

modes to create a mood for the expression of emotions. The modern day's major and minor mode originated from ancient Greek musical practice.

The Major mode is linked with the positive characteristics of being dynamic, determined, defined, and more natural. It expresses joyfulness and excitement; it sounds sharp, hopeful, forward-looking, and happy.

The minor mode's characteristics are passive and depressing; it expresses sorrow, despair, grief, mystery, and melancholy. It sounds low, hopeless and sad – the negative of the major.

Many musicians recognise that using major or minor mode is not the only factor involved in emotion recognition. Preferably, it is the overall effect of music that must be observed, inseparable from rhythm, harmony, melody, intensity, and tempo. Moreover, general emotional feelings cannot be apprehended in a single moment in time but as part of a sequence, influenced by what had immediately been heard previously (Hevner, 1935).

2.3 The style of music

The style of music is a good indicator of emotion; for example, 'blue jazz' provides clues regarding the sadness of early slavery life for black Americans. The style of music is a complicated topic, even for well-trained musicians. It involves not only musicology but also the cultural background of the composers and listeners. For example, a Western European listener may experience significant difficulty appreciating Chinese music. The style is a problematic feature to extract; most musicians instead extract dedicated features and distinct patterns in music to indicate its style.

2.4 Features believed to trigger emotion

The following table lists musical features associated with different emotions. Tempo is typically regarded as the most important, but other factors such as mode, loudness, and melody, can also influence the emotional valence (positive/negative emotion) of music.

Table 1 - Musical Features associated with Emotions

Structural feature	Musical properties	Associated emotions
Tempo	The speed or pace of the music.	Fast tempo: happiness, excitement, anger. Slow tempo: sadness, serenity.
Mode	Type of scale	Major scale: happiness, joy. Minor scale: sadness.
Loudness	The strength (hardness in volume) and amplitude of a sound.	Intensity, power, or anger.
Melody	The incremental/decremental effect of a musical tone.	Gradual and incremental harmonic change: happiness, relaxation, serenity. Sudden harmonic change: excitement, anger, unpleasantness.
Rhythm	The repeating pattern or beat of a song.	Smooth/regular rhythm: happiness, peace. Rough/irregular rhythm: amusement, uneasiness. Varied rhythm: joy.

3 EMOTION MODELS

3.1 The perception of emotion from music

Musical expectancy refers to a process through which a listener can feel emotion when the regular pattern of music is interrupted. Regular progression of pitch (any spatial feature) induces a feeling of calmness and comfort. When this calmness is interrupted, our essential human alertness is triggered to manage the change. This alertness is part of our human emotion. For example, when experiencing fear as a result of perceived danger, we can react through the action of running away or remaining in where we are to face the danger (the flight or fight response). Leonard Meyer (Meyer, Kraehenbuehl & Meyer, 1961), a pioneering music emotion psychologist, linked this psychological behaviour to emotion and meaning in music. Meyer examined the structures and features of music that correspond to emotion. We can observe this from the melodies employed by Samuel Barber, known for creating the sadness of the music. Barber generally adopted simple melodies and lengthened the musical phrase to prolong a feeling of sadness. Then, suddenly, this peace is interrupted by shortening the phrase and introducing a change in feeling. This thrown-off-balance undermines the rational expectations of listeners (Larson, 2018).

Ancient Greek musicians composed music based on modes; they believed each of these modes harmonised musical tunes with the sounds of nature, spiritual feeling, and internal human emotions (James, 1995). These modes are merely arrangements of tonal patterns and tonal intervals (“a concord of tones separated by unequal but carefully proportionated intervals”). Popular musical modes include Dorian, Phrygian and Lydian. These modes are meant to convey different types of mood and thus, elicit emotion.

A standard definition of emotion is that it is human feeling corresponding to an internal or external stimulus. Happiness, love, fear, anger, sadness, and hatred are emotions that can arise because of experience, mood, a personality responding to a situation or scenario, what we observe, or a piece of music that we hear.

Vast research has attempted to define emotions, but to date, there exists no definitive answer. As natural phenomena, it is difficult to capture all the underlying complexities of emotion into a single definition. Charles Darwin argued that emotion is a human skill

of survival (Darwin, 1872). We ‘fear’ danger, and as a result, we engage in the fight or flight response.

3.2 Development of the emotions model

In a music emotion recognition (MER) system, the primary difficulty is the measuring of emotions. To measure human emotion, researchers categorise the basic emotions we have and quantify emotion according to values. Both the categorical model and multidimensional approaches are popularly employed for identifying emotions (Shetty, Kasbe, Jorwekar, Kamble & Velankar, 2015; Shetty et al., 2015).

Psychologists model emotions in two ways: dimensional and categorical/discrete models. The dimensional model (DM) is a more natural approach for quantification, whereas the categorical model (CM) is easier to understand. DM uses quantified values, whereas CM maps to actual words describing a feeling, e.g. ‘happy’, ‘fear’, ‘disgust’, and ‘angry’.

3.3 Discrete model

In 1935 study, Kate Hevner used a list of adjectives to define the distinct types of emotions from the listeners. This list of emotion has since become the standard set of discrete emotions used by emotional studies. In the 1960s, the concept of emotions was further developed into six basic components or 256 types of emotions.

Many argue that emotion is just the composition of the basic components that can be identified as happiness, fear, sadness, and anger.

Silvan Tomkins (Tomkins, 1962) concluded that there are **eight primary** emotions: surprise, interest, joy, rage, fear, disgust, shame, and anguish. Any other emotion Tomkins classified as a combination of basic emotions.

Charles Darwin is a well-known supporter of the basic emotions model. Paul Ekman (Sabini & Silver, 2005) and Carroll Izard (Juslin, 2018) argue that there are various similarities in the ways people across the world produce and recognise the facial expressions of at least six emotions.

3.4 Dimensional model

Valence refers to a feeling of pleasantness (positive) or unpleasantness (negative) about stimuli or a situation. Emotions are commonly associated with a negative value for anger and fear, and a positive value for joy and happiness.

The value of valence is difficult to obtain; how do we evaluate a smile’s valence as being lower than that of laughter? How can the degree of grief be measured using a

quantified negative value? Derived from the observation of expressions made by facial muscle activity, new techniques such as the Facial Action Coding System and modern brain imaging can assist in obtaining the value of valence.

Arousal measures the calmness or excitement level present in valence. In a dangerous situation, our body's physiological reactions cause our palms to become sweaty and our heart to beat faster, situating us in a high-arousal emotional state. Arousal arises from our reptilian brain. The reptilian brain detects whether the situation is threatening or favourable, using our sensory organs (eyes, mouth, skin/hands). This triggers the fight or flight response, which originates from within our survival instincts.

3.5 Combined model

The two-dimensional circumplex model offers a means for combining the discrete and dimensional models in an ordinate-format; this provides a tool for characterising and converting the discrete emotion and the quantised value of emotion.

In 1980, James Russell developed the circumplex model (Russell, 1980). This model suggests that emotions are distributed in a two-dimensional circular space, which includes Arousal and Valence. Arousal represents the vertical axis, and valence the horizontal axis, while the centre of the circle represents a neutral valence and a medium level of arousal. The different type of emotions in the discrete model is evaluated with a Valence and the Arousal Value and placed in the circle of the circumplex model.

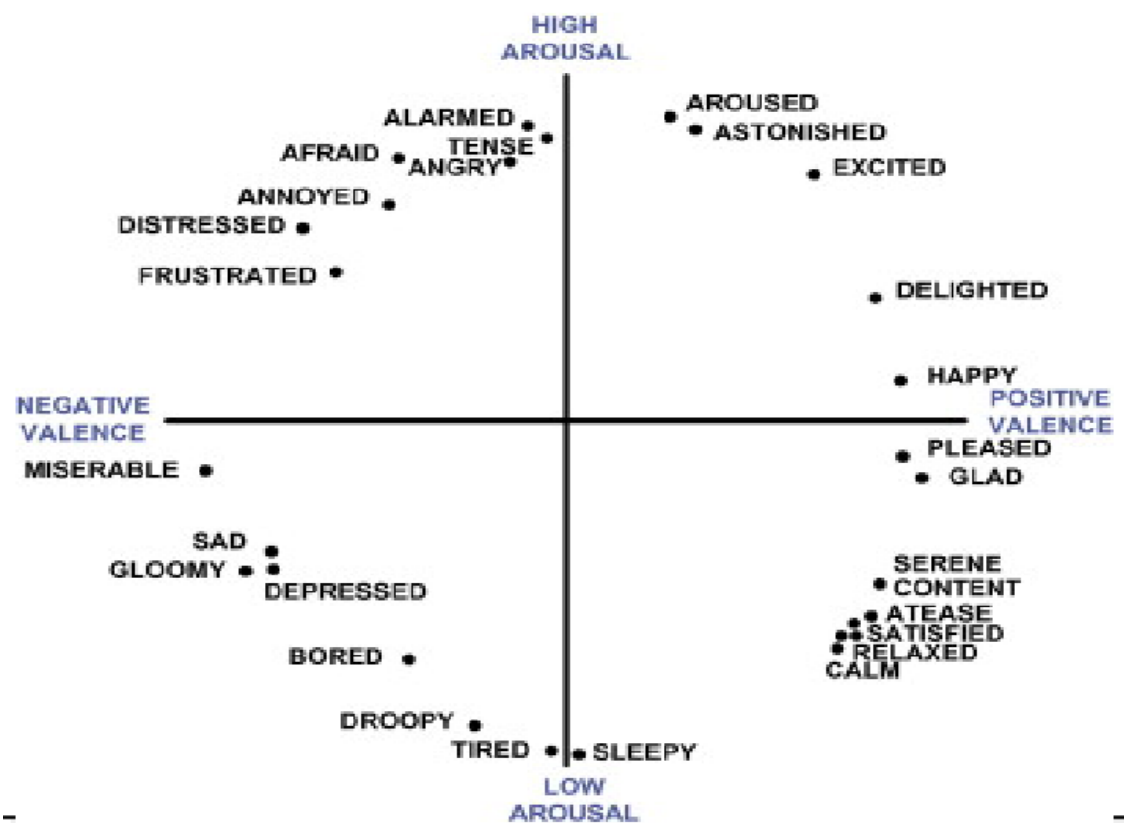


Figure 1 - Circumplex model (James Russell, 1980).

4 MEL-FREQUENCY CEPSTRAL COEFFICIENT

The MFCC is a feature representation popularly used in speech analysis. Davis and Mermelstein introduced the coefficient in the 1980s, and it remains a contemporary technique in the field of sound analysis. The reason for this is that MFCCs can represent sound better modelling the human auditory system.

4.1 The mel-frequency cepstral coefficient MFCC

MFCC is essentially a transformation, and re-grouping process like Cepstrum, except the Frequency for Cepstrum, is rescaled using the Mel conversion; thus, the name the Mel-Frequency Cepstral Coefficients.

MFCC is defined as the following:

$$\mathbf{MFCC}(k) = \mathbf{DCT}(\log_{10}|\mathbf{DFT}(\mathbf{Mel}(k))|^2) \quad \text{Eq. 1}$$

where k is the sample signal

DCT - Discrete Cosine Transformation

DFT - Discrete Fourier Transformation

Mel – Mel scaling

Note that the above formula is almost the same as Cepstrum below, only that IFT(...) in the Cepstrum formula is replaced with DCT(...) and the signal k in DFT() is now Mel(k).

4.2 MFCC is an amalgamation of these concepts/techniques

Power Spectrum

The power spectrum $P(k)$ for the signal $S_i(k)$ is given by:

$$P(k) = \frac{1}{N} |\mathbf{DFT}(S_i(k))|^2 \quad \text{Eq. 2}$$

Fourier Transformation

In acoustic physics theory, a sound wave can be decomposed into a spectrum of other higher frequencies. The standard process to decompose the sound is known as “Fourier Transformation” (FT). This spectrum represents the properties of that sound. Here the sound of the piano, C-Major chord, in a wave format (Figure 2) is decomposed to the spectrum (Figure 3) of different frequencies. We can see the different frequencies below the 5000 Hz, making up most of the C-Major chord.

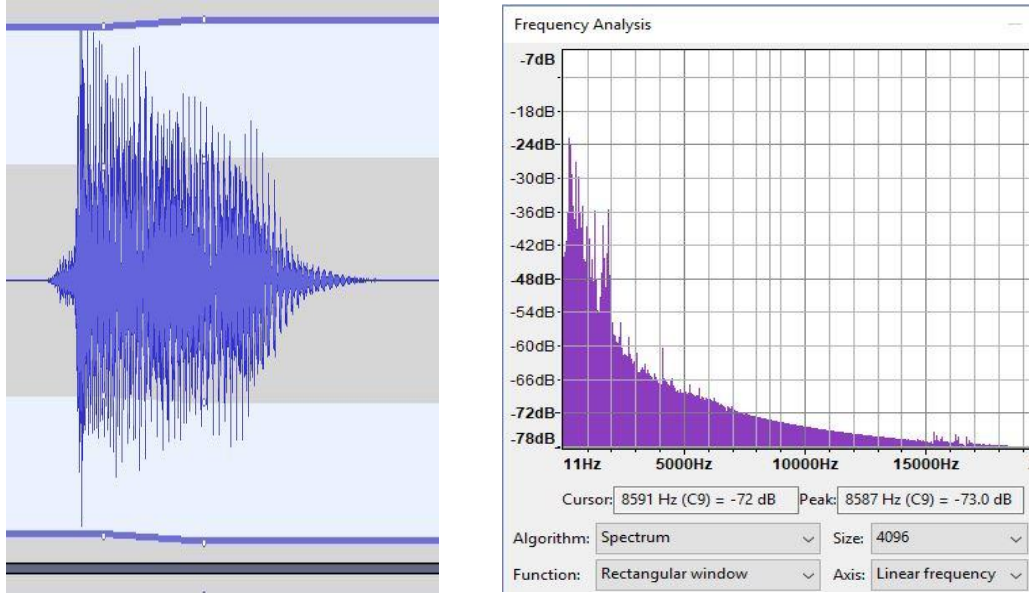


Figure 2 - C Maj Chord sound wave Figure 3- C Major Chord’s spectrum

(Above Figures created with a piano sound to Audacity (audio software))

Cepstrum

In the 1963 paper by Bogert et al., titled with all strange name, “The Quefrency Alanysis of time series for Echoes: Cepstrum pseudo-autocovariance, cross-cepstrum, and Saphe cracking”(B.P. Bogert, 1963), Cepstrum techniques were used for the first time in the detection of seismic echoes, a technique that is currently applied to many sound analysis applications.

A **Cepstrum** is a value, $c(k)$ is the value by Inverse Fourier Transform (IFT) of the logarithm of the spectrum (Discrete Fourier Transformation (DFT)) of a signal, the formula of which is shown below:

$$C(k) = IFT(\log_{10}|DFT(k)|^2)$$

Eq. 3

where

k is the sample signal

IFT = Inverse Fourier Transformation

DFT = Discrete Fourier Transformation

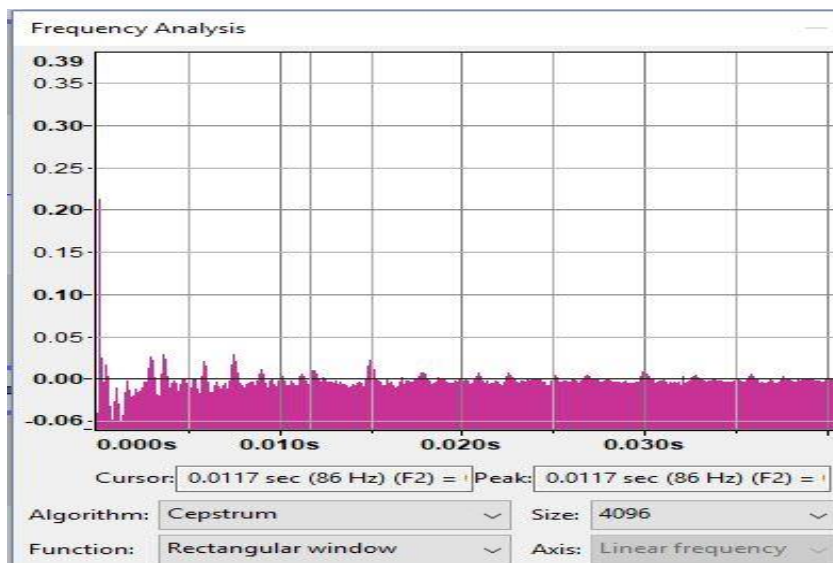


Figure 4 - Cepstrum of the C Major Chord

(Figure is generated by the Audacity by input a piano sound of middle C Major Chord)

Mel scaling

Mel scale formulates the perceptual scale of the non-linearity of human auditory characteristics to the pitches in the human audible range. The human ear exhibits a lazy effect with higher sound pitches, where we are less sensitive to a higher pitch. A feeling of loudness appears to follow a Mel scaling curve as follows. This curve is linear in the low range but logarithmic in the high range. The term ‘Mel’ is derived from the word ‘**Melody**’, to indicate that the scale is based on pitch comparisons.

Essentially, 1000 mels equals 1000 hertz; a frequency lower than 1000 hertz increases linearly to meet the 1000 mel scale, whereas a frequency above 1000 hertz will slowly lose out to the Mel scale (see Figure 2). Four octaves on the Hertz scale above 500 hertz equal roughly two octaves on the Mel scale.

A formula to convert f (Hertz) into m (Mel) is:

$$mel = 2295 \log_{10} \left(1 + \frac{f}{700} \right)$$

Eq. 4

It is plotted as follows:

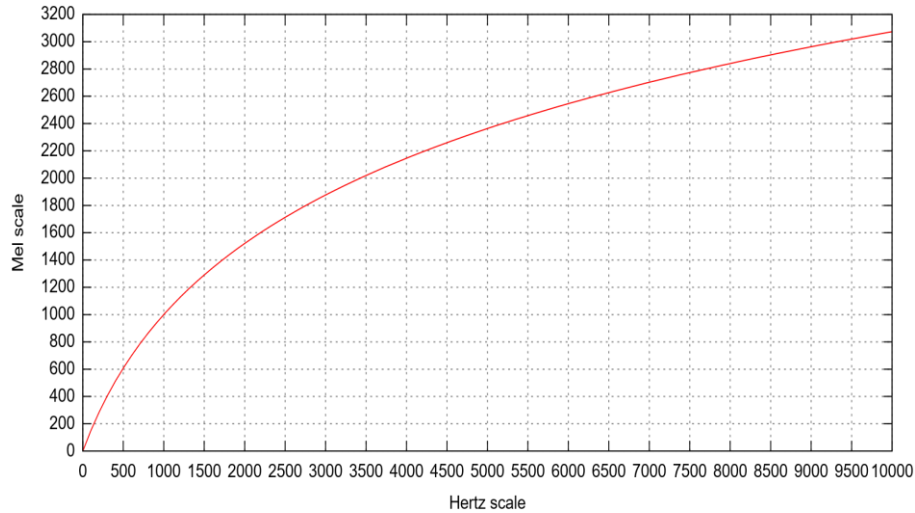


Figure 5 - Mel Scaling curve

(Figure from https://en.wikipedia.org/wiki/Mel_scale)

Because of the Mel-effect of our ear, we take a small window of periodogram bins and sum them up to calculate the energy exists in various frequency regions.

As the frequencies go higher in the filters, we become less sensitive about variations. We are interested in roughly how much energy occurs at each spot. The Mel scale tells us exactly how to space our filterbanks and how wide to make them.

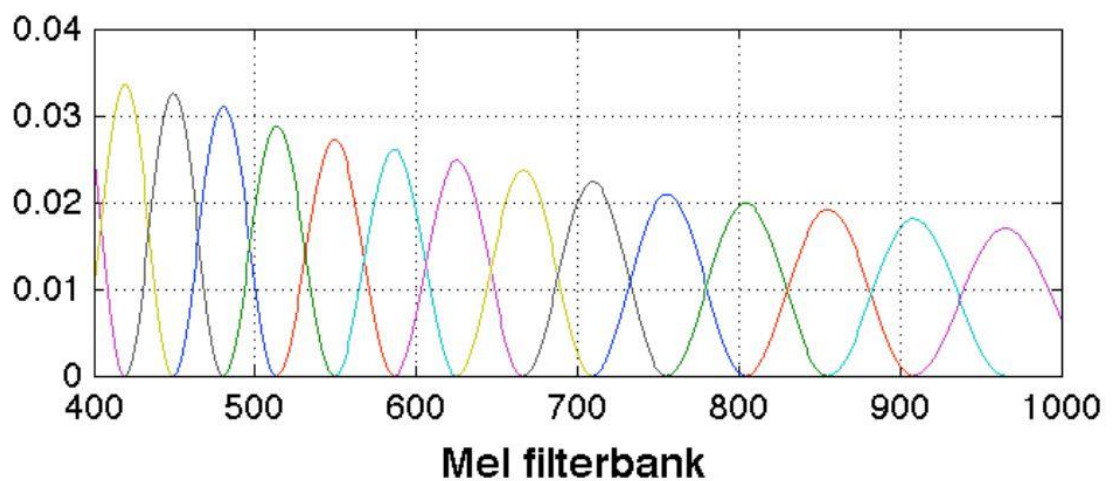


Figure 6 - Mel Filtered Bank bin(Bello, 2013)

In figure 7, we can see how each of the filter banks in capturing the power spectrum in different frequency regions as follows.

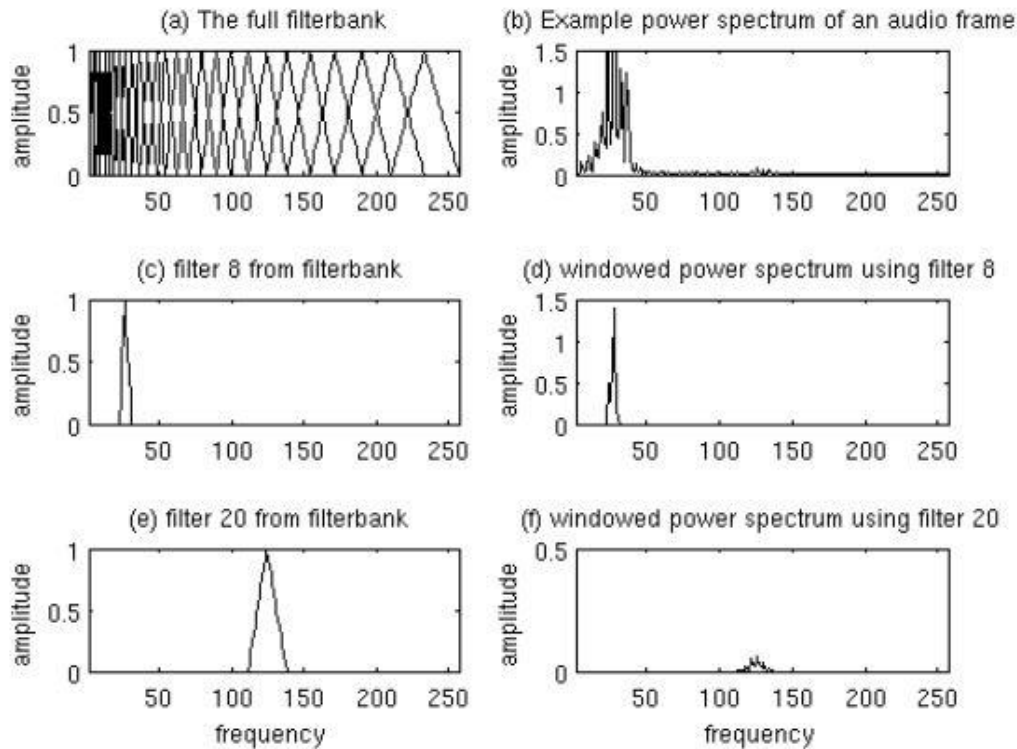


Figure 7- Mel filter Bank bin mask with the Power Spectrum

(Figure from <http://practicalcryptography.com/miscellaneous/machine-learning/guide-melfrequency-cepstral-coefficients-mfccs/>)

Discrete cosine transformation (DCT)

A **discrete cosine transform** is a mathematical transformation process that is commonly used in image processing to compress a large size image into a smaller encrypted form, such as the “jpeg” file.

The final step in MFCC is to compute the DCT of the log filter bank energies. As the Mel-filterbank bins are all overlapped; the filter bank data are strictly correlated with one another. The DCT decorrelates and compresses the data into a few coefficients; this eliminates the redundant information in the data.

DCT is a real-valued transformation, like the Discrete Fourier Transformation as in Cepstrum's IFT. The IFT is replaced with the DCT, which approximates the IFT and compresses the data.

Most of its energy is concentrated on a few low coefficients (effectively compressing the spectrum)

$$X_k = \sum_{n=0}^{N-1} x_n \cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) k \right] \quad k = 0, \dots, N - 1. \quad \text{Eq. 5}$$

Windowing/Framing

Continuity of sound produced a large amount of data and required substantial processing. Each different parts of sound can reflect different properties at a specific instant. So, it is common to use framing techniques on the signal, cut it into smaller pieces, for analysis.

Depending on the circumstances, the overall effect is obtained for all the pieces and then summarised to get an idea about the ultimate effect of the sound.

We can assume that frequencies in a small part of the signal are stationary over a very short period. Therefore, by doing a Fourier transform over this short time frame, we can obtain a good approximation of the frequency contours of the signal by concatenating adjacent frames.

In some cases, we may want to emphasise sections of the signal to reveal to the not-so-obvious feature in the signal. Alternatively, we are interested in a specific part of the signal, different shape of windowing can help to do that.

For example, the use of the Hamming windowing to emphasise the middle section of a signal.

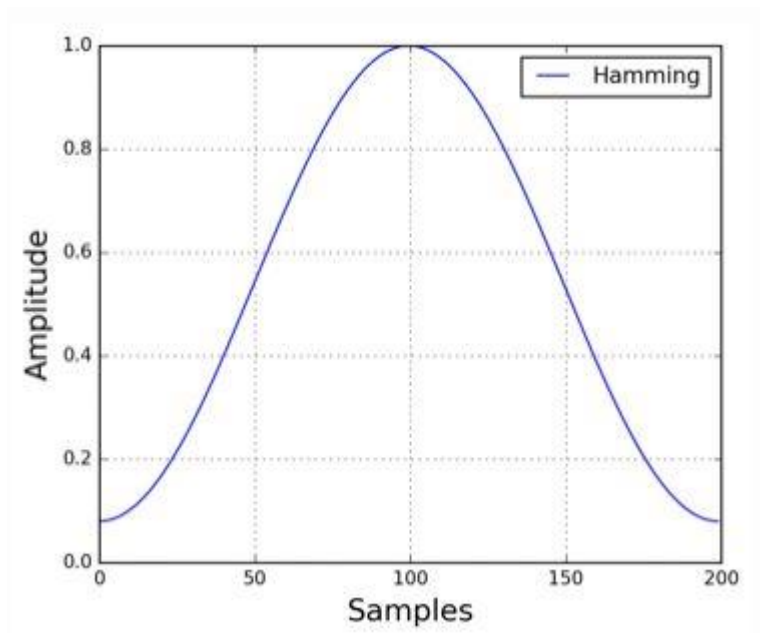


Figure 6 - Hamming Windowing

(Above figure from <https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html>)

4.3 MFCC processes

In practice, the MFCC formula is translated into the following processes:

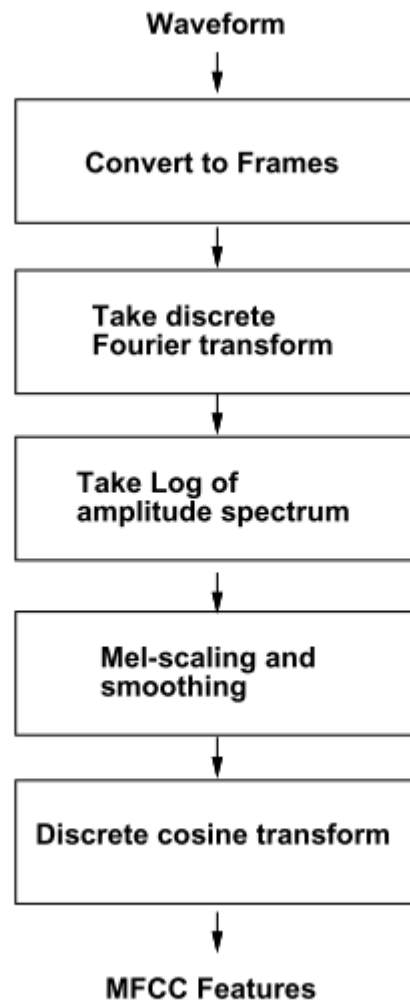


Figure 7 - MFCC Overall process Block Diagram

4.4 Why do we need MFCC?

Features extraction

The selection of features in machine learning (ML) is a critical process. ML relies on salient features data to deduce a predictive model. If the selected features are not reflecting the true nature of the problem, the predictive model is either unusable or can produce the wrong conclusions. However, what is the real nature of the problem? In other words, how to select features related to the conclusion of the problem? In many cases, researchers must use some statistical techniques, such as correlation and covariance, to establish the relationship of features and the designated problems. In this

project, as illustrated, we have based on some historical understanding of music and emotion to decide the features, that is the pitch, tempo and the timbres of music.

Segmentation

Continuity of sound requires substantial processing. Different parts of sound reflect different properties at a specific instant. We need to use the framing techniques on the signal, cut it into smaller pieces, for analysis. The overall effect is obtained for all the pieces; this information is then summarised to get an idea of how much energy or the pattern exists in various sound regions.

Noise reduction and filtering

Noise in the signal represents a significant problem in signal analysis. The traditional approach for reducing the presence of noise in a signal is through filtering. Different types of noise will require different filtering techniques. In Cepstrum, the logarithmic operation is used to suppress the noise level, and to detect an echo. This operation may not be needed, as music timbre is indistinguishable from noise.

The size of the spectrum vector

When sound is decomposed into a spectrum, the size of the frequency vector grows by the range of the spectrum. We want to compress the size of the data but without losing too much of the timbre quality. Although in the compression process, there is some loss of the original data, the crucial data is retained.

5 TEMPORAL AND SPECTRAL PROPERTIES

There is no easy measurement of the music features, such as, “Tempo”, “Melody” and “Rhythm” and “Timbre” from audio file. These terms are somewhat abstract with respect to a scientist who needs to measure more precise information.

Luckily, with the techniques from signal processing, we can extract the temporal and spectral properties from the audio file.

And we know that music sound wave is a superposition of different harmonics wave. As a result, the music sound that we receive is a spectrum (different pitches of sound).

In acoustic physic, we can attribute sound wave with different properties as spectrum distribution or the shape of the spectrum distribution. From the distribution of the spectrum, we can thus interpret the essential music features as listed for further processing.

Now let look at the properties of these Music Features:

Beat is the regular pulse of sound in music, for example, when we count, tap or clap along with the music. Each of the sound pulses is usually similar in loudness.

Tempo is the speed of the Beat in music, that is the Beats Per Minute (**BPM**). For example, at 120 BPM there will be 120 beats in one minute. Tempo governed the style of music terms, such as Slowly, Fast, Allegro, or Largo. A slow tempo is associated with sad expressions (Juslin & Sloboda, 2010).

“**Timbre** depends primarily upon the frequency spectrum, although it also depends upon the sound pressure and the temporal characteristics of the sound” (Acoustical Society of America Standards Secretariat 1994).

The word “**Melody**” in music refer to the sense of “tune”.

And **Rhythm** is referring to the sense of ‘Beat’ and ‘Tempo’.

Tonality in music involves organizing musical structure around a central note. Generally, any Western or non-Western music is periodically returning to a central, or focal tone. More specifically, tonality refers to the particular system of relationship between notes, chords (3 notes together), and keys (sets of notes and chords).

Music depends on both melody and rhythm. Melody adds timbre of music, whereas rhythm adds the pace of the song.

For music to be computable, we have to use the temporal/spectral properties to represent the above properties.

There are more music psychology researches which classify/quantify the relationship between each of music features and the temporal properties these components. We are not looking into the details of this here but just a summary of the effect of the music features.

As we can see in the Existing work, the mixture of the features for use in the research is often not ascertained.

5.1 Temporal features

In the time domain approach, the change of pattern in the loudness or amplitude of sound signifies the properties of sound. It is referred to as the temporal properties.

One example of a temporal property is zero-crossing rate (ZCR), which is the number of times the level of sound crosses over zero. ZCR measures the rate of loudness crossing over to quietness; in other words, it is counting sound pulses. The speed of the pulses is, therefore, the speed of the beats. So, it is a good indicator of Tempo. And we can see from Table 1 that Tempo is vital in expressing emotions.

$$\text{ZCR}(\mathbf{m}) = \frac{1}{2N} \sum_{n=-\frac{N}{2}}^{\frac{N}{2}} |\text{sgn}(\mathbf{x}(\mathbf{n} + \mathbf{mh})) - \text{sgn}(\mathbf{x}(\mathbf{n} + \mathbf{mh} - 1))| \quad \text{Eq.9}$$

$$\text{Where, } \text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

5.2 Spectral features

There are many spectral features; for example, the central frequency (centroid) of the spectrum, the highest frequency of the spectrum, and shape of the spectrum, are generally classified as the spectral properties.

These spectral properties are useful for describing the timbre of music.

5.3 Spectral Centroid

Spectral Centroid (SC) is associated with sound brightness and can indicate the type of sound, e.g. human voice or musical instruments. Human voice trend to cluster in a specific range referencing to a centroid frequency. Centroid can also be a quick summary of the music, such as saying, the G major is the key for a piece of music.

SC is defined as:

$$SC(x) = \frac{\sum_k f_k |X(x,k)|}{\sum_k |X(x,k)|} \quad \text{Eq. 6}$$

5.4 Spectral bandwidth

Spectral bandwidth (SB) is a measure of the range of the pitch in sound and can be used to characterise the asymmetry and distribution of pitches. Each individual instrument also tends to cluster around a specific range of pitches.

SB is defined as:

$$SB(x) = \frac{\sum_k (f_k - SC(x))^2 |X(x,k)|}{\sum_k |X(x,k)|} \quad \text{Eq. 7}$$

5.5 Spectral Flatness

The spectral flatness (SF) of the spectrum indicates a change in pitch, from a high pitch to a lower pitch, and vice versa.

The constant maintenance of pitch usually represents the emptiness (the slow progression) of music. The sudden change of pitches also indicates something is happening. The music pieces make good use of expectancy, sudden change of key and timing, to create rhythmic moods.

SF is defined as:

$$SF(x) = \frac{(\prod_k |X(x,k)|)^{\frac{1}{K}}}{\frac{1}{K} \sum_k |X(x,k)|} \quad \text{Eq. 8}$$

k = band number, K = total number of bands, x = signal

With the advent of artificial intelligence architecture and the promise of self-learning, many music emotion recognition systems are employing artificial neural networks (ANNs) and getting good accuracy in predicting emotions. Malik et al. (2017) combined the recurrent neural network (RNN) (to deal with the time-series nature of music) and the convolutional neural network (CNN) (for feature pattern recognition power). They modified the design of stacked CNN and RNN for continuous prediction of emotion in the valence-arousal space. This modification reduced a significant amount of network parameters, and outperformed the advanced method in 2015 created by Li, Tian, Xu, Ning and Cai, 2016), who used a deep bidirectional short-term memory (DBLTSM) approach.

Many published works often combine MFCC with other features such as temporal and spectral properties. Among the regular features used in MER, there is a search for new features that can recognise musical emotions. Arguably, MFCC has already embedded/summarised all the necessary information of music, in which case there should be no need to combine these musically related features. Researchers often must select between better features or more complicated models. It is the fundamental reason most researchers employ the standard features set (openSMILE, comprising 260 features) to avoid having to choose potentially better features.

The potential of using AI techniques to solve a number of problems is seen as a trend towards adopting more complicated models – from support vector machines (Shetty et al., 2015), to more recently adopting LTSM (X. Li et al., 2016), stacked RNN, and CNN (Malik et al., 2017), and multi-layered feed-forward artificial neural networks (Masood, Nayal, Jain, Doja & Ahmad, 2017).

Despite improvements in the accuracy of these projects, the advanced model structure has given rise to some concerns regarding performance. A CNN model will take hours to train without a graphics processing unit (GPU). Most deep learning neural network architectures are resource hungry, both regarding required memory size and CPU number-crunching requirements. For example, we attempted conducting the handwritten digit recognition sample project (MNIST on TensorFlow) on a CNN (728 input, three hidden layers) with 60,000 images of numerical “digits”. The process took roughly 1.5 hours to train, but with the help of a GPU, took only 15 minutes.

Despite the trend of using complicated neural network structures, Gonzalez (2013) suggests that the structure of the predictive model is not significant. A non-salient feature set will yield poor results, regardless of how good the model is (Gonzalez, 2013).

Gonzalez also points out that the dimension reduction approach (PCA) can be used in machine learning to reduce the full set of features, rather than choosing a smaller set and missing out on vital information. Gonzales chose three alternate feature sets instead of the MFCC, which were less computationally complex and performed better than the MFCC features.

Another paper (Nalini & Palanivel, 2016) considered the use of the residual phase (RP) among MFCC features, which is often omitted. The residual phase is defined as the cosine of the phase function of the signal, derived from the linear prediction (LP) residual. Specific information present in the residual phase is compared to the information present in the current MFCC, increasing the emotion recognition performance (Nalini & Palanivel, 2016). MFCC alone can achieve a good result, above 90%; the use of the residual phase method improved this performance to 96.0%, 99.0%, and 95.0%, using AANN, SVM, and RBFNN, respectively.

7 THE EXPERIMENT

7.1 Introduction

Traditionally, spectral and temporal features have been used to describe the acoustic properties of sound. These features are more accessible to calculate than MFCCs. Additionally, as noted in the previous chapters, we have listed the importance of pitch, tempo, and timbre to express emotion in music; we concluded that emotions are more related to the spectral and temporal properties.

We hypothesize that the complexity of the MFCC may produce some adverse effects in an MER system. In the five MFCC's processes, the logarithmic process might reduce the sensitivity of the music sound. The steps needed to compute Mel-filter banks had been motivated by the nature of human perception of voice signals. There may be no need for Mel-filter banks in music sound processing. The DCT is a complicated function which is sensitive to highly correlated inputs; we doubt it will cause some loss to the music timbres and pitch information. Moreover, the windowing implementation, such as the "Hamming Windowing" for the MFCC, is tailored explicitly for voice applications. Each of the processes adds some complexity to the operations and may cause a loss in the fidelity of the original sound data.

Because the spectral and temporal properties are less complicated in comparison with MFCC, so we think the data distortion is not as bad; therefore, they should perform better in machine learning (regarding predicting emotional value).

In this experiment, we will extract MFCC, spectral, and temporal features for each of the songs in the dataset. The individual features, or a combination of features, will be used to train a simple layer perceptron (a simple artificial neural network). The trained perceptron model will then be used to predict the result, i.e. the emotional value. Then we compare the prediction error of the three features, and the one with the lowest indicated prediction error is considered as the best feature.

There are more complicated neural network configurations suitable for the proposed experiment. However, our primary goal is the comparison of musical features, not the neural network; therefore, we employed the most straightforward neural network. The

simple neural network eliminates the concern of having numerous parameters, as is the case for complex networks.

7.2 Dataset and properties

A 1000-song dataset (Soleymani, Caro, Schmidt, Sha, & Yang, 2013) was used, containing 45-second music clips randomly extracted from complete songs. The 45 seconds of music clips were annotated using arousal and valence levels, on a nine-point scale. A fair share of the genre is selected to guarantee a more even distributed emotional value; because some specific genre is contributed to emotions.

The dataset is annotated continuously and overall for arousal and valence dimensions. We are just using the overall emotion value (Soleymani et al., 2013).

7.3 The annotation of emotion

An important technique used by the 1000 songs dataset creators (Soleymani et al., 2013) is a psychologically-inspired video interface for collecting emotional feeling in multi-dimensional value (valence-arousal) from listeners. The collected emotions were further corrected using a statistical method for any discrepancies. Other researchers, Panda, Malheiro, Rocha, Oliveira and Paiva (2013) created a new dataset by combining information from the content of songs (lyrics, title, comments) as a means for compiling the categorised emotions.

In supervised Machine Learning methodologies, this is called the labelled value. We used this annotated value and the predicted value to calculate the RMSE prediction error.

7.4 Introduction of the Artificial Neural Network (ANN)

Brain-Neurons are the basic information-processing units of the brain, as seen as Fig.8, is essentially consist of the dendrites, axon and terminals.

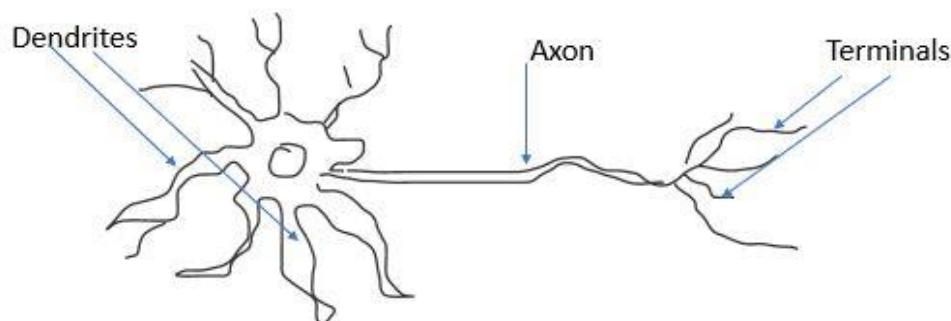


Figure 8 - Simplified Brain Neuron

Like the Brain-Neuron, a **Computational Neuron** is a fundamental unit of the artificial neural network; it has multiple inputs (dendrites), the node (axon) and output (terminals) as shown in following.

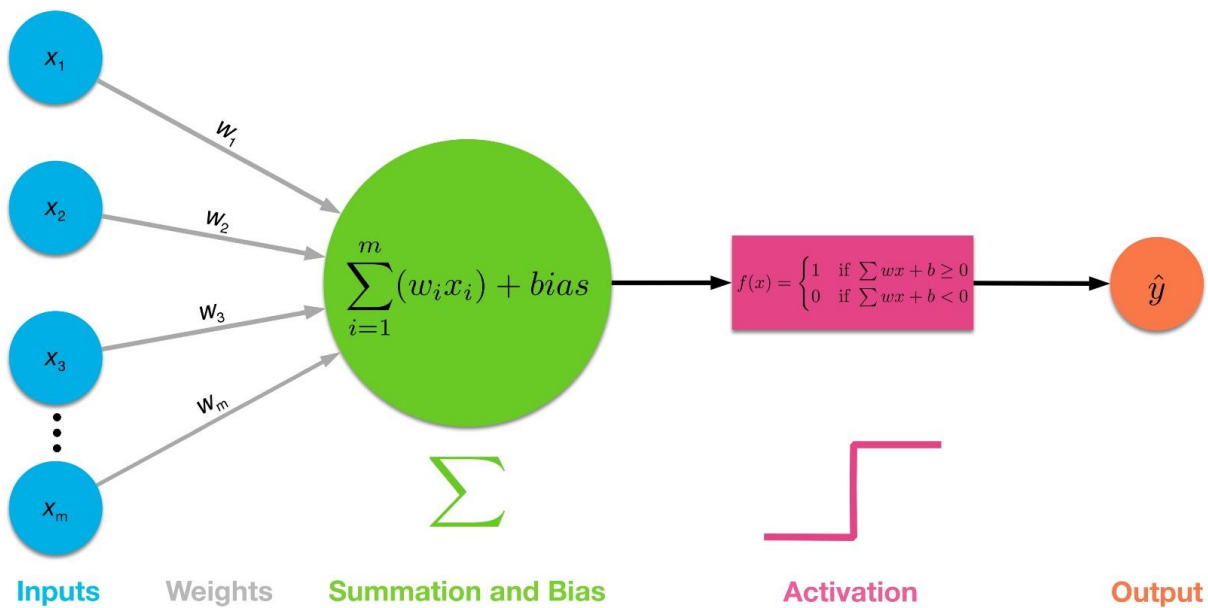


Figure 9 - A computational neuron

(Figure from <https://galaxydatatech.com/2018/06/25/multi-layer-perceptron-model/>)

Artificial Neural Network is a computing concept based on the human brain's neuron structure. A set of neurons are connected in a network fashion, allowing them to send information from one node to another. Neural Network is formed by connecting the output to the input of the next neuron as illustrated in Fig 9 (MLP network).

Neuron/ Perceptron

The concept of **Perceptron** was conceived in the 1950s and 1960s by the scientist Frank Rosenblatt, inspired by earlier work by Warren McCulloch and Walter Pitts on the Neurons. A neuron is the perceptron except the inputs are not weighted.

Each **perceptron** is typically taken several weighted inputs, (x_1, x_2, \dots), and produces a single output. All inputs are fed to an activation function to generate the output.

Generally, its output $O(X)$ is formulated as

$$O(X) = \text{Act} \left(\sum_{i=1}^n (W_i * X_i) + \text{Bias} \right) \quad \text{Eq.9}$$

where Act is the Activation Function of the node.

Activation function controls the output of that node given an input or set of inputs.

The two historically common activation functions are both **Sigmoid**, and **Tanh** is described by

Some common activation functions are:

The sigmoid function is logistic function has been widely used in machine learning basic configuration, especially for the logistic regression and some basic neural network implementations.

$$\text{Sigmoid}(x) = \frac{1}{1+e^{-x}} \quad \text{Eq.10}$$

Tanh function

$$f(x) = \tanh(x) \quad \text{Eq. 11}$$

tanh is also like logistic sigmoid but better. The range of the tanh function is from (-1 to 1). tanh is also sigmoidal (s-shaped). The advantage is that the negative inputs will be mapped strongly negative, and the zero inputs will be mapped near zero in the tanh graph.

Rectified Linear Units (ReLU) function

In most DNNs, ReLUs is used in the hidden layers. A rectified linear unit output 0 if the input is less than 0, and if the input is greater than 0, the output is equal to the input. ReLUs' machinery is more like a real neuron in your body.

$$f(x) = \max(x, 0) \quad \text{Eq.12}$$

ReLU activations are the simplest non-linear activation function to use. When you get the input is positive, the derivative is just 1, so there isn't the squeezing effect you meet on backpropagated errors from the sigmoid function. ReLUs result in much faster training for large networks.

Multi-layer perceptron (MLP)

In Neural networks, nodes are typically arranged in multi-layers, known as Multi-layer perceptron (MLP). It is a type of feedforward artificial neural network which consists of

at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function.

Layers are composed of some interconnected 'nodes' which is an 'activation function'. Input data are presented to the network via the 'input layer', which interface to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output as follows.

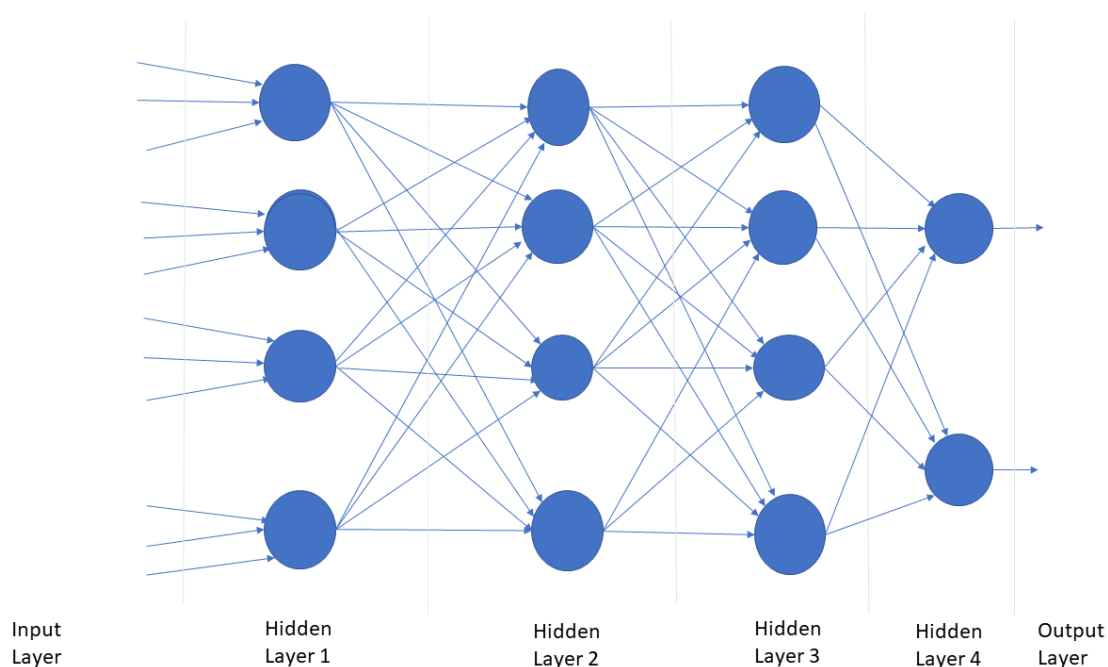


Figure 10 - A Multi-Layer Training process

The typical use of a neural network has the **training** and **prediction** processes.

The neural network, when asked to resolve a problem, is also employing an iteration of try and error fashion to approximate the result, each time the weight is adjusted to produce an output. When the output is wrong, the weight value is revalued, and this is repeated until the output is correct or closer to the value.

And this repetition of weight adjustment is repeated for each set of input until the output value matches the target result. Then, we referred to this process as training. We can use the trained network (the weight and the base are re-value) for prediction with the testing data.

In the training process, there are many initial values of the weight and the based value possible. Like the regression of a mathematical model, the coefficient of the variable is achieved for many iterations of the input data and the output data.

After the network is trained, the node weighted values are used for new input data. This process is called the prediction; this is because the old training data characterize the resultant output of the new data.

Purpose of the Backward Propagation

To solve for the weights and the biases of the neural network become increasingly infeasible as the layers, and the complexity of the network grows. Backward Propagation (BP) is involved in the training phase for a neural network. This BP is used to improve the accuracy of the predicted result by re-adjusting the weights and biases in a backward manner.

In training the neural network, there are the feed-forward phase and backpropagation phase.

In the feed-forward phase we obtained the output of the network. Then, in the backward phase; the error with the predicted output and the expected output is calculated. These discrepancies also called the loss derived with a Cost function. Then we go back and adjust the weights and biases so that we can reduce the error in the next try. As the error reduces the prediction accuracy improves. The error is reduced through what is called the gradient descent process in the **working principle of BP**. In training our network, the goal is to get the value of this cost function as low as possible.

The working principle of BP.

- a. In the feed-forward phase, initial the network with random weight first.
- b. Then we calculate the cost function for the output.
- c. Propagate backward to obtain the **Gradient Descent**
- d. Update the new weight value with a lesser error.
- e. The new output is generated; the cost function is recalculated.

These steps are repeated until the cost function is minimal.

The weight is thus fixed. This is the trained network.

Gradient descent is an update rule for adjusting the weights of the neural network to get us closer to the minimum cost function value we want. The objective of gradient descent is to move the error to the zero levels, we don't want the error to be too positive or too negative.

The new weight value is the old weight subtracting the Gradient value.

$$w_n = w - \alpha \frac{\partial J(W)}{\partial w} \quad \text{Eq. 13}$$

where

the derivative of the Cost Function $\frac{\partial J(W)}{\partial w}$ is gradient,

α is the learning rate; w_n is a new weight.

In typical gradient descent algorithm, the Stochastic Gradient Descent can be used; which prevent the high value of the gradient or diminishing of the gradient as well.

Cost function indicates ‘how good’ the model is in predictions for the output value “a” for a given value of the given set of value “b”.

$$\text{Cost} = \sum_{i=1}^N (y' - y)^2 \quad \text{Eq. 14}$$

Cost denoted as $J(W)$, we need to adjust the weights to achieve a minimum cost function value.

There are many techniques for calculating the cost commonly referred to as the loss function, and they include root mean squared error and cross-entropy among others. The Gradient Descent is the standard optimization algorithms, iteratively work towards their optimal weight value.

There are many algorithms use for the optimising Gradient Descent, for example, ADAM. Adam stands for **Adaptive Moment Estimation**. Adaptive Moment Estimation (Adam) is a method that computes adaptive learning rates for each parameter.

How do Neural Networks Differ from Conventional Computing?

The computational design can be quite different from the conventional sequential execution model. A sequential program can address an array of memory locations where

data and instructions are stored. In a sequential system, the computational steps are deterministic, sequential and logical, and the state of a given variable can be tracked from one operation to another.

ANNs differ from a sequential system that it is non-deterministic. Instead of a complex central process (main program), many simple ones are used, to sum up the weighted inputs from other nodes.

The ANNs network executes an operation for nodes in the network in a parallel manner; when the data present to the inputs, the resultant output is the immediate result of the network. For example, if “101010” present at the input, the result “1” is generated at the output.

What Applications should Neural Networks Be Used For? Why is it better?

Neural networks are universal approximators, and they work best if the system you are using them to model has a high tolerance to error.

Although ANNs have success in the many fields of in the Machine Learning area, namely, pattern recognition. They work very well for non-linear problems, such as

- capturing associations or discovering regularities within a set of patterns;
- where the volume, number of variables or diversity of the data is very significant;
- the relationships between variables are vaguely understood; or,
- the relationships are difficult to describe adequately with conventional approaches.

What are the disadvantages and disadvantages?

Though ANNs have promising result in many severe problems. It has

Disadvantages:

1. Due to the complexity and the non-deterministic nature of the ANNs, the resultant model could sometimes hard to verify. Apart from defining the general architecture of a network and perhaps initially seeding it with a random number, the user has no other role than to feed it input and watch it train and await the output. In fact, it has been said that with backpropagation, "you almost don't know what you're doing".
2. There are many software packages/libraries for users. Some freely available software packages (Tensorflow, Pytorch) do allow the user to examine the

progress of the training at regular time intervals, but the learning itself progresses on its own.

3. Training of the ANNs is a very time-consuming process; the Back Propagation consists of thousands of epochs.

Advantages:

1. ANNs are used in the 'black boxes' approach; users do not need to know the actual structure of the network. This presents the ease of use for users, but it also makes tracing of the training model difficult, and the prediction of the result can be unexpected and unexplainable with the traditional concept.
2. Due to the repetitive pattern of the neural network (each node structure is the same); it is favourable to run ANNs with parallel computing architecture. With each node implemented by a simple computation unit (namely, a summer); recently, the Graphical Processing Unit (GPU) architecture is quite suitable and getting popular.

7.5 Approaches/methods

Features Extraction

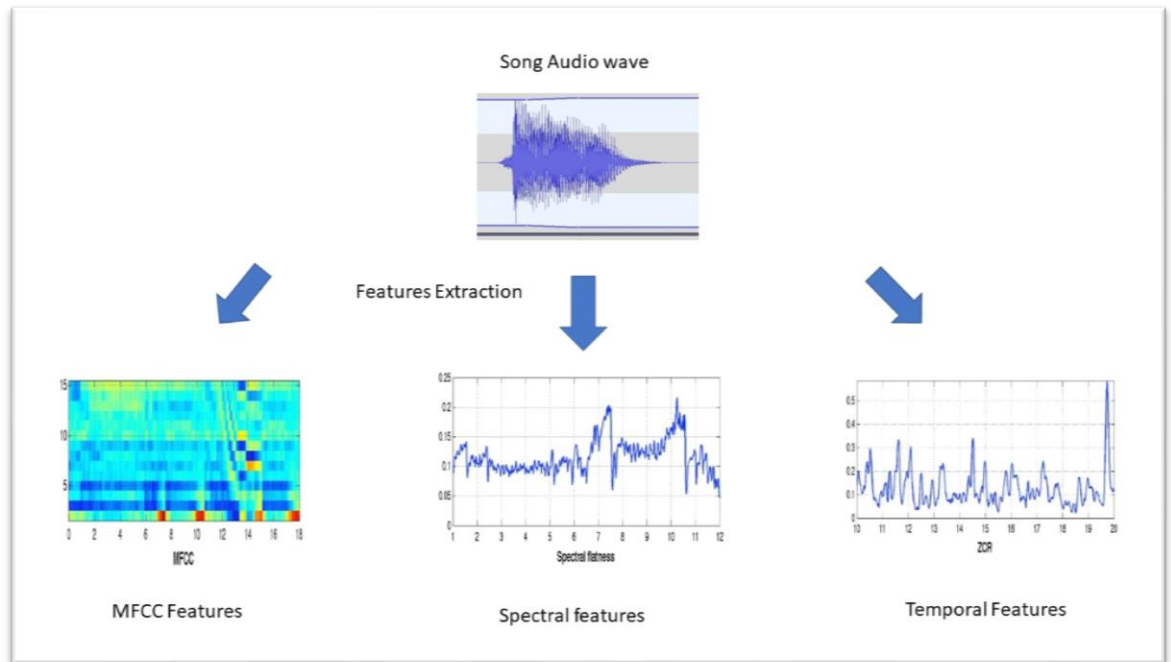


Figure 11 - Extraction different features from a Song

We use the python utility to read the wave file of each song into a vector.

The librosa library provides all the extraction utilities for MFCC features, the spectral features and the temporal features.

Each sound wave signal was divided into roughly 3864 samples; each of a 10 milli-samples window, features of these windows are extracted.

For the centroid features, we used the middle section of the centroid frequency for each of the sound samples: `centroid [0] [300:3500:2]`, which yielded 1600 centroid features.

Similarly, we took the middle section of the zero-crossing rate for each of the sound samples: `ZCR [0] [300:3500:2]` which yield 1600 ZCR.

For the MFCC features, we concatenated the MFCC1 and MFCC2 of the middle section of the audio with 800 each to form the 1600 features.

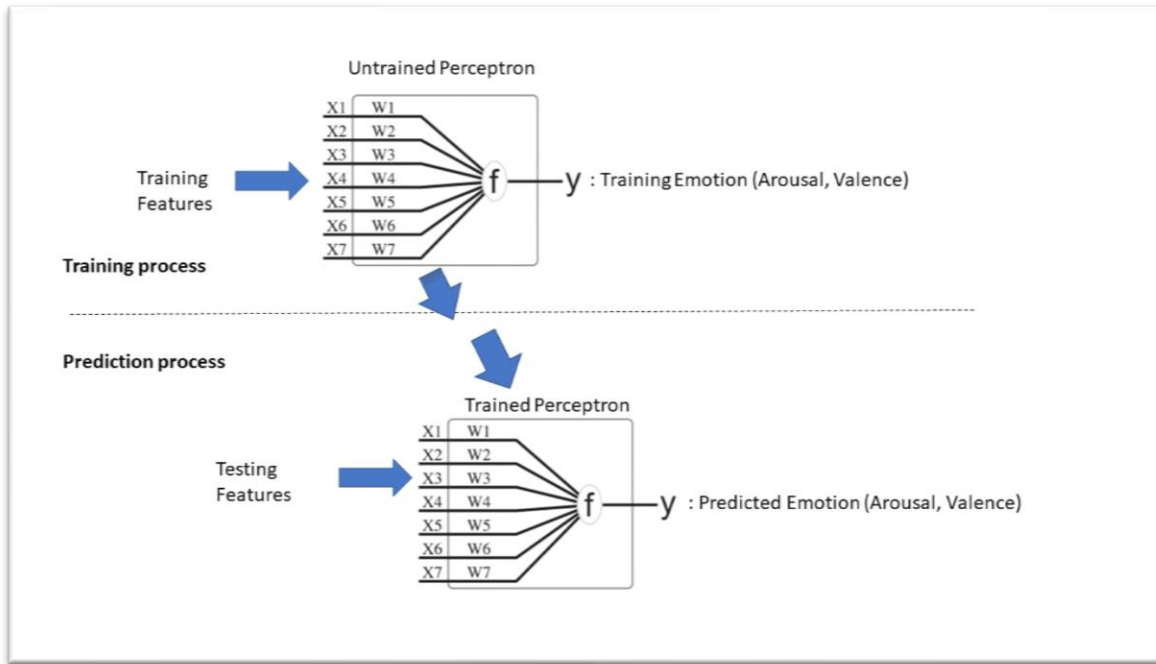


Figure 12 – Machine Learning Models for Training / Prediction of emotion

1. Using Neupy library to design a single layer perceptron (SLP) (Untrained Perceptron in the diagram).
2. Separate the dataset songs into half for training and a half to testing.
3. Extract a feature set from training songs to train the Perceptron.
4. The trained Perceptron is used to predict the emotion value from the testing songs' features set.

The prediction error (PE) is obtained from the labelled emotion value and the predicted emotion value from step 4 above.

The input for the SLP is with 1600 feature values, (the diagram is shown as X1-X7)

We repeated the four steps with each set of features, such as centroid, ZCR and MFCC into the SLP. Then all the PEs are recorded in the result table as follows.

Program and library setup

We used the 'librosa' Python library to extract the features from the audio file. We were able to extract all MFCC, spectral features, and temporal features for our project.

Scikit-learn is a machine learning package in Python, which we used for dimension reduction, e.g. principal component analysis. We used Neupy for the single-layer perceptron (artificial neural network).

The hyper-parameter of the perceptron network

```
net = algorithms.Adam( [
    layers.Input(1600),
    layers.Linear(1),
],
step=0.1,
verbose=True,
show_epoch='4 times',
error='rmse',
shuffle_data=False,
decay_rate=0.01,
addons=[algorithms.WeightDecay]
)
```

Adam is an optimization algorithm is chosen because it provides a clean result in comparison with the traditional gradient descent procedure to update network weights iterative for the training data.

Adam on optimisation has the following benefits for consideration:

- Computationally efficient.
- Little memory requirements.
- Appropriate for non-stationary objectives.
- Appropriate for problems with very noisy/or sparse gradients.
- Require little tuning.

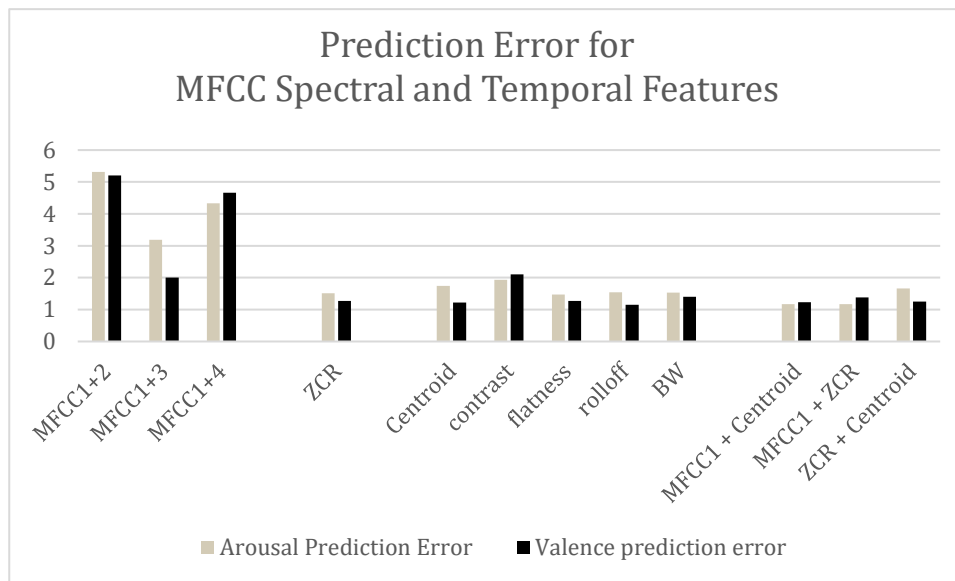
8 RESULTS AND DISCUSSION

8.1 Results

Table 2 - RMSE value of the Prediction Error

Features used	Arousal Prediction Error	Valence prediction error
MFCC1+2	5.3146	5.2021
MFCC1+3	3.1849	2.0
MFCC1+4	4.3308	4.6622
Zero-Crossing Rate (ZCR)	1.5101	1.271
Centroid	1.7425	1.2165
Contrast	1.929	2.1043
Flatness	1.4704	1.2664
Rolloff	1.5448	1.1488
Bandwidth (BW)	1.5316	1.3993
MFCC1 + Centroid	1.1735	1.2309
MFCC1 + ZCR	1.1749	1.3838
ZCR + Centroid	1.6661	1.2531

Table 3 - Prediction Error Comparison



The MFCC features have the highest prediction Error in compare with the other features in both the Arousal and the Valence prediction.

The Temporal features (ZCR) has a similar prediction error as with the spectral features (Centroid, contrast, flatness, Bandwidth).

The MFCC combine with the spectral and the temporal features yield a similar result as the spectral and temporal on its own. Because of the high prediction error of the MFCC, so it has little contribution to the overall effect. Therefore the combination prediction error has a similar value without the MFCC.

Individually, MFCC is not a desirable feature to use. We just used the first four components only, for the sake of direct comparison of features. However, there can be 20 MFCC components; we can use PCA to reduce to a few components; then it may be more useful. Similarly, we can use PCA on many different spectral features.

The first few components (the first four) in MFCC usually held the more important information of the signal this is due to the DCT transformation is more prominent for the principal signal composite. The accuracy or discrepancy in the result is shown that the MFCC1 and is combined with the other MFCCs such MFCC1 with MFCC2 or MFCC4 etc. It cannot explain why MFCC1 + 3 seem to have lower error than the neighbouring two MFCC combination.

8.2 Discussion of the study

The logarithm suppressed spectral information

The logarithm function is introduced to reduce noise, but this will affect the timbre quality of music. One approach to avoid this is the use of temporal or spectral features directly, rather than the spectrum vector being logarithmic-ed, as in the Cepstrum. The most recent trend is to use the time-domain feature directly, without involving complicated processing, and not using the Fourier Transformation at all.

Mel scale may not be suitable for music

The Mel scale may not be the correct auditory transformation to use, as we know that high pitches are generally the signature of an agitated emotion. The aim of music features extraction is not to filter out the more delicate details in sound or distort distinct features.

In Western music, the equal-tempered scale (established by Bach) is generally used, which divides an octave into 12 equally spaced semi-tones. The octave interval corresponds to frequency doubling, and semi-tones that are equally spaced; thus, ascending one semi-tone multiplies the frequency by the twelfth root of two, or, approximately 1.059. The use of Mel scaling will distort the sense of even spacing in the equal-tempered scale.

The dimensional reduction causes some loss

The use of DCT to compress the information by keeping the crucial part of the information; it presents some losses to the details of timbre. Although the data dimension is substantially reduced, the twist and turn of rhythmic and tonal changes are lost. The composer often used these tonal changes to represent the delicate emotions.

Redundant with spectral and temporal

MFCC has been shown to correlate with spectral and temporal features (Logan, 2000). Therefore, it may be redundant to use MFCC with either spectral or temporal features.

The missing part of MFCC

MFCC is only the real number part of Cepstrum; however, there is also a phase part of Cepstrum, which can be useful for music analysis (Nalini & Palanivel, 2016). There are not many projects looking to use phrasal part of MFCC. The audio properties can spread throughout the MFCC components, which can be of 20 components; the first few components (the first four) usually held the more important information of the signal this is due to the DCT transformation is more prominent for the principal signal

composite. The accuracy or discrepancy in the result is shown that the MFCC1 and is combined with the other MFCCs such MFCC1 with MFCC2 or MFCC4 etc.

Human factors

The MFCC may be suitable for machine learning projects that do not need to interpret the meaning of the input features. In MER, it makes more sense to use spectral and temporal features, which are related to musical features. People relate to real musical features (as shown in Table 1), while MFCC has no musical meaning. It will always cast doubt on its validity in the context of music analysis.

Problem with dimension reduction

Many researchers do not believe that MFCC is a Principal Components version of all the critical features; therefore, also, extra features are combined with MFCC in the features engineering process. Researchers have, in the past, using the full set of musical features (OpenSMILE, 260 features), which have overlapping characteristics. They do not contribute to the machine learning's predictive model, but rather, create extra overhead in the machine learning process. The use of principal component analysis to reduce the size of spectral and temporal features can achieve the same effect. In this experiment, we used the PCA to reduce the size of the features set and Python's slicing to reduce the size of the feature. With reduced features, there is no need for a complicated architecture, such as a convolutional neural network, and a simpler prediction model can be used. MFCC can have up to 20 components, and it can be challenging to decide how many to use. For speech, the first few components are generally applied; for music, however, this number is more significant.

The relationship between human voice and music

According to the super-expressive voice theory (Juslin & Sloboda, 2010), what makes music expressive, for example, the sound of a violin, is the fact that it sounds very similar to the human voice; at the same time, it can move far beyond what the human voice can achieve in terms of speed, intensity, and timbre. For example, if human speech is perceived as 'angry' when it has a rapid rate, deep intensity, and a harsh timbre, a musical instrument can sound extremely 'angry' at an even higher speed, louder intensity, and a harsher timbre. The differences in speech and music require different treatments of sound signals. We cannot merely blindly use MFCC in music analysis, as in speech analysis.

Machine learning aspect

Due to the limitation in machine learning, to reduce the numbers of the input features is preferable. The Discrete Cosine Transform (DCT) was used to compress the features data size. Given deep neural networks can accept a more significant number of features, (DCT) is no longer a required step in compression of the data.

Other researchers seemed have better result with MFCC

The argument here is that the MFCC have 5 processes involved; each of these processes would have reduced the fidelity/quality of the salient properties for the later analysis.

Although the MFCC has lost the fidelity of the original properties in the music, it still kept the essential part of the music information for analysis.

This may be the contribution of the Non-linearity of the more complicated Neural Network; whereas the perceptron in this experiment is linear.

In further work, we suggest repeating this problem with a more complicated neural network to pick up the self-learning capabilities of the network.

8.3 Future Works

1. We can further amplify the result obtained by using a complicated neural network. So, we will propose a new set of experiments in which the perceptrons are replaced by a 2-layer neural network with, perhaps, 20 or 50 hidden nodes. This will inspect if MFCC flavour the non-linearity of a more complicated network.
2. To confirm that MFCC is more flavour in speech, we will repeat the same experiment but with speech sound files.
3. Next we will explore the use of the double-stacked CRNNs, (Malik et al., 2017), to predict Valence and Arousal concurrently; this will put the correlation of the two in scope.
4. Further, this approach can be extended into multiple stacked CNNs to include other components of emotion, namely, “motivation” and “jealousy”. Since many have argued that emotion is not merely just Valence and Arousal.
5. Using speech to identify emotions. All the experiment setup here is also for ready investigating emotion in speech as well. As mentioned, MFCC was initially been designed for speech. So, instead of feeding the audio musical wave file as input; the audio speech file can be applied directly.
6. Using facial imaging to identify emotions. The experiment here also is repeated with images files for a different set of features extraction.

9 REFERENCES

- B.P. Bogert, J. R. H. and J. W. T. (1963). The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking. In *Proceedings of the Symposium on Time Series Analysis* (pp. 209–243).
- Bello, J. P. (2013). Low-level features and timbre. *Lecture Notes*, 1–31. Retrieved from http://www.nyu.edu/classes/bello/MIR_files/timbre.pdf
- Dave, N. (2013). Feature Extraction Methods LPC , PLP and MFCC In Speech Recognition. *International Journal for Advance Research in Engineering and Technology*, 1(Vi), 1–5.
- Gonzalez, R. (2013). Better than MFCC audio classification features. *The Era of Interactive Media*, 9781461435, 291–301. https://doi.org/10.1007/978-1-4614-3501-3_24
- Hevner, K. (1935). The Affective Character of the Major and Minor Modes in Music Author (s): Kate Hevner Source : The American Journal of Psychology , Vol . 47 , No . 1 (Jan . , 1935) , pp . 103-118 Published by : University of Illinois Press Stable URL : <http://www.jstor.org/stable/14>. *The American Journal of Psychology*, 47(1), 103–118.
- Hevner, K. (1937). The Affective Value of Pitch and Tempo in Music Author (s): Kate Hevner Source : The American Journal of Psychology , Vol . 49 , No . 4 (Oct . , 1937) , pp . 621-630 Published by : University of Illinois Press Stable URL : <http://www.jstor.org/stable/14>. *The American Journal of Psychology*, 49(4), 621–630.
- Juslin, P. N. (2018). Emotional responses to music : The need to consider underlying mechanisms, (2008), 559–621.
- Juslin, P. N., & Sloboda, J. A. (2010). *Handbook of Music and Emotion: Theory, Research, Applications. Handbook of music and emotion Theory research applications*. <https://doi.org/10.1093/acprof>
- Li, T. L. H., & Chan, A. B. (2011). Genre classification and the invariance of MFCC features to key and tempo. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6523 LNCS, pp. 317–327). https://doi.org/10.1007/978-3-642-17832-0_30
- Li, X., Tian, J., Xu, M., Ning, Y., & Cai, L. (2016). DBLSTM-based multi-scale fusion for dynamic emotion prediction in music. *Proceedings - IEEE International Conference on Multimedia and Expo, 2016-Augus*, It;xocs:firstpage

xmlns:xocs=""/> <https://doi.org/10.1109/ICME.2016.7552956>

- Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. *International Symposium on Music Information Retrieval*, 28(November 2000), 11p. <https://doi.org/10.1.1.11.9216>
- Malik, M., Adavanne, S., Drossos, K., Virtanen, T., Ticha, D., & Jarina, R. (2017). Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition. Retrieved from <http://arxiv.org/abs/1706.02292>
- Masood, S., Nayal, J. S., Jain, R. K., Doja, M. N., & Ahmad, M. (2017). MFCC, Spectral and Temporal Feature based Emotion Identification in Songs. *International Journal of Hybrid Information Technology*, 10(5), 29–40. <https://doi.org/10.14257/ijhit.2017.10.5.03>
- Meyer, L. B. . (2017). Review Reviewed Work (s): Emotion and Meaning in Music by Leonard B . Meyer Review by : David Kraehenbuehl Source : Journal of Music Theory , Vol . 1 , No . 1 (Mar . , 1957), pp . 110-112 Published by : Duke University Press on behalf of the Yale Unive, 1(1), 110–112.
- Meyer, L. B., Kraehenbuehl, D., & Meyer, L. B. (1961). *Emotion and Meaning in Music. Journal of Music Theory*. Chicago, Ill. and London: The univesity of Chicago Press. <https://doi.org/10.2307/843099>
- Nalini, N. J., & Palanivel, S. (2016). Music emotion recognition: The combined evidence of MFCC and residual phase. *Egyptian Informatics Journal*, 17(1), 1–10. <https://doi.org/10.1016/j.eij.2015.05.004>
- Panda, R., Malheiro, R., Rocha, B., Oliveira, A., & Paiva, R. P. (2013). Multi-Modal Music Emotion Recognition : A New Dataset , Methodology and Comparative Analysis. *10'th International Symposium on Computer Music Multidisciplinary Research*, 1–13.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. <https://doi.org/10.1037/h0077714>
- Sabini, J., & Silver, M. (2005). Ekman's basic emotions: Why not love and jealousy? *Cognition and Emotion*, 19(5), 693–712. <https://doi.org/10.1080/02699930441000481>
- Shetty, R., Kasbe, S., Jorwekar, K., Kamble, D., & Velankar, M. (2015). Study of Emotion Detection in Tunes Using Machine Learning. *International Journal of Scientific and Research Publications*, 5(11). Retrieved from www.ijsrp.org
- Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C.-Y., & Yang, Y.-H. (2013). 1000 Songs for Emotional Analysis of Music. *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia - CrowdMM '13*, 1–6.

<https://doi.org/10.1145/2506364.2506365>

Tomkins, S. S. (1962). *Affect, imagery, consciousness, Vol. 1: The positive affects*. Springer Publishing Co. <https://doi.org/10.1037/14351-000>

Urbano, J., Bogdanov, D., & Herrera, P. (2014). What is the effect of audio quality on the robustness of MFCCs and chroma features. *International Society for ...*, (Ismir), 573–578. Retrieved from http://mtg.upf.edu/system/files/publications/025-what-effect-audio-quality-robustness-mfcc-chroma-features_0.pdf