Abstract

The overwhelming amount of music data being recorded on a daily basis and most of the time uploaded to the web, begs the need for finding the most appropriate representation of those signals in a space where a lot of tasks can be achieved with human-like functionality.

One example of those tasks that this paper will be focused on, is the recognition of similarity between sounds coming from different instruments with different playing technique. To achieve this, we will consider two techniques that allow the transformation from temporal representation into a space of features. The first one being the MFCC that is based on an uncorrelated Mel frequency scale [1], and the latter is the scattering representation based on a wavelet transform ([2] and [3]).

this report will present the effort done to get the best out of those two transformations using techniques such as metric learning and study of feature variability between different observations.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motviation

The purpose of this internship is to study the sounds played by different instruments(different playing technique) and present the mathematical factors that helps discriminate between two instruments(two playing technique). This discrimination is done by finding a space where an euclidean distance is maximum for inter-class observations and minimum for intra-class observations. (à revoir cette phrase)The research done in this field does not focus on studying the features in the representing space in a way to increase metrics related to similarity study.Before we start presenting the work done in this internship it will be interesting to look on how this discrimination is naturally done by human.

Our brain processes sounds by analyzing both the frequency components of those signals and the temporal change of amplitude. those components will, based on different factors, help us recognize the loudness, pitch and timbre of the sounds we are hearing. While other factors, that will not be interesting for our study, can help in recognizing the spacial coordinates and other aspect of that sound.

I will be presenting in this report the study of similarity between different audio recordings coming from same and different instruments with different playing technique, and since timber by its most general definition is what allows us to experience, the same note played with the same loudness, in a different way for each instrument.
I will first start by "trying" to define the notion of timbre and then present the problem and its proposed solution.

1.2 What do we know about timbre ?

Since the first explanation of sounds put by Helmholtz in 1863 [4], the definition of timbre has changed, especially with the inability of modern day synthesizers to reproduce the same auditory experience of an instrument based on a primal definition of timbre. After a lot of studies done on the subject by physicists, biologists and musicians we have a clear definition of what timbre is not rather than what it really is. For instance, the American National Standard Institute ANSI in the American National Standard on Acoustical Terminology (*ANSI S1.1-1994*) defines timbre as follows :

> *"Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar." ANSI 1994*

This definition clearly states that timbre is not related to the loudness, pitch or the duration of the tune being played (or spoken). This being said, There are still many factors that help us distinguish between two notes played with the same loudness, pitch and duration, those factors are considered as part of the structure of timbre. Some example of those factors:

- The difference in amplitude of the harmonics.
- Change in harmonics during the attack.
- The deviation of the harmonics from the perfect $n * f$ location where f is the fundamental note and n is an integer.
- The vibrato of the note.
- The sustain and the decay of the note.

The notion of timbre is still an abstract one, thus it will be an obvious step to try to find another solution to solve the problem.

1.3 How to solve the problem ?

In our similarity study, we have to identify the space on which we are trying to classify our data and proceed to proper feature extraction. We have two ways of looking at the feature extraction problem : perceptual or taxonomic [5]. The perceptual approach is

motivated by the biological aspect of the problem, and it aims to explain how we hear the sound by finding a space where the descriptors axis explains the notion of timbre, we will not enter into details in this aspect rather we will consider the taxonomic approach, in which the best feature space is the one that helps in getting the most discrimination between classes.

Based on a state of the art study done on the subject of features extraction for classification purposes, we will be considering two space representation; the first one is the MFCC and the other is the scattering [2], and in both cases we will try to get the highest precision based on the Mean-Average-Precision and Precision-At-5 using techniques of data treatment and metric learning.

Chapter 2

Database and experimental procedure

This section is divided in two parts the first one dedicated to present a structural and musical study of the data present in the SOL(Studio On-Line) database and the other aims to present the experimental procedure and the different steps that forms the project.

2.1 Study of the SOL database

The SOL database is provided as part of the Orchids orchestration system by the team IRCAM. It contains 25119 audio files stored in the 24 bits / 44.1 kHz format played by 16 different instruments which can be subdivided into 32 different classes. There are 498 different combination of instrument plus playing techniques. The type instruments vary from strings (violin, Guitar...), woodwind (Flute, Oboe...) and Brass (Trumpet, Trombone,...) and the playing technique vary from instrument to another and include most of the important techniques for that instrument (Aeolian, crescendo, flatterzunge...) A full list of instruments and playing techniques is provided in [6]. Figure 2.1 shows the distribution of the SOL database along 6 octaves and one non acoustic class, The biggest part is octave 4 where all the instruments are represented in it (with different proportions).

2.2 Experimental procedure

The experimental procedure is divided into two parts, the first one is calculating the features based on one of the two representations: MFCC and scattering(with different

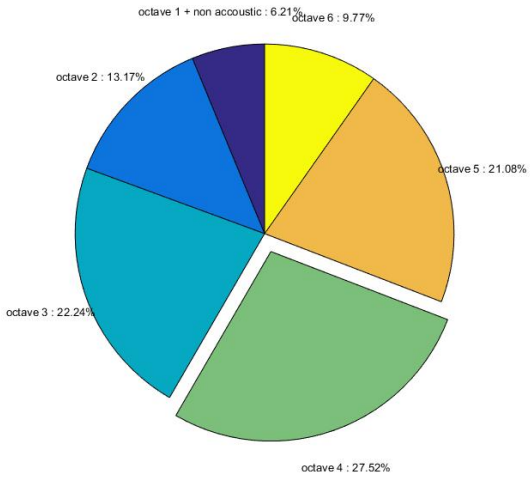

Figure 2.1: Musical octave study of sol database

parameters), and in the second part we calculate the ranking metrics before and after processing the data in the feature space.

For the first part, the use of the two representations is done as follows : using a fixed length window of calculation we obtain a certain number of temporal features based on the length of the signal, then by averaging along the time axis, we obtain one observation by audio file, this will help us stay in a space where the Euclidean distance is enough to calculate the distance between observations.
For the second step, the distance calculation is done using pairwise Euclidean distance measure. Many form of processing in the feature space have been tested and only the one that proved efficient will be shown here and presented in further details in later chapters :

- Standardization.
- High or low variance filtering.
- Projecting into symmetrical data
- Large Margin Nearest Neighboor.[7]

After processing the data, we proceed to calculate the pairwise Euclidean distance, which will be used to calculate the two ranking metrics that will be the basis of our analysis :

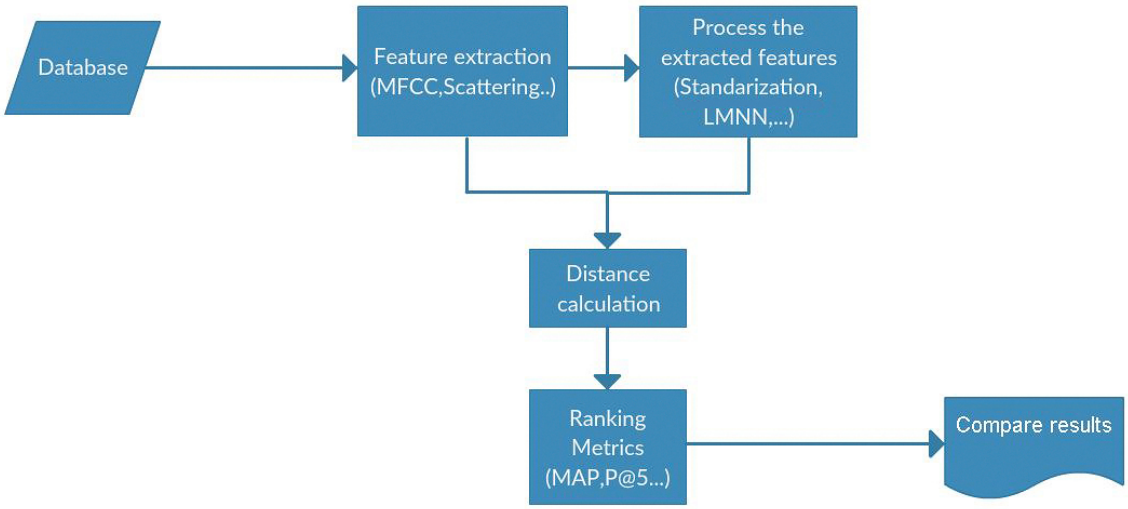

Figure 2.2: Experimental procedure

- Mean Average Precision or MAP.
- Precision At k or P@k

The following chapters we will follow the same scheme in figure 2.2 starting by the state of the art and the feature extraction procedure, following will be the methods of processing in the feature space and at the end using the distance calculation and ranking metrics I will present the results along with an analysis.

Chapter 3

Representations

In this chapter I will be showing for the two space representations used in this paper, the motivation and algorithm of calculation. We will start first by the Mel Frequency Cepstral Coefficients.

3.1 Mel Frequency Cepstral Coefficients MFCC

The Mel Frequency Cepstral Coefficients became famous with the rise of speech recognition systems.Given their ability to module a spoken stream in a compact space while containing most of the information needed for processing spoken audio streams they are still found in most of the speech processing applications. In the musical domain MFCC has been widely used for music genre classification and it is believed to extract a lot of the feature present in the timbre.
The biggest limitation of the MFCC as we will see in details is its inability to module large time scale variation.

3.1.1 Algorithm and motivation

The extraction of the MFCC features is divided into 5 steps as presented in figure 3.1, we will start by explaining each step along with the motivation and the limitation it provides.

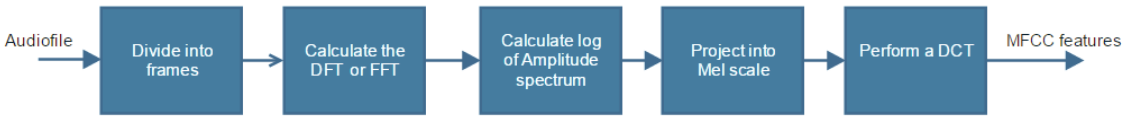

Figure 3.1: Procedure of MFCC feature extraction

1. **Frame division**
   The first step is to divide the waveform into equal time spaced frames, it is obtained by applying a windowing function at fixed time intervals. This step is motivated by the fact that audio signals (music and speech) are non-stationary. The limitation of this step is related to the length of the windowing function. It is proven that the best results are obtained for a window function of 25ms length, this means that the MFCC can not module structures of variation bigger than 25ms.

2. **DFT or FFT calculation**
   The discrete Fourier transform (or fast Fourier transform) is used to project from the time space to the spectrum space. By using the Fourier transform we obtain the amplitude of the spectrum and thus we lose knowledge of the phase. This step is motivated by the fact that for mono recording audio waves are phase independent and thus there is no loss of information.

3. **LOG of amplitude spectrum calculation**
   Given that the perception of sound is logarithmic by nature the third step is to take the log of the amplitude spectrum already obtained. This will leave us with a high dimension space.

4. **Mel scale projection** In this step we aim to reduce the space and smooth the spectrum by applying a mapping from the log amplitude spectrum to the Mel frequency scale. This mapping is algorthmique for high frequencies(higher than 1KHZ) and linear for low frequencies(below 1KHZ). The scale is motivated by the fact that we do not perceive pitch in a linear way. The mapping allow us to reduce the space to n features space, where n can vary up to 40.

5. **DCT calculation** We are now left with 40 highly correlated features and most of them containing unnecessary information. To achieve the decorelation and extract the interesting features a Discrete Cosine Transform is applied and 13 features are extracted.

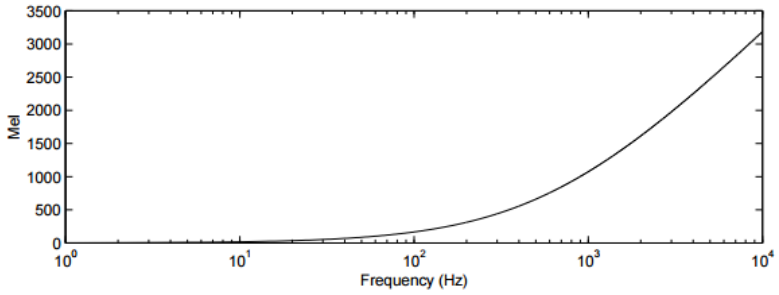

Figure 3.2: Mel scale

3.1.2 MFCC parameter choice

Before we started our study on the MFCC we had to tune the values of different parameters, and since we are dealing with a huge database we extracted a test sample that best represents the data. We proceeded the tuning on the octave 4 (see chapter 2.1) and the default factors for the MFCC were proven to be the best:

- Window length : 25ms
- Mel feature space dimension : 27
- deccorelated feature extraction : 13/27

	MFCC 13/27	MFCC 13/40	MFCC 40/40	MEL 40
Instrument MAP	27.68	25.2	20.24	18.22
Type MAP	14.84	14.63	10.12	7.70

Table 1 : Study on different MFCC tuning for a fixed window of 25ms.

3.2 Scattering Representation

Bibliography

[1] Beth Logan *Mel Frequency Cepstral Coefficients for Music Modeling.* 2000

[2] J. Andén and S. Mallat. *Multiscale scattering for audio classification.*. ISMIR 2011

[3] J. Andén *Time and frequency scattering for audio classification.* January 7, 2014

[4] Hermann Ludwig Ferdinand von Helmholtz *On the sensations of tone as a physiological basis.* 1895

[5] Perfecto Herrera-Boyer and al. *Automatic Classification of Musical Instrument Sounds.* Journal of New Music Research 2003

[6] Yan Maresz and al. *Ircam solo instruments UltimateSoundBank reference guide*

[7] K. Q. Weinberger, L. K. Saul. *Distance Metric Learning for Large Margin Nearest Neighbor Classification.* Journal of Machine Learning Research (JMLR) 2009