Chapter 1

True labeling ranking problem.

1.1 clustering the samples.

In the first part of the study, the following ranking problem is studied. 25119 musical extraction are provided from different solo instruments playing different techniques of play. Those samples will be clustered in three different ways:

1. **16 Labels instruments.**
   In the first labeling type the samples are classified based on the instruments they are played with. Since the database provided contains 16 different instruments, the samples were divided in 16 different clusters.

2. **32 Labels instruments with variation.**
   To add some complexity to the instrument clustering and to further test the stability of the code, an additional variation to the instruments clustering was considered. So for example, a sample played without mute will be given different label than a sample played by the same instrument with mute.

3. **498 Labels of playing techniques.** The last clustering and the most complex ranking problem to solve, is considering that two samples played with different techniques will have different labels even if they are being played by the same instrument. Thus there would be 498 different labels in this clustering.

1.2 Experimental procedure

The experimental procedure is divided into two parts. The first one is computing the features based on one of the two representations: MFCC or scattering(with different parameters). In the second part the ranking metrics before will ve computed before and after processing the data in the feature space.

In the first part, the use of the two representations is done as follows : using a fixed length window of calculationa certain number of temporal features is obtained based on the length of the signal. Then, by averaging along the time axis, only one vector of features by audio file is left.
For the second step, the distance computation 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.[8]

After processing the data, The performance will be evaluated using the two ranking metrics in the euclidean that are the basis of our analysis, MAP and P@k.

The following sections will follow the same scheme presented in figure 1.1. First will be the feature extraction procedure, following will be the methods of processing in the feature space .At the end, The performance will be analyzed using the distance computation and ranking metrics.

1.3 Feature extraction

In this section the two method of feature extraction used in this paper will be presented alongside the motivation and the algorithm of computation. First the Mel Frequency Cepstral Coefficients are presented and then the scattering representation.

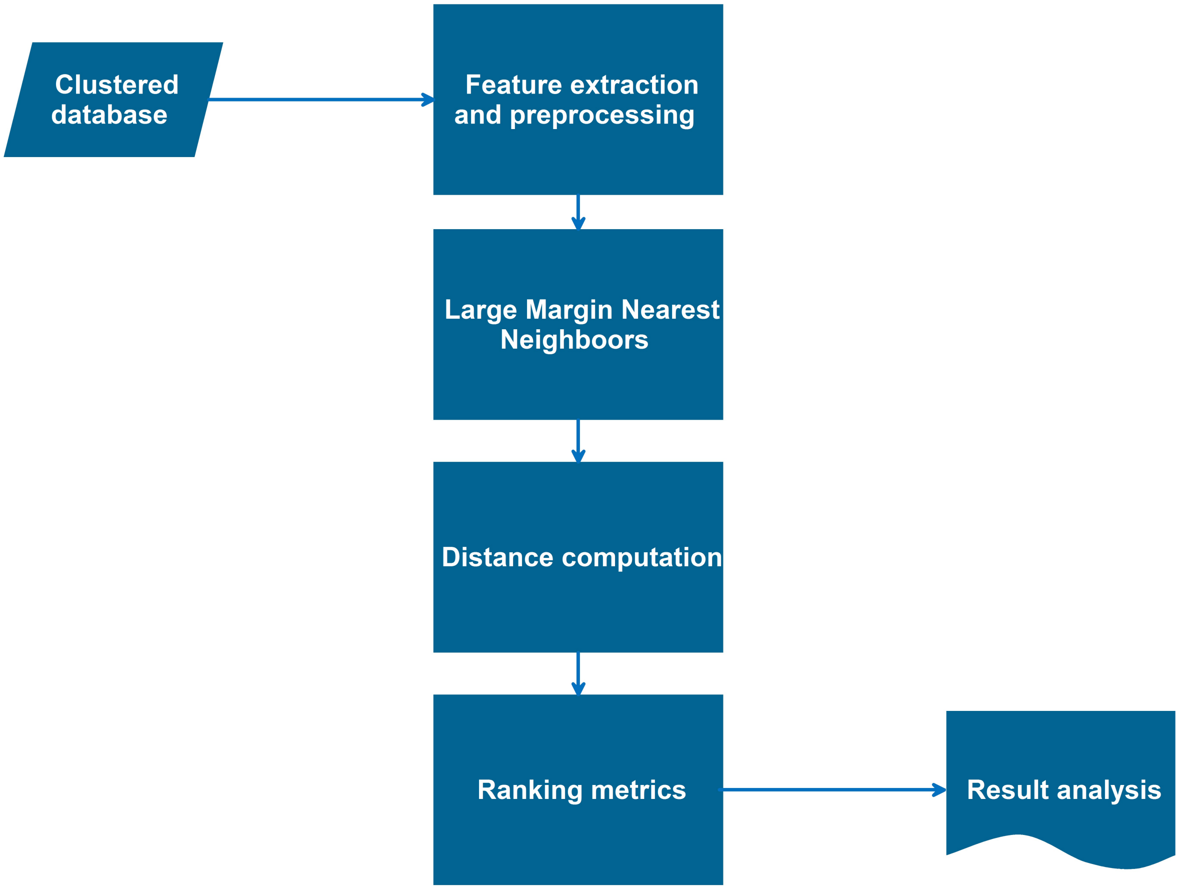

Figure 1.1: Experimental procedure

1.3.1 Mel Frequency Cepstral Coefficients MFCC

The MFCC (Mel Frequency Cepstral Coefficients) became famous with the rise of speech recognition systems. The MFCC are able to module a spoken stream in a compact space. Those coefficients contains most of the information needed for processing spoken audio streams. They are still found in most of the audio processing applications. In the musical domain MFCC has been widely used for music genre classification.
The biggest limitation of the MFCC is its inability to module large time scale variation.

Algorithm and motivation

The extraction of the MFCC features is divided into 5 steps as presented in figure 1.3. In this part, Each step is presented along with the motivation and the limitation it provides.

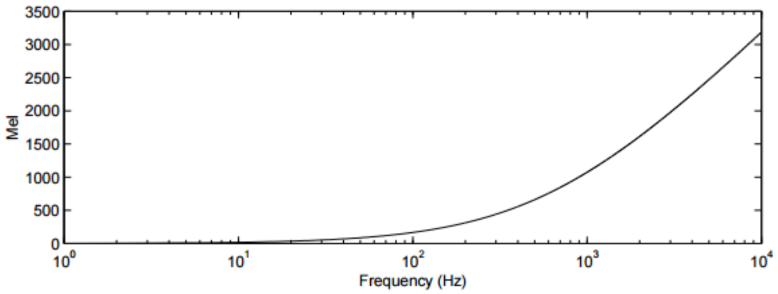

Figure 1.2: Mel scale

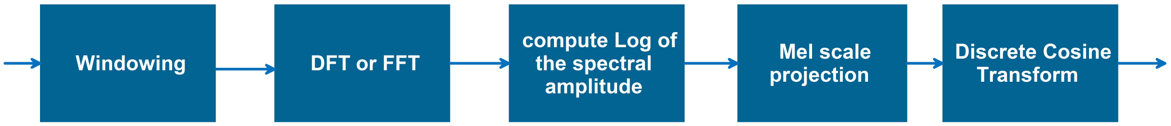
Windowing
DFT or FFT
compute Log of the spectral amplitude
Mel scale projection
Discrete Cosine Transform

Figure 1.3: Procedure of MFCC feature extraction

1. **Frame division**
   The first step is to divide the waveform into equal spaced frames in the time domain. The division is obtained by applying a windowing function at fixed time intervals. This step is motivated by the fact that audio signals (music and speech) can be considered as statistically stationary for a short period of time. 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 length between 25 and 40ms. The longer the frame the bigger the variation of the signal in each frame and this will lead to a loss of information.

2. **DFT or FFT computation**
   The discrete Fourier transform (or fast Fourier transform) is used to project from the time domain to the frequency domain. By using the Fourier transform we obtain the amplitude of the spectrum and thus information related to the phase

will be lost. The motivation of computing the FFT, is that the information provided to the brain by the cochlear is the amplitude of the frequencies(1.3.1).

3. **LOG of amplitude spectrum computation**
   Take the log of the spectrum amplitude obtained by the Fourier transform. This step is motivated by the human perception of the sound loudness. The loudness of the sound is perceived by human in a logarithmic way : 8 time more energy has to be given to achieve a hearing experience of double loudness.

4. **Mel scale projection** In this step the purpose is 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 algorithmic for high frequencies(higher than 1KHZ) and linear for low frequencies(below 1KHZ). The scale is motivated by the fact that human do not perceive pitch in a linear way. The mapping helps in reducing the space to n features space, where n can vary up to 40.

5. **DCT computation** The space now contains 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 only 12 of the 27 features are kept.

MFCC parameter choice

In this sections the results of considering different

Results

The results are divided into three tables with the number of class varying from 16,32 to 498. Two variations of the mfcc are shown with the first one is by taking 12 features out of the 27 and the other is with taking all the 27 features. The last one is by taking the features in the mel space without applying the DCT(40 features).

features	map	pat5
16 classes of instruments		
mfcc 12/27	22.66	85.93
mfcc 27/27	19.72	85.24
mel	16.48	62.90
32 classes of instruments with variation		
mfcc 12/27	20.84	83.98
mfcc 27/27	17.87	82.85
mel	14.72	60.26
498 classes of playing techniques		
mfcc	8.24	43.95
mfcc 27/27	8.85	43.96
mel	5.74	32.90

Table 1.1: Results of the study performed on the MFCC features. Labels taken as 16 class of instruments

The best results were proven to be by taking 12 out of the 27 MFCC coefficients. Those results were expected since by discarding the upper part of the MFCC coefficients the space becomes more representative especially to a problem such as instrument recognition.
The mel coefficients are the one that are highly correlated and it is well seen that for the MFCC what makes it such accurate is the decorrelation step. This is clear by comparing the results of MFCC 27/27 and the mel coefficients.

MFCC features preprocessing

When dealing with a big space of data, Usually a preprocessing technique should be applied before proceeding to use a learning algorithm. This usually proves to be efficient, and there are different techniques of normalization that can be tested for example : The effect of preprocessing the MFCC features was tested using the standardization technique.
Feature standardization : In feature standardization, each feature is taken independently and treated in the following way :
for each dimension we

- Compute the mean.
- Substract the resulting mean.
- Compute the standard deviation.

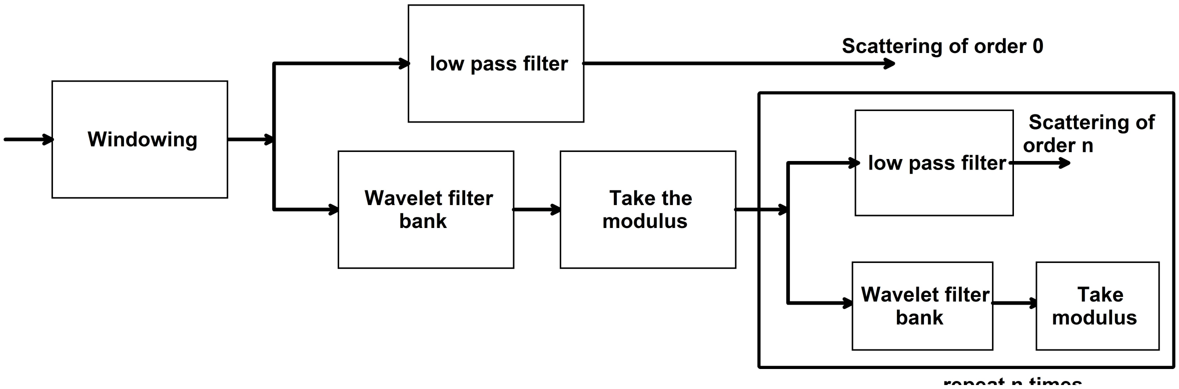

Figure 1.4: Scheme of the scattering transform procedure

- Divide by the standard deviation.

This technique is widely used as a preprocessing for MFCC, since it removes the effect of the DC component that overshadows the space. The result of applying the standardization is shown in the tables below.

features	map	pat5
16 classes of instruments		
raw	22.66	85.93
standardize	24.22	86.89
32 classes of instruments with variation		
raw	20.84	83.98
standardize	22.29	85.12
498 classes of playing techniques		
raw	8.24	43.95
standardize	8.78	45.19

Table 1.2: Comparison between MFCC before and after standardization for the 16 classes of instruments

As it is shown in the tables, the results indicate that standardization yields to a slight performance while performed in the MFCC feature space. As discussed, the standardization effect the variation between observations, and by standardizing this variation, one can anticipate to achieve better performance.

1.3.2 The Scattering transform

The second feature extraction method that was studied is the scattering transform. This method achieves signal decomposition using multiple wavelet transform alongside modulus operators. Figure 1.4 shows a scheme of how the procedure of extracting the scattering coefficients works.

This method first proposed by Stephan Mallat in [5] is relatively new. A lot of research are being made on its application in different domains such as image and audio. In this paper a study on its effectiveness for music classification problems is presented.

1. We first divide the audio samples into frames by using a windowing function. Those windows can vary from multiples of 10 milliseconds to the order of multiples of 100 milliseconds.
2. Apply to the frames a low pass filter. The output of that filter will be the scattering of order n.
3. Apply to the frames a wavelet filter bank. Take the modulus of the output and replace the frames with the output of this step.
4. Repeat the step 2 and 3 n times. For audio signals scattering of order 2 is enough to represent the data.

Reducing variance of the representation

Audio signals contains information that dose not alter the perception of sounds by human. And for tasks such as instrument identification, those information are not important. Such information are :

- Translating an audio signal in time will not alter the perception of this sound by human. Audio signals perception are thus largely time invariant. Discarding the information of time location will not alter the identification of instruments or playing techniques. The representation Φ will thus have the following property $\Phi(x(t-c)) = Phi(x(t))$.
- Other information that can be discarded are the one related to time warping. Time warping can be considered as shifting the signal by a factor that is time dependent. Audio signals are invariant to small scale time warping. What is needed from the representation is not to be time warping invariant $\Phi(x(t-c(t))) =$

$\Phi(x(t))$ rather to be stable to time warping. This can be seen as a lipschitz continuity condition :

$$||\Phi(x(t - c(t))) - \Phi(x(t))|| \leq C||x||||c'||_\infty$$

The scattering transform is based on a wavelet transform. To better understand how the time shifting and stability to deformation is obtained, let us look at each of the figure 1.4 block alone.

Low pass filter : To achieve stability to time shifting, an average in time is applied. This average is achieved by applying a low pass filter. At the first step all the information is lost by this low pass filter, and the result is 0 for scattering of order 0. This same averaging will be applied again before the output of each order to insure time invariance.

High pass filter: Since a lot of the information is lost in the low pass filter, A series of high pass filter is applied to capt the lost information. To insure stability to time warping, The high pass filters used are wavelet functions. The use of wavelet is motivated by the fact that at low frequency they have high frequency resolution and at high frequency the have low frequency resolution. This will guarantee overlapping at high frequencies between a signal warped in time and a signal not warped.

Take the modulus By applying a low pass filter again to the output of the high pass filter, again an average of 0 is obtained. To avoid this, a non linearity should be introduced. In the scattering transform, the best non linearity to be considered is by taking the modulus.

Cascading At each step, only the output of the low pass filter is taken. This means that at each step, a loss of high frequency information is forced. To obtain those information again, a cascade High pass filter and low pass filter is done until the loss of information is not significant anymore. To further demonstrate the importance of second order coefficients the following examples will be discussed.

Examples presenting the importance of second order

In the figure 1.5 two audio samples are presented of an accordion playing the note A3. The first sample is with a soft attack and the second one is with a sharp attack. Since in the first order coefficient, the resolution of high frequency is low, no information related to the attack can be found easily and the two samples can be confused.
By looking at one of the second order coefficients corresponding to one of the high frequencies. The attack can be clearly found. And the difference between the two samples can be easily done.

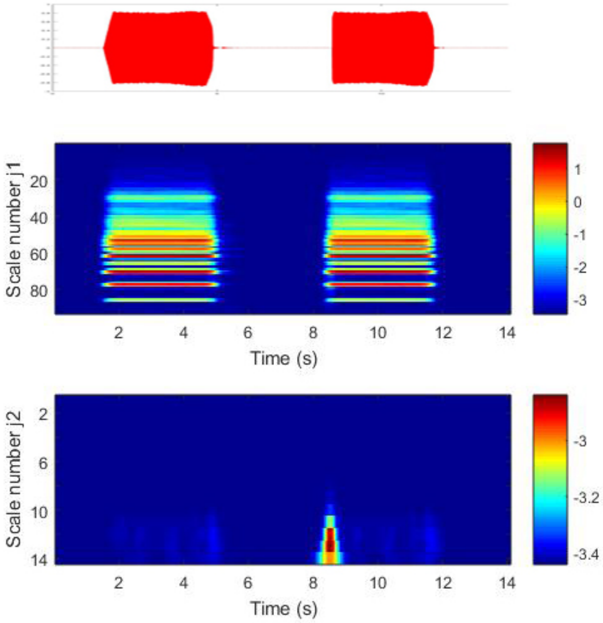

Figure 1.5: [a.] Variation of amplitude in time of two samples one with smooth attack and the second one with sharp attack. [b.] scattering of order 1 of both signals [c.] scattering of order 2 of both signals

In the figure 1.6 the same analysis can be made. This time The first sample present an audio without vibrato, while the other one present the same audio with vibrato. Since the first order coefficients does not have high resolution for high frequencies, the frequency of the vibrato is lost. By taking the second coefficient of one of the first order feature, the frequency of the vibrato is clearly visible as a frequency not varying in time.

Results of varying the length of the windowing function T

The impact of taking smaller window time on the output of the ranking metrics. The test was performed on the 16,32 and 498 labeling variations. The results are given in the three following tables.

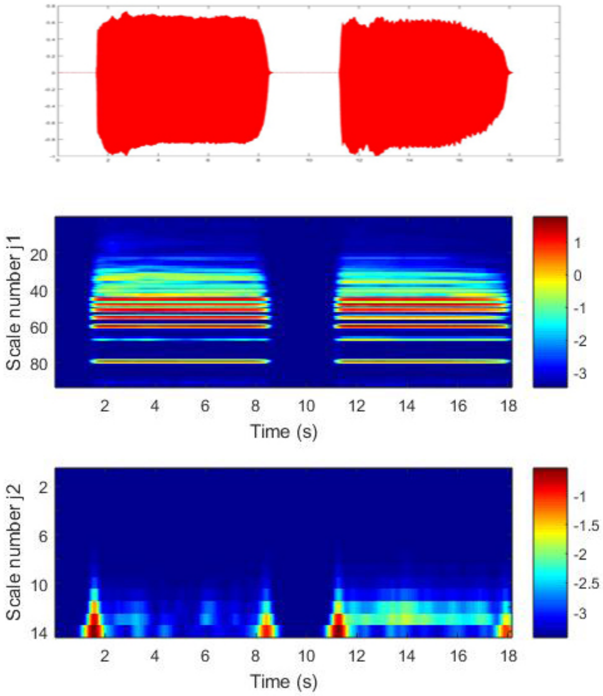

Figure 1.6: [a.] Variation of amplitude in time of two samples one without vibrato and the second one with vibrato. [b.] scattering of order 1 of both signals [c.] scattering of order 2 of both signals

features	map	pat5
16 classes of instruments		
scattering T=25ms	18.07	72.79
scattering T=128ms	16.79	70.47
scattering T=250ms	16.49	70.40
32 classes of instruments with variation		
scattering T=25ms	16.29	70.17
scattering T=128ms	15.18	67.21
scattering T=250ms	14.89	67.05
498 classes of playing techniques		
scattering T=25ms	7.79	40.17
scattering T=128ms	6.66	37.38
scattering T=250ms	6.46	37.07

Table 1.3: Table comparing the results of applying metric ranking to different window length for the scattering transform.

As presented in the tables, the bigger the windowing time the lower the precision. This is due to the fact that the scattering will be discarding information related to the small time variation. In the next part, the effect of preprocessing on the biggest windowing time is presented.

Preprocessing the scattering features

As for the MFCC, the normalization that was tested on the scattering features is the standardization. Another type of preprocessing was applied, that will be referred to as std and median method.

Std and median method : The scattering features present a lot of variability in its factors, with some of the factors being irrelevant to the process. To remove those factors and reduce the space, a study of variance should be done. The variance of each feature alone is computed, and sorted by growing values. The accumulated sum is then computed and either the high or the low variances will be discarded. In both cases an improvement was noticed but the best result was achieved with only leaving 83% of the high frequencies. This is the std part of the method.
The second part is to divide by the vector of median (the median of each feature is computed alone). The feature space is then divided by the median. This will make the space symmetrical and close to Gaussian.
The effect of the standardization and the std and median method is represented in the three tables below.

features	map	pat5
16 classes of instruments		
raw	16.49	70.40
standarize	19.31	78.98
stdandmedian	30.73	93.87
32 classes of instruments		
raw	14.89	67.05
standarize	17.52	75.13
stdandmedian	28.07	90.94
498 classes of playing techniques		
raw	6.46	37.07
standarize	10.69	47.01
stdandmedian	20.41	57.98

Table 1.4: Table comparing the results of applying metric ranking to different technique of preprocessing on a scattering with T=250ms considering 16 class of instruments

For the scattering, the preprocessing technique that will be taken into account is the std and median. By considering the table, applying the std and median technique is proved to be beneficial. It is to be noted that the std and median can not be applied to the MFCC since the coefficients taken are already in a very compact space(space of 12 features).

In the next chapter, we shall see the results of applying the feature extraction on the ground truth labeling problem.

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] Stephan Mallat *Recursive interferometric Representations* 18th European Signal Processing Conference 2010

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

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

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