Musical Instrument Categorization using Spectral- and Cepstral Analysis

Jakob Florén and Joel Torby

Bachelor's thesis
2016:K6

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

Abstract

This thesis suggests and tests two methods for categorizing instruments in a musical signal. The first method involves spectral analysis and the magnitude spectrum, analysing the structure of pitches and overtones and how the amplitude of those changes over time. The second method involves cepstrum analysis and especially the, for speech recognition popular method computing the Mel Frequency Cepstral Coefficients (MFCC). For each method, a weighting procedure using Least Squares is finally used to determine which instrument that are present in a signal and which are not. This is done by comparing the unknown signal to a reference bank of tones. So, even if this thesis is primarily about testing and comparing two methods, it is also about how much a linear Least Squares weighting method can absorb the non-linear calculations of magnitude spectrum and MFCC.

The results for both methods in combination with the Least Squares weighting procedure are promising. For some musical signals the instrument categorisation succeeds very well while for other signals improvements of the methods are needed. These improvements may be achieved by combining features from both methods and adjusting the weighting system.

Contents

1 Introduction

Music, the universal language that passes no one unnoticed. It ignites emotions within us and moves us in certain ways that are often very hard to explain and understand. Some composition of instruments, tones and beats will move some people in a certain way, while leaving others unaffected. Every instrument have different characteristics, these can for instance be shape, what material it is made of and timbre. A trained musical ear, can in most cases distinguish between the instruments that make up a melody and most of us would pretty easy hear the difference between a piano and a guitar playing the same tone. Why is that, and how can it be explained? One of the properties already mentioned is timbre also known as tone color or tone quality. Quoting [1], one simple definition of timbre could be

"In simple terms, timbre is what makes a particular musical sound different from another, even when they have the same pitch and loudness."

One way of trying to explain timbre is to apply spectral analysis to music signals and then analyze the structure of the fundamental pitch and its overtones and then study how they evolve over time. By doing this for different instruments, there can be ways to mathematically explain and separate instruments from each other.

2 Methods

This thesis will test two methods for characterizing instruments in a melody. The first method involves overtone analysis using the Fast Fourier Transform (FFT) and the second method using cepstral analysis and the Mel Frequency Cepstral Coefficients (MFCC) which is the standard method used for voice recognition.

To generalize, given a time frame of data in a melody, both of the above methods will compare that time frame with a corresponding time frame from a set of reference frames. This comparison will differ a bit for both methods and will be explained in detail i section 2.1 and 2.2. As an example, given a time frame from a music mix of, say clarinet, guitar, flute and harmonica, that frame is to be compared with a reference frame for all possible instruments involved in the melody. This process is illustrated in below Figure 1.

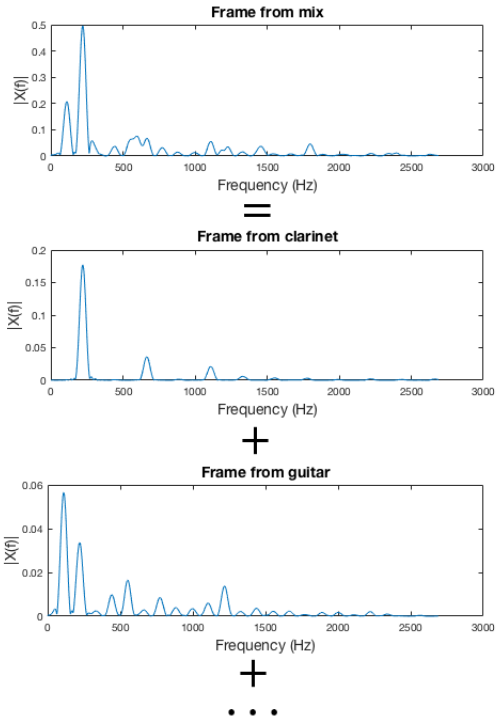

Figure 1: A time frame from a mix is being compared with corresponding time frames from each instrument present in the melody

A weight is then assigned to each instrument resulting in the matrix equation

$$\mathbf{h}^{tot} = w_1\mathbf{h}^{clarinet} + w_2\mathbf{h}^{guitar} + w_3\mathbf{h}^{flute} + w_4\mathbf{h}^{harmonica}, \tag{1}$$

where $\mathbf{h}^{tot}$ is a $m \times 1$ vector holding m, number of coefficients of a sought for feature of the mix frame, such as magnitude, amplitude or power in the frame for a given frequency, $\mathbf{h}^{clarinet}$ holds the corresponding coefficients regarding the clarinet, $\mathbf{h}^{guitar}$ corresponds to the guitar, $\mathbf{h}^{flute}$ for the flute and $\mathbf{h}^{harmonica}$ for the harmonica all with the same dimension as $\mathbf{h}^{\mathbf{tot}}$, that is

$$\begin{pmatrix} h_1^{tot} \\ h_2^{tot} \\ . \\ . \\ . \\ h_n^{tot} \end{pmatrix} = \begin{bmatrix} h_1^{clarinet} & h_1^{guitar} & h_1^{flute} & h_1^{harmonica} \\ h_2^{clarinet} & h_2^{guitar} & h_2^{flute} & h_2^{harmonica} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ h_m^{clarinet} & h_m^{guitar} & h_m^{flute} & h_m^{harmonica} \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix}. \tag{2}$$

Here, the only unknowns are the weights $w_1, ..., w_4$ and these can be estimated using Least Squares. By denoting the matrix by $\mathbf{H}$, the problem becomes to find the vector, say $\mathbf{w}$ that minimizes

$$\left\| \mathbf{H}\mathbf{w} - \mathbf{h}^{tot} \right\|_2, \tag{3}$$

where $\|\cdot\|_2$ denotes the 2-norm. The idea is, that the weights will now say how much of a certain instrument is present in a particular frame from a mix.

Using ordinary Least Squares will result in negative weights but since a contribution from an instrument can not be negative the use of Non-Negative Least Squares[2] is preferable. So, the problem instead becomes, find the vector $\mathbf{w}$ that minimizes

$$\left\| \mathbf{H}\mathbf{w} - \mathbf{h}^{tot} \right\|_2 \;\; \textit{subject to } \mathbf{w} \geq 0. \tag{4}$$

The implementation used in this report for solving the Non-Negative Least Squares problem is the implementation from matlab, which is itself built upon [2].

2.1 FFT, Fast Fourier Transform

In order to find timbre in instruments, the first approach is simply by calculating the Fast Fourier Transform (FFT)[8] of a signal and determine the power of the peaks in the magnitude spectrum (= absolute value of the FFT). Since a tone played by an instrument is said to be non-stationary the signal is divided into time frames. By taking this into account the fact that the overtones might change over time is considered.

To get a better view of what is meant by overtone structure study Figure 2 and compare the power proportions for one instrument. For the flute playing $D6$, the fundamental frequency (1176Hz) have a very strong amplitude and for the first and second overtone (2349 Hz and 3520 Hz) the amplitude is low for both the acoustic and the digital flute. For the guitar playing $A3$, the fundamental tone (220 Hz) and the first overtone (440 Hz) are dominating the signal for both the acoustic and the digital guitar.

The idea on which this report builds upon, is that two different instruments of the same category (e.g. two guitars) playing the same tone will have a similar timbre which in this case means that they will have the same overtone structure. Hopefully the overtone structure will be unique in some way for every instrument category so that it will be possible to separate two different instrument categories.

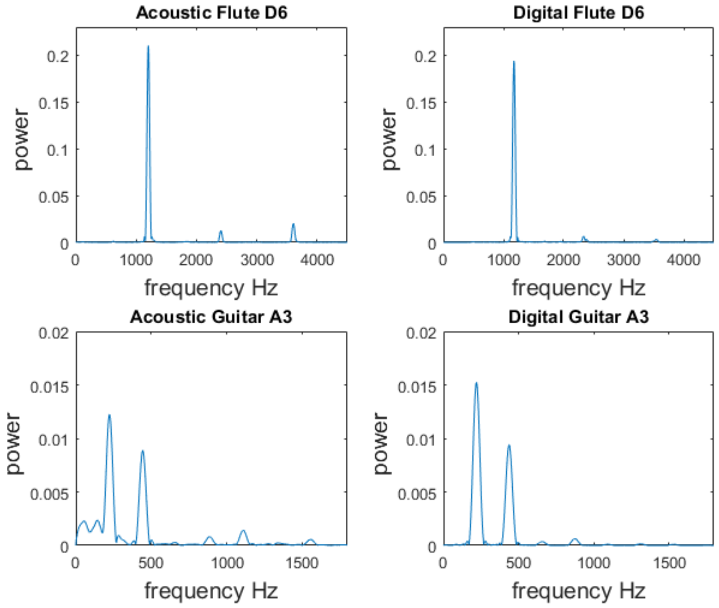

Figure 2: Overtone structure of two different flutes and two different guitars

In order to categorize a tone played by an unknown instrument, the overtone structures are calculated for all frames. This is compared with a bank of overtone structures from known reference signals. By using Non-Negative Least Squares the weights of each reference signal's overtone structures are calculated for all frames. That is

$$\mathbf{h}_l^{tot} = w_1\mathbf{h}_l^{clarinet} + w_2\mathbf{h}_l^{guitar} + ... \quad , \tag{5}$$

where $\mathbf{h}$ is the overtone structure in the frame l (for different instruments) and where $\mathbf{h}^{tot}$ is the unknown signal consisting of one or more instruments. w_1, w_2,... are the weights of the references.

Practically the overtone structure for a time segment of the signal is calculated by:

1. The fundamental frequency is identified. This is done e.g. by using the absolute value of the FFT and localize the first significant peak, see Figure 3.

2. The signal is divided into r frames of 40 ms (which becomes $K = 1764$ data points if the signal is sampled at sampling frequency, $f_s = 44100$ Hz) and use 75% overlap. The signal is then multiplied with a Hanning window

$$W(n) = 0.5(1 - cos(2\pi\frac{n}{K})), \quad 0 \leq n \leq K - 1, \tag{6}$$

   resulting in the new signal y. Now y is zero-padded up to $N = 2^{14} = 16384$ datapoints. The magnitude spectrum is, for each time frame l, given by

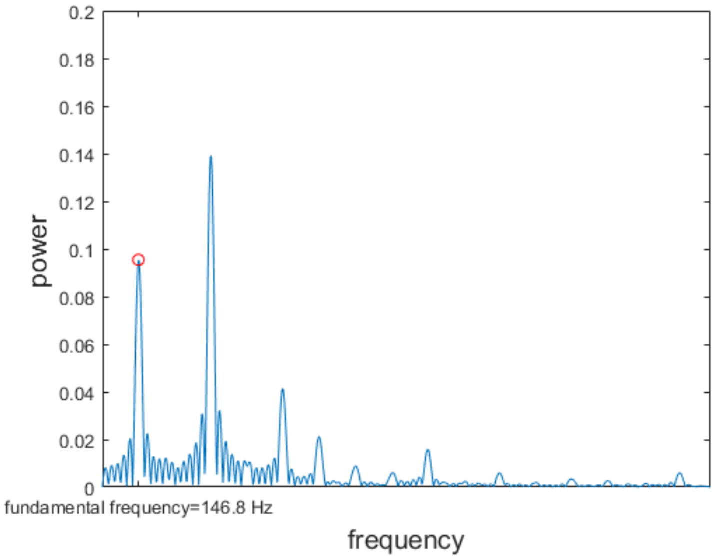

Figure 3: Finding fundamental frequency from the absolute value of the FFT of an unknown signal. The tone $D3 = 146.8$ Hz is found.

$$S_l(k) = \left| \sum_{n=0}^{N-1} y_l(n) e^{\frac{-2\pi ikn}{N}} \right|, \quad k = 0, 1, ..., N-1, \tag{7}$$

where S is the magnitude spectrum of the discrete Fourier transform of y.

Using the power spectrum (the squared magnitude spectrum) was considered but discarded since the spectrum raised the highest peak and reduced all other peaks. This contributed to that the overtone structure became less significant for the different instruments which would deteriorate the results.

3. Overtones are localized. These are found as peaks in the magnitude spectrum close to the integer multiples of the fundamental frequency, see Figure 4.

4. The identified amplitudes of the peaks are saved in the vector

$$\hat{\mathbf{p}}_l = [\hat{p}_0 \quad \hat{p}_1 \quad ... \quad \hat{p}_{m-1}], \tag{8}$$

where l is the time frame, $\hat{p}_0$ is the amplitude of the fundamental frequency and $\hat{p}_1$,..., $\hat{p}_{m-1}$ are the amplitudes of the first $m-1$ overtones. How to choose an optimal m for music instruments is not investigated in this project.

Since the overall power of the amplitudes vary from different recordings (for example depending on loudness and position of the instrument compared to the recorder) only the proportions

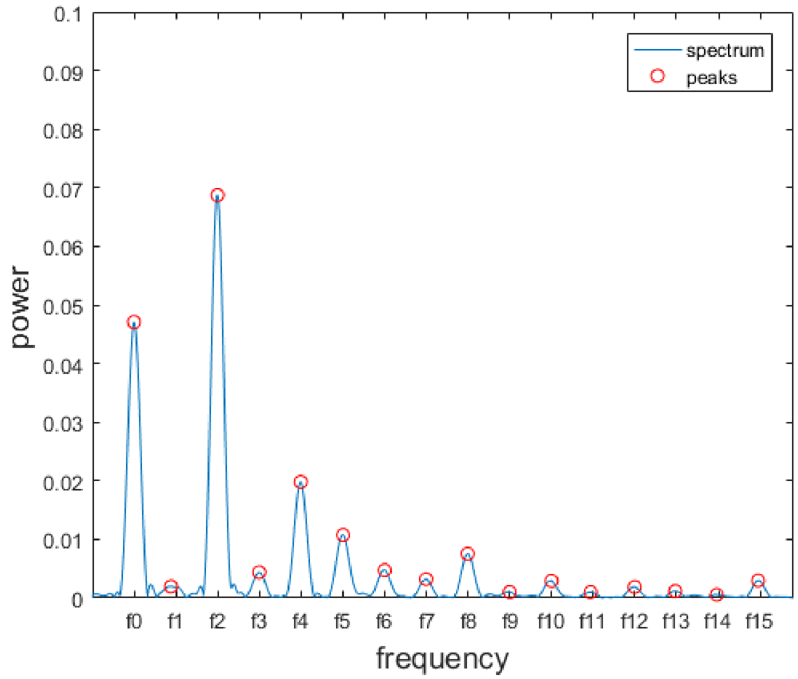

Figure 4: Magnitude spectrum with localized peaks

of the amplitudes are investigated. Therefore the amplitudes are divided by the norm forming a normalized amplitude vector

$$\mathbf{p}_l = \frac{\hat{\mathbf{p}}_l}{\|\hat{\mathbf{p}}_l\|_2} \tag{9}$$

for every time frame l.

5. Since a tone played by an instrument is non-stationary, the overtone structure is calculated for all time frames in the unknown signal. The vectors $\mathbf{p}_l$ results in a matrix $\mathbf{Q}$ (of size $r \times m$ where r is the number of time frames and m is the number of coefficients in $\mathbf{p}_l$)

$$\mathbf{Q} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_r \end{bmatrix} \tag{10}$$

that describes how the normalized amplitudes, of the fundamental frequency and the $m-1$ first overtones, changes over time. By continuing the example with the unknown signal, the corresponding Q-matrix is shown in Figure 5. In this case the fundamental frequency and the first 4 overtones are shown. The matrix $\mathbf{Q}$ shows how the normalized amplitudes of the overtones vary over time. For the clarinet playing $D3$ which is shown in this example, the normalized overtones are quite stationary.

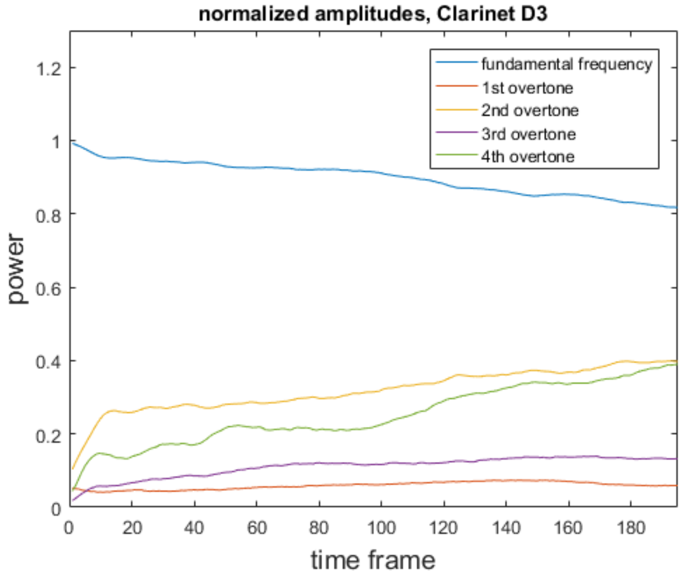

Figure 5: The matrix $\mathbf{Q}$, showing the normalized amplitudes of the overtones over time

The same algorithm is used for calculating the overtone structure for the reference signals. After that a Non-Negative Least Squares calculation of the $\mathbf{p}$ vectors of the references are made for every time frame l. Here the general feature vector $\mathbf{h}_l$ in equation (5) is replaced by the normalized amplitude vector $\mathbf{p}_l$ for the unknown signal and every reference. Finally the reference corresponding to the highest weight is said to be of the same instrument category. In Figure 6 it's clearly seen that the unknown signal matches the clarinet $D3$ reference the best.

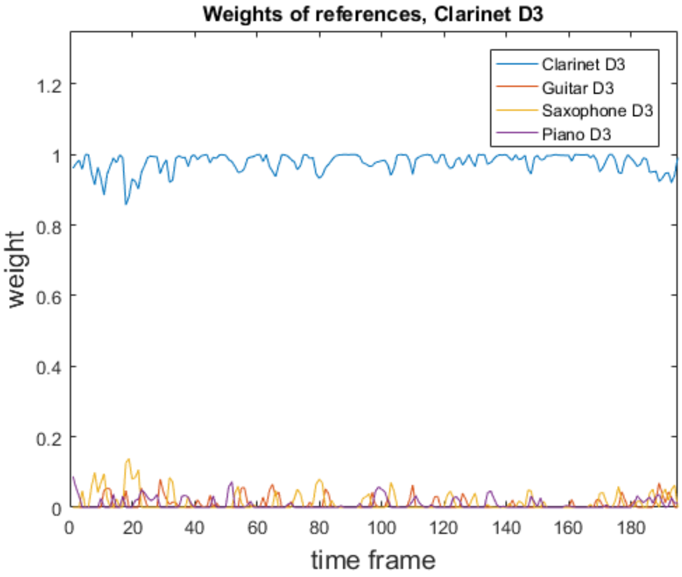

Figure 6: Weights of references

2.2 MFCC, Mel Frequency Cepstral Coefficients

For speech signals and especially for voice recognition Mel Frequency Cepstral Coefficients, MFCC is widely used and Malcolm Slaney's matlab library [3] is an accepted library in the area of speech recognition. Since this method works well in determining the timbre in voiced speech, it might be possible to apply it to music and use it to distinguish different instrument categories. In this method, the algorithm for determining the MFCC for voiced speech will be applied to music signals and then the obtained coefficients will be used to determine which instruments are included in a tune by again using Non-Negative Least Squares. That is, **h** in equation (1) now holds the MFCC.

To find the MFCC of a signal, the following steps are performed:

1. The signal x is pre-emphasized by

   $$x(n) = x(0) + \sum_{n=1}^{M} x(n+1) - \alpha x(n), \tag{11}$$

   where the standard value of α is 0.97 and M is the total number of samples. This is a high-pass filter and if the signal x would have contained a trend this step would have eliminated it.

2. The signal is divided in the same way as for the FFT-method (step 2). The magnitude spectrum S_l is calculated for each of the time frames x_l, where $1 \leq l \leq r$. Using equation (7), now with a Hamming window

   $$W(n) = \begin{cases} 0.54 - 0.46cos(\frac{n\pi}{K}) & ,0 \leq n \leq K \\ 0 & ,\text{otherwise} \end{cases} \tag{12}$$

   of length K. Doing this for all frames will result in a $r \times K$ matrix **S** containing spectrums for all frames.

3. A mel-scale filter bank containing overlapping triangular windows is constructed. At low frequencies the human ear detects changes in frequency rather easy but for higher frequencies a change is more difficult to detect. This means that for low frequencies the perception can be regarded as being linear whereas for higher frequencies the perception is more logarithmic. So, after converting to mel-scale the human ear can detect changes in high mels just as good as for low mels. To possibly be able to describe the timbre noticeable by the human ear in mathematical terms this conversion may be needed.

   For speech the number of total filters commonly used is between $26 - 40$ [4] and the frequency range is in between $f_{min} = 100$ Hz and $f_{max} = 6800$ Hz, [5]. For music on the other hand the frequency range is a lot wider, especially, since the structure of the overtones is interesting. The highest f_{max} that could be chosen is $f_s/2 = 44100/2 = 22050$ Hz, the Nyquist frequency. The lowest tone on a piano, $C0$, corresponds to 16.35 Hz which is not far from zero. For this implementation $f_{min} = 0$ Hz and $f_{max} = 22050$ Hz are set. Since music instrument has a wider range of frequencies than speech, more filters in the filter bank is likely to be needed.

The number of linear spaced filters is set to be 40 and the number of logarithmic spaced filters is also set to be 40 resulting in a total number of $L = 80$ filters in the filter-bank. The length of each filter is K and by storing all filters in a matrix, $\mathbf{F}$, this matrix will now be of dimension $K \times L$. The higher overtone considered, the less significant it tends to be and therefore it's more effected by noise. A logarithmic decaying of the amplitudes of the triangular filters is therefore applied for the higher frequencies.

For an example of how the filter-bank could look like, see Figure 7. There are many different conversion from Hz to mels. Which conversion that would be optimal to use is not scientifically proven.

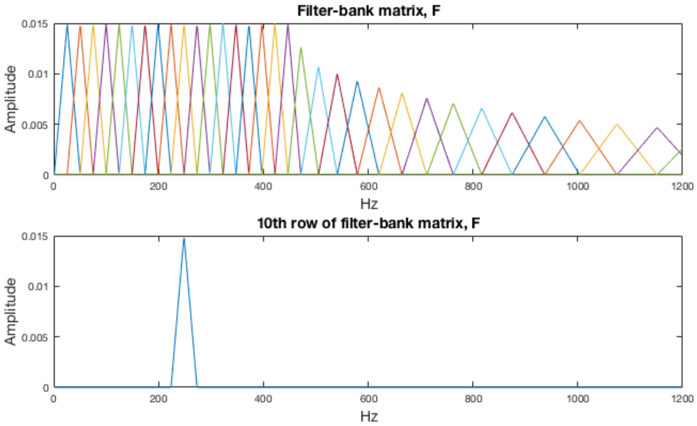

Figure 7: Example of a filter-bank matrix $\mathbf{F}$ consisting of 20 linear spaced triangular filters with constant amplitude and 20 logarithmic spaced triangular filters with decaying amplitude

4. The matrix $\mathbf{S}$ (with dimensions $r \times K$) is multiplied with the filter-bank matrix $\mathbf{F}$ (with dimensions $K \times L$), resulting in a $r \times L$ matrix $\mathbf{D}$ which is logarithmized,

$$\mathbf{D} = log(\mathbf{S} \cdot \mathbf{F}). \tag{13}$$

5. The Discrete Cosine transform (DCT) is applied. Since the filters in the filter-bank are overlapping, the energies in each row in $\mathbf{D}$ are correlated. A way to de-correlate them is to use the DCT. So for each row $\mathbf{d}$ in $\mathbf{D}$, the DCT is applied. The coefficients in $\mathbf{c}$ now received is the sought MFCC. That is

$$\mathbf{c}_n = h_n \sum_{l=0}^{L-1} \mathbf{d}_l cos\left(\frac{\pi}{2L}(2l-1)n\right), \quad n = 0, 1, ..., L-1, \tag{14}$$

where

$$h_n = \begin{cases} \frac{1}{\sqrt{L}} & , \quad n = 0 \\ \sqrt{\frac{2}{L}} & , 1 \leq n \leq L-1 \end{cases} \tag{15}$$

is applied to normalize each vector, yielding the row vectors $\mathbf{c}_n$ of the resulting MFCC coefficent-matrix $\mathbf{C}$ to be othogonal. Here $\mathbf{C}$ is now a $r \times L$ matrix

$$\mathbf{C} = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,L} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ c_{r,1} & c_{r,2} & \cdots & c_{r,L} \end{bmatrix}. \tag{16}$$

In speech recognition the only used coefficients in MFCC are the numbers 2-13. The rest of the coefficients are discarded since they don't provide any significant information. Applying this to the matrix $\mathbf{C}$ results in the final $r \times 12$ MFCC coefficient-matrix.

As an example, in Figure 8 and Figure 9 these $2-13$ MFCC coefficients are plotted for two different frames. In the first figure, the MFCC of the acoustic guitar matches the MFCC of the digital guitar a bit better than the MFCC of the clarinet. In the second figure that connection is no longer that clear. Another way to analyze the MFCC is to study how the energy at certain frequencies develop over time. This is illustrated in Figure 10 where the first coefficient is plotted over the first 100 frames for each signal. This way it can be seen how the low frequencies contents changes over time. The signals of the guitars more or less exhibit the same pattern up to just a scaling factor of the magnitude. The clarinet on the other hand develops in a totally different manner. Also, in this case it seems that our combination of filter-banks is preferable, making it easier to distinguish between the instruments.

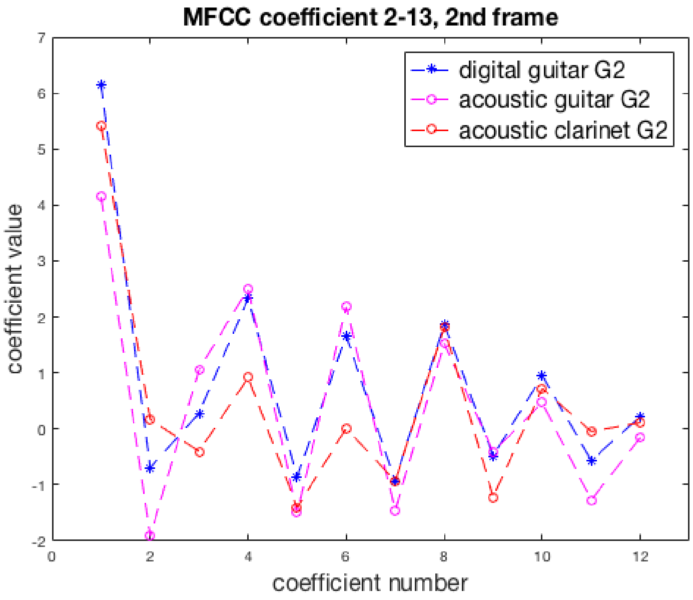

Figure 8: $1st$ frame of signal

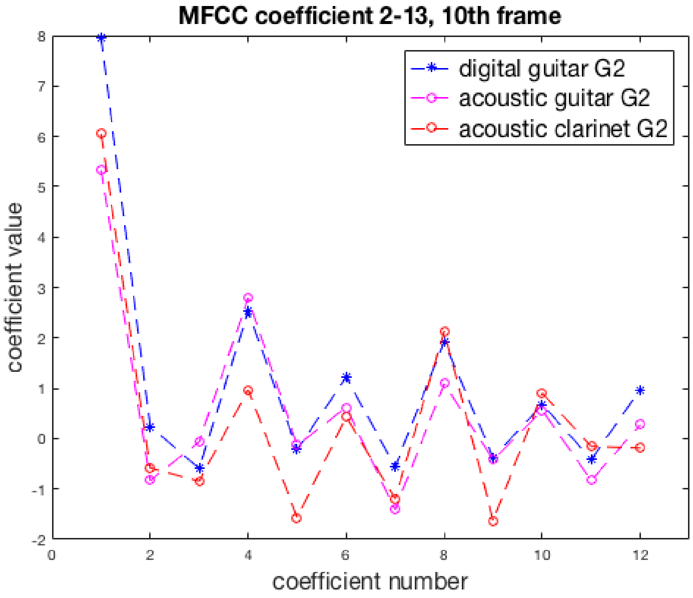

Figure 9: $10th$ frame of signal

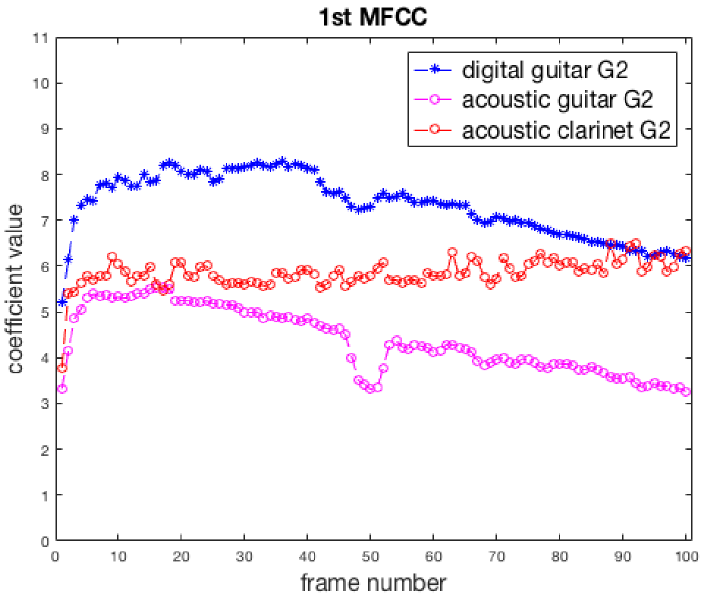

Figure 10: $1st$ MFCC and how it changes over time

3 Tests

For the tests, some assumptions are made regarding the unknown signal that is about to be tested. Firstly, the instruments possibly present are known and can be a single tone or mix of either, clarinet, guitar, saxophone and piano. Also, the tones that are played are assumed to be known. The tones are also sorted in octaves meaning that for example $A2$ and $A3$ are considered to be different tones and not just an A.

The data that will make up our reference bank for the tests are digital tones and mixes created in Garageband [6] and Audacity [7]. Tones from acoustic instruments will also be used and these are home recordings. For the tests where mixes of acoustic tones are needed, these are just added together. All signals, digital or acoustic are sampled at sample frequency 44100 Hz.

For both methods the following tests are performed:

1. Given one known tone played by one unknown instrument, determine the instrument.
2. Given two known but different tones from two known instruments, determine which instrument that play which tone.
3. Given one known tone played by two unknown instruments, determine the instruments involved.

The tests are designed in a way that hopefully will cover as many scenarios as possible. From cases where overtones do not overlap to cases where the tones are the same.

Each of the above tests is itself divided in two different cases. The first case will have both the references and the unknown signal made up from the same digital tones. Case 1 will therefore cover the more theoretical aspect of our methods. The second case will have digital references and an acoustic mix of unknowns. Case 2 will therefore deal with a more realistic scenario. The main reason why each test is divided in two cases comes from the discovery that for some tones, not all, there is a rather big difference in acoustic and digital tones. In particular the lower tones for the clarinet seem to differ more when the digital tones and the acoustic tones are compared.

For every case two signals are tested. For test 2 where two different tones are played the combination of tones have been chosen so that for one signal, many of the overtones overlap and for the other signal only a few overtones overlap.

Totally 2 methods are tested in 3 different ways with 2 cases where each case is tested for 2 different signals. The total amount of signals tested becomes 24 ($2 \cdot 3 \cdot 2 \cdot 2$). This is far from enough for being able to statistically say if a method works for a specific test or not. The tests instead gives indications of what might work and what kind of problems that might appear.

In section 3.1 and 3.2 the results of the tests will just be presented. In section 4 the tests will be commented and analyzed.

3.1 Tests for FFT-method

As explained in section 2.1 this method depends on selecting a number of overtones. This number is hard to set for all instrument and it depends on how much information about an instruments that is present in the higher overtones and this is hard to tell. For the tests of the FFT-method the fifth overtone of the highest known tone in the unknown signal will be set as an upper bound on the total amount of frequencies included. This is chosen to be enough upon just a few observations and a more thorough analysis of this matter may be valuable but not done here.

Test 1

Case 1

The first two tests for different single-tone signals are shown in Figure 11. Here, the unknown signals are a clarinet playing the tone $B4$ (left) and a guitar playing $D3$ (right). Here both the unknown signal and the references are digital tones. The plot shows how the estimated weights changes over time.

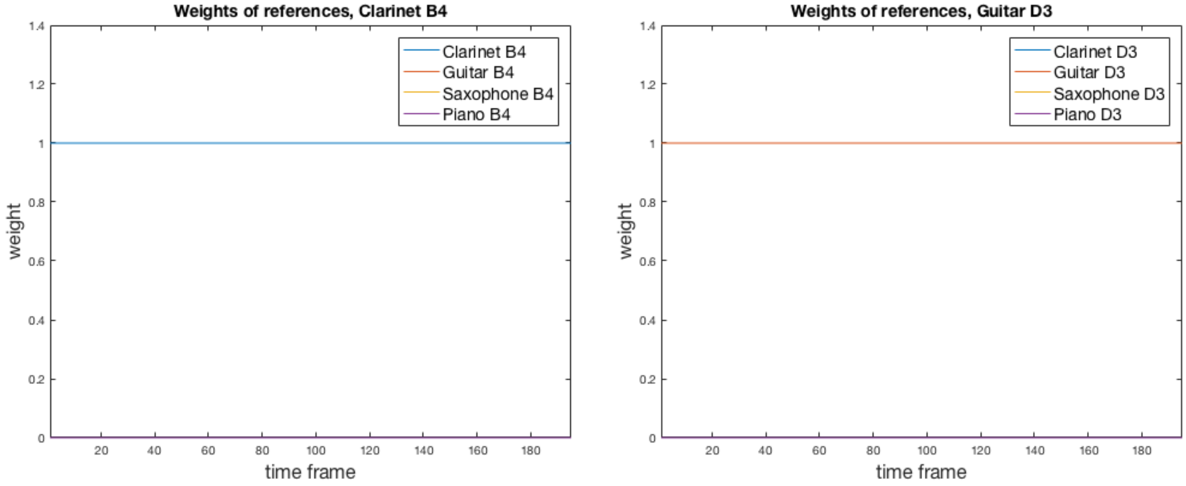

Figure 11: FFT-method. Test 1, case 1

The test result of this case is not very interesting since the solution in equation (4) for every time frame l, the weight vector $\mathbf{w}$ becomes

$$\underset{\mathbf{w}\geq 0}{\operatorname{argmin}} \left\| \left(w_1 \mathbf{p}_l^{clarinet} + w_2 \mathbf{p}_l^{guitar} + w_3 \mathbf{p}_l^{saxophone} + w_4 \mathbf{p}_l^{piano} \right) - \mathbf{p}_l^{tot} \right\|_2, \tag{17}$$

where $\mathbf{p}$ is the normalized amplitude vector. This is trivial since the signal of the unknown instrument is the same as one of the references.

Case 2

For this case the references are the same as in case 1 but the unknown signal is acoustic. The tests are plotted in Figure 12.

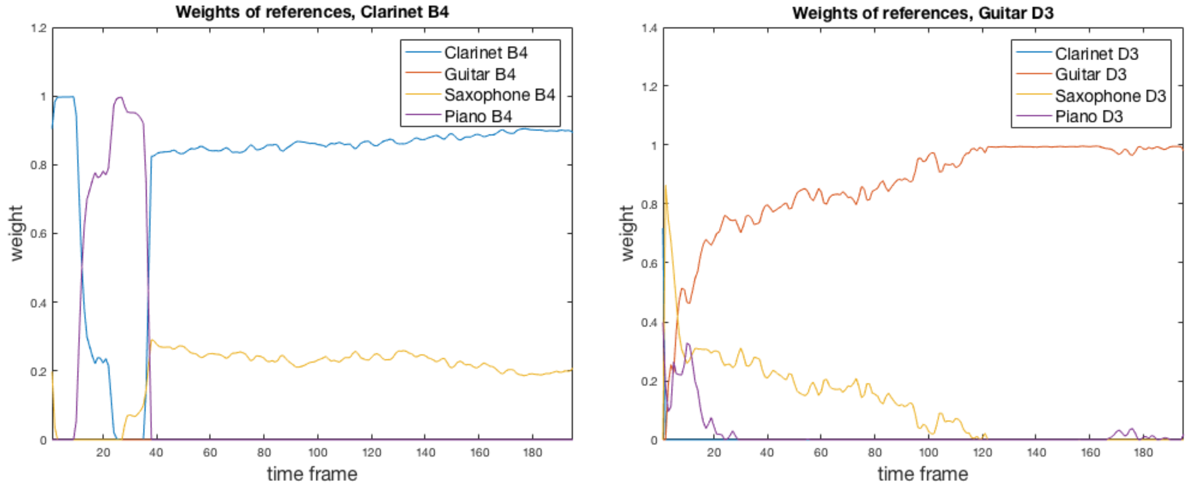

Figure 12: FFT-method. Test 1, case 2

In the left plot, the weight for the clarinet (blue) is dominating for all time frames, except somewhere between time frames $10 - 40$ where contribution from the piano is estimated to be present. In the right plot the estimated weights for the guitar dominates in all time frames except in the beginning of the signal where the saxophone is the dominating one.

The normalized mean of these weights where the sum of all the means are equal to 1, can then be used as a measure of the likelihood of each instrument participating in the unknown signal. Table 1 shows a calculation of the likelihood for the left plot. The corresponding likelihood for the right plot is shown in Table 2. From these tables it is clearly seen that the correct instruments are estimated to be significant.

Clarinet $B4$	Guitar $B4$	Saxophone $B4$	Piano $B4$
0.7177	0	0.1789	0.1034

Table 1: Likelihood of each instrument for left plot in Figure 12

Clarinet $D3$	Guitar $D3$	Saxophone $D3$	Piano $D3$
0.0037	0.8576	0.1180	0.0207

Table 2: Likelihood of each instrument for right plot in Figure 12

Test 2

The second test is performed just as the first but since this test is about determining what instrument playing what tone, in a rewritten equation (5) where $\mathbf{h}_l$ is replaced by $\mathbf{p}_l$ resulting in

$$\mathbf{p}_l^{tot} = \sum_{j=1}^{s}\sum_{k=1}^{t} w_{k,j}\mathbf{p}_{l,j}^{instrument_k}, \tag{18}$$

where s indicates the number of tones in the unknown signal and t indicates the number of instruments used. As an example, let's say that the unknown signal consists of the two different tones played by either a clarinet or a guitar. This yields

$$\mathbf{p}_l^{tot} = w_{1,1}\mathbf{p}_{l,1}^{clarinet} + w_{1,2}\mathbf{p}_{l,2}^{clarinet} + w_{2,1}\mathbf{p}_{l,1}^{guitar} + w_{2,2}\mathbf{p}_{l,2}^{guitar}. \tag{19}$$

Each case for test 2 is performed on two signals, where the first is a mix of a clarinet tone $D4$ and a guitar tone $A4$. The second signal is a mix of a clarinet playing $B4$ and a guitar playing $D3$.

Case 1

The results are plotted in Figure 13.

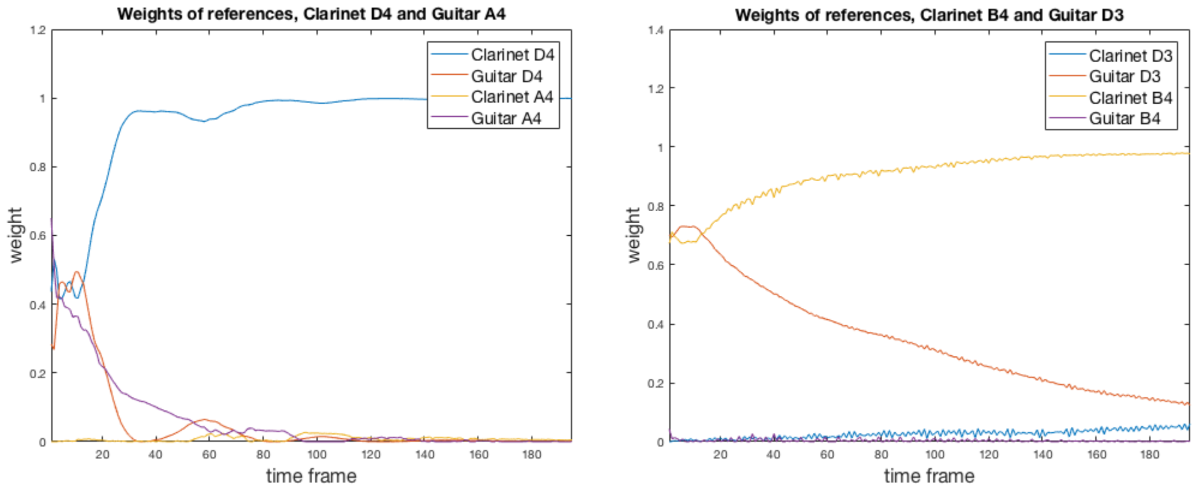

Figure 13: FFT-method. Test 2, case 1

The left plot has a strong contribution from the correct clarinet $D4$ throughout all time frames whereas the guitar is wrongfully indicted as playing both the correct tone $A4$ and the wrong tone $D4$. In the right plot, the two significant tones (both correctly estimated) are the $B4$ from the clarinet and $D3$ from the guitar. The likelihoods are found in Table 3 and 4. By Table 3, the likelihood of the correct guitar $A4$ tone is biggest but not significant from $D4$

Clarinet $D4$	Guitar $D4$	Clarinet $A4$	Guitar $A4$
0.8802	0.0509	0.0077	0.0611

Table 3: Likelihood of each instrument for left plot in Figure 13

Clarinet $D3$	Guitar $D3$	Clarinet $B4$	Guitar $B4$
0.0290	0.2612	0.7096	0.0003

Table 4: Likelihood of each instrument for right plot in Figure 13

Case 2

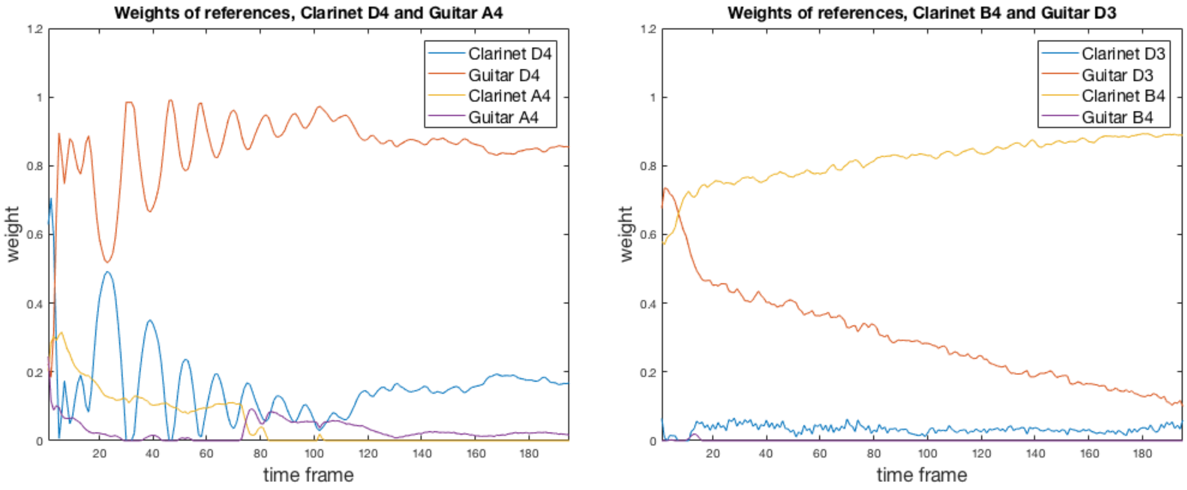

Figure 14: FFT-method. Test 2, case 2

In the left plot of Figure 14 and also from Table 5 the wrong guitar tone $D4$ dominates significantly followed by clarinet $D4$. Even though the signal actually involves a guitar tone $A4$, this tone is calculated to be the least significant. In the right plot and also from Table 6 the correct tones are the ones appearing as significant.

Clarinet $D4$	Guitar $D4$	Clarinet $A4$	Guitar $A4$
0.1448	0.7780	0.0500	0.0271

Table 5: Likelihood of each instrument for left plot in Figure 14

Clarinet $D3$	Guitar $D3$	Clarinet $B4$	Guitar $B4$
0.0290	0.2612	0.7096	0.0003

Table 6: Likelihood of each instrument for right plot in Figure 14

Test 3

Each case for test 3 is performed on two signals, where the first is a mix of a clarinet and a guitar, both playing the tone $B4$ and in the second test, both are playing $D4$.

Case 1

The test results are shown in Figure 15 and in Table 7 and 8.

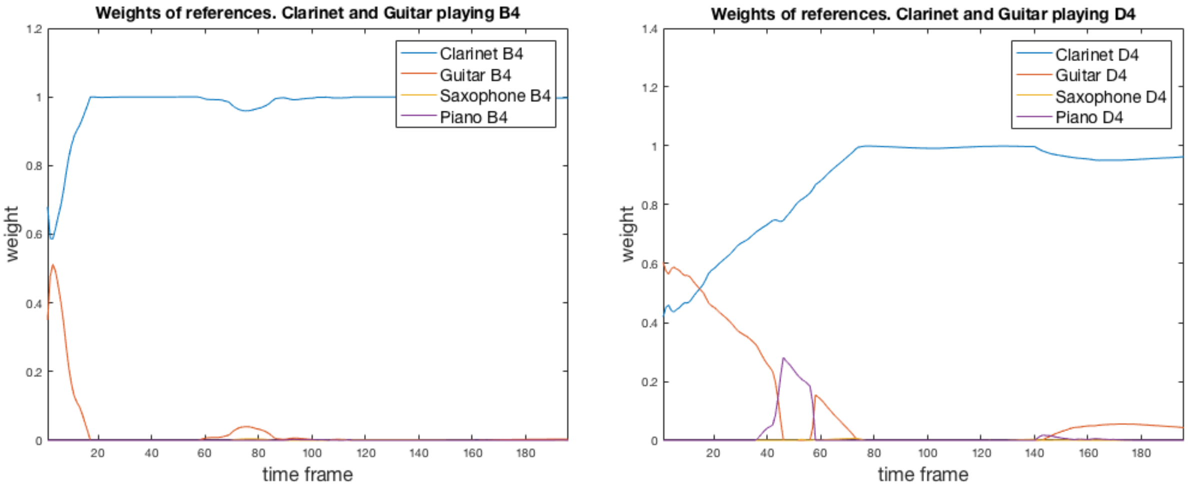

Figure 15: FFT-method. Test 3, case 1

From the plots in Figure 15 and also from Table 7 and 8 it can be concluded that the correct instruments are estimated to be most significant. Even though the contribution from the guitar in the left plot is estimated to be small, it is still a lot bigger than for the saxophone and piano.

Clarinet $B4$	Guitar $B4$	Saxophone $B4$	Piano $B4$
0.9744	0.0249	0.0005	0.0001

Table 7: Likelihood of each instrument for left plot in Figure 15

Clarinet $D4$	Guitar $D4$	Saxophone $D4$	Piano $D4$
0.8675	0.1143	0.0005	0.0177

Table 8: Likelihood of each instrument for right plot in Figure 15

Case 2

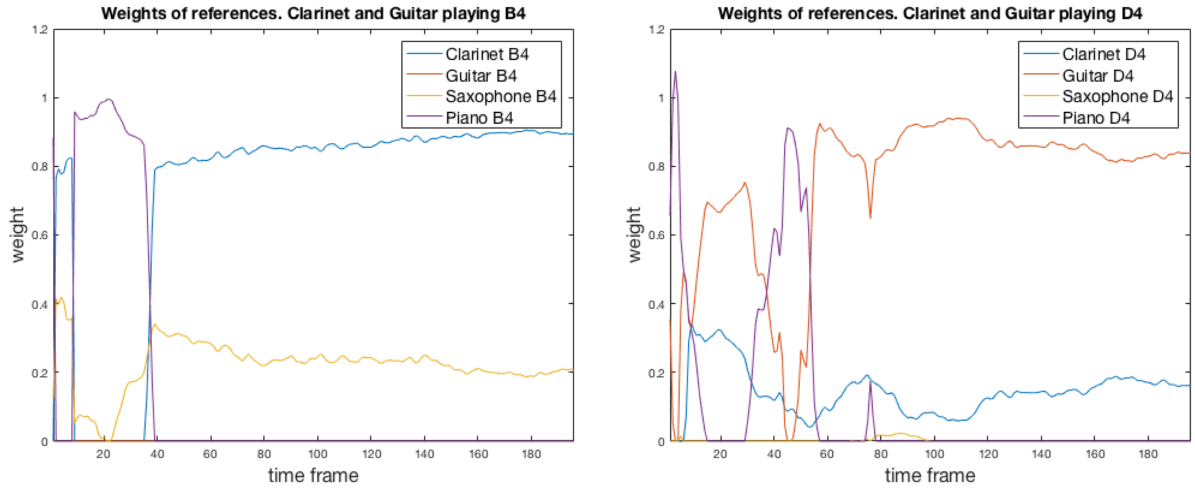

Figure 16: FFT-method. Test 3, case 2

The left plot of Figure 16 shows a correct dominating clarinet but the guitar that is supposed to be present is estimated to be zero as seen in Table 9. The right plot although estimated to hold a large contribution from the piano, shows the correct estimated instruments. This becomes clear when studying Table 10.

Clarinet $B4$	Guitar $B4$	Saxophone $B4$	Piano $B4$
0.6685	0	0.2028	0.1287

Table 9: Likelihood of each instrument for left plot in Figure 16

Clarinet $D4$	Guitar $D4$	Saxophone $D4$	Piano $D4$
0.1678	0.7230	0.0026	0.1065

Table 10: Likelihood of each instrument for right plot in Figure 16

3.2 Tests for MFCC-method

Test 1

Now the same tests that was done for the FFT-method will be done using the MFCC-method. The first test is for one unknown instrument playing one tone.

Case 1

The unknown signal is in this case the same as one of the references.

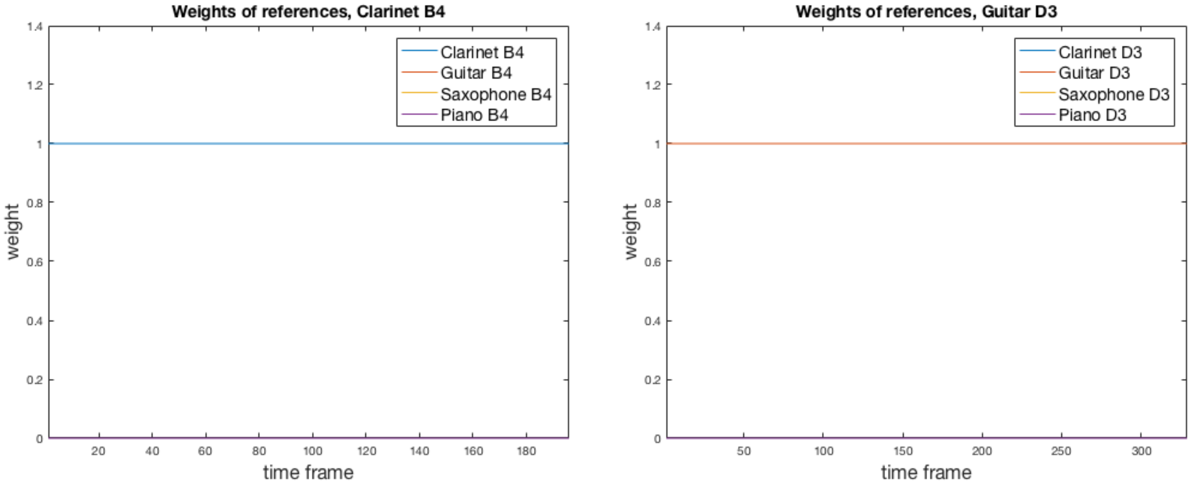

Figure 17: MFCC-method. Test 1, case 1

Since the unknown signals is the same as one of the references the optimization problem of equation (17) is the same except using the MFCC vector **c**.

Case 2

Doing the same test but instead using acoustic signals, different from the references, as unknown signals results in Figure 18.

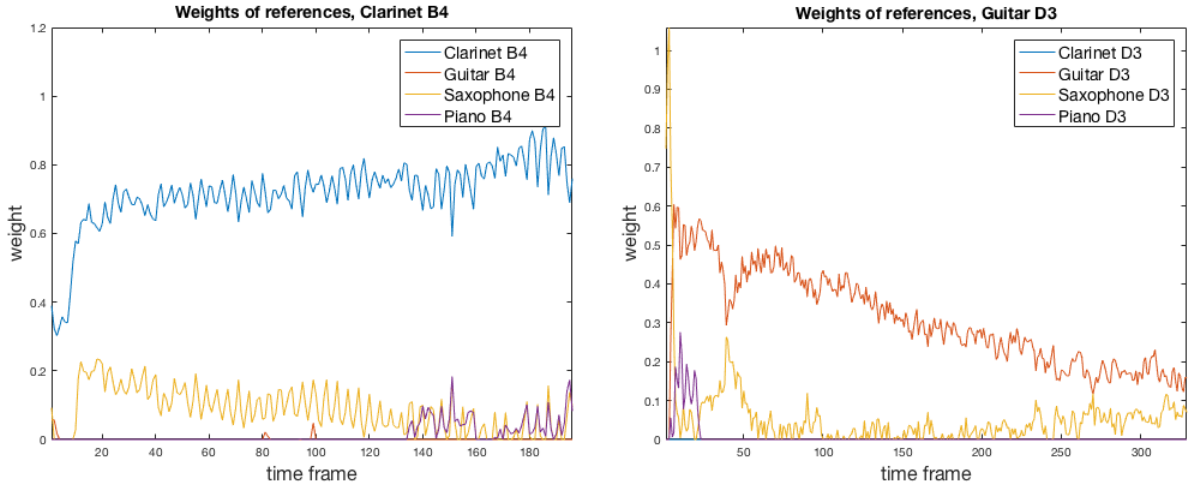

Figure 18: MFCC-method. Test 1, case 2

In the left plot the correct clarinet tone $B4$ is dominating in all time frames. In the right plot the correct guitar tone $D3$ is the most significant instrument for almost all time frames.

Clarinet $B4$	Guitar $B4$	Saxophone $B4$	Piano $B4$
0.8765	0.0013	0.1046	0.0176

Table 11: Likelihood of each instrument for left plot in Figure 18

Clarinet $D3$	Guitar $D3$	Saxophone $D3$	Piano $D3$
0	0.8309	0.1498	0.0193

Table 12: Likelihood of each instrument for right plot in Figure 18

Test 2

In the next test the two instrument used and the tones played are assumed to be known. The MFCC-method will now determine which instrument playing which tone.

Case 1

For the first case the unknown signals are the same as in the references.

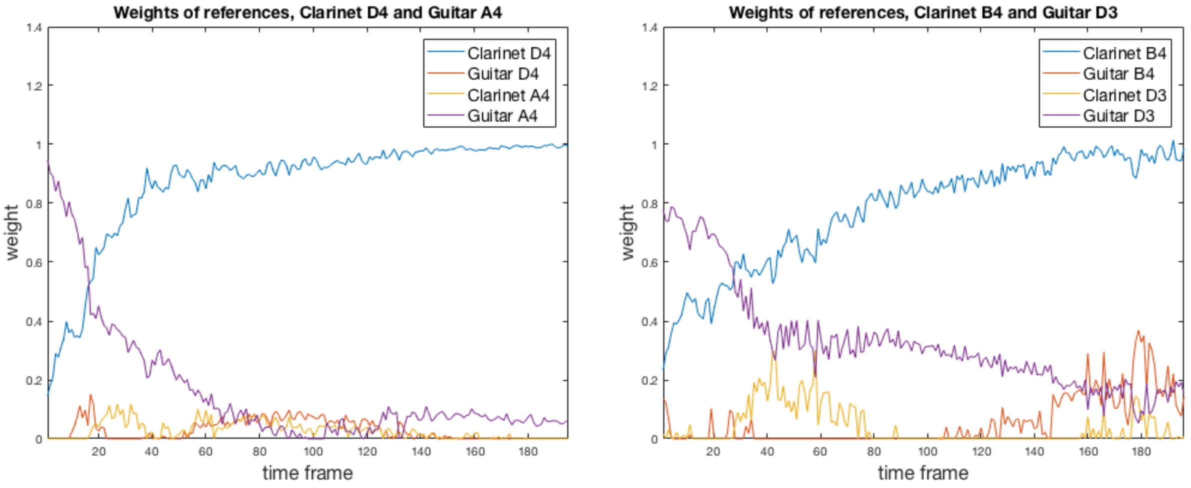

Figure 19: MFCC-method. Test 2, case 1

For the left plot the correct guitar tone $A4$ is dominating the weights for the first 20 frames and then the correct clarinet tone $D4$ is the most dominant. The right plot have a similar look but with a little bit more influence of the wrong guitar and the wrong clarinet.

Clarinet $D4$	Guitar $D4$	Clarinet $A4$	Guitar $A4$
0.7963	0.0265	0.0251	0.1521

Table 13: Likelihood of each instrument for left plot in Figure 19

Clarinet $B4$	Guitar $B4$	Clarinet $D3$	Guitar $D3$
0.6482	0.0466	0.0326	0.2725

Table 14: Likelihood of each instrument for right plot in Figure 19

Case 2

The result from testing for different reference signals and unknown signals is shown in Figure 20.

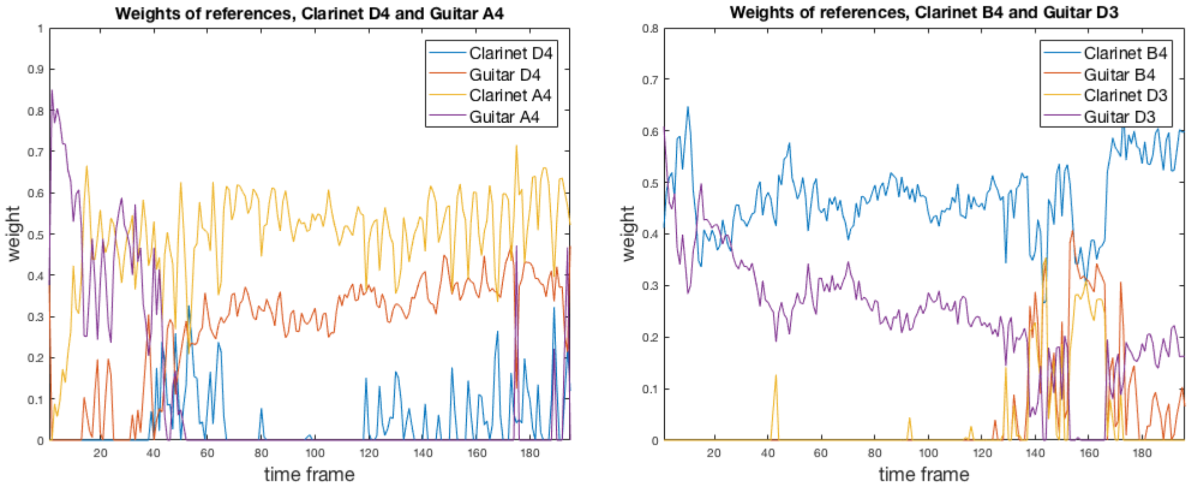

Figure 20: MFCC-method. Test 2, case 2

As seen in the left plot and in Table 15, the opposite tone for both instruments are estimated to be significant. In the right plot and in Table 16 the correct tones are estimated.

Clarinet $D4$	Guitar $D4$	Clarinet $A4$	Guitar $A4$
0.0447	0.2934	0.5381	0.1239

Table 15: Likelihood of each instrument for left plot in Figure 20

Clarinet $B4$	Guitar $B4$	Clarinet $D3$	Guitar $D3$
0.5964	0.0619	0.0419	0.2999

Table 16: Likelihood of each instrument for right plot in Figure 20

Test 3

Just as before, test 3 is about finding the correct instrument in a mix where two instruments are present, playing the same tone.

Case 1

See Figure 21 and Table 17 and 18 for the results of case 1, that is when reference and unknowns are both digital signals.

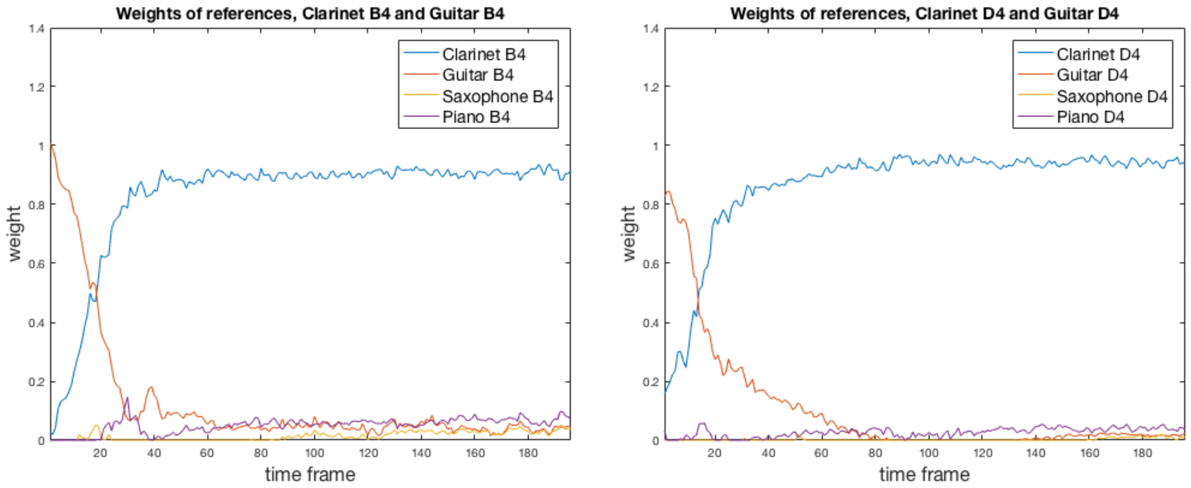

Figure 21: MFCC-method. Test 3, case 1

Studying both plots and the tables, there is in this case no doubt about which instruments that are estimated to be present.

Clarinet $B4$	Guitar $B4$	Saxophone $B4$	Piano $B4$
0.8089	0.1274	0.0141	0.0496

Table 17: Likelihood of each instrument for left plot in Figure 21

Clarinet $D4$	Guitar $D4$	Saxophone $D4$	Piano $D4$
0.8684	0.1031	0.0020	0.0264

Table 18: Likelihood of each instrument for right plot in Figure 21

Case 2

The case when the unknown signals come from acoustic recordings is plotted in Figure 22.

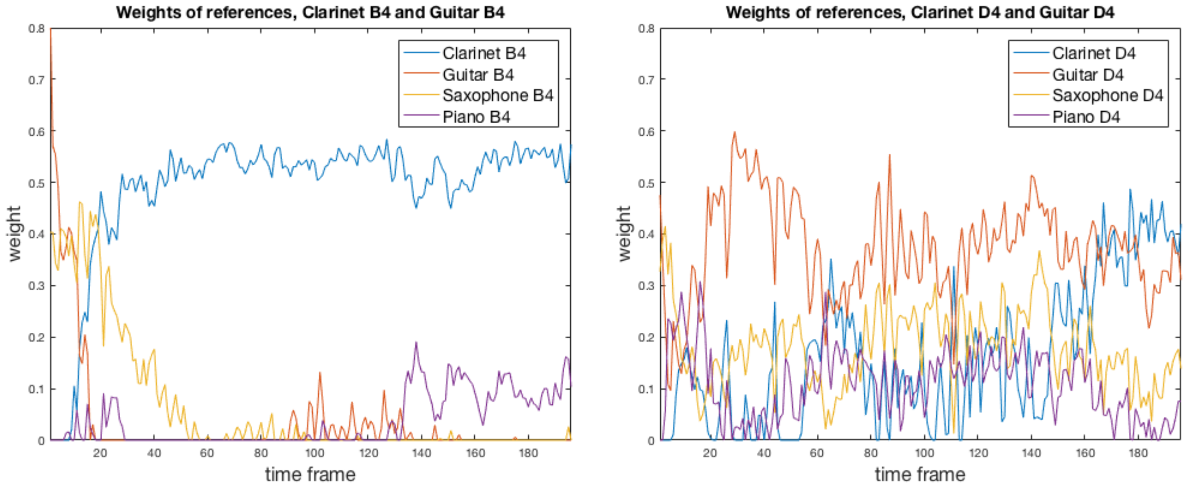

Figure 22: MFCC-method. Test 3, case 2

By Table 19, the instruments present are estimated to be the clarinet (correct) and the saxophone (incorrect). Studying the right plot and Table 20 makes it quite hard to tell the difference between the clarinet and the saxophone. The instrument estimated as the most significant is though the guitar, which is one of the two correct instruments.

Clarinet $B4$	Guitar $B4$	Saxophone $B4$	Piano $B4$
0.7779	0.0585	0.1047	0.0589

Table 19: Likelihood of each instrument for left plot in Figure 22

Clarinet $D4$	Guitar $D4$	Saxophone $D4$	Piano $D4$
0.2003	0.4507	0.2177	0.1312

Table 20: Likelihood of each instrument for right plot in Figure 22

4 Results and analysis

In this section the tests will be commented and analyzed. First the results will be judged only be studying the plots and after that the results will be analyzed on a higher level to see if they are aligned with theory.

Plots of test 1

First the test sections for both the FFT-method and MFCC-method in case 1, Figure 11 and Figure 17, will be commented. For this case the unknown signal and one reference are the same which gives a trivial Least Squares calculation (equation (17)). The reference vector minimizing the error is of course the clarinet vector and the guitar vector respectively. This yields the weights of the corresponding references to be equal to 1. When it comes to the second case Figure 12 and Figure 18 the references are no longer the same as the unknown. Both methods yields clear significant estimations of the correct instrument even though contributions from other instruments are partly present. As mentioned, digital tones and acoustic tones can be quite different. Therefore it might not be optimal having a digital reference when testing on acoustic signals, even though the results becomes pretty good. If the reference signal would have been more similar to the unknown signal, the result would have been even better.

Analyzing test 1

The set up for the test for the FFT-method was

$$\mathbf{p}_l^{tot} = w_1\mathbf{p}_l^{clarinet} + w_2\mathbf{p}_l^{guitar} + w_3\mathbf{p}_l^{saxophone} + w_4\mathbf{p}_l^{piano}, \tag{20}$$

where $\mathbf{p}$ is the normalized amplitude vector for each signal in time frame l. Calculating the magnitude spectrum involves taking the absolute value of the Fourier Transform. Since the number of unknown instruments in this test is known to be just one, taking the absolute value will not disrupt the linearity in equation (20) so the test results are reliable. The MFCC-method involves not only the non-linear operation of taking the absolute value but also it involves taking the logarithm. In particular the logarithm is taken of the summed up energies for each filter bank (see step 4, section 2.2). Equation (20) can be rewritten using $\mathbf{d}$ instead of $\mathbf{p}$. Here $\mathbf{d}$ is the summed up logarithmized energies for each filter bank of the l'th time frame. The DCT is then calculated (step 5, section 2.2) and since the DCT is a linear transformation the linearity is kept. By similar arguments as for the FFT-method and again using the knowledge that there is just one unknown instrument. The conclusion for the MFCC-method is that also here the results are aligned with theory and therefore reliable.

Plots of test 2

Judging from the right plot of the first case, Figure 13 and Figure 19, both FFT-method and MFCC-method estimates very well the correct instruments for the signal when the overtones do not overlap at any frequency in the regarded range. For the left plot in the same figures the result for the the MFCC-method is also good but the FFT-method estimates both the guitar $D4$ and $A4$ to be present. So even when the optimal references are used, the FFT-method fails here for some reason. The MFCC-method on the other hand interestingly succeeds in determining the correct instruments also for the signal in the left plot. In the second case where the reference signals and the unknown signals are different, see Figure 14 and Figure 20, both methods works very well when the overtones does not overlap (see right plot in the figures). Comparing the left plot in Figure 20 with the left plot in Figure 19 the MFCC-method seems to fail only due to the fact that the reference is not optimal. The same comparison for the FFT-method gives that it would have failed even if the reference would have been optimal.

Upon creating the mixes used for the tests, each instrument was set to be equally loud. Therefore one may suspect the weights of each instrument to be equal throughout the signal but this is not always the case. The amplitude of a guitar tone is high in the beginning after which it is decaying. The clarinet is more or less high during the whole tone. The weights of the mix will therefore be somewhat equal, close to the onset. Then the weights of the guitar will decrease and the weights of the clarinet will increase. The behavior shown in the left plot of Figure 19 for both instruments is therefore the one expected. Using the mean as the likelihood measure may therefore underestimate the presence of the guitar.

Analyzing test 2

There is a problem about the idea of adding magnitude spectrums of the references in order to get the magnitude spectrum of an unknown signal. If for example the unknown signal is a combination of a clarinet and a guitar at time frame l, the mixed signal is

$$x_l^{clarinet+guitar} = x_l^{clarinet} + x_l^{guitar} \tag{21}$$

$$\Leftrightarrow X_l^{clarinet+guitar} = X_l^{clarinet} + X_l^{guitar}, \tag{22}$$

where X is the FFT of the signal x for the corresponding instrument. We have equivalence since the Fourier transform is a linear transform. X is now a complex signal which is hard to deal with. To be able to continue doing any further analysis, real values are needed. Therefore the absolute value is taken for each X. By taking the absolute value, the equivalence is lost and by the triangular inequality, it instead becomes

$$\Rightarrow |X_l^{clarinet+guitar}| \leq |X_l^{clarinet}| + |X_l^{guitar}|. \tag{23}$$

This fact will definitely cause problems to this model since the model assume an equality. But, if it would be possible to say that

$$|X_l^{clarinet+guitar}| \approx w_1|X_l^{clarinet}| + w_2|X_l^{guitar}|, \tag{24}$$

then it would be possible to continue as if there was an equality instead of an inequality and just have this generalization in mind as a potential error source. In Figure 23 this generalization is investigated. Here the signals of a clarinet playing $D4$ and a guitar playing $A4$ are added to become a mix of the two instruments. In the left plots the magnitude spectrum for the separated guitar and clarinet is shown. In the right plots the magnitude spectrum of the mix and the magnitude spectrum of the sum of the separated instruments are shown. The two top plots correspond to time frame 20 and the two bottom plots correspond to the time frame 40.

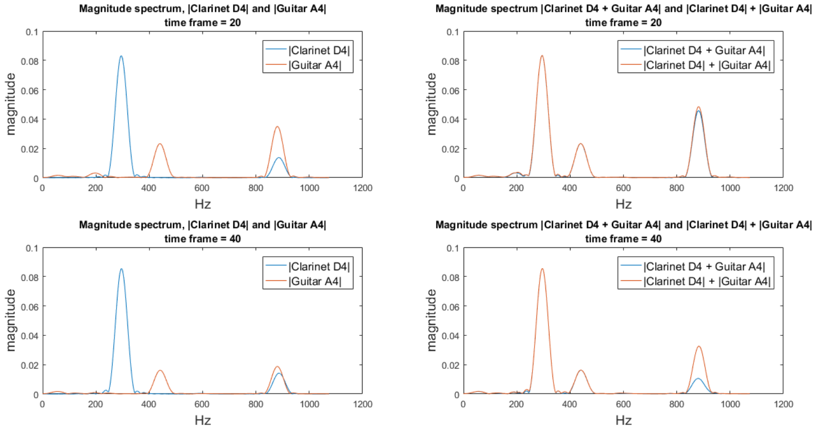

Figure 23: Example of a clarinet playing $D4$ and a guitar playing $A4$. The magnitude spectrum of the mixed signal is plotted together with the magnitude spectrum of the sum of the two separated instruments

Studying the top right plot in Figure 23 one can see that the two graphs almost coincide with each other perfectly. By instead looking at the bottom right plot the graphs coincide almost perfectly except at the peak around 880 Hz where the graphs differ a lot. By analyzing more time frames (not shown in this report) the same behavior occur with coinciding graphs except at 880 Hz. The main reason for this is because the tones $D4$ and $A4$ both have an overtone at 880 Hz which in the mixed signal will be overlapping. If the mixed signal will have the same amplitude as the summed separated signals or not mostly depends on the phase. If the wave-formed separated signals would have their peaks at the same time when they are mixed together, then the sum of the separated absolute signals would be the same. This might be the case in the top right plot. On the other hand if the signals are out of phase when they are mixed, meaning that one signal have a peak at the same time as the other signal has a negative peak, then the amplitude of the mixed signal will be zero. This might partly be the case in the bottom right plot. Due to small differences in the tune, the phase can change between different time frames making it almost impossible to know the amplitude at an overlapping overtone. In the above example it would have been possible to find the phase-shift by calculating the cross-correlation between the signals in the mix. This is not possible in a realistic example though, since then the signals in the mix would be unknown.

In the top right plot of Figure 23 the graphs almost coincide also at 880 Hz. This appearance was also seen in some other time frames. One reason why the graphs sometimes coincide at 880 Hz and sometimes not could be that the instruments may be close to correctly tuned but not perfectly. An other reason could be that noise disturb the signal, and therefore resolution from the FFT won't be perfect. An other factor that could affect is that a tone wobble a bit around the correct tone, this phenomenon have especially been seen for string instruments. The fact that the equality in equation (24) holds for non-overlapping frequencies but not for overlapping, explains why the FFT-method might succeed more often for mixed signals when there are few overlapping overtones and the method might be less robust for mixed signals with a lot of overlapping overtones.

Considering the MFCC of a signal consisting of more than one instrument is a bit different compared to the FFT-method. In the FFT-method the non-linear operation involved was taking the absolute value. The MFCC-method will involve the non-linear operations of taking the absolute value and the logarithm.

A way to explain how this non-linearity would affect the calculation a mixed signal consisting of a clarinet and a guitar is considered for one time frame l

$$x_l^{clarinet+guitar} = x_l^{clarinet} + x_l^{guitar} \tag{25}$$

$$\Leftrightarrow X_l^{clarinet+guitar} = X_l^{clarinet} + X_l^{guitar}, \tag{26}$$

where X is the Fourier transform of x. The magnitude spectrum is calculated by taking the absolute value. Due to that the absolute value is a non-linear operation the equality disappears and instead there is an inequality

$$\Rightarrow |X_l^{clarinet+guitar}| \leq |X_l^{clarinet}| + |X_l^{guitar}|. \tag{27}$$

In the above analysis regarding the FFT-method it was in some sense shown that this can be rewritten as

$$\Leftrightarrow |X_l^{clarinet+guitar}| \approx v_1|X_l^{clarinet}| + v_2|X_l^{guitar}|, \tag{28}$$

where v is the weight of each magnitude spectrum. The magnitude spectrum is denoted as S and S_l is a row vector.

$$\Leftrightarrow S_l^{clarinet+guitar} \approx v_1 S_l^{clarinet} + v_2 S_l^{guitar}. \tag{29}$$

The magnitude spectrum is then multiplied with a filter bank where all the energy of each filter is summed, resulting in

$$\Leftrightarrow S_l^{clarinet+guitar} \cdot F \approx v_1 S_l^{clarinet} \cdot F + v_2 S_l^{guitar} \cdot F. \tag{30}$$

The next step in the MFCC algorithm is taking the logarithm. Since the logarithm is a non-linear operation there is an assumption made that the approximation will hold if the references are weighted once again

$$\begin{aligned} \Rightarrow log(S_l^{clarinet+guitar} \cdot F) \approx\ & w_1 log(v_1 S_l^{clarinet} \cdot F) \\ & + w_2 log(v_2 S_l^{guitar} \cdot F), \end{aligned} \tag{31}$$

where w is a weight, similarly as for the absolute value, calculated by Least Squares. By using logarithm rules this can be written as

$$\begin{aligned} \Leftrightarrow log(S_l^{clarinet+guitar} \cdot F) \approx\ & w_1 log(v_1) + w_1 log(S_l^{clarinet} \cdot F) \\ & + w_2 log(v_2) + w_2 log(S_l^{guitar} \cdot F) \end{aligned} \tag{32}$$

$$\begin{aligned}\Leftrightarrow d_l^{clarinet+guitar} \approx\ & w_1 log(v_1) + w_1 d_l^{clarinet} \\ & + w_2 log(v_2) + w_2 d_l^{guitar} \end{aligned}, \tag{33}$$

where $d_l = log(S_l \cdot F)$. The expression $wlog(v)$ becomes an error that is made. If the reference is good then v should be close to 1. Since $log(1) = 0$ the error can be assumed to be small. The formula can therefore be rewritten as

$$\Rightarrow D_l^{clarinet+guitar} \approx w_1 D_l^{clarinet} + w_2 D_l^{guitar}. \tag{34}$$

The final step in the MFCC is taking the DCT which is a linear transform.

How the weighting idea for the MFCC-method works will now be investigated. A clarinet playing $D4$ and a guitar playing $A4$ are combined to a mix and the separated signals are used as references. In equation (31) the assumption is made that the filter bank energy in the mixed signal can be described as a linear combination of the separated signals. The filter bank energy for the mixed and the separated signals for one time frame is shown in Figure 24.

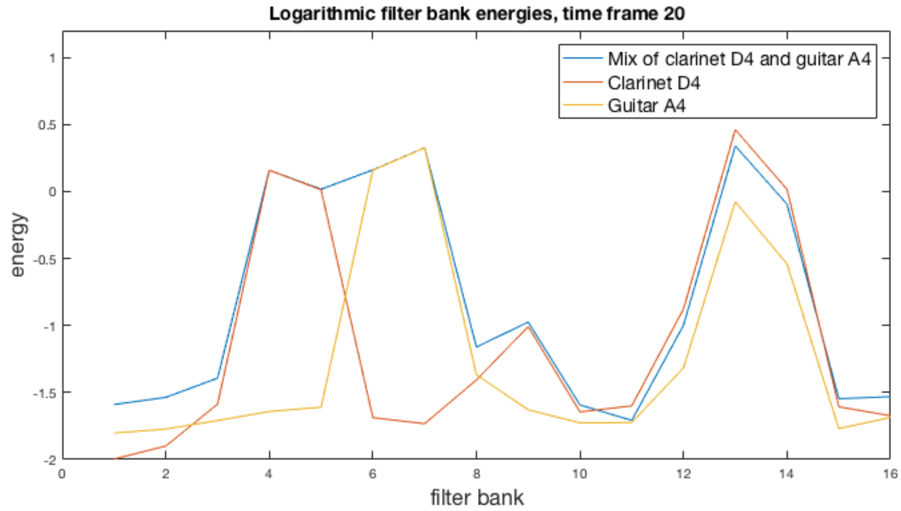

Figure 24: Filter bank energies for clarinet $D4$, guitar $A4$ and a mix of both in the time frame 20

As seen in Figure 24 the energy of the mixed signal follows the energy of the first three peaks from the separated signals well which is a similar result as have been seen also for the FFT-method. For the forth peak (filter bank 13), which corresponds to an overlapping overtone for the separated signals, the mixed signal is almost as the highest peak. If the MFCC-calculation would have been perfectly linear then the fourth peak of the mixed signal would have had the height of the sum of the clarinet and the guitar. By studying more time frames (which will not be shown in this report) the intuition is that the mixed signal follows the highest peak for each filter bank of the separated signals.

Continuing in Figure 24 the y-axis is in the logarithmic scale. Since the logarithm is non-linear it is not possible to take the sum of the filter bank energies of the clarinet and the guitar in order to get the energy of the mix. By for example looking at filter bank 4, if the weights w would have been 1 for both the clarinet and the guitar, then the approximated mixed energy would have been $\approx 1 \cdot 0.2 + 1 \cdot (-1.7) = -1.5$ which is a really bad estimate of the mixed peak which should be 0.2. The best estimate would instead be setting $w_1 = 1$ and $w_2 = 0$. But for example in filter bank 7 the best estimate would have been achieved setting $w_1 = 0$ and $w_2 = 1$. By using Least Squares for the whole time frame the weights are set to $w_1 = 0.4$ and $w_2 = 0.7$. Studying more time frames the weights seems to become some kind of mean of the separated signals since both signals influence the mixed signal at different filter banks. It can be concluded that this weighting does not give a very good estimation of the energy of the mixed signal.

The Least Squares calculate the weights of each instrument that yields the smallest error from the unknown signal. Even if the error becomes large by using a combination of the correct references the error might still be the smallest compared to using wrong references. This could be the reason why the MFCC-method succeeds even if the theory says that it shouldn't.

Plots of test 3

In all of the four tests for case 1, seen in Figure 15 and Figure 21 the methods succeeds. The guitar in left plot for the FFT-method is maybe estimated a bit low but this is the only minor detail. When it comes to case 2 (see Figure 16 and Figure 22) the analysis becomes a bit harder for all four tests. Starting with the right plots and the corresponding Table 10 and 20, it might seem that the estimation of the guitar succeeds quite well for both methods. This is probably not correct though, as the typical pattern of the guitar tone should be decaying and look more like the one in Figure 21. This makes the results doubtful. The results in the corresponding left plots is a bit better. Here the estimated clarinet follows the overall pattern a clarinet tone is supposed to have but the guitar is for the FFT-method non-existing and for the MFCC-method estimated to be non-significant.

Analyzing test 3

The analysis of test 3 is similar to the one done for test 2. The difference is that for all signals in test 3, all the overtones overlap. The analysis for test 2 concluded that problems with non-linearity almost only occur when the overtones were overlapping. With this into consideration the results from test 3 would be expected to be worse than for test 2. That was also the the way it turned out for case 2 where no results was totally correct. For test 3 case 1 all the tests succeeded which is a little bit confusing since one test for the FFT-method test 2 case 1 failed. Test 2 are supposed to be better than test 3 but that seems to not always be the case. To be able to do any further analysis about this, more tests would have been needed to insure that this was not a random outcome.

By studying both plots in Figure 16 and the left plot in Figure 22 the pattern of the weights for both the correct and the incorrect instruments are shifting a lot for the first 60 time frames. After that the weights becomes more stationary. One reason for this might be that, right at the beginning of a tone when the tone is struck, the signal changes a lot and the variance is high. The signal therefore becomes more non-stationary for these time frames and first after a while the tone becomes more stationary. The FFT needs stationary signals to calculate appropriate spectrums. Therefore the varying weights for the early time frames in the figures might come from incorrect spectrum estimates.

5 Discussion and future improvements

In this project, characteristics for different musical instruments have been described mathematically by analyzing the energy and structure of pitches and overtones. Two methods using different approaches, the first using magnitude spectrum analysis and the second using cepstrum analysis have been proposed and tested. The test results, as seen, have been varying in reliability and accuracy leaving room for future improvements and modifications. For the aspect of only finding one unknown instrument playing one tone, both methods works fine. When dealing with signals consisting of two instruments the analysis becomes harder, especially when overtones overlap. None of the methods have been working perfectly for these signals and none of the methods have been far better than the other.

The FFT-method is very similar to the MFCC-method in many ways but there are some differences. One is that the FFT-method is using the magnitude spectrum and the MFCC-method is, in some sense, using the logarithmized magnitude spectrum. The logarithmized magnitude spectrum is good to use if smaller peaks, which often occur at higher frequencies, are of interest. If only the first overtones are of interest then the non-logarithmized spectrum would be better. It is not investigated how many overtones are needed to categorize an instrument in the best way. Depending on how many overtones that are needed to do a good categorization, it should be decided whether to use logarithmized or non-logarithmized magnitude spectrum.

One main advantage with the FFT-method is that it only uses information in the spectrum that belongs to certain wanted peaks. The MFCC-method uses all information in the spectrum, even the space between these peaks, which is just unnecessary. When comparing the unknown and the reference signals for the MFCC-method, both the error at the peaks and the space between the peaks are minimized. It would be much better to minimize the error at the peaks only.

On the other hand one advantage with the MFCC-method is that it uses filter banks. By applying filter banks, information not only at the peaks, but also around the peak is retrieved. This is good, since mixing two tones with overlapping overtones will far from always result in a linear addition of the overlapping peaks but instead the power may be distributed around the maximum of the peak.

If the tones are known, the filter banks used in the MFCC-method could be optimized by taking into account features of the FFT-method. Firstly, the filters in the filter bank can be placed exactly where the pitches and overtones are located, resulting in fewer filters. By this approach the energy between the peaks is ignored. If the tones and overtones are well separated the need of overlapping filters may also be eliminated. Secondly, the limits of the frequency range for the filter banks could using known pitches be set in an optimal way.

As seen in case 1 for almost all tests, having a perfect reference will give a correct weight estimation. Extending the amount of reference signals would improve the instrument categorization a lot. Every acoustic instrument have in some way a sound of it's own. By having many references of the same instrument category the chance that an unknown instrument would be similar to one of the references increases.

When the instruments are assumed known and it comes to determine which instrument playing which tone (test 2) one possible improvement could be to add up the reference signals for all possible combinations before doing the weight calculation. This reduces the problem to be of the form in test 1 where non-linear operations does not corrupt the result, that is

$$\begin{aligned}\mathbf{p}_l^{tot} &= w_1\left(\mathbf{p}_{l,1}^{clarinet}+\mathbf{p}_{l,2}^{clarinet}\right)+w_2\left(\mathbf{p}_{l,1}^{clarinet}+\mathbf{p}_{l,2}^{guitar}\right)\\ &+w_3\left(\mathbf{p}_{l,1}^{guitar}+\mathbf{p}_{l,2}^{clarinet}\right)+w_4\left(\mathbf{p}_{l,1}^{guitar}+\mathbf{p}_{l,2}^{guitar}\right)\end{aligned}, \tag{35}$$

where term one and four in the right-hand side disappears if the instruments are known to be different. This will hold also for the MFCC-method, where in equation (35), **p** is replaced by **c**. Of course, as the amount of instruments and tones increases the number of combinations in the right hand side also increases, making this approach ineffective.

In order to determine if a reference instrument is significantly present in an unknown signal or not, the mean has been used. This is probably not an optimal measure since the weights are sometimes expected to vary over time. For example the guitar use to have a strong contribution to a mixed signal in the beginning and then decaying to be almost zero. An optimal measure about significance level would in some way take into account how the weights of the references are expected to vary over time.

The idea about using Least Squares and calculate the weight for every instrument in the reference bank is good if no assumptions of the instruments in the unknown signal can be done. But if it is known that the unknown signal only consists of one instrument, then it is a bad idea to combine four references trying to match the unknown signal. A better idea would instead be to rescale each reference with Least Squares to match the unknown signal and then calculate the mean square error (MSE). The reference with the smallest MSE would then be the best guess. For the case when the unknown signal consists of two instruments they would be weighted with Least Squares and then added together before the MSE is calculated.

If the tones are different and the instruments are known, another possible improvement for the FFT-method could be to apply an iterative process starting from the lowest fundamental frequency in the spectrum of the unknown signal. By analyzing and comparing each of the peaks one by one and see how the amplitude changes over time it might be possible to within just a narrow range of peaks be able to determine one or more of the instruments. This is illustrated in Figure 25. Here the fundamental tone and a few of the overtones from a guitar and a clarinet is plotted. As seen, there is a big difference in how the pitch is evolving over time between the instruments and this can be used in some cases.

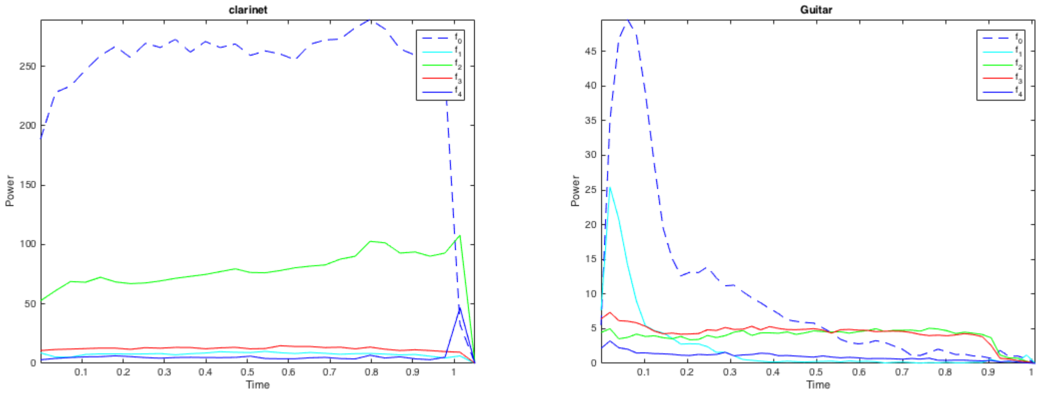

Figure 25: Overtones and how they change over time

6 Bibliography

[1] https://en.wikipedia.org/wiki/Timbre

[2] C.L. Lawson, R.J. Hanson "Solving Least Squares Problem" Prentice Hall, 1974, p. 158-161

[3] https://engineering.purdue.edu/ malcolm/interval/1998-010/

[4] http://practicalcryptography.com/miscellaneous/machine-learning/guide mel-frequency-cepstral-coefficients-mfccs/

[5] S.K. Kopparapu, M. Laxminarayana, "Choice of mel filter bank in computing MFCC of a resampled speech" *10th International Conference on Information Science, Signal Processing and their Applications* p. 121-122, 2010

[6] http://www.apple.com/se/mac/garageband/

[7] http://www.audacityteam.org

[8] G. Lindgren, H. Rootzén, M. Sandsten "Stationary Stochastic Processes For Scientists And Engineers" CRC Press p. 129, 2014

Bachelor's Theses in Mathematical Sciences 2016:K6
ISSN 1654-6229

LUNFMS-4016-2016

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/