KTH - Royal Institute of Technology

Pattern Recognition (EQ2340) - Exercise Project
A.2 - Feature extraction - Speech Recognition System

Alessio Russo - alessior@kth.se (911103-T192)
Lars Lindemann - llindem@kth.se (891113-4131)

October 2015

Contents

1 Time analysis **2**

2 Spectrograms **4**

3 Comparison: Spectrogram and Cepstrogram **5**

4 Correlation Analysis **9**

5 Dynamic Features **10**

6 Additional thoughts **11**

1 Time analysis

First our signals are analysed in the time domain, where it is possible to see some patterns in the signals. The code used to plot the time series is the following one:

```
[y_female,fs_female]    = audioread('female.wav');
[y_male,fs_male]        = audioread('male.wav');
[y_music,fs_music]      = audioread('music.wav');

%time domain vectors, scaled with the corresponding time period.
t_female                = 1/fs_female*(0:length(y_female)-1)';
t_male                  = 1/fs_male*(0:length(y_male)-1)';
t_music                 = 1/fs_music*(0:length(y_music)-1)';
%%
%plot of the music signal and female speech in time
figure(1)
subplot(2,1,1); plot(t_female,y_female);
ylabel('Signal amplitude'); xlabel('Time in seconds');grid;
title('Female voice signal');
subplot(2,1,2); plot(t_music,y_music);
ylabel('Signal amplitude'); xlabel('Time in seconds'); grid;
title('Music signal');

%plot of the music signal zoomed from 0.62 to 0.76 seconds
figure;
indexes_music = t_music > 0.62 & t_music < 0.76;
plot(t_music(indexes_music),y_music(indexes_music));
ylabel('Signal amplitude'); xlabel('Time in seconds'); grid;
title('Music signal - Armonics pattern');
%plot of the female speech zoomed with voiced/unvoiced patterns

%unvoiced pattern,corresponds to s ( i shot...)
indexes_female_unvoiced = t_female > 0.17 & t_female < 0.26;
%voiced pattern, corresponds to i ( i ...)
indexes_female_voiced = t_female > 0 & t_female < 0.9;
figure;
subplot(211);
plot(t_female(indexes_female_unvoiced),y_female(indexes_female_unvoiced));
ylabel('Signal amplitude'); xlabel('Time in seconds');grid;
title('Female voice signal - Unvoiced pattern');
subplot(212);
plot(t_female(indexes_female_voiced),y_female(indexes_female_voiced));
ylabel('Signal amplitude'); xlabel('Time in seconds');grid;
title('Female voice signal - voiced pattern');
```

The results are shown in the following figures: 1,2,3 .
At first glance, without zooming in, both signal do not have significant patterns, but it is clear to see that the music signal is more uniform than the speech signal.
This is so because of the harmonic structure of the music signal: it consists of several harmonics, which are an integer multiple of the fundamental frequency. The same can be said for some sounds in the human speech: letters like a,i,v,b, etc... are *voiced* sounds, which means that those are sounds similar to the one made by musical instruments.
On the other hand, letters like s,c, f are *unvoiced* sounds. The main difference between *voiced* and *unvoiced* letters is that in the latter the person does not make any vocal cord vibrate: because of that we don't have an harmonic struc-

ture in *unvoiced* sounds, being then a more noisy signal (it is only air passing through the larynx).

The music signal, if we zoom in a smaller time patch, has a clear harmonic structure (it is easy to see the sinusoidal pattern), and this is shown in figure: 2.

For the female speech we can look for example at the start of the signal, where we have both a voiced and an unvoiced signal: *i* and *s*. In figure 3 we can see in the top graph the time series of the unvoiced sound, which starts after 0.1 seconds, and corresponds to the letter *s*. In the bottom graph we have the voiced sound, which is the first letter being pronounced by the woman, *i*.
Notice how the unvoiced signal just looks like random noise, whilst the voiced signal has a clear harmonic structure.

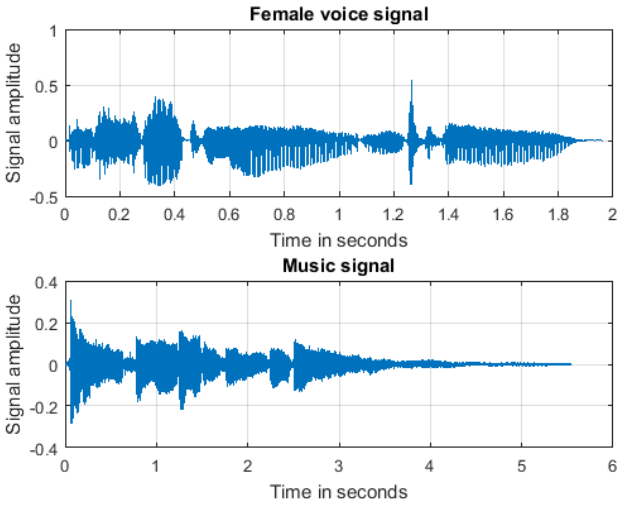

Figure 1: Time series of the female speech and music signal

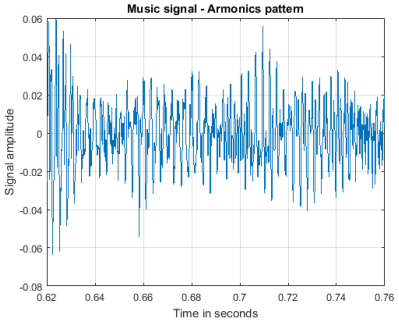

Figure 2: Harmonic structure of the music signal

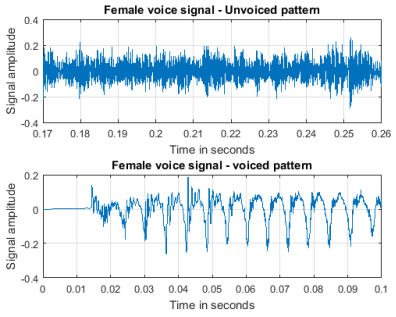

Figure 3: Plot of unvoiced and voiced sounds of the female speech

2 Spectrograms

The code, that has been used in order to investigate music and female speech signal, is the following:

```
window_length = 0.03; %in ms
num_cepstral  = 13;

[mfccs_f,spectgram_f,f_f,t_f]       = GetSpeechFeatures(y_female,...
                                fs_female,window_length,num_cepstral);
[mfccs_mu,spectgram_mu,f_mu,t_mu]   = GetSpeechFeatures(y_music,...
                                fs_music,window_length,num_cepstral);
[mfccs_ma,spectgram_ma,f_ma,t_ma]   = GetSpeechFeatures(y_male,...
                                fs_male,window_length,num_cepstral);


figure(2)
imagesc(t_f,f_f,log10(spectgram_f)); colormap jet
xlabel('Time in seconds'); ylabel('Frequency in Hz'); title('Female speech');
annotation('textarrow',[0.2 0.25],[0.85 0.9],'String','Voiced');
annotation('textarrow',[0.15 0.2],[0.65 0.7],'String','Unvoiced');
figure(3)
imagesc(t_mu,f_mu,log10(spectgram_mu)); colormap jet
xlabel('Time in seconds'); ylabel('Frequency in Hz'); title('Music signal');
annotation('textarrow',[0.2 0.25],[0.83 0.88],'String','Harmonics');
```

The results are shown in figure 4. Within the music signal, the harmonics are clearly visible as well as the fact, that the fundamental frequency is changing over time. The same characteristics as in the music signal can be found for voiced sounds within the female speech and is marked by an arrow. The spoken sentence is *I shot an arrow into the air* - the first part of shot, namely *sh*, is unvoiced (feel the larynxgeal prominence). This unvoiced sound occurs after approximately 0.1 seconds and lasts for nearly 0.15 seconds and are characterised by a broad spread over the frequency range. Two other occurrences at 0.45 (the *t* within sho**t**) and 1.25 (the *t* within in**t**o) the seconds are unvoiced. As already discussed, this is also visible within the time domain signal.

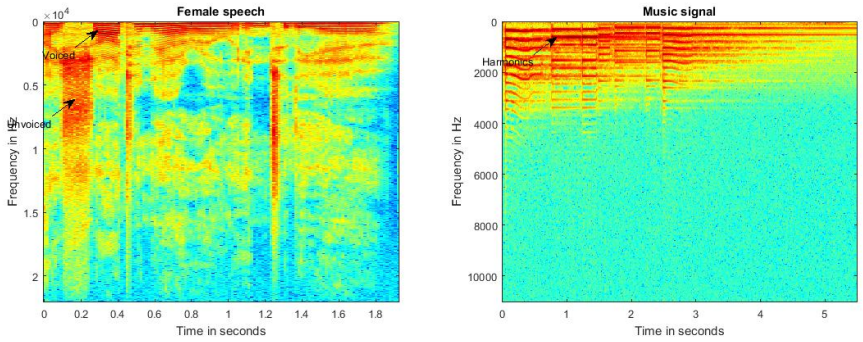

Figure 4: Spectrogram of femal speech (on the left) and spectrogram of music signal (on the right)

3 Comparison: Spectrogram and Cepstrogram

The comparison of spectrogram and cepstrogram has been performed for the combination of female speech versus music and for female speech versus male speech. The code is given by:

```
window_length = 0.03; %in ms
num_cepstral  = 13;

[mfccs_f,spectgram_f,f_f,t_f]       = GetSpeechFeatures(y_female,...
                                fs_female,window_length,num_cepstral);
[mfccs_mu,spectgram_mu,f_mu,t_mu]   = GetSpeechFeatures(y_music,...
                                fs_music,window_length,num_cepstral);
[mfccs_ma,spectgram_ma,f_ma,t_ma]   = GetSpeechFeatures(y_male,...
                                fs_male,window_length,num_cepstral);


% mean the signal
mfccs_f = mfccs_f-repmat(mean(mfccs_f,2),1,size(mfccs_f,2));
mfccs_mu = mfccs_mu-repmat(mean(mfccs_mu,2),1,size(mfccs_mu,2));
mfccs_ma = mfccs_ma-repmat(mean(mfccs_ma,2),1,size(mfccs_ma,2));
% normalize variance to 1
norm_f   = (std(mfccs_f.')).';
mfccs_fn = mfccs_f.*repmat((1./norm_f),1,size(mfccs_f,2));
norm_m   = (std(mfccs_mu.')).';
mfccs_mn = mfccs_mu.*repmat((1./norm_m),1,size(mfccs_mu,2));
norm_ma   = (std(mfccs_ma.')).';
mfccs_man = mfccs_ma.*repmat((1./norm_ma),1,size(mfccs_ma,2));

figure(7)
subplot(2,1,1); imagesc(t_f,f_f,log10(spectgram_f)); colormap jet
xlabel('Time in seconds'); ylabel('Frequency in Hz'); title('Female voice - Spectrogram')
subplot(2,1,2); imagesc(t_mu,f_mu,log10(spectgram_mu)); colormap jet;
xlabel('Time in seconds'); ylabel('Frequency in Hz'); title('Music - Spectrogram')

figure(8)
subplot(2,1,1); imagesc(t_f,1:num_cepstral,mfccs_fn); colormap jet
xlabel('Time in seconds'); ylabel('MFCCS'); title('Female voice - Cepstrogram')
subplot(2,1,2); imagesc(t_mu,1:num_cepstral,mfccs_mn); colormap jet
xlabel('Time in seconds'); ylabel('MFCCS'); title('Music - Cepstrogram')

figure(9)
subplot(2,1,1); imagesc(t_f,f_f,log10(spectgram_f)); colormap jet
xlabel('Time in seconds'); ylabel('Frequency in Hz'); title('Female voice - Spectrogram')
subplot(2,1,2); imagesc(t_ma,f_ma,log10(spectgram_ma)); colormap jet;
xlabel('Time in seconds'); ylabel('Frequency in Hz'); title('Male voice - Spectrogram')

figure(10)
subplot(2,1,1); imagesc(t_f,1:num_cepstral,mfccs_fn); colormap jet
xlabel('Time in seconds'); ylabel('MFCCS'); title('Female voice - Cepstrogram')
subplot(2,1,2); imagesc(t_ma,1:num_cepstral,mfccs_man); colormap jet
xlabel('Time in seconds'); ylabel('MFCCS'); title('Male voice - Cepstrogram')
```

The results for the music and female speech comparison are shown in figure 5 and 6. For a human, the spectrogram representation is easier to interpret since the all-time occurring harmonics within the music signal are very characteristic. Also, the distinction between voiced and unvoiced sounds in the speech signal are obvious. As a human, you also have kind of a physical notion for these real physical quantities that relate to certain frequencies. On the other hand,

the cepstrogram is low dimensional in the sense that the amount of data has been reduced significantly, which makes it easier to handle the data within a computer. But for a human, the understanding of the pattern are much harder here. The normalised cepstrogram seems to be random on a first glance. It is hard to distinguish between voiced and unvoiced sound by looking.

The other comparison is shown in figure 7 and 8. The spectrograms are similar for a human since they have kind of the same behaviour. Voiced and unvoiced parts are visible, but they are a bit blurred within the male signal. The first and third unvoiced parts are visible within the male signal, whereas these characteristic is not visible for the second unvoiced part. Also, the harmonics are more or less blurred. Hence, there are differences between the signals, but the similarities are visiblie for a human (this goes along with the fact that we know that it represents the same phrase beforehand). Due to the high amount of data and also the difference in output power and pitch between those signals, it could be hard for a computer to make this statement. The cepstrogram comparison seems random at first, but having a deeper look into it, it becomes clear that the cepstrograms share some properties, which can be handled easily by the computer. The pitch has been removed such that the harmonics don't differ anymore and also the general behaviour when comparing the coefficients are similar. For instance the third coefficient has high (red) values around 1.4 seconds and low (blue) values around 0.4 seconds for both male and female signal. In general it seems like the computer has an advantage here compared to a human to interpret these pictures.

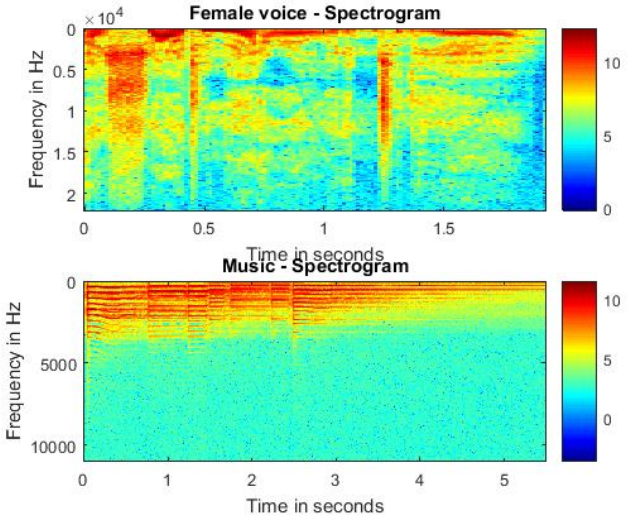

Figure 5: Spectrogram comparison: female speech vs. music signal

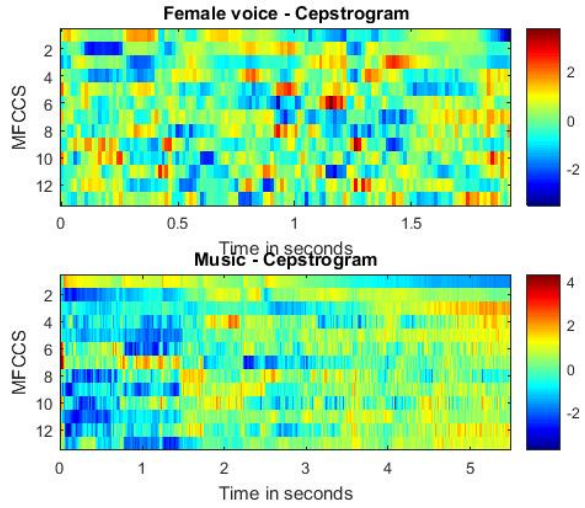

Figure 6: Cepstrogram comparison: female speech vs. music signal

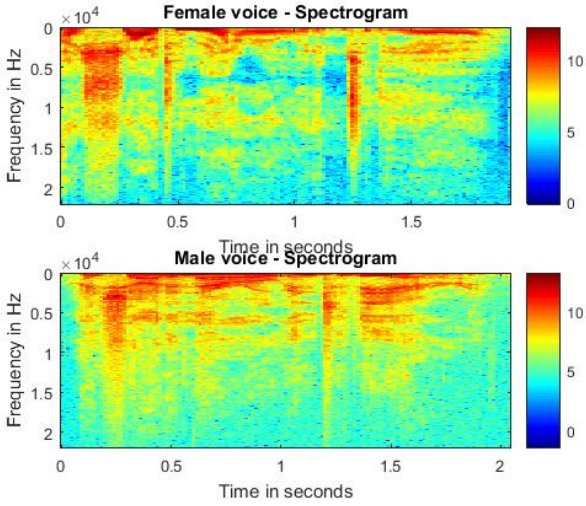

Figure 7: Spectrogram comparison: female speech vs. male speech

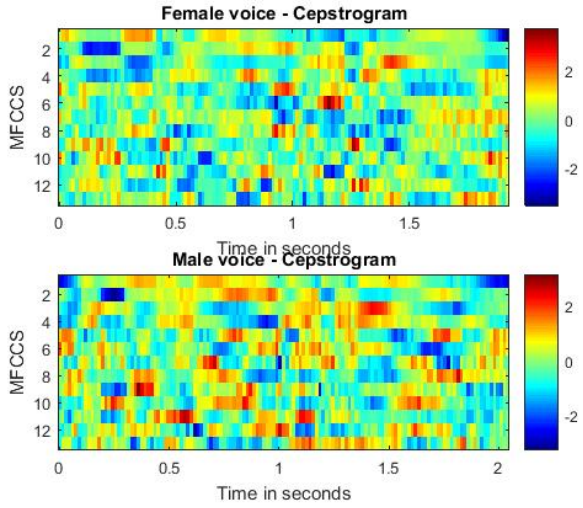

Figure 8: Cepstrogram comparison: female speech vs. male speech

4 Correlation Analysis

A correlation analysis between the spectrogram and the cepstrogram coefficients has been performed by the following code

```
frame = 100;

[c_spectra,rho1] = corr(log10(spectgram_f(:,1:frame)).');
[c_cepstra,rho2] = corr(mfccs_f(:,1:frame).');
figure(11)
subplot(1,2,1); imagesc(abs(c_spectra)); title('Correlation - Female voice');...
    xlabel('Frequency bins'); ylabel('Frequency bins');
subplot(1,2,2); imagesc(abs(c_cepstra)); title('Correlation - Female voice');...
    xlabel('Cepstral Coefficient'); ylabel('Cepstral Coefficient');
```

Figure 9 shows the absolute values and it is visible, that the MFCC coefficients are less correlated since the matrix is more diagonal (the area outside the diagonal is much darker in the cepstogram, which means weaker intensity). While elements adjacent to the diagonal elements have a correlation value between approximately 0 and 0.4, the spectral correlation have mostly a much higher correlation, especially for the frequency bins 150 to 400. Figure 10 shows the same picture with the *gray colormap*, which supports the previous statement.

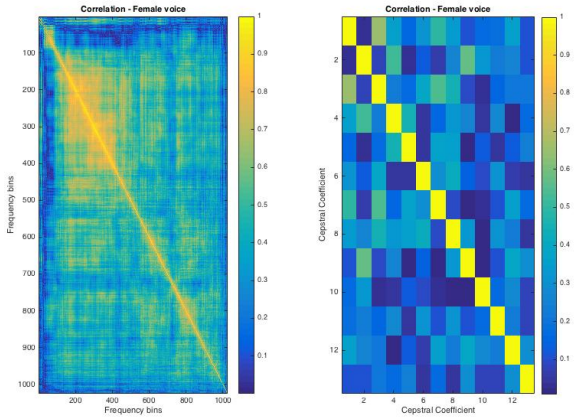

Figure 9: Coefficient correlation analysis

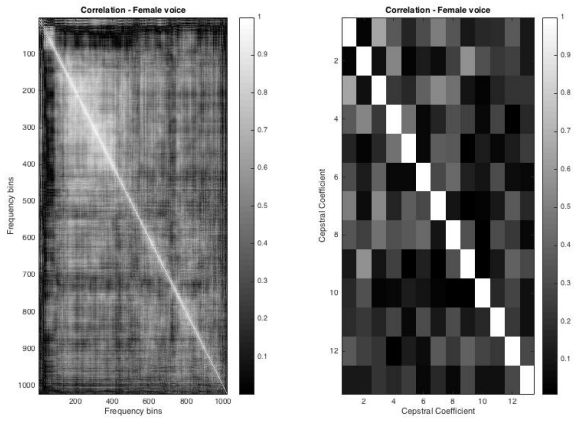

Figure 10: Coefficient correlation analysis (gray colormap)

5 Dynamic Features

In order to get better classification results, dynamic features have been implemented by a simple function using the Matlab *diff* command. This functions allows it to get information about the relative differences between the sampling instances.

```
function [MFCC,f,t] = mfcc_dyn(signal,fs,window_length,num_cepstral)

% Calculate Cepstral
[mfccs,~,f,t]       = GetSpeechFeatures(signal,...
                                fs,window_length,num_cepstral);
% Make the signal have 0 mean and unit variance
mfccs       = mfccs-repmat(mean(mfccs,2),1,size(mfccs,2));
norm_f      = (std(mfccs.')).';
MFCC        = mfccs.*repmat((1./norm_f),1,size(mfccs,2));
% Calculate dynamic features
delta       = diff(MFCC.');
delta_f     = [delta(1,:) ;delta];
deltadelta  = diff(delta);
deltadelta_f= [deltadelta(1,:); deltadelta(1,:); deltadelta];
% Concatenate dynamic features into whole feature vector
MFCC        = [MFCC;delta_f.';deltadelta_f.'];
end
```

It should be noted here, that the *diff* command outputs one values less than it gets as an input, e.g. if the input has 120 samples, the output has 119 samples. This is why the output of the function has been manipulated (first value has been copied) in order to retain the same size. The function can be used in the following way:

```
window_length = 0.03; %in ms
num_cepstral  = 13;
[mfcc,f,t] = mfcc_dyn(y_female, fs_female, window_length, num_cepstral);
```

The file is attached to this report.

6 Additional thoughts

1. *Can you think of a case where two utterances have noticeable differences to a human listener, and may come with different interpretations or connotations, but still have very similar MFCCs?*

To answer this question first of all we should ask ourselves which features of the speech signal are considered/removed by the MFCCs.
First we do a Cepstral analysis, which captures the spectral envelope of the signal, connecting all the formants [1] of the signal. By doing so we can say that the pitch is being discard.
Consider for example a voiced phenomena: if you were to analyse the spectrum of the sound *ah* with different pitches you would still see the same envelope though the fundamental frequency changes. This reasoning does not make sense for unvoiced sounds since it is more noisy and there is no harmonic structure, though the spectral envelope exists also for unvoiced sounds.
Then we apply a bank of filters to approximate the human perception: humans concentrates on certain regions of the spectral envelope.

The main point is that by discarding pitch we are mainly losing information for *voiced phenomena*. Consider again the sound *ah*, it can be used with different meanings based on the pitch: screaming, surprise, etc... though the envelope is still the same. So the MFFCs does not capture this information and a pattern recognition system may fail to recognise the correct situation (surprise, anger,...).

Despite that, this method is well suited for those scenarios where the system needs only to recognise the letters and not whether the person is angry or happy. For example, if you want to call someone and you do not want to digit the numbers you can spell all the numbers and the MFCCs feature extraction still would work since we don't care about the pitch in this case.

2. *What about the opposite situation: are there two signals that sound very similar to humans, but have substantially different MFCCs?*

Signals that are noisy to humans may have a pattern if analysed with frequency/cepstral analysis. For example, consider the sound produced by a fly and an audio sample of the vuvuzelas sound: they sound almost identical to humans, but the spectral envelope is different. Vuvuzela have high constant spectrum at low and high frequencies, and low constant spectrum at medium frequencies, whilst flies produce a sound which is similar to random noise and is flat up to high frequencies. Check figure 11 for the frequency and cepstral plots.

[1] A formant is a region of the frequency spectrum where there is a local/absolute maxima.

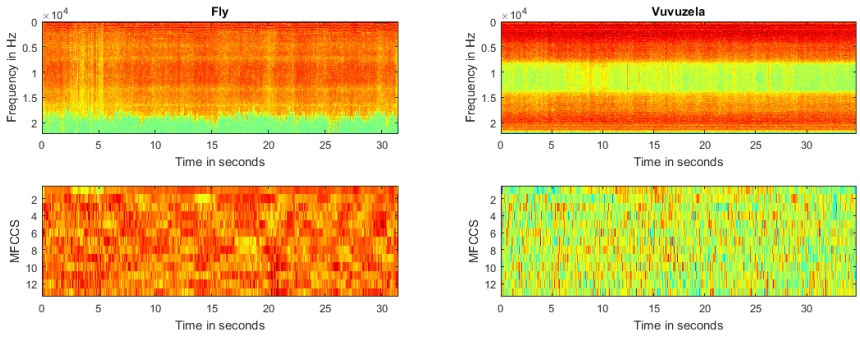

Figure 11: Spectrogram and Cepstral analysis of the fly (left) and vuvuzelas (right plot) signals.