Assignment 2, Pattern Recognition (EQ2340)

Sebastian Bujwid, bujwid@kth.se

Martin Hwasser, hwasser@kth.se

October 8, 2017

Task A.2.2

> **Question 1.** A copy of your MatLab code that plots the female speech and the music signal over time and also zooms in on representative signal patches illustrating oscillatory behaviour, as well as voiced and unvoiced speech segments. All code should be attached either in one or more separate m-files, or as a zip archive.

The code plotting is in *plot_sound.m*. Figure 1 presents the sounds signals for the full files and figure 2 shows selected 20ms snippets.

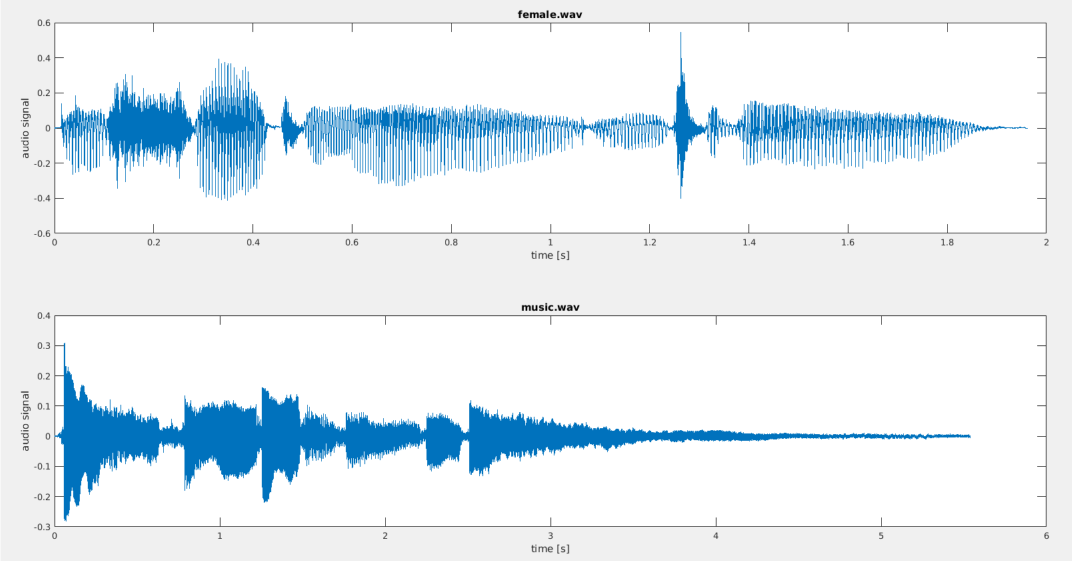

Figure 1: Sound signals

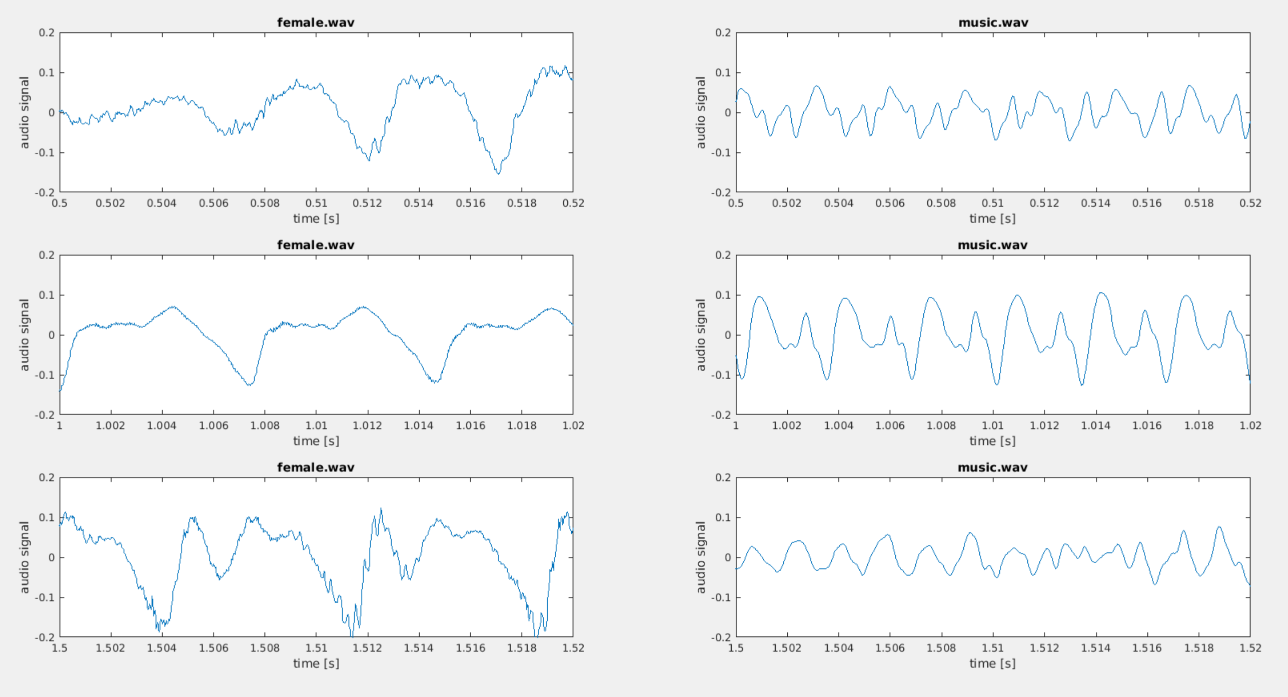

Figure 2: 20ms snippets of sound signals

Question 2. A copy of your MatLab code that plots spectrograms for the same two signals. Put markers in the graph using annotation to demonstrate that you have identified the occurrence of harmonics in the music sample, as well as voiced and unvoiced segments in the speech.

The code for this part is in file *plot_spectrograms.m*. On the figure 3 we present spectrograms with annotated regions with harmonics.

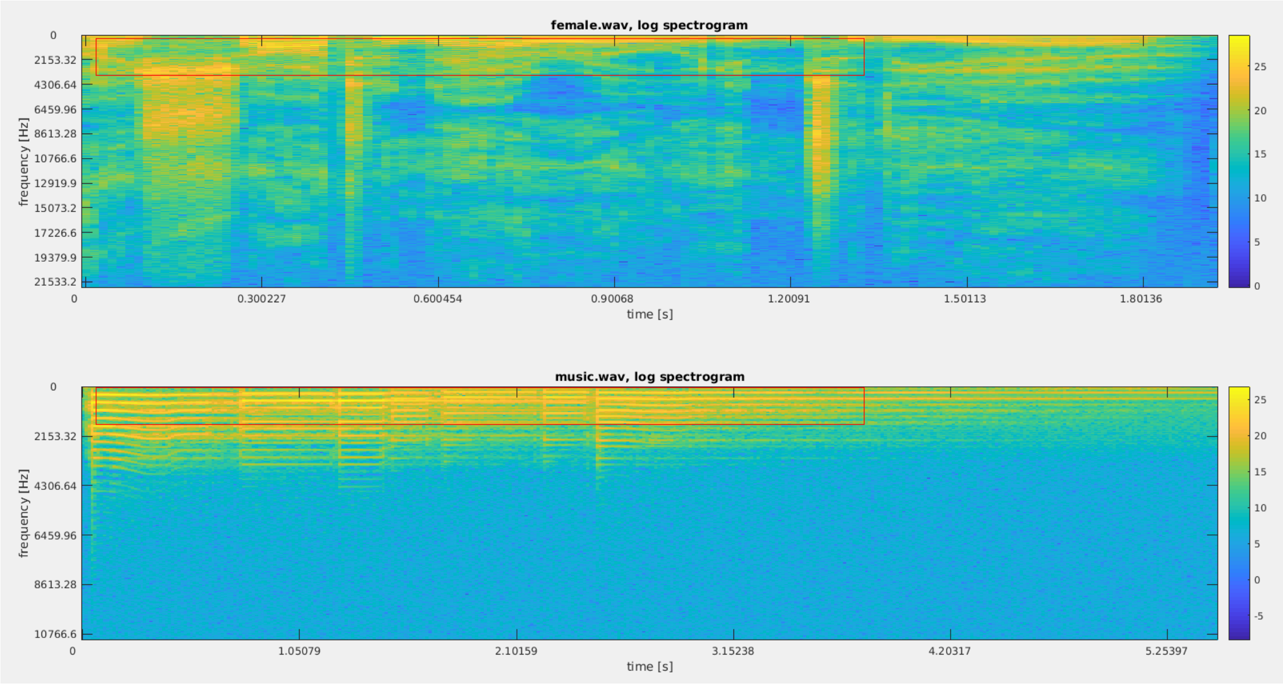

Figure 3: Spectrograms with annotated fragments with harmonics (in red)

Question 3. A copy of your MatLab code that compares the spectrogram and (normalized) cepstrogram representations of the two signals above, and the same for a female and a male speaker uttering the same phrase.

The code for this part is in file *plot_spectra_cepstra.m*. Figures 4, 5, 6 show both spectrograms and normalized MFCCs for *female.wav*, *male.wav* and *music.wav* correspondingly.

Regarding the question from the lab instruction, we think that spectrograms are much easier to interpret for people since we are able to identify different frequencies. MFCC are difficult to interpret, because the coefficients are independent and abstract, so it is hard to see any structure in it.

When comparing spectrograms of *female.wav* and *male.wav* it seems possible to identify that they correspond to the same phrase and we believe that computer could also recognize it, since we know that there exist high quality speech recognition systems taking pure spectrograms as an input. We would not necessarily say that MFCC features come correspond to the same phrases by just looking at the plots and we are not used to interpret such features visually, but possibly computer could perform much better by using MFCCs, since they extract relevant information from the signal and get rid of the noise, leaving independent features which are easier to model statistically.

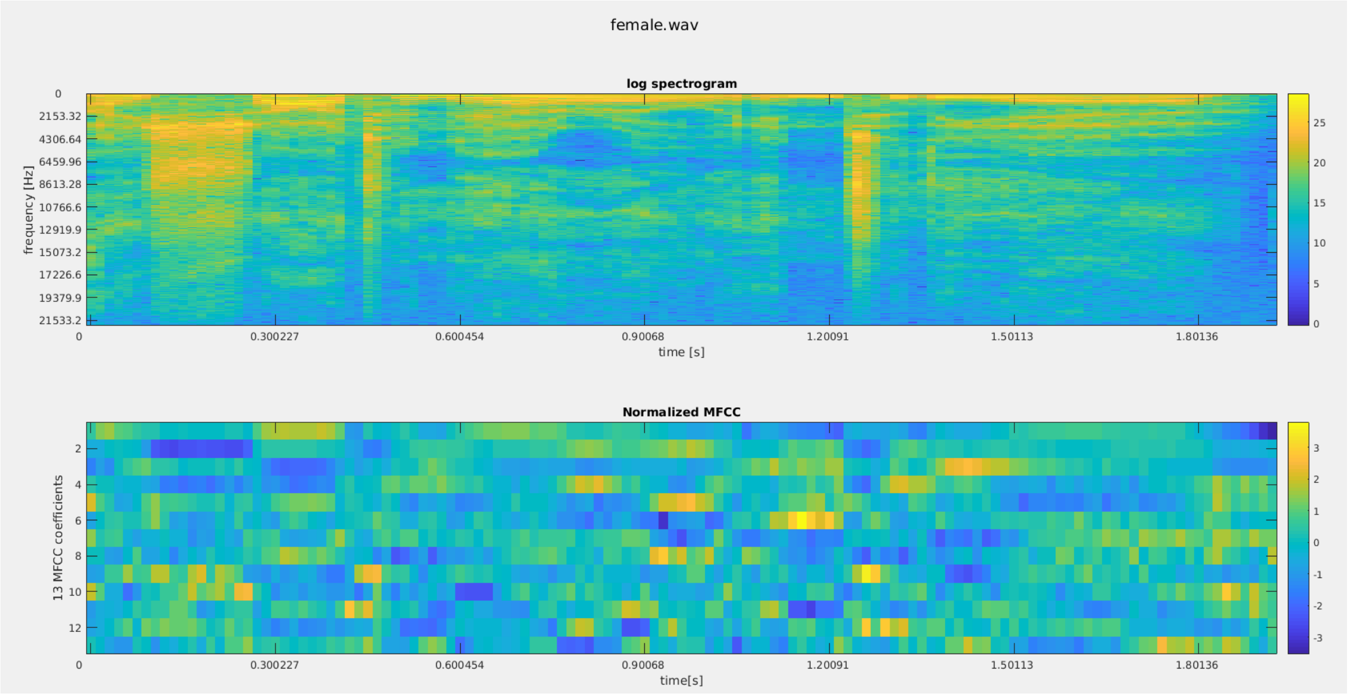

Figure 4: Spectrogram and normalized MFCC for *female.wav*

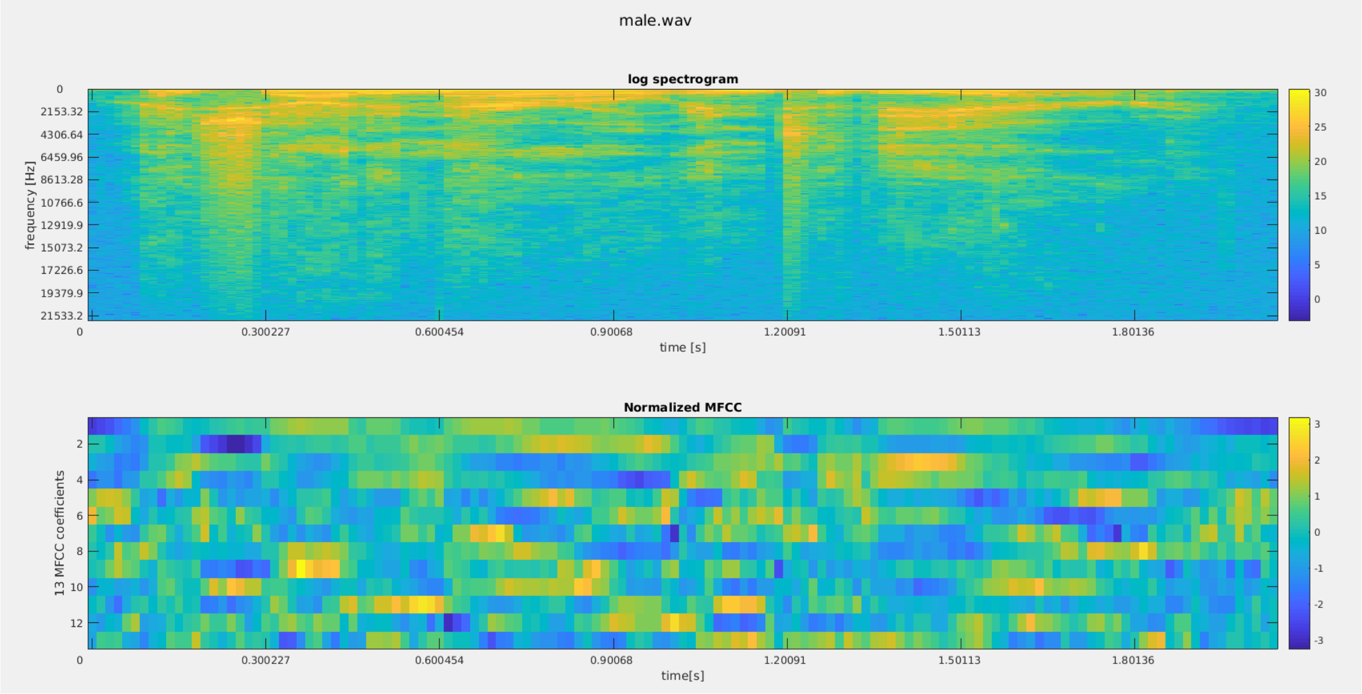

Figure 5: Spectrogram and normalized MFCC for *male.wav*

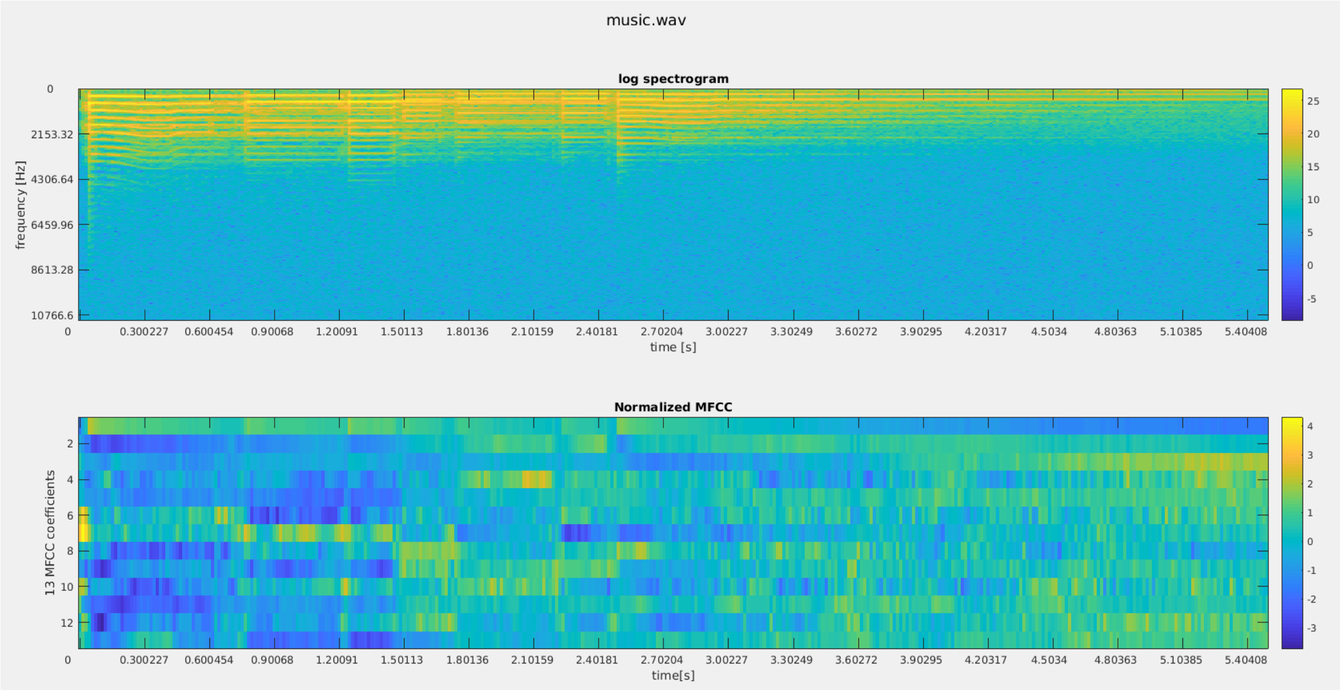

Figure 6: Spectrogram and normalized MFCC for *music.wav*

Question 4. A copy of your MatLab code that plots and compares correlation matrices for the spectral and cepstral coefficient series.

The code for this part is in file *plot_corr.m*. Figures 7 and 8 show correlation matrix of absolute values of MFCC features and spectrogram correspondingly.

Correlation matrices of MFCC features look much more diagonal, since non-diagonal elements of the matrices are close to zero. When we look at correlations of spectrogram values on figure 8 we see that there are values which are very correlated (close to 1).

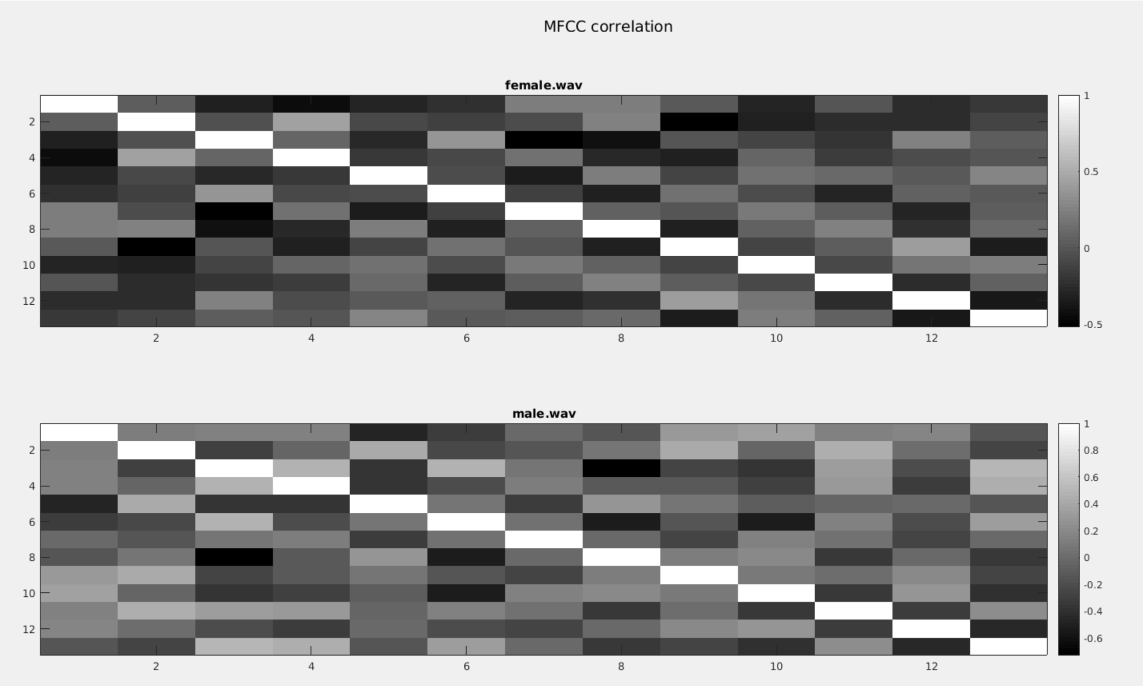

Figure 7: Correlation matrices of absolute values of MFCC features.

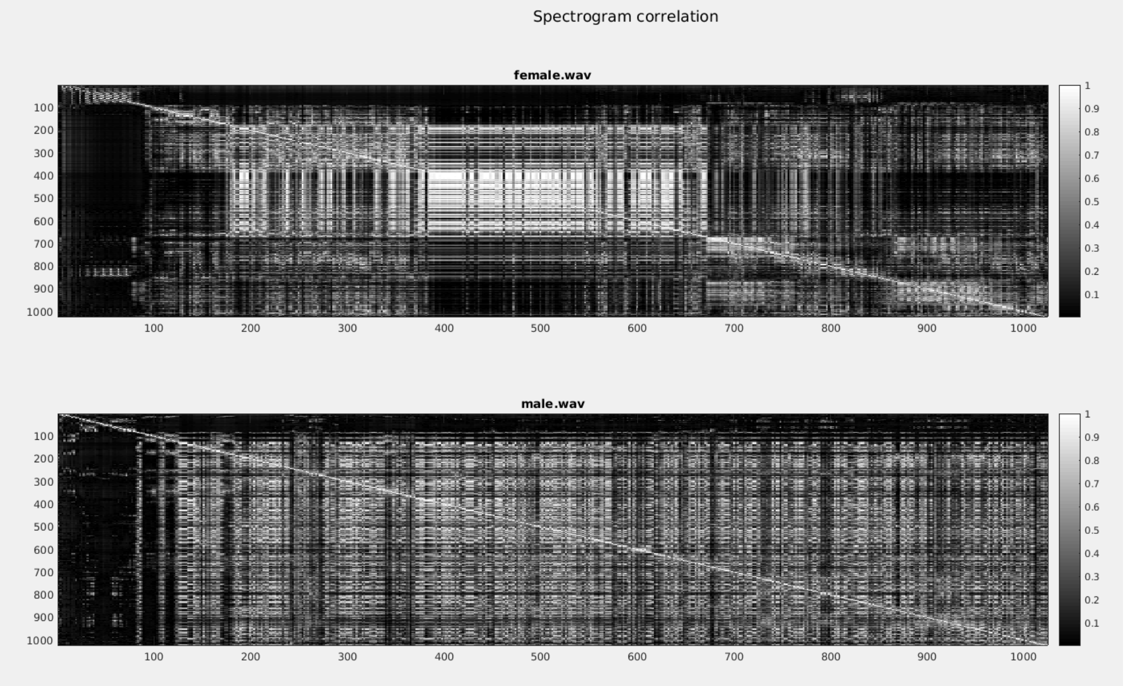

Figure 8: Correlation matrices of absolute values of spectrogram.

Question 5. Answers to the questions in the text associated with the plots.

All the questions have been answered in corresponding sections.

Question 6. A working MatLab function which computes feature vector series that combine normalized static and dynamic features, if you wish to use this technique in your recognizer.

A function calculating dynamic MFCC features is presented below. It returns in total 39 features: 13 MFCC features, corresponding 13 first derivatives and 13 second derivatives.

```
function mfcc_dynamic = mfcc_features(file_path, sampling_frequency)

wave=audioread(file_path);

win_length=0.03;
ncep=13;

[mfcc, spectr, f, t] = GetSpeechFeatures(wave', sampling_frequency, win_length, ncep);

mfcc_dynamic=zeros(13*3,size(mfcc,2));
deltas1=diff(mfcc')';
deltas2=diff(deltas1')';
mfcc_dynamic(1:13,:)=mfcc;
mfcc_dynamic(14:26,2:end)=deltas1;
mfcc_dynamic(27:39,3:end)=deltas2;

end
```

Question 7. Some thoughts on the possibility of confusing the MFCC representation in a speech recognizer. 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? (Hint: Think about what information the MFCCs remove.) What about the opposite situation – are there two signals that sound very similar to humans, but have substantially different MFCCs?

The low-order cepstral coefficients are mostly independent of pitch after applying a Fourier transform. This can impact utterances where tonality is important, for example the Swedish word ”tomten”. If the intonation focuses on the first syllable ”tom”, the meaning is ”a plot of land”, whereas if the intonation is on the second syllable ”ten”, the word would mean ”Santa Clause”. Humans excel at recognizing these tones.

On the other hand, for humans it can sometimes be hard to differentiate between fricatives such as ”th” and ”f”. In figure 9 we see the normalized MFCCs for the words ”three” and "free" respectively. We see that the MFCCs look quite different although the audio files sound very similar to a human ear.

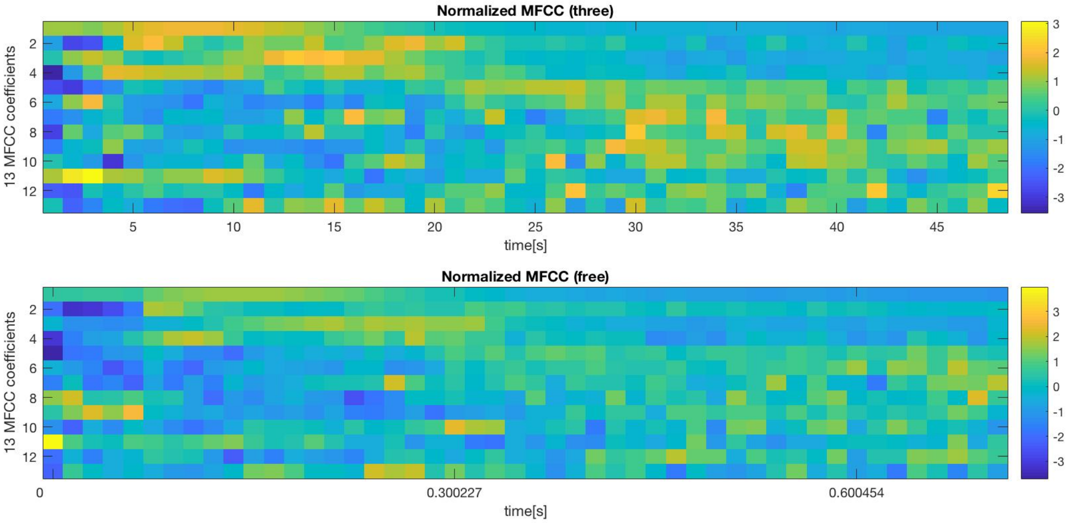

Figure 9: Normalized MFCCs of the words "three" and "free".