Assignment 2

Per Emil Hammarlund, Albert Öst

2019-04-26

Visualization of the sound files

Figure 1 shows the momentary amplitude over time for each of the two audio files.

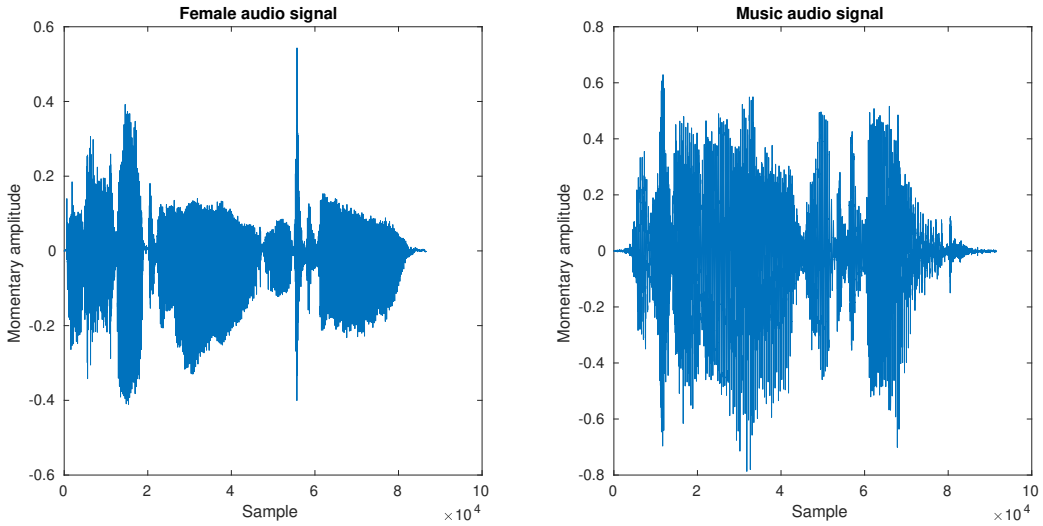

Figure 1: Momentary amplitude over time

Next, we zoomed in on a range of 20 milliseconds, this corresponded to $44100 * 20 \cdot 10^{-3}$ samples. This gave the following observation:

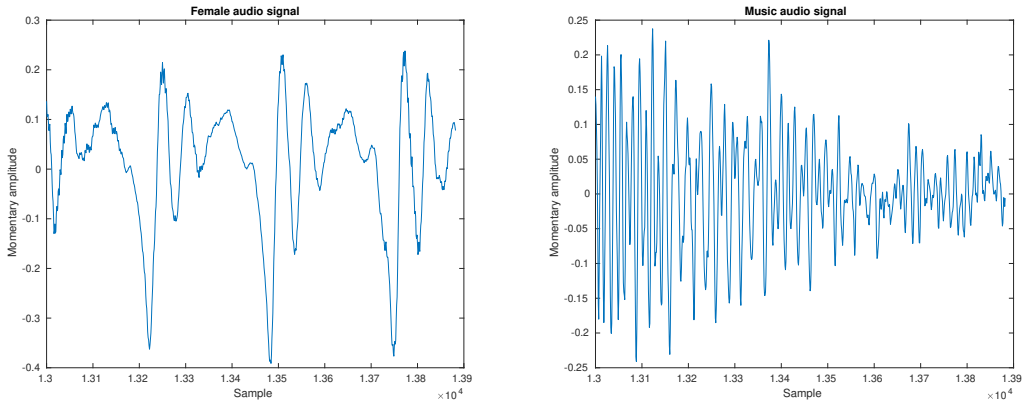

Figure 2: Momentary amplitude in 20ms timeframe

As can be seen in in figure 2 above, the behavior of the sound signal from the two audio files has an oscillatory behavior.

Voiced and unvoiced speech segments

Figure 3 below shows a plot of the momentary amplitude for the "sh" sound from the female voice audio file.

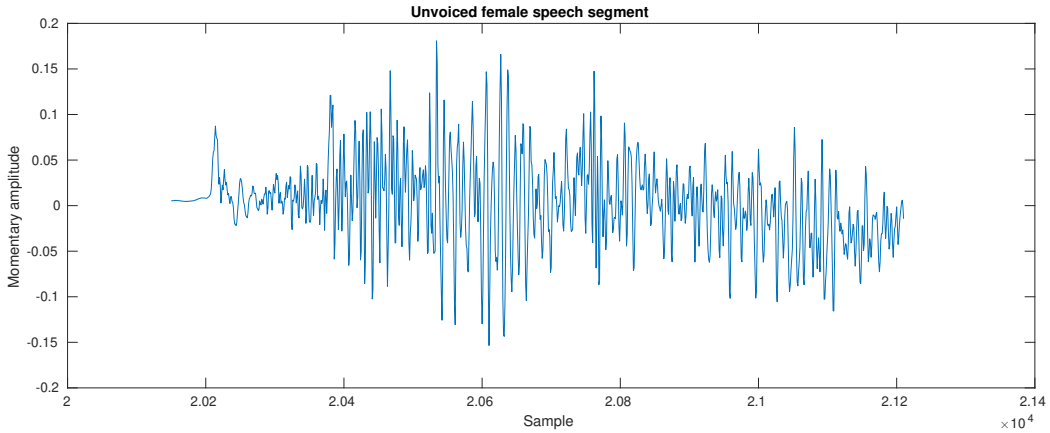

Figure 3: Visualization of unvoiced segment from female voice sound file

As can be seen in figure 3, the unvoiced signal is not harmonic and looks more like random noise.
Figure 4 shows a plot of the momentary amplitude of the voiced "O" sound.

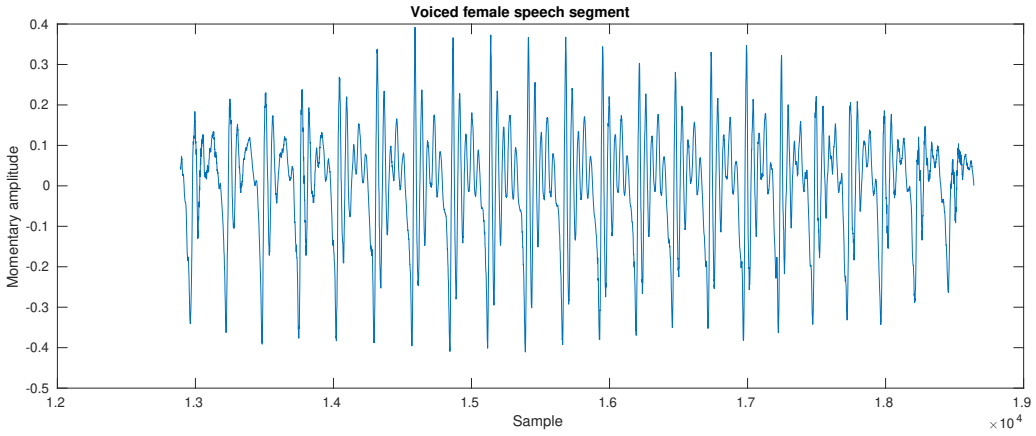

Figure 4: Visualization of voiced speech segment from female voice sound file

From figure 4, we can see that the voiced sound has a harmonic behavior and follows a distinct pattern.

Spectrogram of sound files

Figure 5 points out the unvoiced and voiced segments of the female voice file and shows the harmonics of the music file.

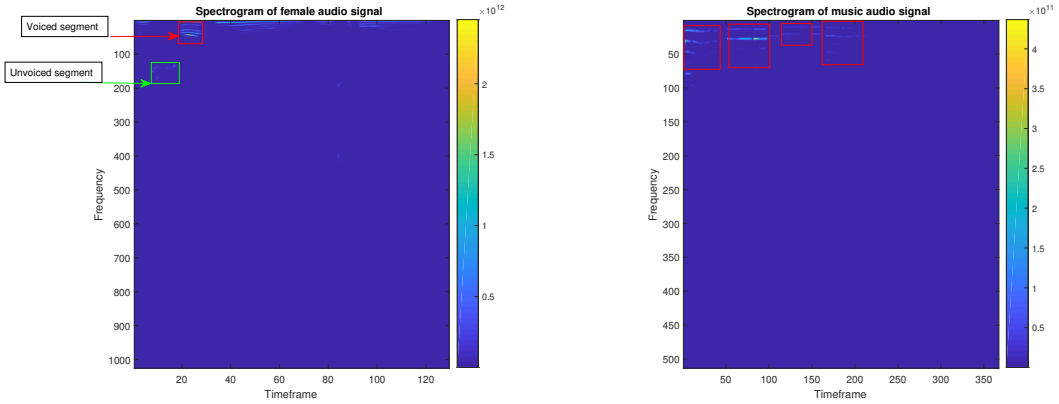

Figure 5: Annotated spectrogram of audio files

In the two spectrograms, the red boxes point out audio segments with harmonic behavior.

Figure 6 shows the logged version of the spectrograms:

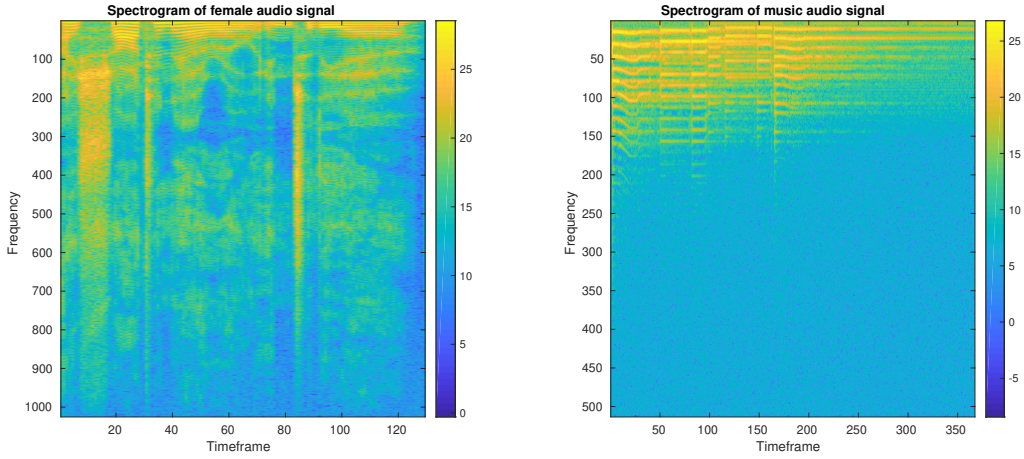

Figure 6: Logged version of the spectrograms for the two audio files

Comparison between ceptrogram and spectrogram

Figure 7 and 8 show comparisons between the spectrogram and ceptrogram representation of the audio files:

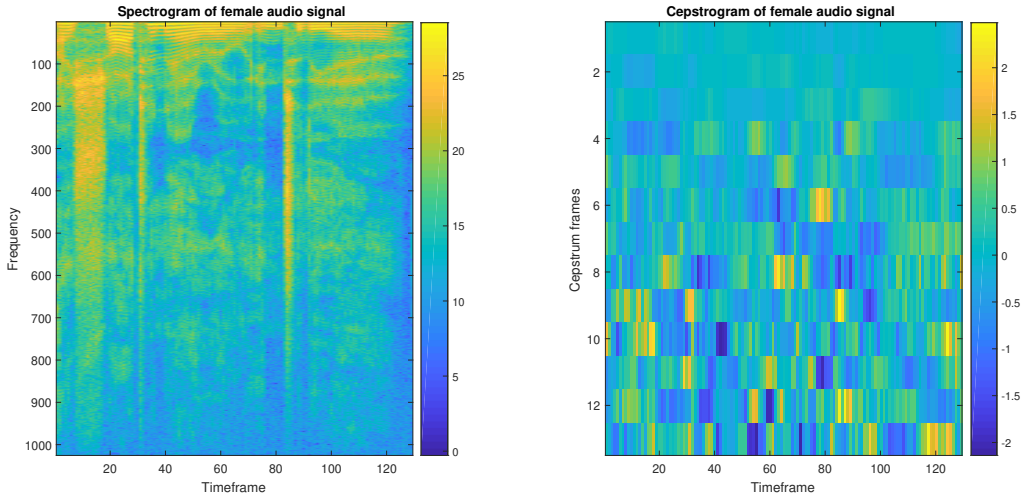

Figure 7: Spectrogram and ceptrogram representation of the female audio file

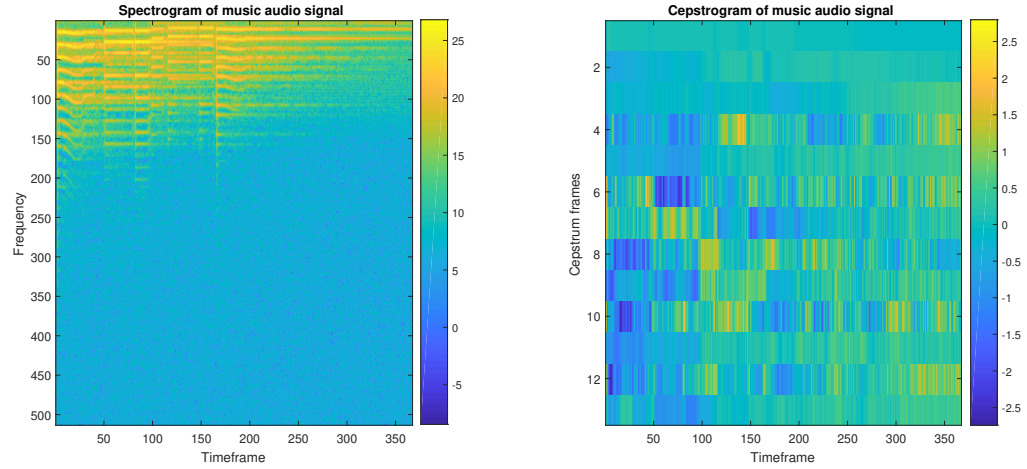

Figure 8: Spectrogram and ceptrogram representation of the music audio file

When comparing the two representation in the visualization of the female voice. Voiced and unvoiced segments are easily identifiable in spectrogram, and hard to identify as a human being in the ceptrogram. However, it is easier to see the intensity of each band pass (cepstrum frame) in the cepstrogram.

Comparison between male and female audio

Figure 9 shows the spectrogram for the female and male audio signals.

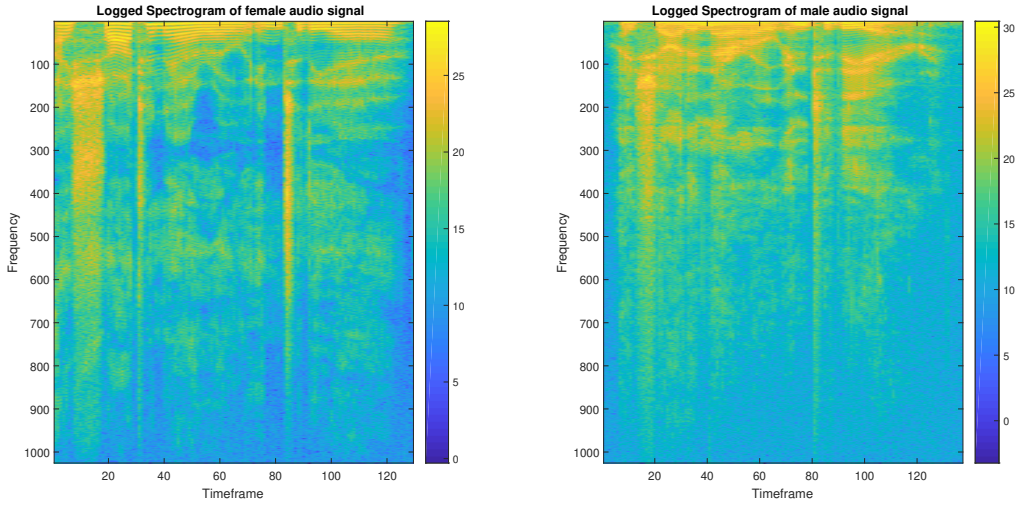

Figure 9: Spectrograms of female and male audio signals

Looking at figure 9, it is visible that the female and male audio signals have roughly the same form. However, this would be hard for a computer to see due to the high complexity of comparing the two signals.

Figure 10 shows a comparison between the male and female audio files.

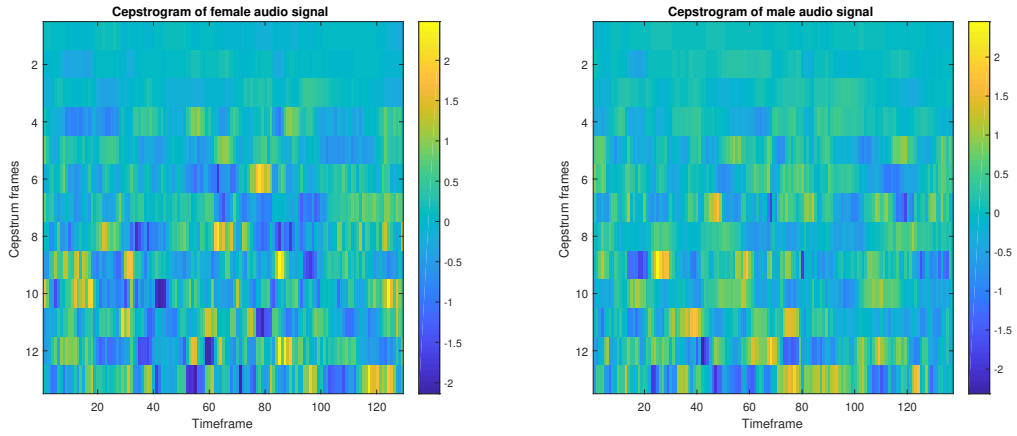

Figure 10: Cepstrograms of male and female audio signals

As can be seen in figure 10 above, the cepstrograms of the female and male audio files roughly have the same intensity at the same points. This make sense since cepstrograms are supposed to be pitch invariant. This would be far easier for a computer to infer as well due to the decreased dimensionality of the representation of the audio signals.

Figure 11 and 12 below show the correlation matrices between the male and female audio files.

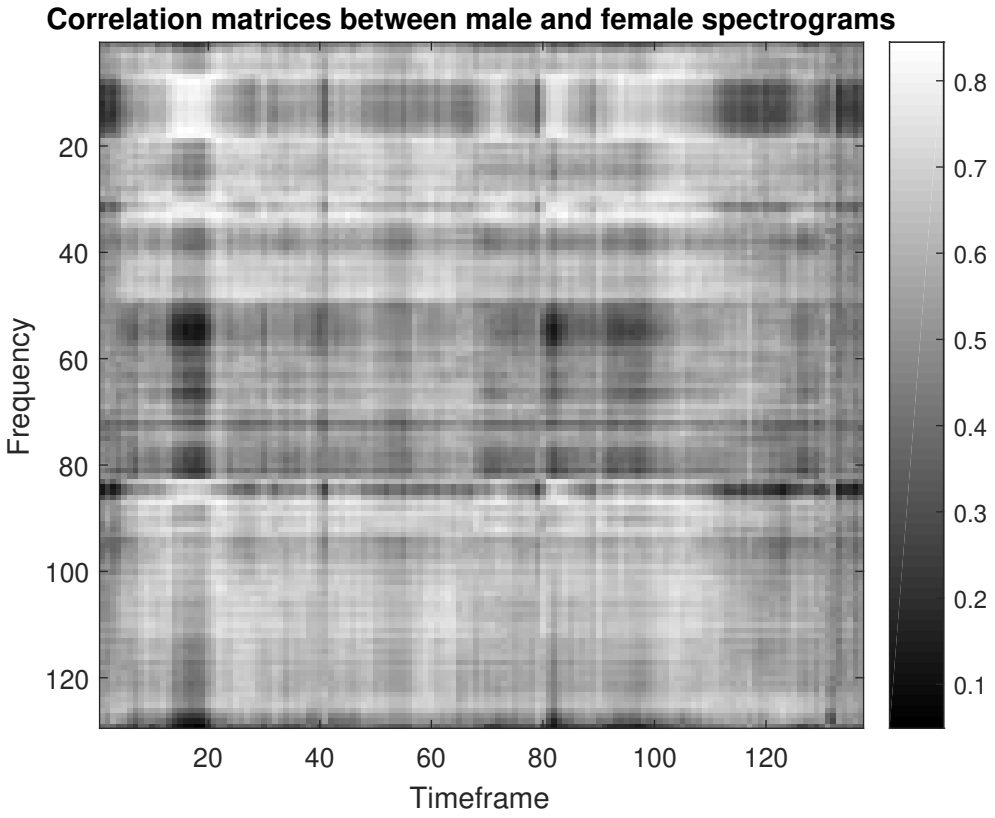

Figure 11: Correlation between male and female spectrograms

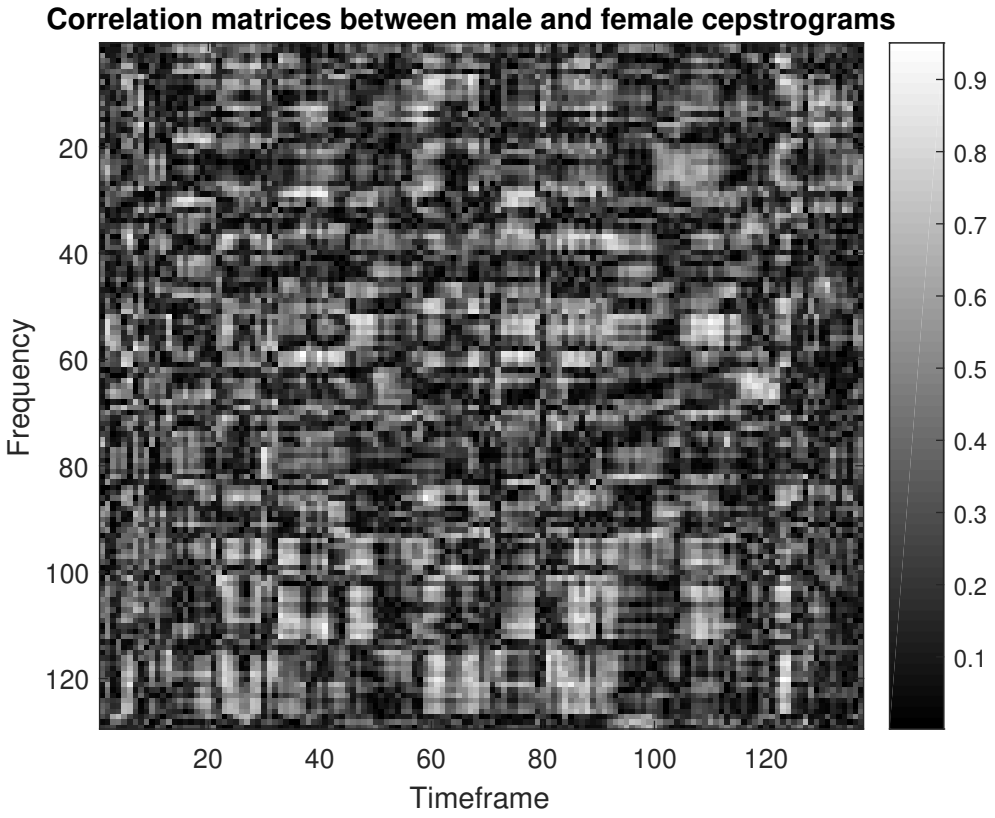

Figure 12: Correlation between male and female cepstrograms

From which it can be seen that the spectrograms are more correlated. The correlation between the cepstrograms of the male and female audio is far smaller and is therefore the more diagonal of the two.

Extraction of dynamic features

The dynamic features where extracted by writing the following ocde (see assignment2.m for original source):

```
% Extraction of dynamic features
dyn_female_cepgram = zeros(ncep * 3, size(female_mfccs, 2));
dyn_female_cepgram(1:ncep, :) = female_mfccs;
dyn_female_cepgram(ncep + 1 : 2 * ncep, 1) = zeros(ncep, 1);
dyn_female_cepgram(ncep + 1: 2 * ncep, 2:end) = diff(female_mfccs, 1, 2);
dyn_female_cepgram(2 * ncep + 1: end, 1) = zeros(ncep, 1);
dyn_female_cepgram(2 * ncep + 1 : end, 3:end) = diff(female_mfccs, 2, 2);

dyn_male_cepgram = zeros(ncep * 3, size(male_mfccs, 2));
dyn_male_cepgram(1:ncep, :) = male_mfccs;
dyn_male_cepgram(ncep + 1 : 2 * ncep, 1) = zeros(ncep, 1);
dyn_male_cepgram(ncep + 1: 2 * ncep, 2:end) = diff(male_mfccs, 1, 2);
dyn_male_cepgram(2 * ncep + 1: end, 1) = zeros(ncep, 1);
dyn_male_cepgram(2 * ncep + 1 : end, 3:end) = diff(male_mfccs, 2, 2);
```

Some thoughts on the possibility of fooling the MFCC

The sounds “ch” and “sh” are easily distinguishable from each other for humans. But the MFCC will only see these as unvoiced utterances with same spread. The same text with a strong dialect would be seen as a different utterance in the feature extractor, but a human listenser with a keen ear would have little trouble understanding. Humans also have the ability to filter out ambient noise (such as a load crowd) and hear what is being said. The feature extractor on the other hand, will be unable to filter out to the desired source.

Since the MFCC feature extractor removes information, it will not be able to find differences that can not be found in the original sound.