Pattern Recognition - Report for Assignment 2

Navneet Agrawal, navneet@kth.se
Lars Kuger, kuger@kth.se

I. Background and Problem Formulation

Nowadays speech recognition systems are a common feature in many devices as mobile phones or cars. These speech recognition systems are often built using *Hidden Markov Models* (HMM) which require that their input only contains relevant data. To do this feature extraction, *mel frequency cepstrum coefficients* (MFCC) are commonly used [1]. In this report, the underlying hypothesis that MFCCs serve well as extracted features is investigated.

II. Methodology

In order to investigate the aforementioned problem, several steps will be conducted. First of all, the music and speech audio files will be plotted in time domain and analyzed graphically in order to obtain some knowledge about how these audio files are constituted and how that relates to the human hearing theory.

In a second step, the spectrograms of these audio files are generated and plotted so that information about the frequency contents can be used. To generate a spectrogram, the original signal is multiplied with a window function of length T_w. The resulting non-zero signal is called a *frame*. This is done several times while the window function is shifted to the right by exactly T_w on the x-axis after each multiplication so that many frames are obtained. Each frame then is Fourier transformed separately. Note that $T_w = 30$ms and the Hanning window was used in this case.

Thirdly, the spectrogram of the audio files can be compared graphically to the cepstrogram which is the visualization of all MFCCs as time series. This way it is possible to draw conclusions about how suitable cepstrograms are for being processed by machines. The goal is to obtain the spectral envelope independently of the spectral details since the spectral envelope contains most of the relevant information whereas the spectral details mostly contain irrelevant information [2]. Thus, a cepstogram is calculated by smoothing the Fourier transform of the frames that were already used in the spectrogram according to the bark or mel scale. This operation is supposed to represent the human hearing system [1]. After that, the results of this operation are logarithmically transformed which will transform the multiplication of spectral envelope and spectral details into an addition. Fourier transforming this expression will separate the two parts so that we can use the information of the spectral envelope independently of irrelevant information. A more detailed description of this can be found in [2].

To determine if the MFCCs are more suitable for speech recognition than the usual spectrograms the correlation properties of both can be compared. Uncorrelated coefficients are desirable since this means that they do not depend on the pitch of the speech.

III. Results

A. Music and speech in time domain

In Fig. 1 a female voice signal and a music signal in time domain are shown. It can be seen that the music signal as well as the voiced ”aaa” sound look like waves and are quite similar. In contrast, the unvoiced ”sshh” sound differs remarkably from the other plots shown in the figure. There is no clear periodic wave visible and the signal looks rather noisy.

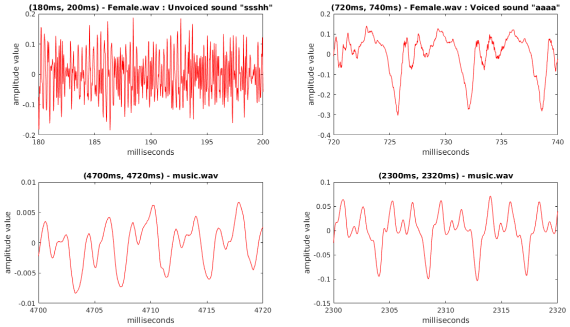

Fig. 1: A female voice signal and a music signal in time domain are displayed. Note that the plots are zoomed in so that excerpts of 20ms are shown.

B. Spectrogram

In the previous section a voice signal as well as a music signal were shown in time domain and differences in their characteristics could be noted. In Fig. 2 the spectrograms of the respective signals are shown. In the spectrogram of the speech signal, the difference in voiced and unvoiced sounds can be seen clearly. The voiced ”aaa” sound produces harmonics in the lower frequencies whereas the power in higher frequencies is considerably less. On the contrary, the unvoiced ”sshh” sounds spreads the power across a large frequency range and therefore resembles noise. Furthermore, the music signal consists of clearly visible harmonics one of which is marked by an arrow in the figure.

C. Cepstrogram

From the spectrograms calculated and shown in the previous section, the cepstrograms can be generated. As can be seen in Fig. 3 the cepstrogram keeps and represents the differences between female speech and music that are seen in the spectrogram.

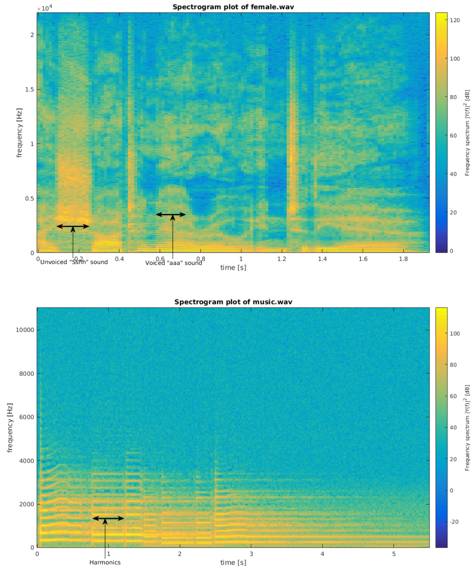

Fig. 2: The spectrograms of a female voice signal (top) and a music signal (bottom) are displayed. The arrows mark exemplary intervals that show certain characteristics such as the harmonic nature of the music signal and the difference between unvoiced and voiced sounds in female speech.

In contrast, the cepstrograms are fairly similar for a sentence spoken by a male and a female speaker as shown in Fig. 4. This is a desirable result since we aim at identifying the content of what has been said and not the pitch of the speaker's voice. Therefore, it can be assumed that it is easier for a machine to identify a spoken sentence from a cepstrogram as compared to a spectrogram.

D. Correlation properties of spectrogram and cepstrogram

The calculation of the cepstrogram is supposed to decrease the correlation of low-order cepstral coefficients such that these are independent of the pitch of the speech [1]. This can also be seen in Fig. 5. It is visible that the correlation matrix of the cepstral coefficients is much more uncorrelated (darker) than the correlation matrix of the spectrum (lighter). Hence, the spectrogram is more pitch dependent than the cepstral coefficients are.

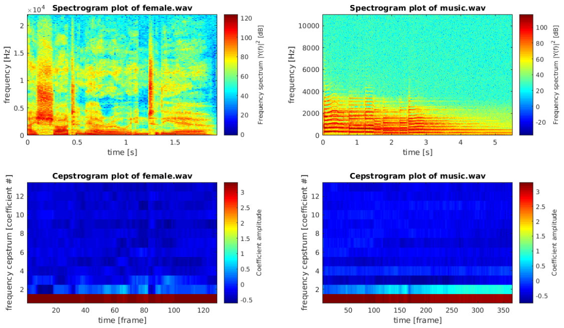

Fig. 3: Spectrograms (top line) and cepstrograms (bottom line) of female speech (left column) and music (right column). Both spectrograms and cepstrograms differ clearly when comparing female speech and music.

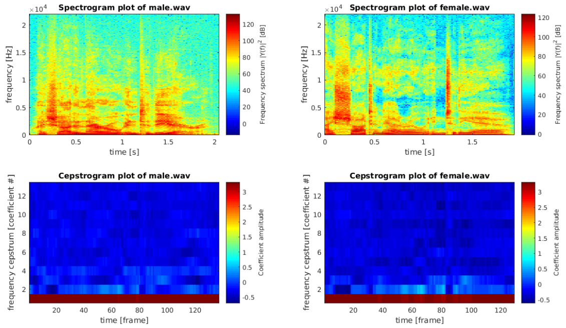

Fig. 4: Spectrograms (top line) and cepstrograms (bottom line) of male (left column) and female speech (right column). Although the spectrograms differ the cepstrograms are clearly more similar.

E. Possible Problems of MFCCs

As shown in the course of this paper, MFCCs remove the pitch of speech to a large degree. This might cause problems when this pitch is important, i.e. in cases of a question where the voice is usually raised to the end of the sentence as opposed to statements where the voice is lowered to the end. A distinction between a question and a statement therefore could be difficult.

Another problem that might occur is that MFCCs can look quite different even though the spoken phrase might have been the same, i.e. when a person speaks with an accent and pronounces words slightly different.

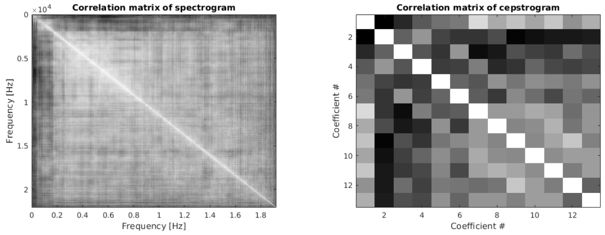

Fig. 5: Correlation matrices for spectrogram frame (left) and cepstrogram coefficient series (right). The darker the gray the less correlated the frames or coefficients are.

IV. Conclusions

In conclusion, the MFCCs have desirable properties for speech recognition since they keep relevant differences, e.g. between music and speech, but remove disturbing information, e.g. the pitch of the voice. Furthermore, the complexity of the MFCCs is much lower than the complexity of spectrograms and MFCCs have more favorable mathematical properties since they are less correlated than the spectrogram frequencies. To further improve the performance of speech recognition a function has been developed to take the so called dynamic features, that means the derivative and its derivative of the MFCCs, into account.

References

[1] Arne Leijon and Gustav Eje Henter, *Pattern Recognition - Fundamental Theory and Exercise Problems*, KTH School of Electrical Engineering, 2015

[2] Kishore Prahallad, *Speech Technology: A Practical Introduction. Topic: Spectrogram, Cepstrum and Mel-Frequency Analysis*, Carnegie Mellon University and International Institute of Information Technology Hyderabad, Retrieved on October 1, 2016 from
http://www.speech.cs.cmu.edu/15-492/slides/ and name of file: mfcc.pdf