EN2202 Pattern Recognition
Assignment 2 - Feature Extraction

Fernando J. Iglesias García
fjig@kth.se

Bernard Hernández Pérez
bahp@kth.se

September 12, 2012

- The feature extraction step is very important in any pattern recognition system. Usually the input includes a lot of data and we need to focus on the signal aspects we are interested in. This depends on the task, in this report we will focus on the features extraction for Speech Recognition. Before extracting the data we need to understand speech signals and the features used in speech recognition.
  - Sound is usually represented by discrete samples of fluctuations in air pressure, forming "sound waves". For a better understanding of this concept take a look at the representation of the female voice at Figure 1 and its augmented section at Figure 2.

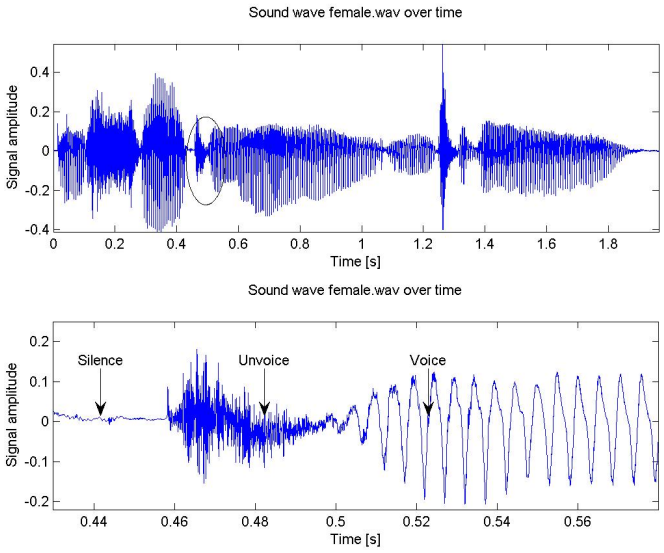

Figure 1: Female signal voice representation.

If we want to extract information about frequency we can use the Fourier transform, calculated computationally in a efficient manner using the fast Fourier transform, FFT. The problem with this method is that we lose all time information, and as the hear does, we want a compromise between time and frequency information. For this reason we split the signal in several time intervals, and do the Fourier transform of each. With this method we can obtain the intensity representation for each frequency over time, that is called spectrogram.

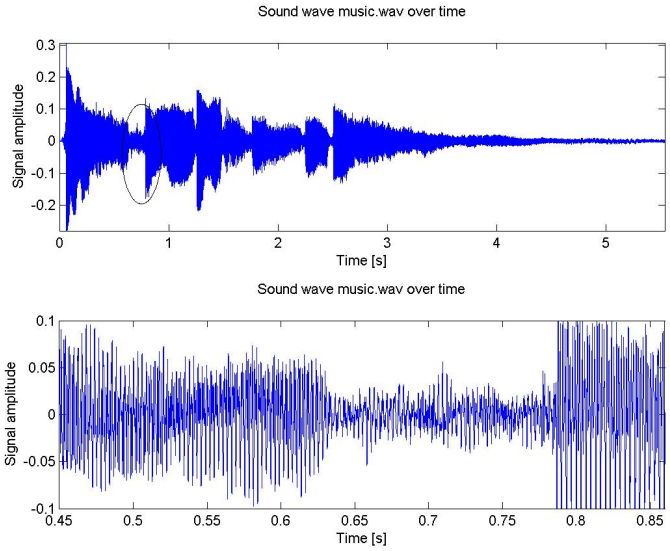

Figure 2: Augmented section of the female signal representation. Its possible to split in three different regions that represent silence, unvoiced sounds and voiced sound respectively.

The speech, and in our case the female voice consists of voiced and unvoiced sounds. The voiced sounds have harmonics and usually corresponds with vowels. Unvoiced sounds have no single determining frequency of pattern bus contains a big amount of energy spread over almost all frequencies. Some examples of both different regions are marked in Figure 3].

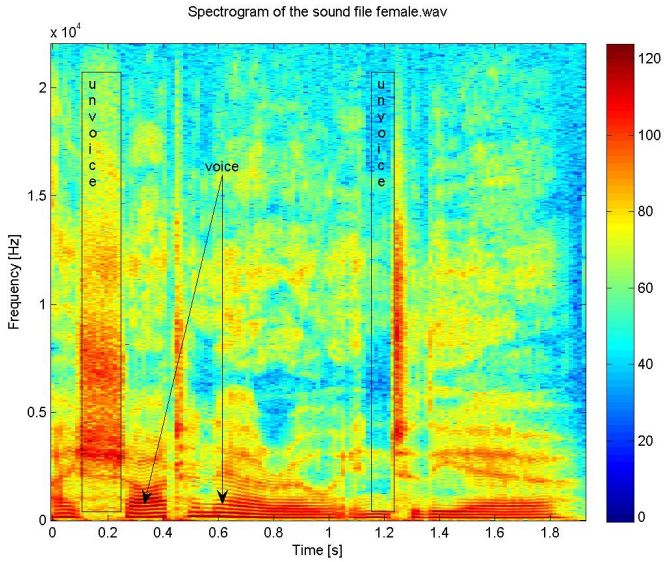

Figure 3: Voice and unvoice regions for female voice signal.

The music signal is composed by harmonics. Harmonics are components whose frequency is a multiple of the basic frequency f. And they are represented on the spectrograms as bands of signals with

high intensity moving up and down in unison. Some of there are marked in Figure 4].

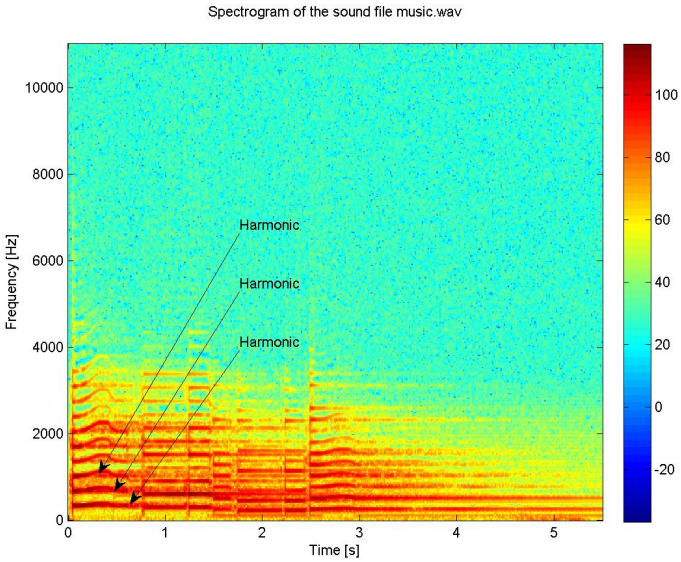

Figure 4: Harmonics annotation for the music signal spectrogram.

  - Spectrograms representation gives us a better impression of how sound is perceived, but is not well suited for speech recognition. The problem is that still have to much high rate, so another Fourier transform is applied to the data tocalculate the
    mel frequency cepstrum coefficients MFCCs.

- After the explanation of both methods and how they works, it may be instructive four our task, speech recognition, to make some comparations between female voice and male voice signals, and spectrograms and cepstrograms representations.

  - *Which representation do you think is the easiest for you, as a human, to interpret, and why?.*

    Looking at the pictures .....

  - *Can you see that they represent the same phrase?. Could a computer discover this?. Why/why not?. What about computer?.*

    Looking at....

  - *Which matrix, spectral or cepstral, looks the most diagonal to you?*

    Looking at...