3. SPEECH RECOGNIZER

We implemented a speech recognizer using a modified version of the DTW algorithm. Using DTW and speech processing techniques (i.e. Mel Frequency Cepstral Coefficients), our algorithm is able to detect which action we want to perform: open the google search or open our facebook webpage.

This speech recognizer takes as input a sound file and compares it with some template words: *google* and *facebook*. In order to make our recognizer more robust, we have used two template words for each sound. After comparing the input sound with each template word, the algorithm chooses the one with the shortest path.

3.1 Speech Processing

Since each sound has a different length, DTW is a good algorithm for comparing them. However, sound Waves are sinusoids and cannot be properly compared on the time domain (amplitude vs seconds). The reason is simple, a sound wave of a word can change radically from a person to another, and even the sound waves of a word generated by a single person can be very different.

If we observe the sound waves of our sounds, we see their differences.

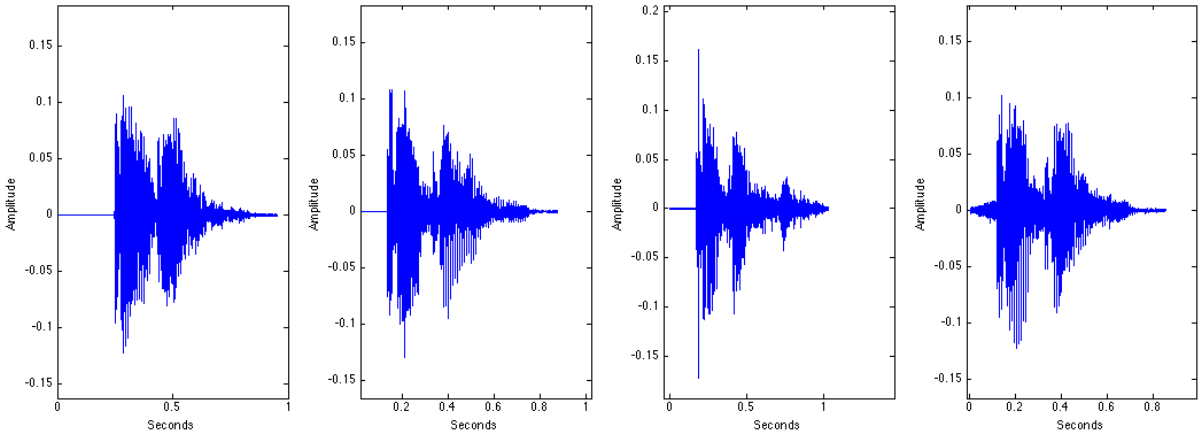

Figure 1. Sound Waves of the word Google.

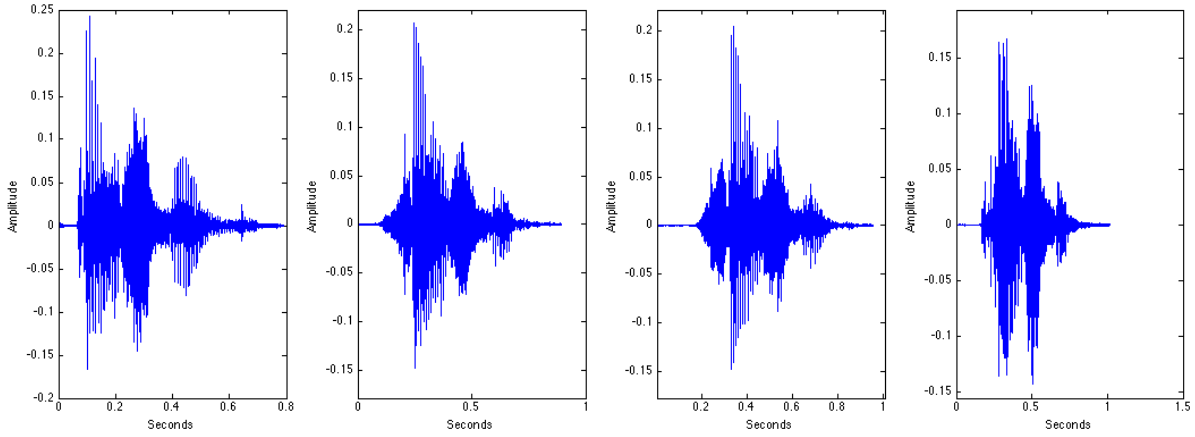

Figure 2. Sound Waves of the word Facebook.

An extended approach is to transform the sound waves into the frequency domain and then, delte the unnecessary frequencies and extract some coefficients ([1], [2]). In this project, we tried with two different coefficients: Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coefficient (LPC). Unfortunately, we didn't obtained good results with LPC and we decided to use only MFCC.

1. Pre-emphasizing: In the process to digitalize audio, it passes through a low-pass filter and the high frequencies are attenuated. In this step a pre-emphasizing filter is applied to emphazise the higest frequenciy components of the speech, that is increase the energy in the higher frequencies.

$$y(n) = x(n) - 0.95x(n-1)$$

2. Framing: From the previous images, we can clearly see that sound signals are not stationary. However, we can assume that if we take a small sample called frame (approximately of 10ms or 20 ms) it will be stationary. We created a set of frames of 16ms.

3. Windowing: In order to make the transition between frames smoother, a window is multiplied with each frame. The most used window for speech procesing is a *Hamming Window*:

$$w(n) = 0.54 - cos(\frac{2\pi n}{N-1})$$

where N is the length of the window, in this case the length of the frame.

4. Mel Frequency Cepstral Coefficient (MFCC): Mel scale is based on human perception of frequencies. The basic idea of MFCC is to only keep the most relevant frequencies and ignore the others. The process is the following:

1. Convert the frames into the frequency domain using the Discrete Fourier Transform (DFT).
2. Apply the Mel Scale filter bank, which is a set of triangular filters, and take the logarithm.
3. Finally, to transform the results into the time domain again the Discrete Cosine Transform (DCT) is applied.

From this process we obtain K coefficients, in our case K = 11.

At the end, we end up having M frames and each frame contains a vector of 11 coefficients.