Report Lab 2 Preparation

Pattern Recognition NTUA
9th Semester

Efthymios Tzinis 031212007
Konstantinos Kallas 03112057

November 17, 2016

Introduction

In this work we prepared steps 1 till 9 from the Laboratory Exercise 2. The main purpose of this Exercise includes the implementation of an automatic speech recognition system which is dedicated to individual cognition of spoken digits from 15 different people. The features used for this purpose are some mel-frequency-cepstrum-coefficients (MFCC's) which were extracted with Matlab as described below. MFCC's are definitely proven to be substantial features, see [1], for the representation of any utterance and as a result, we portend to be useful for our cause too.

Step 1

We extracted for every utterance of our data a respective signal of voice which was saved into a structure of a cell in order to eschew problems with the diversity of dimensions throughout the data set.

Step 2

We preemphasized every speech signal using a transition function in Z space of the form:

$$H(z) = 1 - 0.97z$$

Function filter() in Matlab is offered for this specific purpose. After we applied the specific filter we tested some of the derived data arbitrary and concluded that our filter indeed smoothed the noise contained in the initial data set.

Step 3

In this step we windowed every preemphasized signal by using the hamming() function in Matlab. Some conventions were followed according to the given sample frequency of $Fs = 16kHz$:

- Duration of every window in time: $T = 25\ msec$ and thus in sample space $NT = 400\ samples$
- Duration of every window in time: $T_overlap = 10\ msec$ and thus in sample space $NT_{overlap} = 160\ samples$
- The number of windows applied were found with the following expression: $N = \frac{n}{NT - NT_{overlap}}$ where n = number of samples of the specific utterance. Some zero padding was needed in the end of our data utterances in order to take the apt number of samples in every windowed signal.
- We applied the function of a typical hamming window: $w(n) = 0.54 - 0.46cos(\frac{2\pi n}{NT-1}), 0 <= n <= N - 1$
- The windowed signals are now $s_i(n) = s_p(n+(i-1)*(NT-NT_{overlap}))w(n)$

Step 4

We implemented a triangular filter bank sequence depended on the mel scale of $Q = 24$ central frequencies. To begin with, we speculated that information below 300 Hz in frequency space is useless as it could not be produced by a human. In the same concept the frequencies above 8kHz are also useless for our task, as mentioned before $Fs = 16kHz$ and the Nyquist criteria for anti aliasing is in use. According to the given transform from frequency to mel:

$$f_c^i = 2595log(1 + \frac{f^j}{700}), j = 1, ..., Q$$

we computed the respective mels to the aforementioned frequency boundaries $(300Hz, 8000Hz)$. We added two extra dummies frequencies in order to aptly divide the mel space linearly. As a final step, we computed the differences between 2 adjacent frequencies and as a consequence the bandwidth of every triangular filter b^j. The highest point of the triangular is always 1 $H^j(f_c^j)$.

In figure 1 we can see the full filterbank of our system. By the form of our filterbank we can understand that our implementation is actually valid. Because triangles in lower mels are more dense in the lower frequencies rather than the higher ones. Every triangle was implemented according to the specifications and thus many of them are not isosceles.

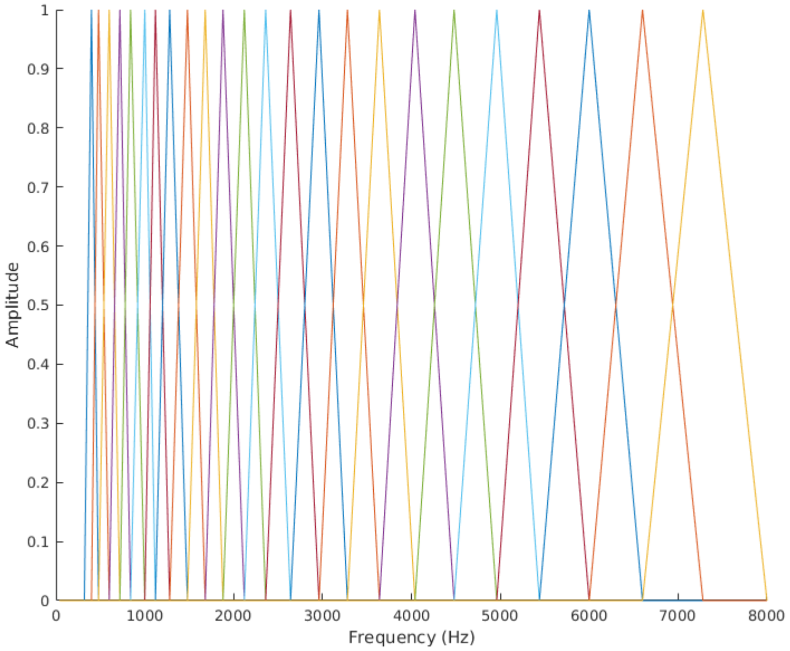

Figure 1: Full Filterbank

Step 5

For every utterance we have all the windowed signals. For every one of them we apply the filter bank individually and we compute the energy $E_i(j)$ for every filter j for the input $s_i(j)$. In figure 2 we can see the energy coefficients $E_i(j)$ for utterance (digit:5,speaker:5) the differences between the samples 19 and 27. The theoretical explanation of this phenomenon is based on the nature of digit 5 while it is spoken. Five demands huge release of energy in initial samples and it's denigrating while the time passes.

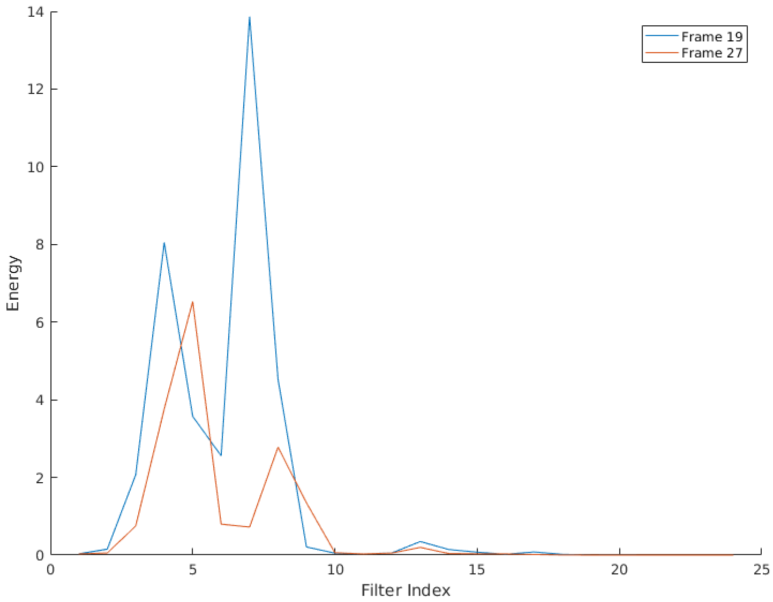

Figure 2: Energies Comparison for Utterance of Speaker 5 for Digit 5

Step 6

Now we have computed the energy coefficients for every filter and for every frame of all the utterances. A simple computation of $G_i(j)$ according to the expression:

$$G_i(j) = log(E_i(j)), i = 1, ..., Q$$

Step 7

According to the DCT transform we computed the Cepstrum Coefficients from the previous computed $G_i(j)$. For the computation of the values n_1, n_2, k_1, k_2 we used the ID *03112057*, thus $d_1d_2d_3d_4 = 2057$ and $(n_1, n_2, k_1, k_2) = (12, 5, 7, 5)$. In figure 3 we demonstrate the histograms of the Cepstrum Coefficients $C_i(n1)$ and $C_i(n2)$ for digits k_1, k_2.

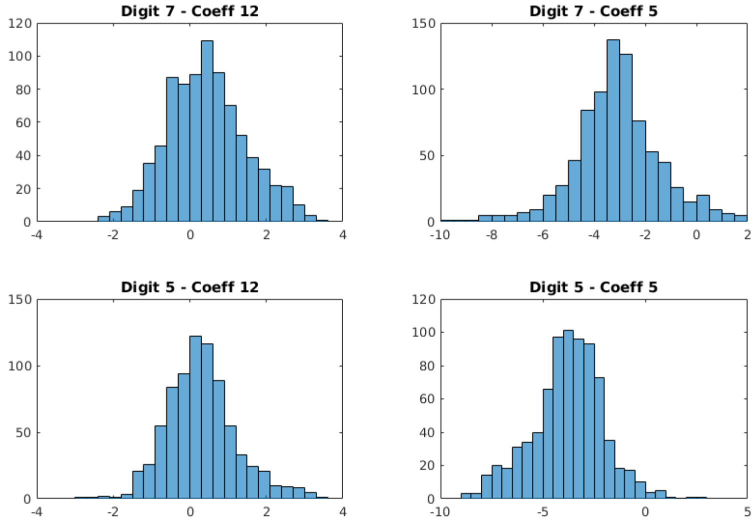

Figure 3: Histograms for Cepstrum Coefficients $C_i(n1)$ and $C_i(n2)$ for digits k_1, k_2

Step 8

In figure we demonstrate the initial energy coefficients and the reconstructed ones for the digit 5, speaker 5, for frames 19 and 27 respectively. We can see that the reconstructed Energy coefficients which are produced by first taking the inverse DCT transform and then the exponent to reconstruct the energy coefficient $\hat{E}_i(j)$ represent quite good the initial coefficients. Some diminishing of data amplitude in higher frequencies is expected but do not change our perspective of the Energy in any case. So the derived curvature is an adequate representative for the energy coefficient.

Step 9

For the digit k_1 we computed the mean values throughout all the frames for the Cepstral Coefficients $C_i(n1)$ and $C_i(n2)$. The repetition of the task for every speaker and for every digit gave us the figure 5. First the minutiae symbols in the background represent the mean value for every speaker when the prominent symbols shown in the legend represent the universal mean values for every speaker and digit. We can see that some of the digits have quite similar characteristics like "three" and "eight" while the digit "five" seems to be very alienated from all the other digits. Moreover, we can speculate that the MFCC's characteristics can discriminate our digits up to a certain point but a higher dimension representation and clustering seems to be needed in order to achieve high results. Relevant work in the field as in [2] showed that MFCC's can be used in order to achieve a speaker independent speech recognition system which is our purpose. It is worth mentioning that the MFCC's contain in their prosody the

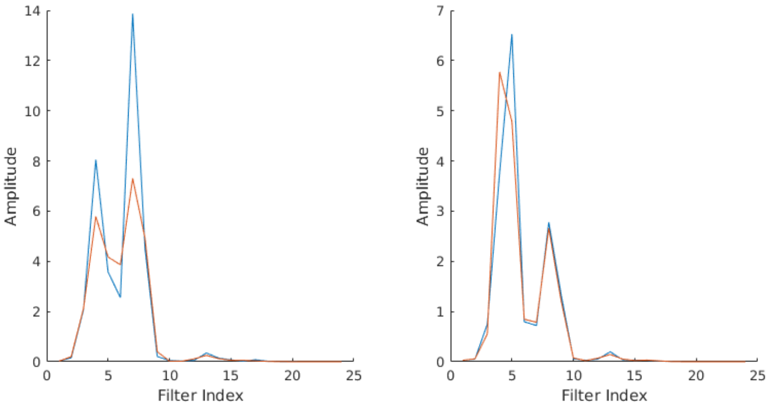

Figure 4: Energy Coefficients initial (Blue) and reconstructed (Red) for the frames 19 and 27 of speaker 5, digit 5, left and right respectively

characteristics of the speakers. Therefore, a vocal tract length normalization should be endorsed in order to achieve a transformation of data with speaker independent information through vector representation of MFCC's. We should not be oblivious to the fact that the linguistic model should be also simulated somehow in order to achieve good results in our task of separating the different digits through voice and consequently achieving a better classification ratio.

We demonstrate also in figure 6 the results of the mean values of all digits for a different pair of MFCC's $C_i(0)$ and $C_i(1)$. Obviously the separation pattern for the digits is quite different than this in the figure 5. Thus, the fusion of all these MFCC's could be very useful for the proper separation of data and probably lead us to a linear separable data set if combined properly.

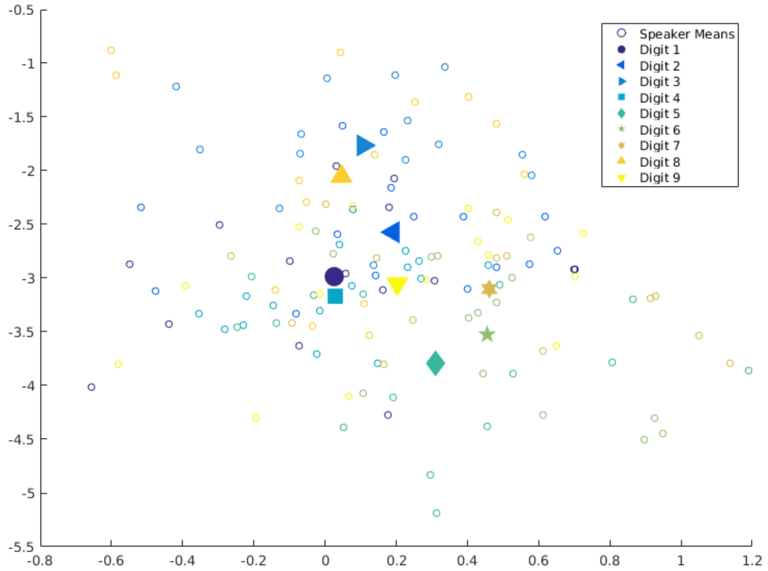

Figure 5: Mean Values for all Digits ($C_i(n1)$ and $C_i(n2)$)

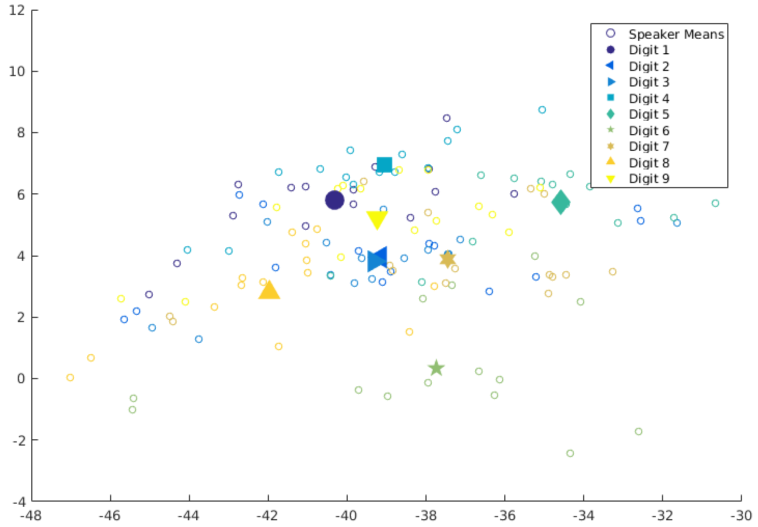

Figure 6: Mean Values for all Digits ($C_i(0)$ and $C_i(1)$)

References

[1] Koustav Chakraborty, Asmita Talele Prof. Savitha Upadhya. *Voice Recognition Using MFCC Algorithm.* International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Volume 1 Issue 10 (November 2014)

[2] Kaberpanthi, Neeraj; Datar, Ashutosh. *Speaker Independent Speech Recognition using MFCC with Cubic-Log Compression and VQ Analysis*. International Journal of Computer Applications 95.26 (2014)