DT2119 Lab 1

Evangelia Gogoulou
gogoulou@kth.se

Velisarios Miloulis
miloulis@kth.se

April 12, 2018

1 Tasks

1.1 Mel Frequency Cepstrum Coefficients

The intermediate steps of computing the Mel Frequency Cepstrum Coefficients for the utterances "one" and "five" for man speaker are illustrated below:

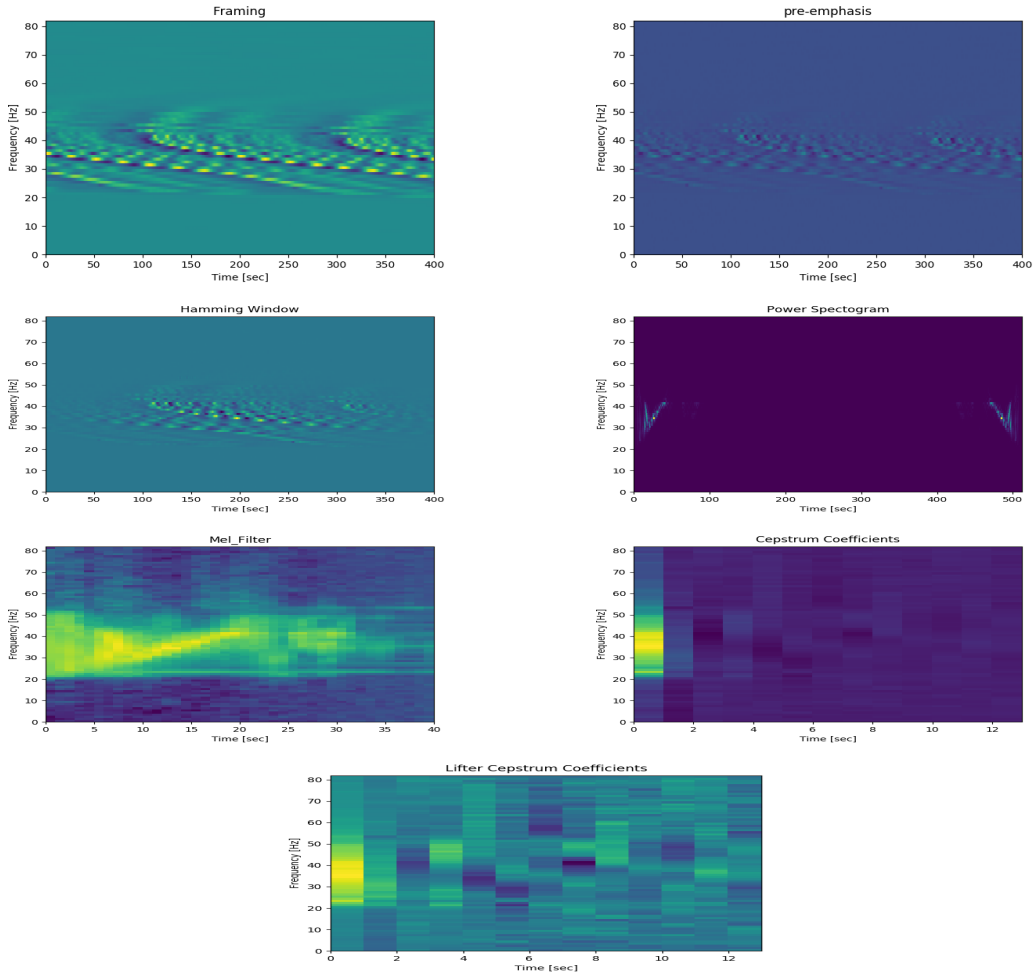

Figure 1: MFCC process for utterance 'One' - man speaker

Pre-emphasis coefficients: The numerator is a 2-D array [1, -p] and the denominator is 1-D [1]. Hamming window:For smoothing the signal, by zeroing values outside the window's interval

In the figures above we can clearly see that the intermediate results of MFCC computation differ for each utterance. However, the difference is much more clear after the application of liftering transform. After comparing the two graphs, we can see that the overall filterbank energy is higher for utterance "one" than utterance "five". However, in both cases coefficient one has the higher energy.

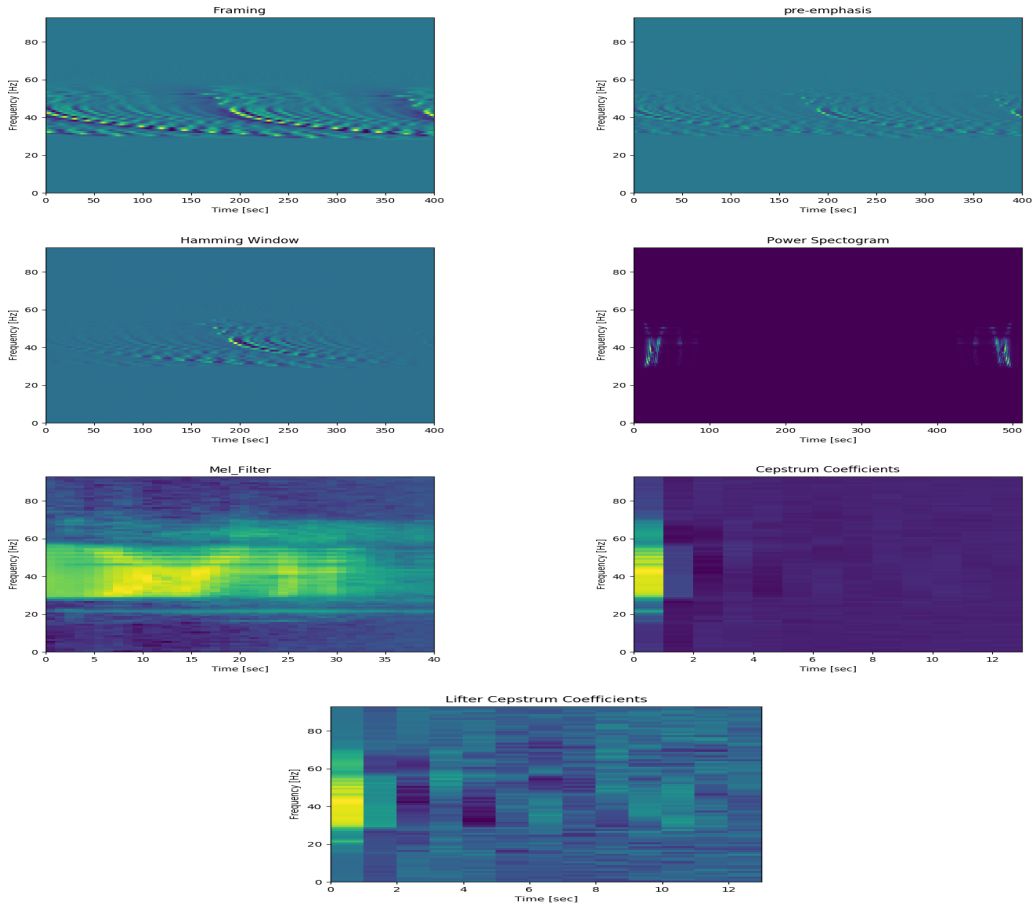

Figure 2: MFCC process for utterance 'Five' - man speaker

1.2 Feature Correlation

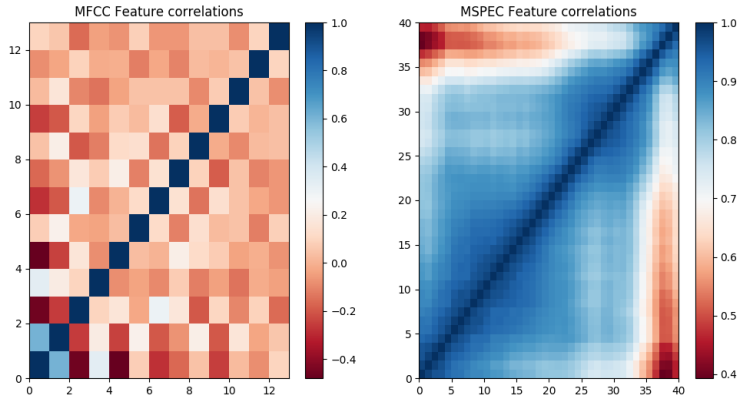

Figure 3: The correlations between the MFCC and `Mel` filterbank features.

Fig.3 shows the correlations of the between the MFCC features and the Mel filterbank features. From the visualization, we can see that the MFCC features show very low levels of correlation (except for the diagonal, of course). There are some cases of higher negative correlation, but overall the positive correlations can be assumed to be non-existent. Therefore, the assumption of diagonal covariance matrices, i.e modelling the samples of each utterance with independent Gaussians, is justified.

In contrast, the Mel filterbank features are highly correlated to each other. That is to be expected, since we obtain those features before applying the Discrete Cosine Transform to them. The DCT is used to decorrelate the Mel filterbank features, while dropping all the coefficients after the 13th. It is worth noting that the MFCC scheme uses the DCT instead of another transform (a DFT for example) because of the high spectral compaction it exhibits, gathering all the spectral information in the first few coefficients, and thus achieving some degree of compression of the signal, allowing for dropping the coefficients 14-40. This loss of information and decorrelation has been criticized in the context of Speech Recognition with ANNs (for example,

Abdel-Hamid et al, 2014).

1.3 Comparing Utterances

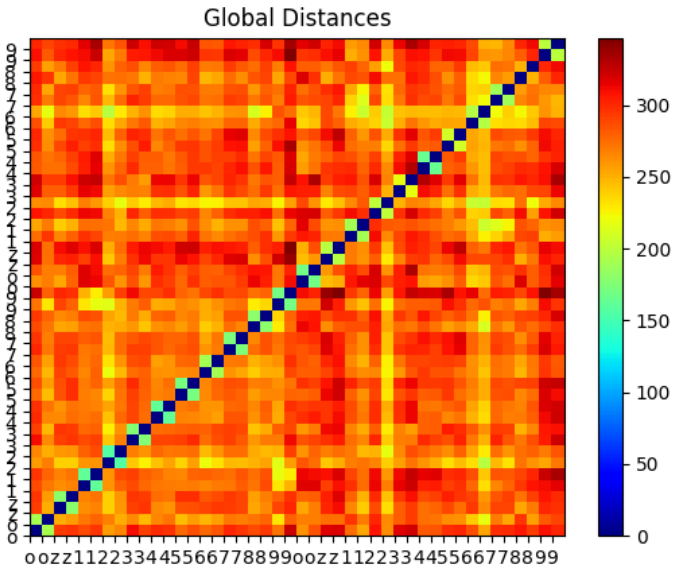

Figure 4: The global, pairwise distances between utterances, computed using DTW using pre-computed local Euclidean distances between two utterances. The first 22 digits from the start of the axes represent the male speaker, while the later 22 the female speaker.

Fig.4 show the global distances between utterances. As expected, the distances between the same utterances are zero. The first and second utterance of the same digit by the same speaker are also very close, but there is a measurable difference. The female speaker samples exhibit greater differences between two consecutive utterances of the same digit, with the digit `8` being as different between the consecutive utterances of it as any other completely different digits.

The lower left segment of the figure corresponding to the distances between the digits uttered by the male speaker exhibits more consistent distances between different digits. While an as clear differentiation cannot be said to exist in the upper right segment, corresponding to the distances between digits uttered by the female speaker, the lowest distances between the utterances for the female speaker are found there.

The distances between the digits among the different speakers are generally quite high. This leads to the conclusion that the Euclidean distance as a metric could separate the digits, conclusion reinforced by Fig.5. We can see that by and large the same digits uttered by the same speaker are grouped together at the leaves of the dendrogram, while the clusters formed by merging at the parent level do represent phonetic similarities between the different digits.

1.4 Exploration of Speech Segments with Clustering

For this question, we created Gaussian Mixture Models with various numbers of components. In all cases, a diagonal covariance matrix was used, motivated by the weak correlation of the mfcc coefficients.

One example of the posteriors generated for mfcc coefficients of the uttered word "seven" delivered by a woman speaker is illustrated in Fig.6

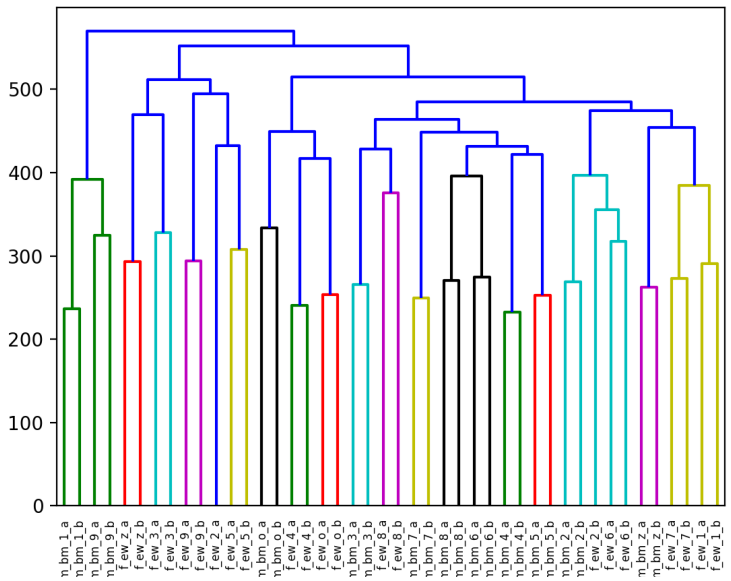

Figure 5: The dendrogram representing the hierachical clustering performed o the global distances between utterances.

In each of the graphs in Fig.6, each color corresponds to the posterior generated by a different gaussian component. In all cases, there are coefficients with only one non-zero posterior, which means these coefficients belong to a specific gaussian component. However, there are also coefficients which are generated by more than one gaussian components.

The resulting labels of gaussian components which correspond to the maximum posterior are illustrated in Fig.7. Each graph corresponds to the case of the uttered word "seven" delivered by a single speaker (man or woman).

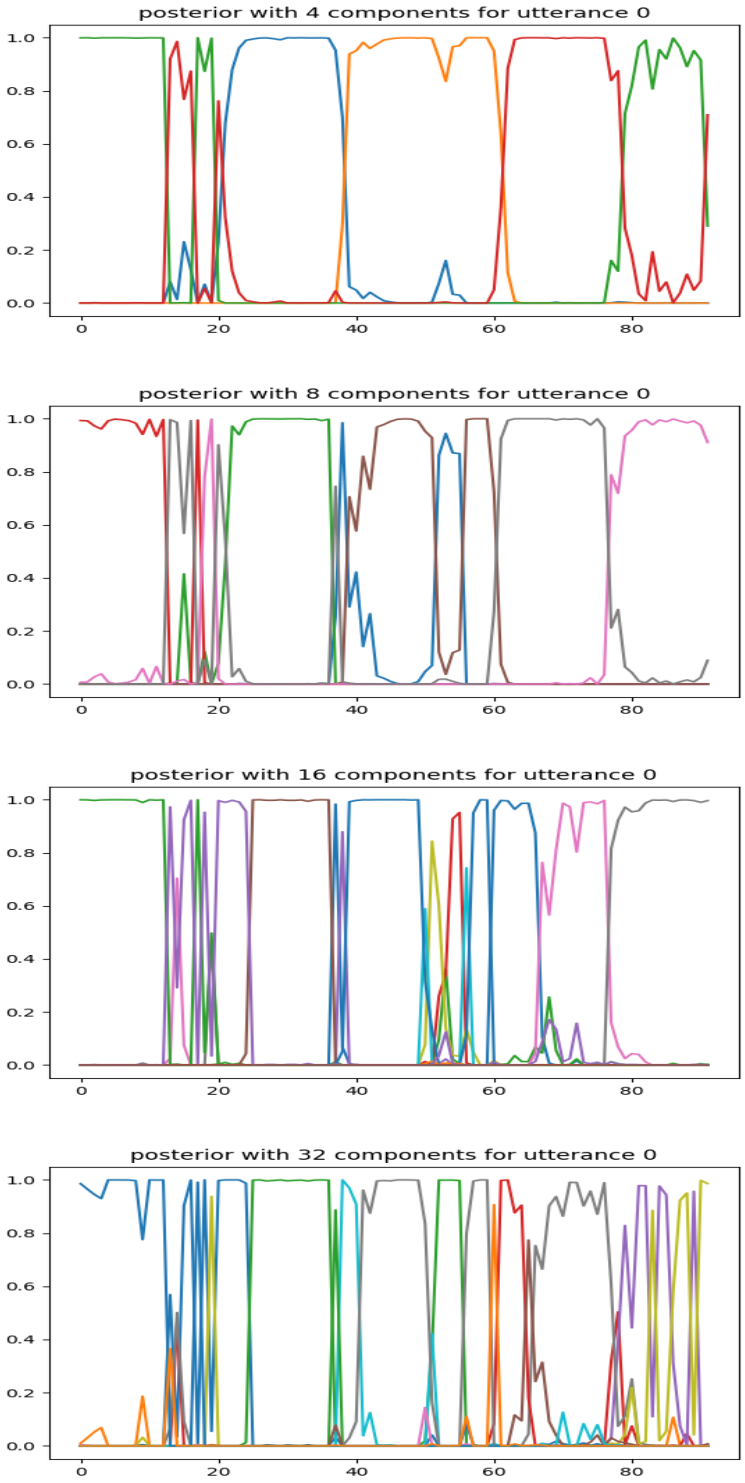

Figure 6: Posterior probabilities generated by GMMs with various numbers of components. In all cases, the examined utterance contains the word "seven" delivered by a woman speaker.

Each line in the graphs in Fig.7 corresponds to one trial of a single speaker (man or woman) who utters the word "seven". In both graphs, both posteriors follow the same trend. This result was expected, since both the uttered word and speaker remain the same.

The graph in Fig.8 shows the labels assigned by the GMM to the mfcc coefficients. Each line corresponds to the same uttered word delivered by a different speaker (man or female).

Comparing the two cases, we can identify some common patterns in the trend of the posterior generated by our GMM. This is due to the common characteristics between all human voices, independently of gender. Looking at each mfcc coefficient seperately, we can see that it belongs to a different component when the speaker changes.

The case of different utterances delivered by the same speaker is studied below:

9 illustrates the GMM labels assigned to mfcc coefficients of different utterances. In some cases, the assigned gaussian component for different utterances remains the same. This observation is related to the phonetic similarities between digits "five" and "seven".

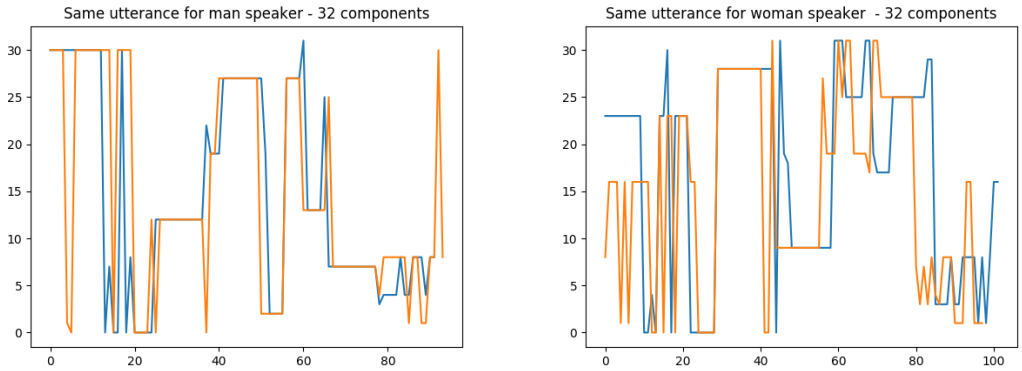

Figure 7: Labels assigned by a GMM with 32 components. Each graph corresponds to the case of a different speaker, when the examined utterance remains the same.

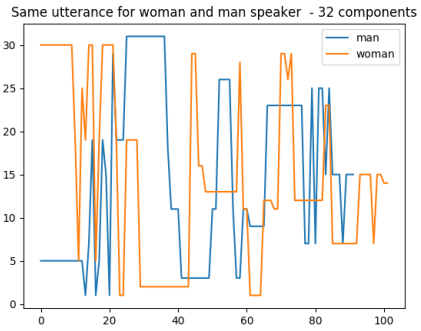

Figure 8: Labels assigned by a GMM with 32 components. Each line corresponds to the case of a different speaker, when the examined utterance remains the same.

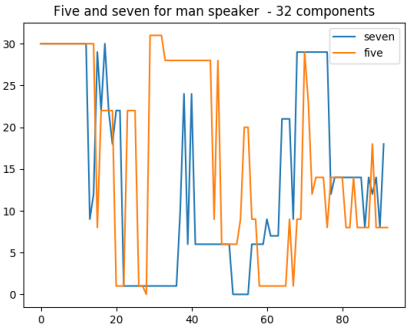

Figure 9: Labels assigned by a GMM with 32 components. Each lines corresponds to the case of a different utterance (five, seven), when the speaker remains the same.