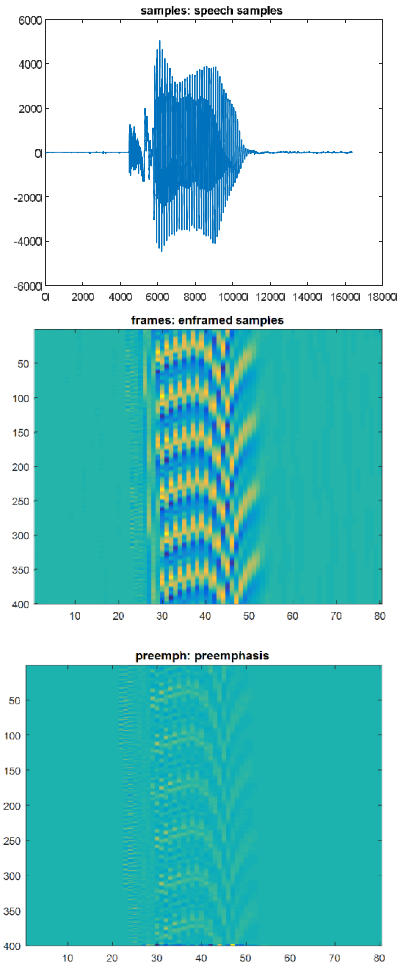
DT2118 Lab1: Feature extraction

Masoumeh Poormehdi Ghaemmaghami

Mateusz Buda

1. Mel Frequency Cepstrum Coefficients step-by-step

Steps of computing MFCCs are implemented. First the example array to double check that our calculations are right, are used. The results for example array are shown in Fig.1.



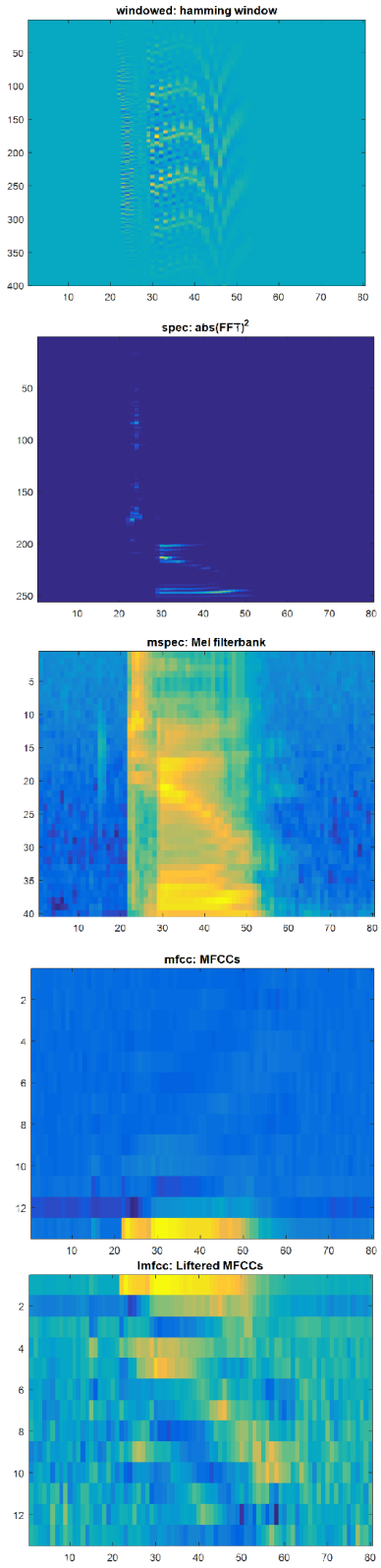


Fig.1: Evaluation of MFCCs step-by-step

1.1 Enframe

First step is framing the speech signal. We assume window length of 20 milliseconds and shift of 10 ms for the utterance example ['samples'].

$$\text{window length} = \text{winlen\_time} \times \text{sampleRate}$$
$$\text{shift} = \text{shift\_time} \times \text{sampleRate}$$
$$\text{sampleRate} = 20000$$

1.2 Pre-emphasis

In order to flatten speech spectrum, we used a pre-emphasis filter with pre-emphasis coefficient equal to 0.97, which is a high-pass FIR filter described in Eq. (1).& (2)

$$H(w) = 1 - 0.97w^{-1} \qquad (1)$$

$$Y[n] = X[n] - 0.97X[n-1] \qquad (2)$$

This filter is applied to each frame in the output from the previous part.

Hamming Window

In order to reduce the discontinuities of the speech signal at the edges of each frame, a tapered window is applied to each one. The most common used window is Hamming window, described in Eq. (3) and shown in Figure 2.

$$\mathrm{W(n)} = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right), 0 \le \mathrm{n} \le \mathrm{N} \qquad (3)$$

The window length is :$L = N + 1$, N is number of samples in each frame.

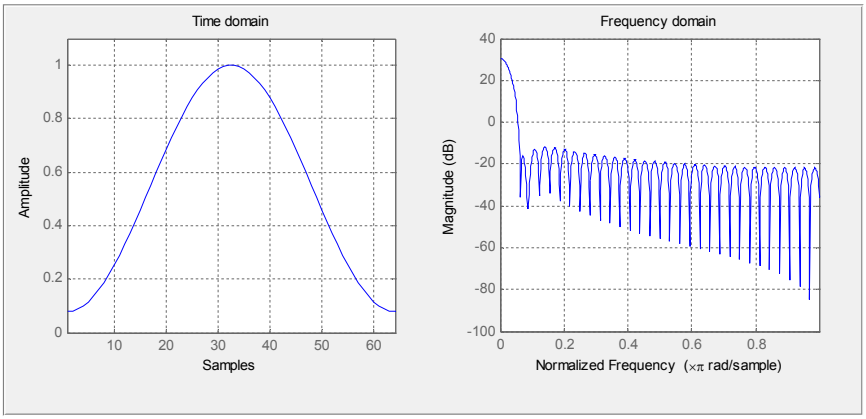


Fig. 2: Hamming window

1.3 Fast Fourier Transform

In order to convert each frame of N samples from time domain into frequency domain, the Fast Fourier Transform (FFT) of the data is computed and then the squared modulus of the result is computed. FFT length of samples is 512. According to the Sampling Theorem, f_{max} is defined as Eq. (4)

$$f_{max} = \text{sampleRate}/2 \qquad (4)$$

1.4 Mel filterbank log spectrum

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale as shown in figure 3 is then performed. A set of overlapping triangular bandpass filter, which contain 13 linear and 27 logaritmic filters is used. The filters are applied to the output of the power spectrum from the previous step and for each frame the natural log of the result is taken.

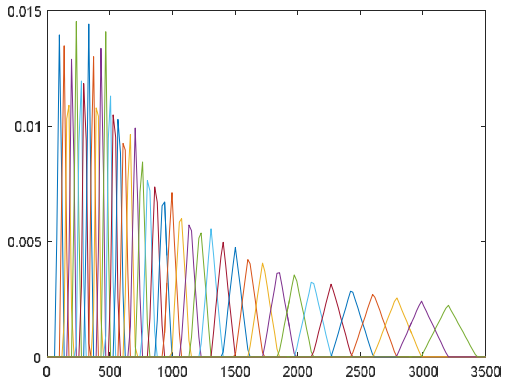


Fig.3: Mel filterbank

1.5 Cosine Transofrm and Liftering

In this step, the log filter-bank amplitudes are transformed into cepstral coefficients. The result is called Mel Frequency Cepstrum Coefficient (MFCC). In this way, the Discrete Cosine Transform is applied to the outputs of the filterbank. Coefficients from 0 to 12 (13 coefficients) are used. Then range of the coefficients is corrected by using liftering.

2. Feature Correlation

All the MFCC frames from all utterances in the tidigits array into a big feature $N \times M$array are concatenated. N is the total number of frames in the data set and M is the number of coefficients. Correlation coefficients between MFCC features and also Mel filterbank features are calculated and illustrated in Fig. 4 and 5 respectively.

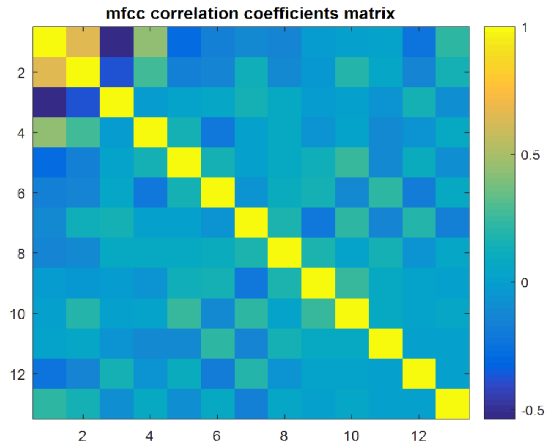


Fig. 4: Correlation coefficients between MFCC features

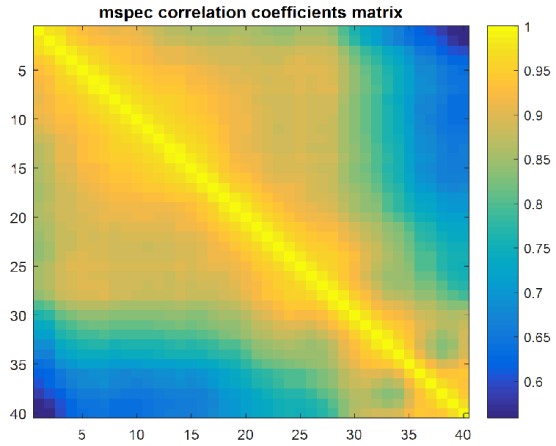


Fig. 5: Correlation coefficients between Mel filterbank features

By comparing the results of Correlation coefficients between Mel filterbank features and between MFCC features, it is clear that Mel filterbank features are more correlated. Also it is clear that assumption of diagonal covariance matrices for Gaussian modelling is justified. A Gaussian model's shape is seen in the Correlation coefficients between Mel filterbank features in Fig.5.

3. Comparing Utterances

3.1 Dynamic Time Warping (DTW)

Given two utterances of length N and M respectively, accumulated Euclidean distances between MFCC vectors in the first and second are calculated. For this purpose local Euclidean distances between MFCC vectors in the first and second utterance is computed and it is illustrated in Fig.6 which is related to word ‘o’ with two speakers. Then global distance between utterances with the Dynamic Time Warping algorithm is calculated. Also Accumulated Euclidean distances for word ‘o’ for two speakers is illustrated in Fig.7. Global distance between utterances is shown in Fig8.

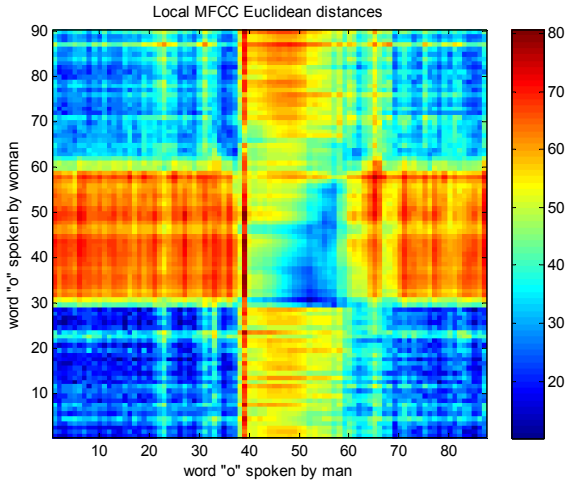


Fig. 6: Local MFCC Euclidean distances for word ‘o’ with two speakers

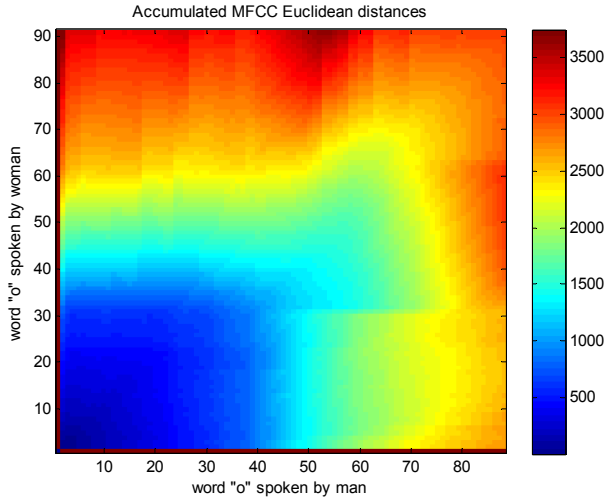


Fig. 7: Accumulated MFCC Euclidean distances for word ‘o’ with two speakers

This picture illustrates that, distance separates digits well even between different speakers. Also it shows that, speaker man is faster than speaker woman.

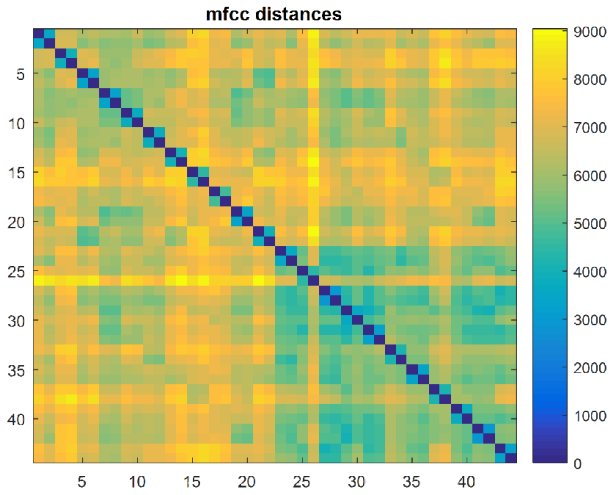


Fig. 8: Global MFCC distance between utterances

As it is clear, in the global distance, the diagonal of the distance matrix is zero (dark blue points) because it is the distance of the same utterance. Also the second up and down diagonal of the distance matrix (bright blue) are too small, because they are related to repetition of the digit. In the clustering small distances are related to same cluster.

3.2 Hierarchical clustering

Hierarchical clustering on the distance matrix using the linkage function from Matlab is done. The function dendrogram in Matlab is used for displaying the result. This result is illustrated in the Fig.9.

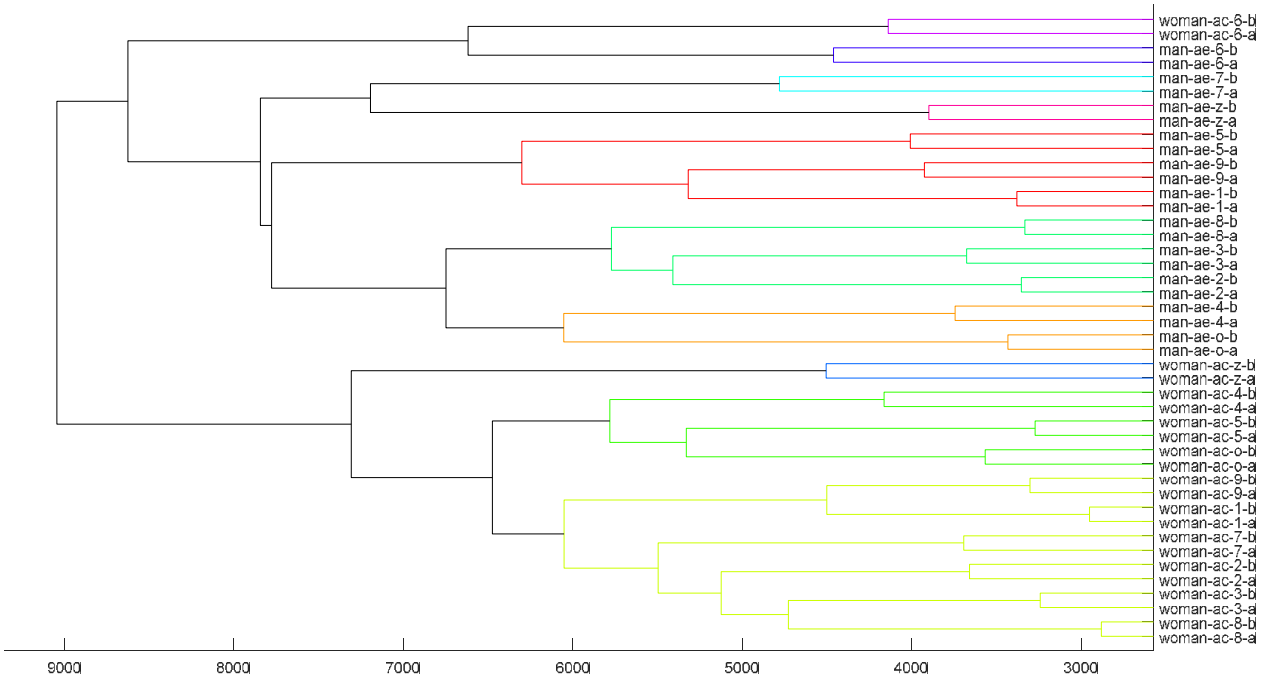


Fig. 9 clustering tidigits database

The result shows that utterances with same gender, speaker and digit are clustered correctly in a specific cluster and the repetition is in the next level. In the other words, at the first level of clustering, we got two utterances of the same digit and speaker. At the next level we have a clustering of man and woman utterances. Only digit 6 was correctly classified between speakers and based on the digit.