Project 2 - Speech recognition

Péter Á. Bánkuti

2019. 05. 15.

ISEL - Instituto Superior de Engenharia de Lisboa

Author's Note

This project was created in context of the Speech Processing subject of Instituto Superior de Engenharia de Lisboa, held by prof. Carlos Eduardo de Meneses Ribeiro, and the main source of information was his book "Processamento Digital de Fala", for further information on the formulas and theory, consult the book.

Introduction

The aim of this project was to create a simple speech recognizer from a given audio file set and test it, with the use of a different test audio set. Where, the audio set is composed of a short audio files each containing a single vocalized letter or number. Furthermore, each letter or number has multiple pronunciatons from different speekers.

To acheive this goal the audio files had to be preprocessed. The inital step was locating the portion in the file where speech might be occuring. Afterwards, the speech had to be segmented and for each of those segments the autocorrelation had to be calculated. With the use of these autocorrelation, the "distance" of the audio files could be estabilished. Which in turn could be utilized to recognize the content of an audio file. The recognition would happen in a way where the unknown audio file's distance to each of the files stored in the training set is calculated, and from those, the lowest values would point to the match with the highest probability.

Stages

1. Speech extraction form audio file

In the first stage of the project the speech content is extracted from the audio file. The main similarities within the audio files are the content in the middle, the leading and trailing white noise.

The speech is extracted by checking the elevation of the amplitude of the signal. The extraction starts by checking wether the amplitude is higher than a set of increasing treshold values if all the tresholds are passed by than the extraction is valid, and we wait until the the signal dies bellow the lowest threshold. By doing this only would result on fragmented extraction. In consequence, firstly the resulting speech fragments are concatonated if thay were close enough, and removed if they were short.

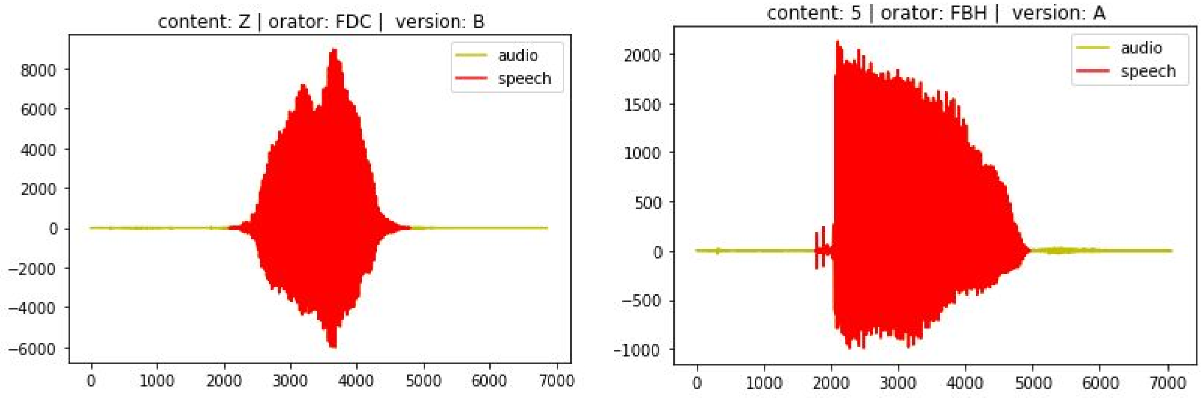

Fig 1.1-1.2.: *Files of letter Z and number 5*

2. Extracting the autocorrelation signiture of the speech

After the speech is extracted, it is iterated over in a fixed step size and a longer window size. Which will result in overlaping segments of equal length.

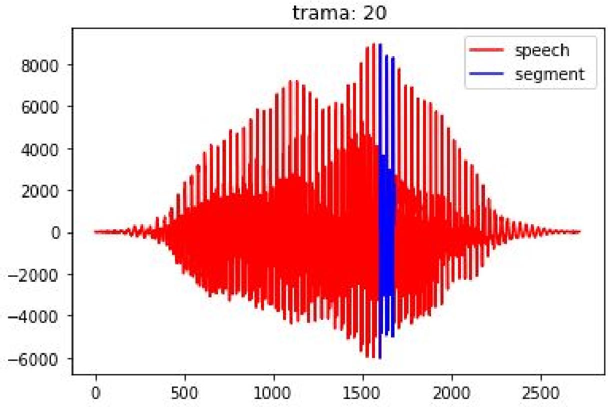

Fig 2.1.: *The 20th segment of the letter Z*

Following, each segment has its autocorrelation of order p calculated and the set of these autocorelation will represent the signiture of the speech.

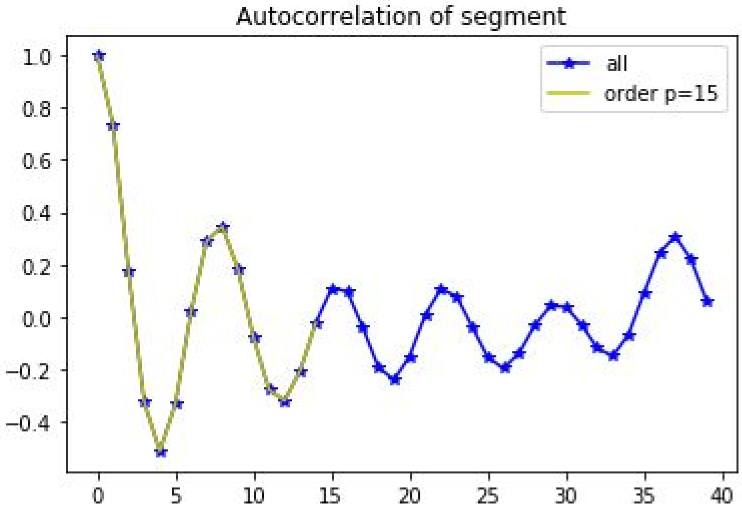

Fig 2.2.: *Autocorrelation of the 20th segment of letter Z*

3. Distance calculation

The “distance” between two speeches gives a single numeric value representing the similarities between those speeches, consequently it can be used for classification.

The first step is having the autocorrelation signiture of each of the speeches, than calculate the distance between each segment’s autocorrelation against the other speech’s segment’s autocorrelation. However, not every combination has to be calculated (the last segments of one speech is not comparable to inicial segments of the other).

The figure bellow (*fig 3.1*) shows the distance map between the same file while the other (*fig 3.2*) shows the distance map of the between two different files with the same speech content, spoken by the same person. It can be seen that the more similar the speech is the lower is the distance among the respective segments.

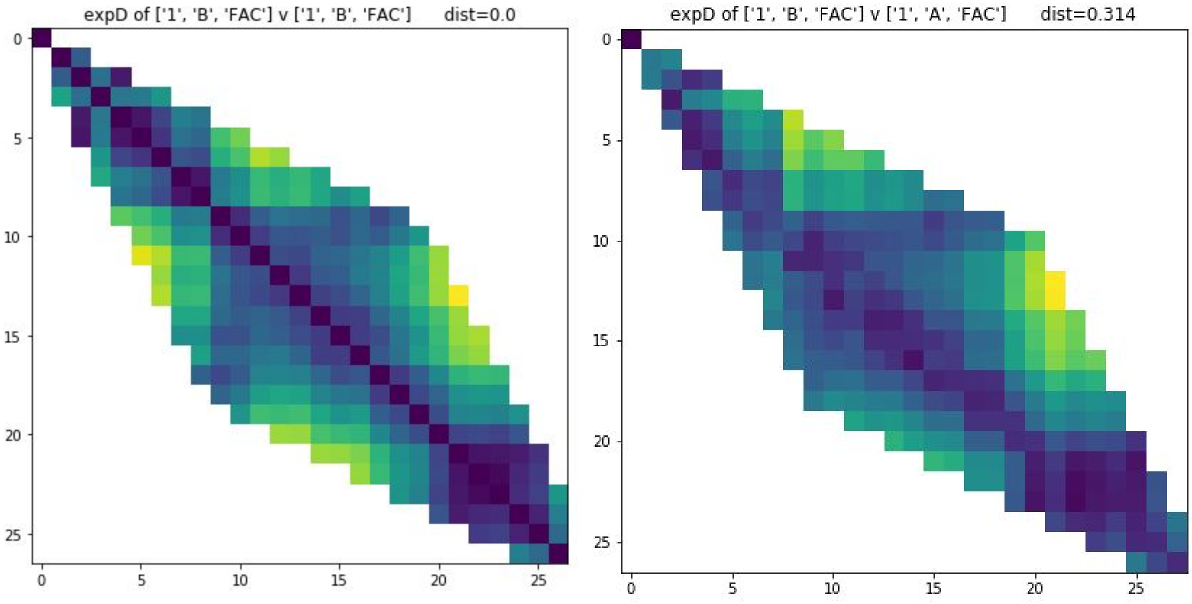

Fig 3.1-3.2.: *Distances between the same file, and between the same content*

While the figure bellow shows the distance map between the letter Z and number 5.

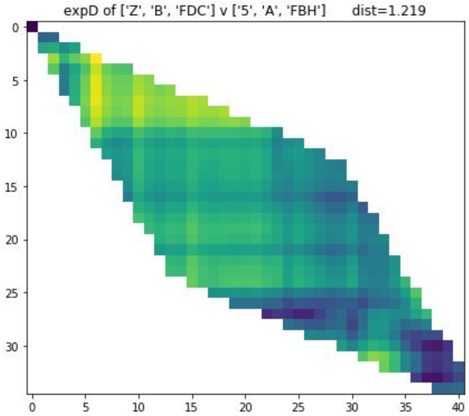

Fig 3.3.: *Distance map between Z and 5*

To calculate the distance is important find the route with the minimal cummulative value. This happens in a stepwise manner. The route starts at (0,0)=0 coordiante and the next posible step can be towards (0,1)=inf, (1,1)=value and (1,0)=inf. From these the one with lowest numeric value will be selected, than the position is updated, the value is added to the comultive and the step count is increased 1. In the end the normalized distance is calculated the commulative value by the step count.

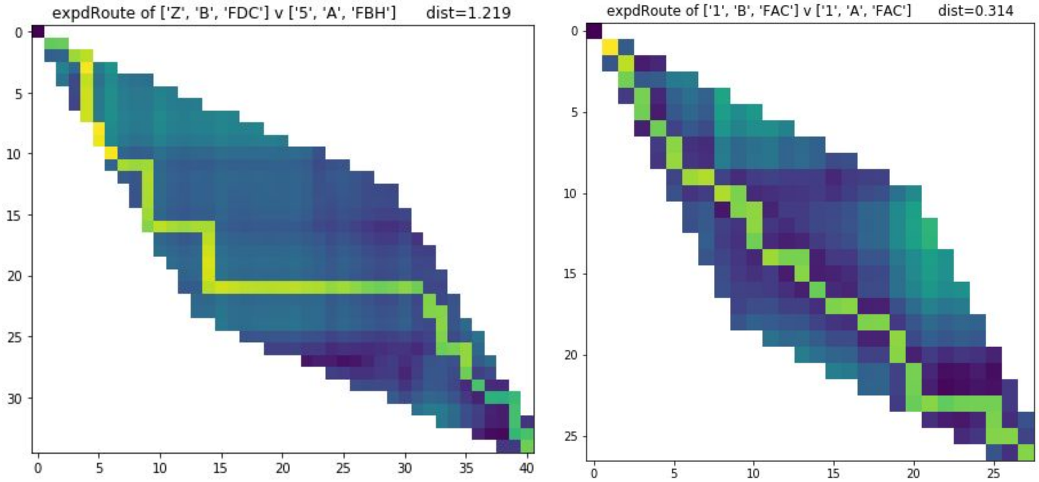

Fig 3.4-3.5.: *Distance map between Z and 5, and two different 1s*

4. Classification

The classification happens by comparing the distances of the unknown speech to the labeled ones. On the figures bellow, the first 20 files of the test set is classified against the first 20 of the training set. The figure to the left shows the distances between those 20-20 files, while the right image represents the lowest distance (classification), yellow if is a match or blue if is a mismatch.

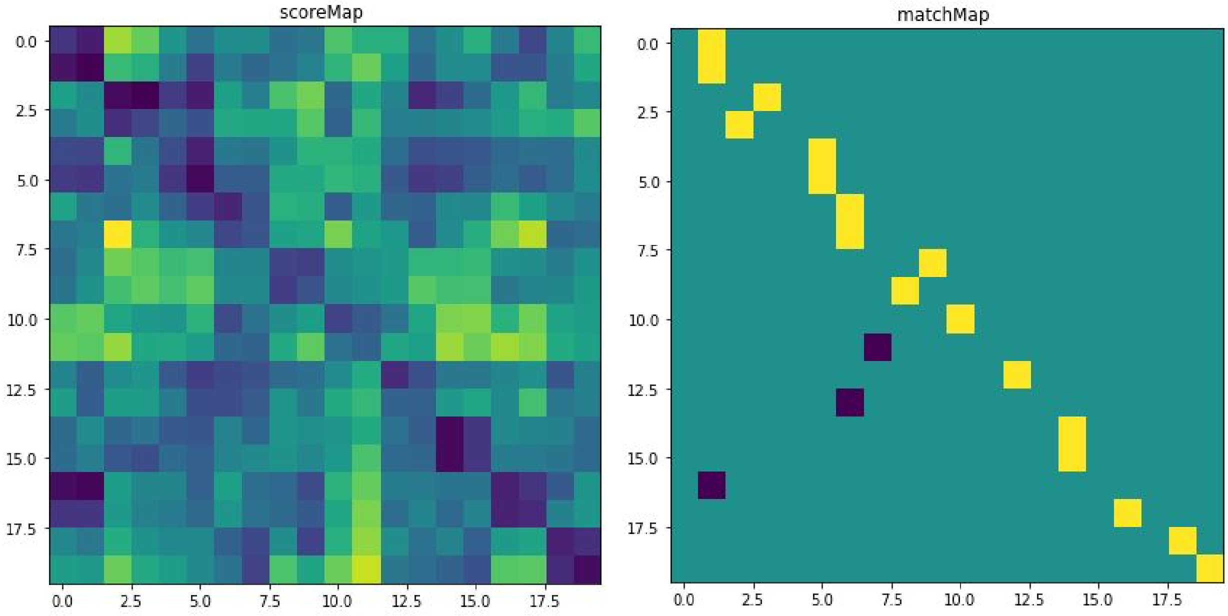

Fig 4.1-4.2.: *Distances of the training set and test set.*

Results

By using all the files of the training and test set, the correct match rate of the classification is 82.95%.

References

1. Carlos Eduardo de Meneses Ribeiro.: Processamento Digital de Fala. Instituto Superior de Engenharia de Lisboa