MUSIC SYNCHRONIZATION

Authors: An Nguyen, Afolaranmi James

Tables of content

1/ Introduction

The human auditory system is very effective in identifying a musical piece performed by different people (this could also be different instrumental renditions) at a different pace. The aim of this project – music synchronization- is to replicate this human ability using audio processing tools.

Dynamic Time Warping is an algorithm used in measuring the similarities between two signals (which might be of different speeds). DTW is employed in this project to find the warping path between the extracted features of the given audios of different lengths (ie different speeds). The Librosa library has a function that can carry out DTW efficiently.

The aim of this project is to find the similarities between the reference audio and the comparing audios then set the pace of the other audio to that of the reference audio.

2/ Implementation

a) <u>Stages of implementation</u>

1. Choose reference signal: the faster (ie shorter) signal autti.wav is chosen as the reference signal. ****In this report the reference signal is also called the synthesis frame and the signal is compared to as analysis frame.**

2. Load audio: the signal to be compared are loaded into the coding environment using the Librosa library.

3. Extract feature: the important feature that characterizes the audio signal is extracted either as chroma or MFCC. Both methods were employed but the chroma gave a better result compared to MFCC.

4. Align extracted feature: in this stage, the DTW is used to find the similarities between the two audio signals of different lengths. After the DTW is applied a warping path is obtained in a cost matrix.

5. Time stretch: The stretch factor R for different frames in the cost matrix is obtained by taking the ratio of the synthesis frame to the analysis frame. This was done to account for the inconsistency in the speed of the singers in the audio signal given. The stretch factor R could also be obtained by taking the ratio of the total length of the reference signal to the length of the comparing audio but this will only assume that the singers in the audio maintain a constant speed throughout the audio.

6. Realign time-stretched version: the DTW of the time-stretched version of the second audio between the reference audio is calculated and the warping path is compared to the initial path obtained.

b) <u>Assumptions</u>

1. Noise was added to the audio signal to prevent NaN values in the cost matrix. The noise compensated for the quiet portions of the signal.
2. The signal obtained after time-stretching is not the exact size as the reference signal (about 0.2% audio signal is lost), this is due to the one-to-many and many-to-one mapping of the DTW.
3. Different time-stretch factors R for the frames in the cost matrix to account for the inconsistent speed of the singers.

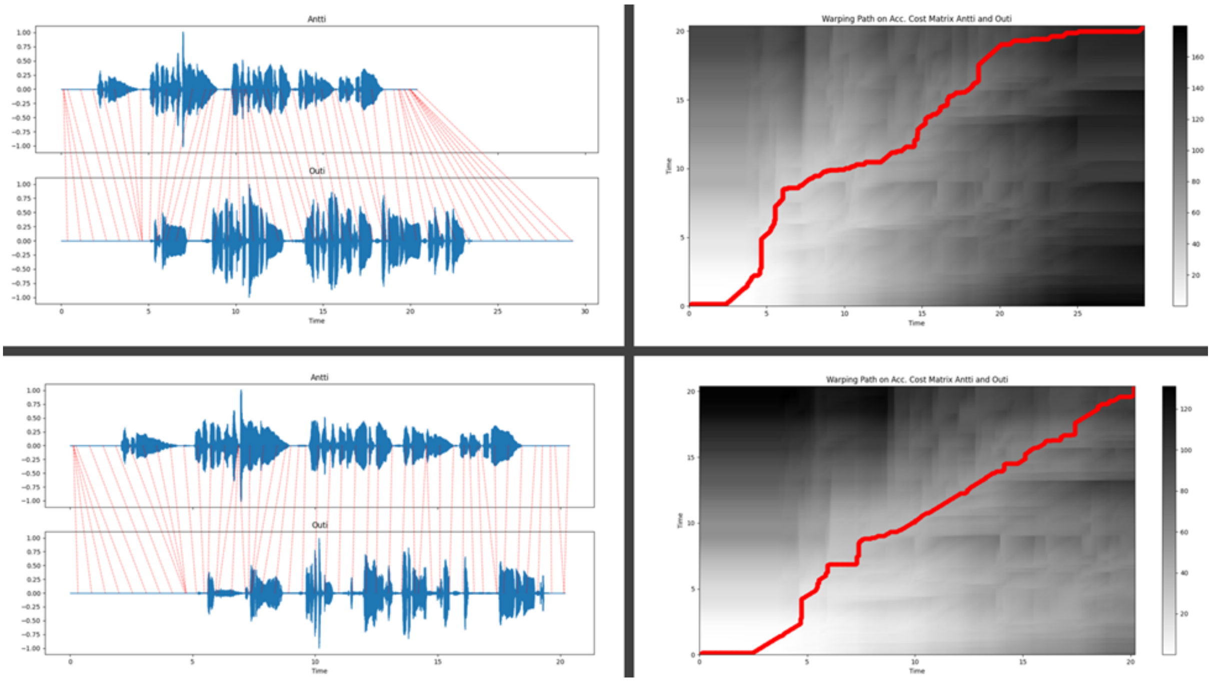
c) Evaluation

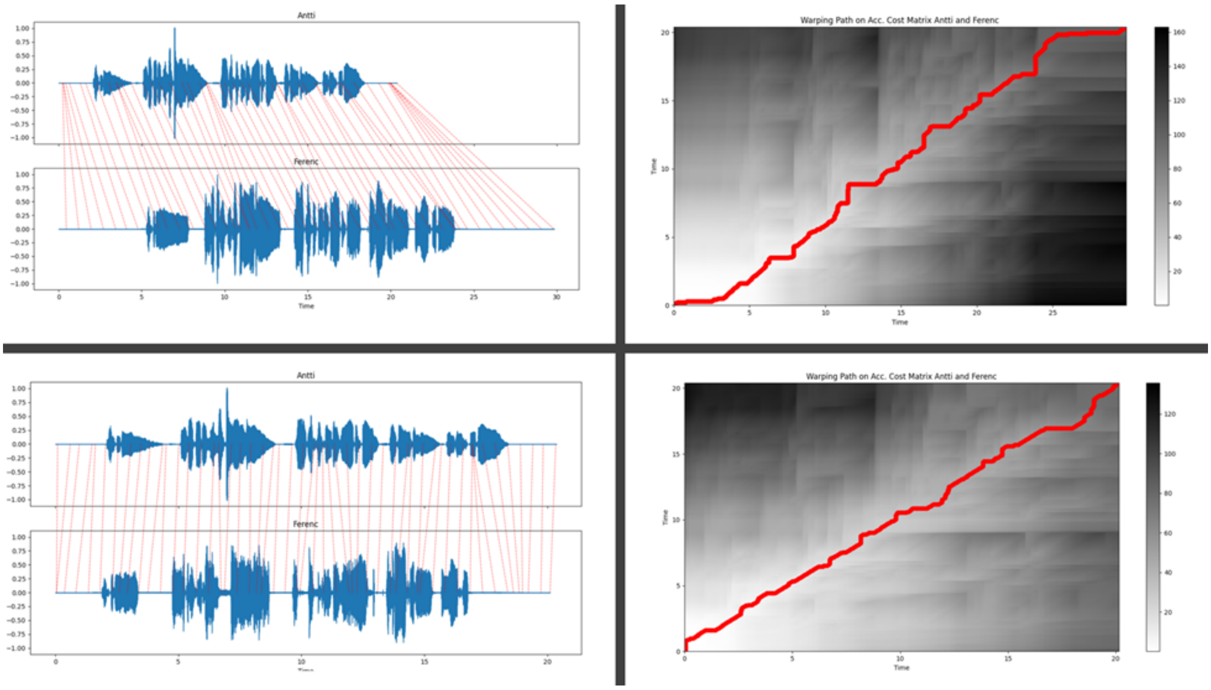
The evaluation includes the difference between the initial warp path and warp path after processing. Secondly, the plot of the connection in the time-domain so that we know if the signal happens where it should be or not. Thirdly, listen and compare 2 audio files to see if a syllable happens at the same time or not.

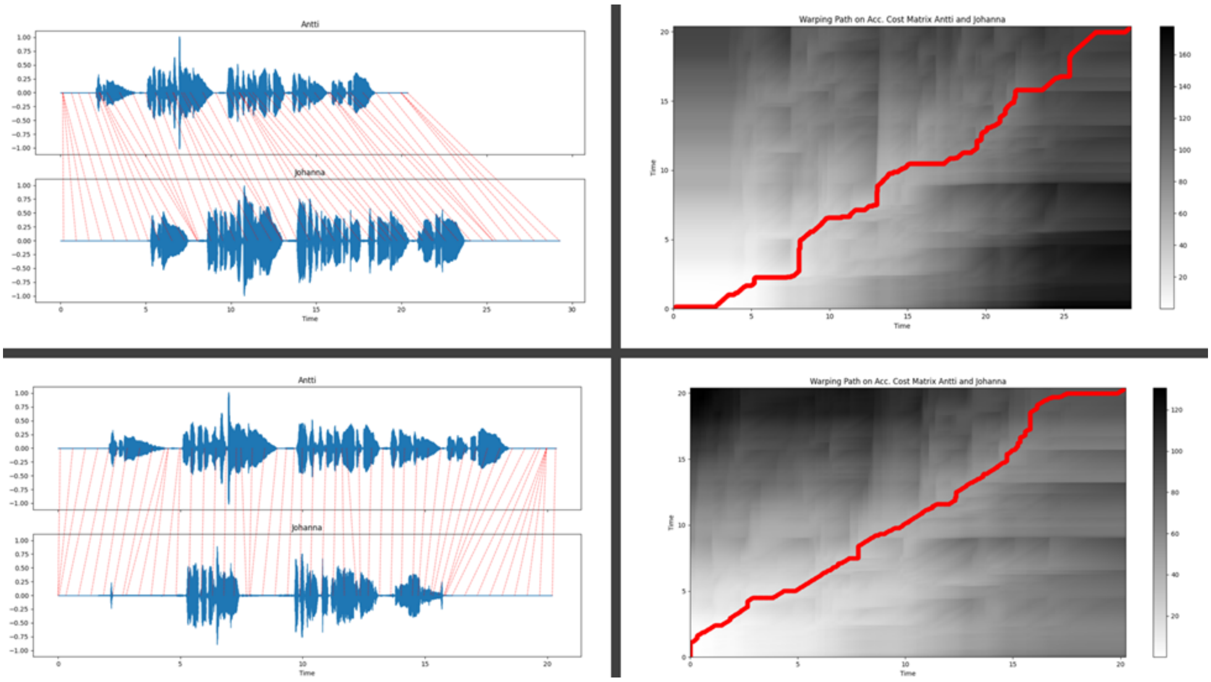
4/ Conclusion

The synthesis audio we chose is antti_yesterday.wav (called A) has a big difference in length compared to other audio.

- The following graph in the layout is 2*2, where the top left is (1,1). We have: (1,1) and (1,2) are time-amplitude plots when connect visually before time-stretch. The remaining are plots after time-stretching







We can see that the time-stretch process makes the warping path much better and the visual connection between the graph is more correct (according to the sound packets-the singing part). When we listen to the output audio, the corresponding sound is played at the correct timing as in A but it sounds very lagging because the time-wrap takes some part too fast or too slow.

Therefore, we can conclude that this approach may be better for an instrument than normal singing.