Assignment report

Music Synchronization

Trung Van | H293914 | trung.van@tuni.fi

Ara Jo | 468955478 | ara.jo@tuni.fi

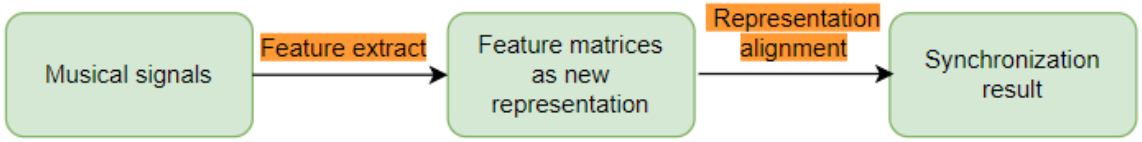
I. Introduction

This project is the final project assignment for COMP.SGN.120-2021-2022 Introduction to Audio Processing at TAU. The chosen topic is topic3 Music Synchronization. The task can be described as synchronizing different perfomances of the same musical piece (Yesterday-The Beatles). After finishing the project, students are expected to revise knowledge from lectures and to have some hands-on experience on:

- Examine audio data and select suitable method(s) for processing
- Audio signal annotation, feature extracting (chroma-graph, MFCCs), time-stretching, and sequences alignment
- Formalize conclusion based on collected results.

II. Theoretical background

In this project we applied the simplest version of music synchronizatio which includes two distinct steps which can be seen from the figure below:



The following content is our summarized version based on our understanding of methods used in this project

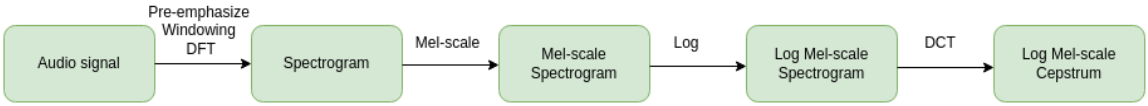
1. Feature extraction

Feature extraction methods considered and selected in this project are based on *Chroma based feature* and *Mel Frequency Cepstrum* due to the following reasons:

- Signals are musical hence, Chroma feature (especially with Constant Q Transformation) which has proven its efficiency in this topic should be implemented in the first attempt.
- The given audio signals contains purely human voice without any additional instrumental signal. Therefore, we would like to try MFCC as an attempt to treat the audios like normal talks to extract the features then apply the alignment.

Chroma based feature's main idea is to aggregate the coefficients of the sample belonged to the same pitch class profile. Using this method, one would be able to highlight the visibility of presented notes (quantized into 12 bins) along the time axis at the cost of signal level information. One of the great contribution for the effectiveness of chroma based feature is the universal rule and notation applied on musical content. This restriction allows the compression in signals because narrowing down the scope into chroma based structure does not largely change the musical identity of the signals.

MFCC focuses on the energy distribution of a signal taking into account the critical band of the auditory system and log-power perception. The typical procedure to form an MFCC can be seen from the figure below:



The main idea of cepstrum is to apply another DFT , or preferrablely DCT, over the log spectrum of a signal. Cepstrum analysis is motivated by the observation of harmonic structrure in the spectrogram which is contributed largely by the resonance in the vocal tract. By capturing this macro structure, one can obtain descriptive information about vowels in an audio signal. The soure-filter model is applied in cepstral domain with glottal excitation as soure and the vocal tract as filter part.

2. Representation alignment

Dynamic Time Warping is chosen for the representation alignment step. DTW based alignment can be summarized with the following steps, for a given pair of sequences:

1. Pre-calculate the feature vectors for each sequence
2. Generate the cost matrix
3. Form the warping path with minimal aggregated cost

III. Implementation

1. Describe given data

There are files of four people singing the song 'yesterday' by the Beatles. Each song has a different length. They are singing the same part with the same lyrics with different speeds and pitches. When we hear the original sound file, we can recognize that file 1 antti's yesterday is very close to the original song. For the file 2 ferenc's yesterday, the pitch sounded lower than the antti's one. File 3 Johanna's yesterday sounded very similar to ferenc's one. It sounded transpose down compared to antti's one. File 4 outi's yesterday sounded transposed up. It sounded very high compare to any of the other files. With these four different files, we will sync the files with the time stretching method, MFCC, chroma features.

2. Signal processing

For signal processing, we used 4 different processes to analyze the audio files. We divided processes into preprocessing, feature extraction, alignment, visualization.

First, we preprocess the audio files. Since all the files have different starting and ending points we first remove the silence of the first and end of the files. When we plot the audio signal after we trimmed, we could observe similar patterns for each lyric.

For the next step, we used three different methods to extract the feature. We used chroma features, MFCC features, and time stretching. For extracting chroma features, we used the librosa library (`librosa.feature.chroma_cqt, librosa.sequence.dtw`) to extract chroma features, and perform dynamic time warping. We found the optimal path to reduce the cost matrix and calculate the right time window for each file. We observed the optimal path on the chroma feature graph.

For MFCC features, we also use librosa library (`librosa.feature.mfcc`) to extract MFCC features. After extracting features, we applied dynamic time warping and identify the optimal path. We also discovered the optimal path for MFCC features of files.

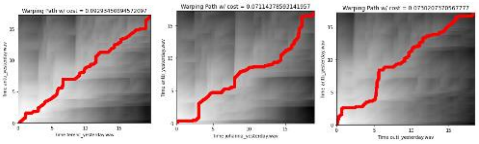
For time-stretching, we use antti's yesterday for base audio. We used (`librosa.effects.time_stretch`) for time-stretching. Then apply MFCC and chroma to see the difference of applying a time-strecthing function. Since we applied both MFCC and chroma features, we could compare the optimal paths to the previous unstretched version.

3. Evaluation

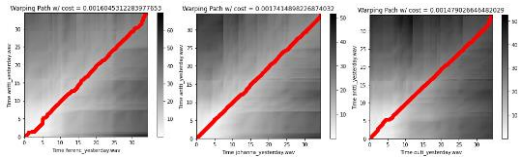
For chroma feature extraction, we used chromagram to see which chroma that each file has. By using dynamic time warping, we aligned these chroma features. Librosa library has the function to assign dynamic time warping and calculating cost function. By comparing two files' cost functions, we can find the optimal path of the two files. A similar method is applied to MFCC as well.

For chroma feature alignment, when we compare the same chroma feature window it synced well. For antti's yesterday and ferenc's yesterday they were synced 85% of each chroma feature window. However, when we compare antti's yesterday with Johanna's yesterday or outi's yesterday their sync

was not very good compared to ferenc's one. This is because even though it was the same melody, they sang with a different pitch which means they transposed down or up the song. This results in the chroma features not aligning very well. The following figures are plotting optimal warping path for each song comparison.



For MFCC the library use mel-frequency which can be more perceived by the human ear. This results in more natural consequence than chroma feature alignment. We can detect a significant difference when we compare cost functions. The optimal path was very linear, unlike chroma feature cost functions. Besides, when we plot the audio signals with a warping path, we could see the audio signal shape of the warping window is matching. The following figures are plotting optimal warping path for each song comparison.



For stretched signals, it was similar to the original signal. For chroma feature aligning, they were synced but only part of the signal. Once the warping path is not matched there is no way to match each windows again. For MFCC, the cost function was very linear as well.

4. Role allocation

We researched all the topics to choose one topic. After we chose topic 3. For chroma feature extraction and aligning we did separately and selected the optimized code. For analyzing the MFCC features, Trung implemented the function. For testing each warping window, I implemented the code to test the files are syncing in certain points.

IV. Conclusion

From our observation, due to the nature of given signal that contains only human singing excerpts. Both chroma based – from musical audio approach and MFCC- from the speech recognizing perform well as a task of feature extraction. The alignment with DTW provides quite promissing results with low costs which are consistant to listening re-assessment by us.