Assignment 1: (CSC3160)

Introduction

In this assignment, you will work on various speech processing tasks using the 'librosa' library. 'librosa' is a Python package for music and audio analysis, providing the building blocks necessary to create music information retrieval systems. For more information, visit the librosa documentation. The total score is 100 marks and it will be converted into **6%** for your final grades.

Library Installation

To begin, you need to install the necessary libraries. Use the following commands to install 'torch', 'torchaudio', and 'librosa':

```
!pip install torch>=1.2.0 # Install torch
!pip install torchaudio # Install torchaudio
%matplotlib inline       # Sets the backend of matplotlib to the 'inline'
    backend
!pip install librosa    # Install librosa
```

1. [20 marks] Extract MFCCs

In this part, you need to extract MFCCs (Mel-frequency cepstral coefficients) with different frame sizes and steps. The speech signal is at a 16 kHz sampling rate. Please use the audio sample in the assignment folder, `reference.wav`.

1. Extract MFCCs with frame sizes of 20ms, 30ms, and 40ms.
2. Plot the MFCCs of the 100th frame for each frame size.

2. [10 marks] Record yourself

Record yourself reading the following script and save it as `myspeech.wav`. Then, resample your recording to 16 kHz and 24-bit depth.

> "The examination and testimony of the experts enabled the Commission to conclude that five shots may have been fired."

You can use your microphone. Do not skip this question because you need this file for the following questions. FIY, if you use an iPhone to record your voice, you can convert the m4a file into a wav file by the following code.

```
import librosa
from scipy.io.wavfile import write as write_wav
fs = 16000
x, _ = librosa.load("./myspeech.m4a", sr=fs)
write_wav("./myspeech.wav", fs, x)
```

3. [20 marks] Pitch Estimation

Extract the pitch (fundamental frequency, F0) of the provided speech sample and your recording. Both audio samples must be at 16 kHz, and the hop length should be 200.

1. Plot the pitch trajectory of the provided speech sample (`reference.wav`).
2. Plot the pitch trajectory of your own recording (`myspeech.wav`).

You can use the YIN algorithm from librosa.

4. [50 marks] Dynamic Time Warping

Now you have the reference speech (`reference.wav`) and your recording (`myspeech.wav`). Use Dynamic Time Warping to align your recording to the reference speech.

1. Visualize the aligned mel-spectrograms (you can use any other acoustic features like MFCC or spectrograms) of the reference speech and your recording.
2. Explain how to obtain the time stamp of the word "Commission" in your recording using the result of the DTW algorithm and the time stamp of the word in the reference audio[1] (without listening to your recording). Implement it.

You can use the FastDTW tool to align two mel-spectrograms.

[1] "Commission" is spoken in [3.6s, 4.1s] of `reference.wav`