Lecture 4: Analysis of Audio Signals

Introduction

Topic of this week's lecture was analysis of audio signals. The lecture covered FFT, the most important algorithm in modern audio signal processing. Besides that it went over approaches we need to take in order to perform feature extraction. This included envelope detection, pitch detection, spectral centroids, beat tracking, etc. This diary will briefly cover mentioned topics and provide an alternative approach to envelope estimate computation.

FFT (Fast Fourier Transform)

Fast Fourier Transform is an algorithm for computing discrete Fourier transform efficiently. As DFT (Discrete Fourier Transform) performs same operations multiple times, FFT's approach involves reusing previously computed values and as a result decreases the number of multiplications needed from O(N^2) to O($NlogN$) which in case of a 1024-point DFT results in a 100 times faster computation time.
However, there are several issues with FFT such as rounding errors which occur due to imperfect nature of floating point arithmetic. Moreover, if our signal is not periodic within the sampled interval, i.e. we compute DFT over a non-integer number of periods, it causes spectral leakage where energy of one frequency is spread over several adjacent frequencies. Besides that, if we compute a DFT over a signal that started abruptly, we will encounter spectral smearing which results in desired frequency peaks being much harder to distinguish from the rest.
In order to alleviate these issues we can use windowing and zero padding which suppress the effects of spectral leakage as they ensure a smoother fade ins and fade outs. Besides that, in order to analyze signal both in time and frequency domain we can opt for Short Time Fourier Transform which instead of taking in the whole signal performs a sequence of FFTs using a smaller sample count and a fixed hop size. This enables us to create spectrograms which contain both temporal and frequency data at the same time.

Audio Signal Features

Audio features that are relevant to humans include duration, loudness, pitch and timbre. While the first three are a bit more straight-forward, timbre is a more complex, multi-dimensional feature affected by a multitude of factors including brightness, balance between odd and even harmonics, noisiness of the sound, envelope and inharmonicity.

Envelope Detection

Envelope detection of a signal can be done by taking absolute value of our signal and performing temporal averaging using a leaky integrator. Besides that, I found another interesting method in SciPy documentation[1] which uses analytic signal[2] and Hilbert Transform[3]. The necessary code and its result can be found down below. The signal used can be represented using the following equation:

$$f(t) = sin(100\pi t)cos(2\pi t)$$

```python
import numpy as np
import matplotlib.pyplot as plt
import scipy.signal as signal

duration = 2
fs = 200
samples = fs * duration
t = np.arange(samples) / fs
#generating signal with modulation
original_signal = np.sin(100*np.pi*t) * np.cos(2*np.pi*t)

# performing envelope estimation
analytic = signal.hilbert(original_signal)
envelope = np.abs(analytic)

plt.figure(figsize=(10, 5))
plt.grid()
plt.plot(original_signal)
plt.plot(envelope)
plt.xlabel("Sample Number")
plt.ylabel("Amplitude")
plt.legend(["Signal", "Envelope"])
plt.savefig("envelope.png")
plt.show()
```

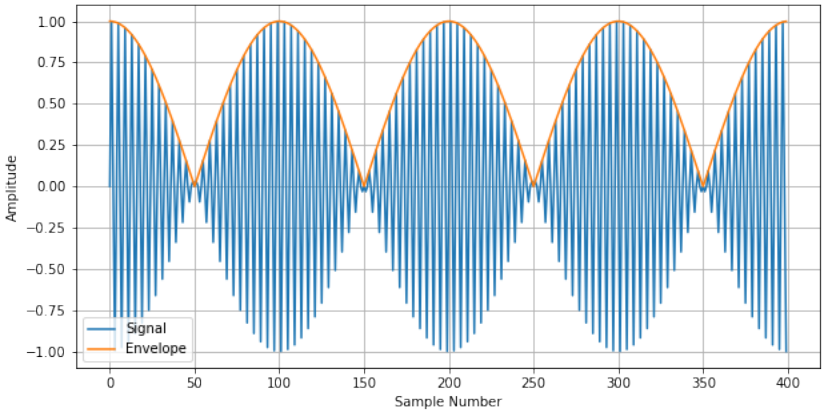

Figure 1: Envelope of a generated signal.

Loudness

As we are aware, humans don't perceive loudness of different frequencies equally. Therefore, loudness estimation needs to be performed. This can be done in several different ways, some of which are running RMS value or converting to decibel scale as human hearing sensitivity roughly matches the logarithmic scale. Besides those methods, it is also possible to develop more complex auditory model of loudness perception to acquire more accurate results.

Pitch

Pitch is defined as the perceived fundamental frequency. While fundamental frequency f_0 is a physical quantity, pitch is subjective. However, for sine waves pitch is equal to f_0. Moreover, humans have the ability to perceive pitch clearly both for complex harmonic and inharmonic tones. In fact, our auditory system attempts to assign pitch to all sounds. In order to extract pitch information we can use various methods. However, most popular method which was originally developed for speech processing purposes is autocorrelation method which correlates the signal with lagged version of itself in order to find the place where it matches the best and extracts the frequency value from the lag at that point.

Spectral Centroid

Spectral centroid method is used to describe the brightness of an audio signal. In order to determine its value we compute center of gravity of its magnitude spectrum using the following equation:

$$c = \frac{f_s}{N} \frac{\sum_{k=0}^{\frac{N}{2}} k|X(k)|}{\sum_{k=0}^{\frac{N}{2}} |X(k)|}$$

References

[1] SciPy - Hilbert Transform

[2] Analytic Signal

[3] Hilbert Transform