Mel Spectrograms and STFT Spectrograms

Ming Chun (Jim) Wei

Fourier Transforms

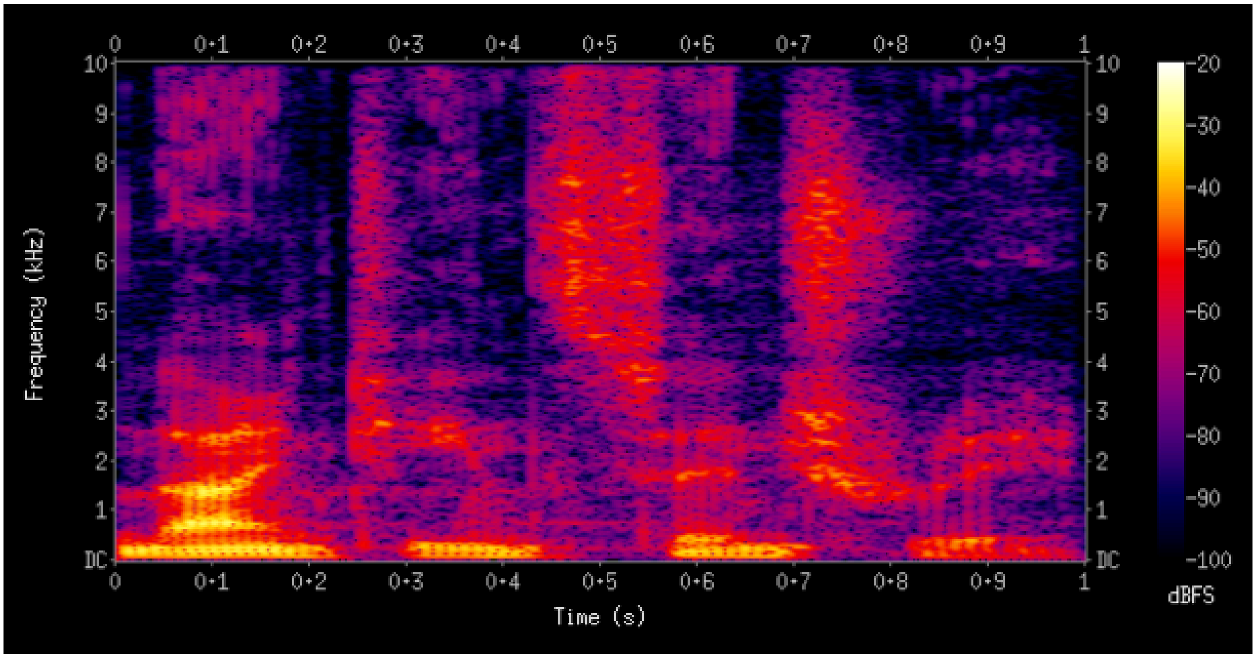
Fourier transform is a mathematical tool to break down signals into its structures. In other words, it is used to examine what a signal is made out of. A signal can be built by combining simple sinusoidal waves with different properties such as frequencies, amplitude, and phase. The Fourier Transform takes a function from one domain and expresses it in terms of another domain, which is often, but not always, the frequency domain.

Short-Time Fourier Transform

Short-Time Fourier Transform (STFT) is a fourier transform that converts a signal by splitting it into short, overlapping frames and applying fourier transform to them. One could imagine it as taking fourier transforms of a sliding window on a signal spectrogram. Starting from time 0, fourier transform is applied to the signals in the window, then, the window is shifted to the right by a certain amount before repeating.

Some parameters of STFT include window size, frame size, and hop size. The window size is essentially the size of each short clip of the signal. The frame size is the size of the frame that the fourier transform is applied to. When the frame size is equal to the window size, we are simply performing a fourier transform to the signals within the window. When the frame size is greater than the window, we simply zero pad the signals. Finally, the hop size is simply how much to slide the window each time.

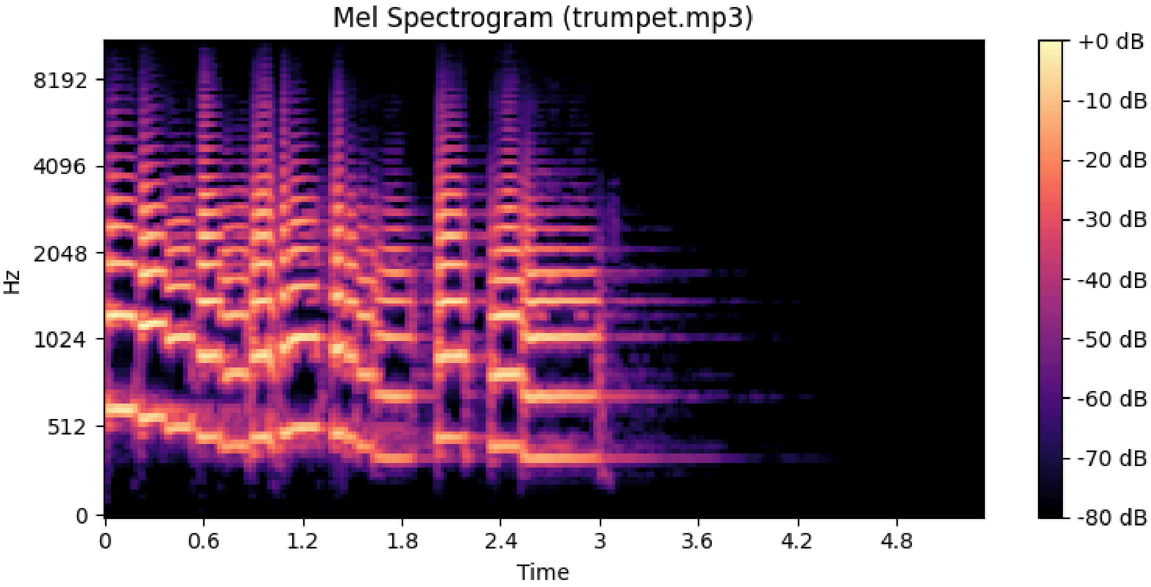
Like every algorithm, STFT has its trade offs. Larger window / frame sizes capture more frequency resolution but lowers the time resolution (smooth out rapid changes). Smaller hop sizes increase time resolution but increases calculation and redundancy. Finally, the window function — the function that extracts the signal in the window balances sharpness in time vs frequency. Some common functions are Hamming and Hann.



An example of STFT spectrograms. Showing how the frequencies of a signal change over time, with time on the x-axis, frequency on the y-axis, and color showing the intensity of each frequency at each moment.

Mel Spectrograms

A Mel Spectrogram is simply a transformation of an STFT spectrogram where the frequency axis is mapped onto the Mel scale, which mimics the way humans perceive pitch. The Mel scale is logarithmic at higher frequencies and linear at lower frequencies.



An example of a Mel spectrogram. The x-axis represents time, the y-axis represents frequency mapped onto the Mel scale (closely aligned with human perception of pitch), and the color intensity indicates the strength (amplitude or power) of each frequency band at each moment in time.

Relation to Deep Fake Audio Detection

In order to perform detection, extraction of audio features is required. In this case, there are three main features that are useful. The first feature is time frequency representation — information about how audio frequencies change in time. The second feature is perceptually-relevant amplitude representation. Humans perceive frequency logarithmically rather than linearly — changes in frequency are not perceived linearly by the human ear — therefore perceptually-relevant amplitude representation helps with this issue. Finally, the third feature is Perceptually-relevant frequency representation. The mel spectrogram contains all three of these features.

Tensor Shape

Given a **batch_size** of audio samples of the same length, a mel spectrogram representation in tensor format would have the shape: (**batch_size, n_channels**, **n_mels**, **time_frames**). **n_channels** represents the number of channels in an audio (typically 2 — left and right). **n_mels** represents the number of frequency bins, determined when extracting the mel spectrogram. **time_frames** represents the number of time frames, determined by the sample rate, number of samples in a **Fast Fourier Transform (FFT)** window, length of the audio clip, and the hop size. A single example (x, y, z, q) would be the numeric feature of the **x-th** audio sample, in the **y-th** channel, **z-th** mel frequency bin, and **q-th** time frame.