Speech Recognition using DL  
Pytorch

Sridhara Manideep

horizontal line

Jun 27, 2022

# [Building End-to-End Speech Recognition Model using DL](https://www.assemblyai.com/blog/end-to-end-speech-recognition-pytorch/)

## [Speech Processing for Machine Learning](https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html?undefined)

* Mel-Frequency Cepstral Coefficients (MFCCs) were very popular features for a long time; but more recently, filter banks are becoming increasingly popular.
* The signal goes through a pre-emphasis filter; then gets sliced into (overlapping) frames and a window function is applied to each frame; afterwards, we do a Fourier transform on each frame (or more specifically a Short-Time Fourier Transform) and calculate the power spectrum; and subsequently compute the filter banks.
* To obtain **MFCCs,** a Discrete Cosine Transform (DCT) is applied to the filter banks retaining a number of the resulting coefficients while the rest are discarded. A final step in both cases is mean normalization.

### Pre-Emphasis

* Pre-emphasis is a very simple signal processing method which increases the amplitude of high-frequency bands and decreases the amplitudes of lower bands. In simple form, it can be implemented as **yt=xt−αxt−1**
* A pre-emphasis filter is useful in several ways:
  + Balance the frequency spectrum since high frequencies usually have smaller magnitudes compared to lower frequencies
  + Avoid numerical problems during the Fourier transform operation and
  + May also improve the Signal-to-Noise Ratio (SNR).

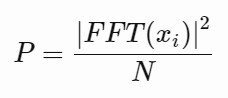
### Framing

* After pre-emphasis, we need to split the signal into short-time frames.
* Frequencies in a signal change over time, so it doesn’t make sense to do FT across the entire signal as we would lose the frequency contours of the signal over time.
* Therefore, by doing a Fourier transform over this short time frame, we can obtain a good approximation of the frequency contours of the signal by concatenating adjacent frames.
* Typical frame sizes in speech processing range from 20 ms to 40 ms with 50% (+/-10%) overlap between consecutive frames. Popular settings are 25 ms for the frame size, frame\_size = 0.025 and a 10 ms stride (15 ms overlap), frame\_stride = 0.01.

### Window

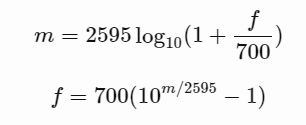
* After slicing the signal into frames, we apply a window function such as the Hamming window to each frame. A Hamming window has the following form:
* There are several reasons why we need to apply a window function to the frames, notably to counteract the assumption made by the FFT that the data is infinite and to reduce spectral leakage.

### Fourier-Transform and Power Spectrum

* We can now do an N-point FFT on each frame to calculate the frequency spectrum, which is also called Short-Time Fourier-Transform (STFT), where N is typically 256 or 512, NFFT = 512; and then compute the power spectrum (periodogram) using the following equation

### Filter Banks

* The final step to computing filter banks is applying triangular filters, typically 40 filters, nfilt = 40 on a Mel-scale to the power spectrum to extract frequency bands.
* The Mel-scale aims to mimic the non-linear human ear perception of sound, by being more discriminative at lower frequencies and less discriminative at higher frequencies.
* We can convert between Hertz (f) and Mel (m) using the following equations:



### MFCC Coefficients

* To obtain MFCCs, a Discrete Cosine Transform (DCT) is applied to the filter banks retaining a number of the resulting coefficients while the rest are discarded.