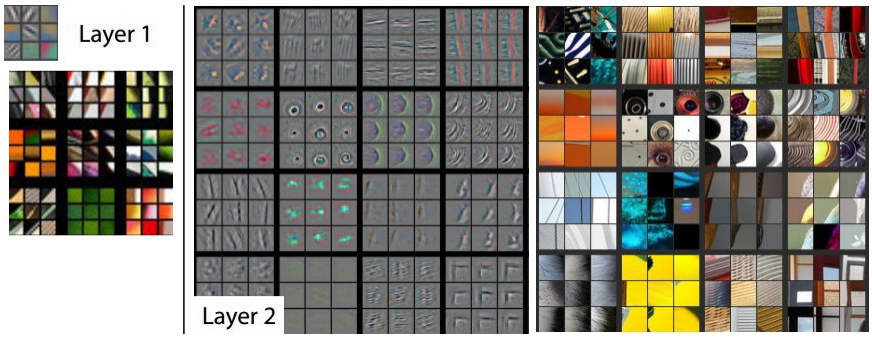
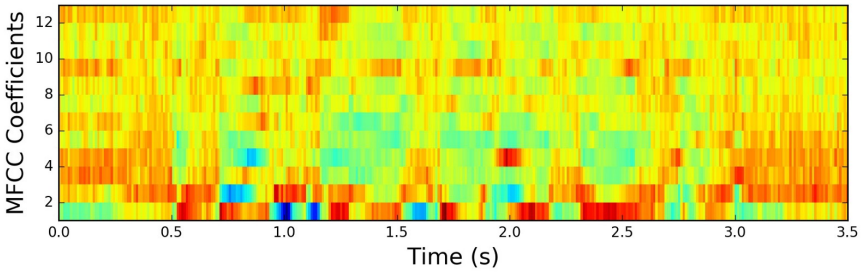
Pratheeksha Nair Follow
Jul 24, 2018 · 4 min read

> *Disclaimer 1 : This article is only an introduction to MFCC features and is meant for those in need for an easy and quick understanding of the same. Detailed math and intricacies are not discussed.*

Never having worked in the area of speech processing myself, harking upon the word "MFCC" (quite often used by peers) left me with the inadequate understanding that it is the name given to a particular kind of "feature" extracted from audio signals (similar to edges that constitute a kind of feature extracted from images).



Features extracted by a CNN from images

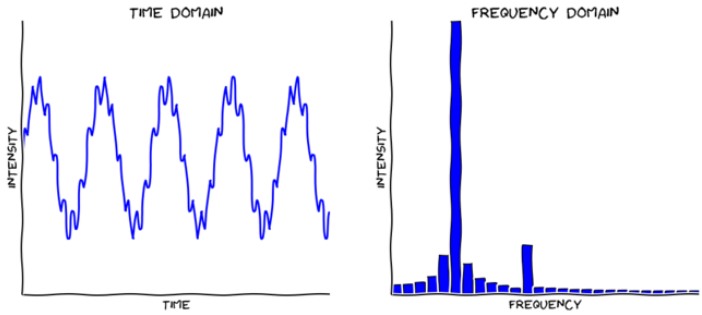


Features extracted from speech signals. Pretty huh?!

It took me quite a bit of reading from multiple sources to grasp the novice's understanding of what MFCC features are. So I decided to help out fellow humans in need by compiling the information I collected in an easy-to-understand manner.

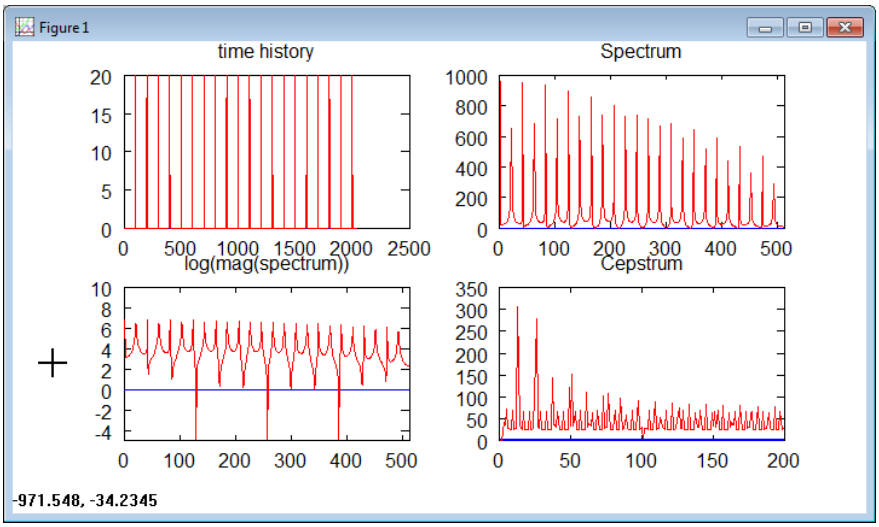
Let's begin by expanding the acronym MFCC—*Mel Frequency Cepstral Co-efficients*.

Ever heard the word *cepstral* before? Probably not. It's *spec*tral with the *spec* reversed! Why though? For a very basic understanding, cepstrum is the information of rate of change in spectral bands. In the conventional analysis of time signals, any periodic component (for eg, echoes) shows up as sharp peaks in the corresponding frequency spectrum (ie, Fourier spectrum. This is obtained by applying a Fourier transform on the time signal). This can be seen in the following image.



On taking the log of the magnitude of this Fourier spectrum, and then again taking the spectrum of this log by a cosine transformation (I know it sounds complicated, but bear with me please!), we observe a peak wherever there is a periodic element in the original time signal. Since we apply a transform on the frequency spectrum itself, the resulting spectrum is neither in the frequency domain nor in the time domain and hence Bogert et al. decided to call it the *quefrency domain*. And this spectrum of the log of the spectrum of the time signal was named *cepstrum* (ta-da!).

The following image is a summary of the above explained steps.



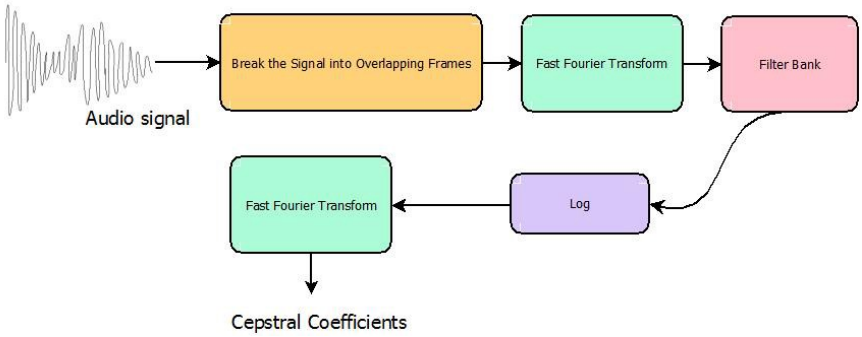
> *Cepstrum was first introduced to characterize the seismic echoes resulting due to earthquakes.*

Pitch is one of the characteristics of a speech signal and is measured as the frequency of the signal. *Mel scale* is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear (humans are better at identifying small changes in speech at lower frequencies). This scale has been derived from sets of experiments on human subjects. Let me give you an intuitive explanation of what the mel scale captures.

The range of human hearing is 20Hz to 20kHz. Imagine a tune at 300 Hz. This would sound something like the standard dialer tone of a land-line phone. Now imagine a tune at 400 Hz (a little higher pitched dialer tone). Now compare the distance between these two howsoever this may be perceived by your brain. Now imagine a 900 Hz signal (similar to a microphone feedback sound) and a 1kHz sound. The perceived distance between these two sounds may seem greater than the first two although the actual difference is the same (100Hz). The mel scale tries to capture such differences. A frequency measured in Hertz (f) can be converted to the Mel scale using the following formula :

$$\mathrm{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

Any sound generated by humans is determined by the shape of their vocal tract (including tongue, teeth, etc). If this shape can be determined correctly, any sound produced can be accurately represented. The envelope of the time power spectrum of the speech signal is representative of the vocal tract and MFCC (which is nothing but the coefficients that make up the *Mel-frequency cepstrum*) accurately represents this envelope. The following block diagram is a step-wise summary of how we arrived at MFCCs:



Here, Filter Bank refers to the mel filters (coverting to mel scale) and Cepstral Coefficients are nothing but MFCCs.

> *TL; DR—MFCC features represent phonemes (distinct units of sound) as the shape of the vocal tract (which is responsible for sound generation) is manifest in them.*

Disclaimer 2 : All images are from Google images.