1 Mel Frequency Cepstral Coefficient (MFCC) method

Method is used to extract features from audio files. Introudced by Davis and Mermelstein in the 1980's. Before MFCC, used methods were Linear Prediction Coefficients (LPC) and Linear Prediction Cepstral Coefficients (LPCC). But, MFCC is state-of-the-art ever since.

1.1 Steps of MFCC method

1.1.1 Divide signal into short frames

Signal is divided in short frames because of various reasons. Probably most important reason is that we want every frame to have specific frequency component. It is not favorably to to collapse all audio information (e.g. word 'hello') into one single spectrum (FFT collapses temporal data). Next reason could be complexity of computation.
Audio signal varies over time, but is statistically stationary over short segments. Signal is still changing in short intervals, but we assume it is stationary. Audio signal is usually framed into 20-40 ms frames. With this length we get reliable spectral estimate - signal is simple enough to be described efficiently by few coefficients. Important is that frames must overlap. Frame step is usually 10 ms.
Overlapping frames are important firstly because of negative result of window function (next subsection), second overlapping minimizes poorer time domain resolution which requires longer window.

Example:
Frame length is 25 ms, signal has 16 kHz sample rate and frame step is 10 ms. Then we have 0.025 * 16000 = 400 samples per frame.
Fist 400 sample frame contains samples from 0 to 400, next 400 sample frame contains samples from 160 to 560, and so on.

Next steps are applied to every single frame. Set of n (default n=13) MFCC coefficients is extracted for each frame.

1.1.2 Applying windowing function on the frame

Segmenting longer audio signal into shorter actually has the same effect of rectangular windowing. Problem with implicit rectangular windowing is spectral leakage.
Therefore non-rectangular window function is applied. Usually Von Hann or Hamming. Unfortunately window function can be lossy at the edges of window (frame), but with overlapping frames that problem is solved.

1.1.3 Calculating DFT on the frame

Audio signal is now devided in T* overlapping frames. Now we want to analyse spectrum of signal. Time aspect of signal is often not suitable as input to classifying algorithms, except maybe Neural Networks. Spectrum of audio signal will be gained with DFT, FFT.
If audio is $s(n)$, then $i-th$ frame is $s_i(n)$. In $s_i(n)$, n varies over $[1, n*]$, where n* is number of samples in frame (e.g. n* = 400).
TODO: dft formula
Where $h(n)$ is an N sample long analysis window (e.g. Hamming window).
$S_i(k)$ is result of DFT taken from $s_i(n)$. Finally $P_i(k)$ denotes periodogram of power of spectrum $S_i(k)$. Periodogram is estimation of spectral density. In this case we will look at estimate of power spectrum.
TODO: peridogram formula
FFT is preformed for K points, usually K = 512. We keep K1 coefficiens, usually K1 = 257.
Motivation: human cochlea.

1.1.4 Mel-spaced filterbank

Now we have power spectrum of every frame $P_i(k)$. In this part we will 'group' this spectrum with Mel-spaced filterbank. Grouping is done as a way of compressing. Maybe different ways of grouping could be done, but MFCC is motivated by human hearing system. Hearing system can not easy differ difference between two closely spaced frequency(motivation: cochlea).
Mel-bank filter is set of 20-40 triangular filters. This filters are spaced according to Mel-scale. Mel scale relates perceived frequency. Humans (motivation: cochlea) are better at differing small changes in pitch (quality that makes it possible to judge sounds as "higher" and "lower" in frequency) in low then high frequency. Formula for converting to mel scale is:
TODO: MEL converting
Mel bank filters are very narrow at low frequency and they get wider as frequency rises, because we get less concerned about variations at high frequencies. Filterbank comes in form of 26 vectors with length of 257. Where 26 is number of filters and 257 is number of DFT coefficients (periodograms of power spectrum). Filterbanks are calculated in a way that each triangle is multiplied with periodogram of power spectrum and coefficients that are covered by triangle are summed. Result is 26 numbers where each number represents power in every triangular filter in filterbank.

1.1.5 log and DCT applied to the values of powers in filterbank

Finally we take cepstral coefficients. Definition if cepstral coefficient is inverse DFT of log power spectrum.
TODO: formula for cepsral
MFCC method is specific in a way that it takes logarithm of power spectrum,

but then applies DCT to log power spectrum. In this way cepstral coefficients are gained. But DCT is taken because of overlapping frames. As frames are overlapped, then we have correlation of power coefficients. Correlation is broken down with DCT transform instead of IDFT. Now covarinace matrix could be used to model feature
Logarithm values is taken because human hearing motivation. Humans do not hear loudness (power of signal) in linear scale. Log is also taken because of cepstral definition, and it allows cepstral mean substraction - channel normalisation technique.
Now we have 26 DCT coefficients, cepstral coefficients. We take 2-12 of them. Higher DCT components represent fast changes in filterbank powers. These fast changes are degrading (or could degrade) performance of model using this features, so dropping them could bring better results.
Finally, coefficients we take are called MFCCs.

1.1.6 Additional enhancement on MFCC

Fist is delta and delta-delta features. Used alongside of MFCCs. Next is liftering of MFCC.

References

[1] https://dsp.stackexchange.com/questions/6499/help-calculating-understanding-the-mfccs-mel-frequency-cepstrum-coefficients

[2] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[3] https://dsp.stackexchange.com/questions/31/how-do-i-interpret-the-dct-step-in-the-mfcc-extraction-process