Mel-Frequency Cepstral Coefficients (/feature-extraction/mfcc)

1. Introduction

The most commonly used feature extraction method in automatic speech recognition (ASR) is Mel-Frequency Cepstral Coefficients (MFCC) [1]. This feature extraction method was first mentioned by Bridle and Brown in 1974 and further developed by Mermelstein in 1976 and is based on experiments of the human misconception of words [2].

To extract a feature vector containing all information about the linguistic message, MFCC mimics some parts of the human speech production (/speech/speech-production) and speech perception (/speech/sense-of-hearing). MFCC mimics the logarithmic perception of loudness and pitch of human auditory system and tries to eliminate speaker dependent characteristics by excluding the fundamental frequency and their harmonics. To represent the dynamic nature of speech the MFCC also includes the change of the feature vector over time as part of the feature vector [3,4].

2. Implementation

The standard implementation of computing the Mel-Frequency Cepstral Coefficients is shown in Figure 1 and the exact steps are described below [3].

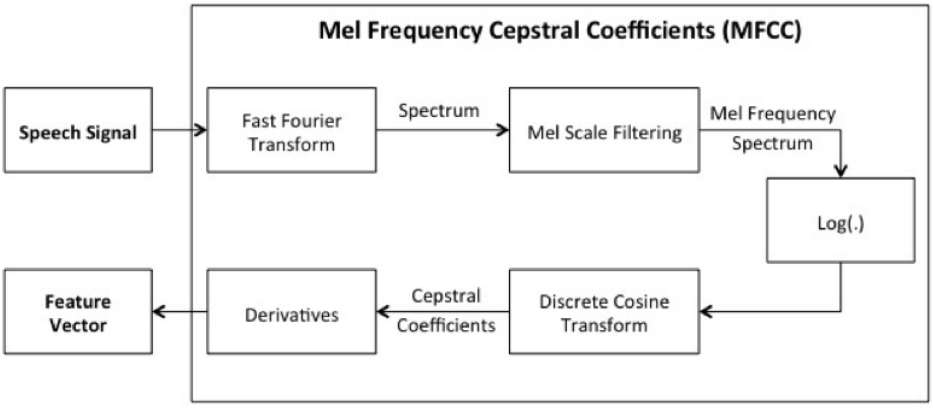

Figure 1: Block diagram of the MFCC algorithm

The Input for the computation of the MFFCs is a speech signal in the time domain representation with a duration in the order of 30 ms.

2.1 Fourier Transform

The first processing step is the computation of the frequency domain representation of the input signal. This is achieved by computing the Discrete Fourier Transform (http://en.wikipedia.org/wiki/Discrete_Fourier_transform).

$$c_{\tau,k}^{(1)} = \left| \frac{1}{N} \sum_{j=0}^{N-1} f_j \; exp\left[-i \; 2\pi \; \frac{jk}{N} \right] \right| \qquad k = 0, 1, ..., (N/2) - 1$$

Where N is the number of sampling points within a speech frame and the time frame τ. For implementations the Fast Fourier Transform (http://en.wikipedia.org/wiki/Fast_Fourier_transform), which is a variation of the Discrete Fourier Transformation (http://en.wikipedia.org/wiki/Discrete_Fourier_transform) optimized for speed, is used. [3]

2.2 Mel-Frequency Spectrum

The second processing step is the computation of the mel-frequency spectrum. Therefore, the spectrum is filtered with N_d different band-pass filters (http://en.wikipedia.org/wiki/Band-pass_filter) and the power of each frequency band is computed. This filtering mimics the human ear because the human auditory system uses the power over a frequency band as signal for further processing. This processing step can be described by

$$c_{\tau,j}^{(2)} = \sum_{k=0}^{N/2-1} d_{j,k} \; c_{\tau,k}^{(1)} \qquad j = 0, 1, ..., N_d$$

, where d is the amplitude of the band-pass filter (http://en.wikipedia.org/wiki/Band-pass_filter) with the index j at the frequency k. The filter bank with the band-pass filters (http://en.wikipedia.org/wiki/Band-pass_filter) cannot mimic the ear because the ear can use any frequency as center frequency. For ASR N_d equidistant band-pass filters (http://en.wikipedia.org/wiki/Band-pass_filter) on the mel scale are used. The mel-scale is a non-linear scale that is adapted to the non-linear pitch perception of the human auditory system (http://www.recognize-speech.com/speech/sense-of-hearing) (For more information about the mel scale see Sense of Hearing (/speech/sense-of-hearing)). The number, the shape (triangular, trapezoidal rectangular) and the center frequency of the band-pass filters (http://en.wikipedia.org/wiki/Band-pass_filter) can be varied [3]. Figure 2 shows a typical filter-bank with 25 triangular band-pass filters (http://en.wikipedia.org/wiki/Band-pass_filter). Some research suggests that to few and to many band-pass filters (http://en.wikipedia.org/wiki/Band-pass_filter) have a negative impact on the classification performance and that overlapping rectangular shaped filters achieve a better performance compared to triangular shaped filters [5].

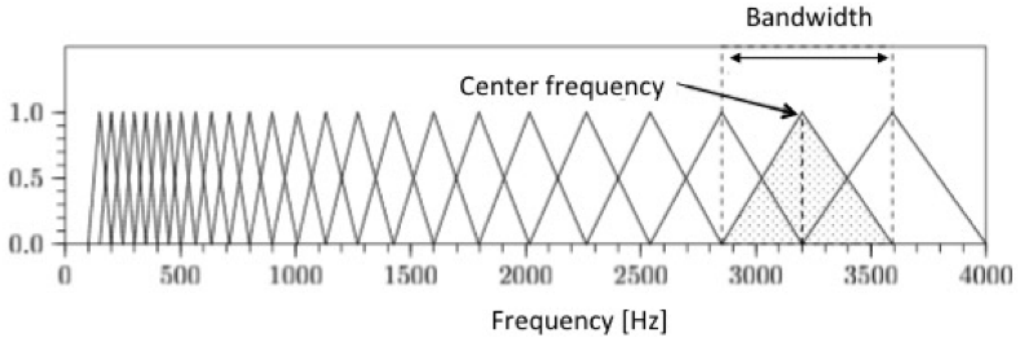

Figure 2: Filterbank with 25 triangular bandpass filters to compute the mel frequency spectrum. [4]

2.3 Logarithm

The third processing step computes the logarithm of the signal, to mimic the human perception of loudness because experiments showed that humans perceive loudness on a logarithmic scale [3].

$$c_{\tau,j}^{(3)} = log(\, c_{\tau,j}^{(2)} \,) \qquad j = 0, 1, ..., N_d$$

2.4 Cepstral Coefficients

The fourth processing step tries to eliminate the speaker dependent characteristics by computing the cepstral coefficients. From the Source-Filter model (/speech/speech-production#SourceFilterModel) is known, that the signal is the convolution of the speaker dependent source signal and the filter signal. To suppress the source signal the cepstrum is computed. The cepstrum can be interpreted as the spectrum of a spectrum. Therefore, the speaker dependent harmonics of the fundamental frequency are transformed to one higher order cepstral coefficient under ideal conditions (highlighted bar in Figure 3b). The inverse transformation of the lower cepstral coefficients show the frequency response of the vocal tract (Figure 3c) and the inverse transformation of the higher order cepstral coefficients show the frequency spectrum of the source signal. Therefore, the speaker dependent harmonics are suppressed by taking the lower order cepstral coefficients for further processing. The cepstrum of a signal is computed by

$$F^{-1}\left\{log(\,F\left\{f_n\right\}\,)\right\}$$

, where f is the input signal and F is the Fourier Transformation (http://www.recognize-speech.com/feature-extraction/wavelet-based-features#FourierTransform) [6]. The computation of the logarithm can be omitted because the logarithm of the signal was computed in the previous processing step 2.3. Instead of the Fourier Transform the discrete cosine transform (http://en.wikipedia.org/wiki/Discrete_cosine_transform) can be used because the absolute value of the spectrum, respectively the periodic continuation of the signal, is real and symmetric. The cepstral coefficients are computed by

$$c_{\tau,j}^{(4)} = \sum_{j=1}^{N_d} c_{\tau,j}^{(3)}\ cos\left[\frac{k\ (2j-1)\pi}{2N_d}\right] \qquad k = 0, 1, ..., N_{mc} < N_d$$

where N_{mc} is the number of chosen cepstral coefficients for further processing [3]. Typically N_{mc} is in the range of thirteen to twenty.

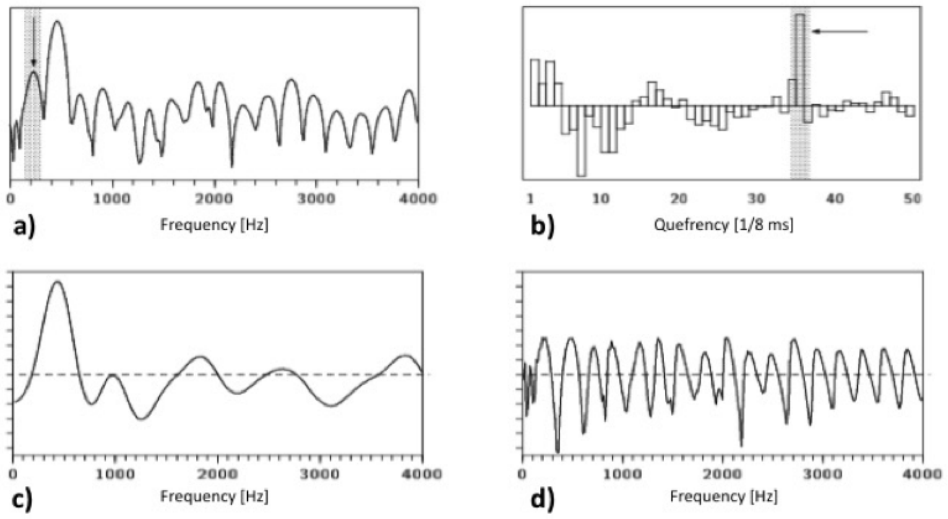

Figure 3: a) Logarithmic power density spectrum of a speech signal. The highlighted frequency is the speaker dependent fundamental frequency. b) Cepstral coefficients of a speech signal. The highlighted quenfrency is the transformed fundamental frequency and the corresponding harmonics. c) Logarithmic power density spectrum of the inverse transformation of the low-pass filtered cepstral coefficients. Cutoff quefrency q= 20ms d) Logarithmic power density spectrum of the inverse transformation of the high-pass filtered cepstral coefficients. Cutoff quefrency q= 20ms. [4]

2.5 Derivatives

All previous processing steps included information about the current signal frame. To represent the dynamic nature of speech the first and second order derivatives of the cepstral coefficients extend the feature vector [6].

$$\Delta c^{(4)}_{\tau,j} = c^{(4)}_{\tau+1,j} - c^{(4)}_{\tau-1,j} \qquad \Delta\Delta c^{(4)}_{\tau,j} = \Delta c^{(4)}_{\tau+1,j} - \Delta c^{(4)}_{\tau-1,j}$$

The final feature vector is

$$c_\tau = \left[c^{(4)}_{\tau,j} \ , \ \Delta c^{(4)}_{\tau,j} \ , \ \Delta\Delta c^{(4)}_{\tau,j} \right]$$

. A typical MFCC feature vector would be calculated from a window with 512 sample points and consist of 13 cepstral coefficients, 13 first and 13 second order derivatives. This example would reduce the dimensionality from 512 to 39 dimensions.

3. Limitations

Even though MFCC feature vectors are commonly used in ASR systems, the MFCC feature vectors have some limitations. Most of these limitations arise from the computation of the cepstral coefficients.

One critical assumption of the cepstral coefficients is that the fundamental frequency is much lower than the frequency components of the linguistic message. This assumption is needed because otherwise the exclusion of the fundamental frequency and the harmonics is not possible while including all information about the linguistic message. However, many female speaker do not fulfill this assumption. Therefore, it is unknown if the speaker dependent characteristics can be suppressed for all speakers. [6]

Another limitation of the cepstral coefficients their lack of interpretation. Only the first two cepstral coefficients c_0 and c_1 have a meaningful interpretation. c_0 is the power over all frequency bands and c_1 is the balance between low and high frequency components within the signal frame. The other cepstral coefficients have no clear interpretation other than they contain the finer detail of the spectrum to discriminate the sounds. Due to this lack of interpretations the reaction of MFCC features to accents or noise is unknown. As a consequence the feature vector distributions for each speaker have to be merged, which yields greater variances and could reduce the separability of the classes. [1]

Furthermore cepstral coefficients apply an equal weight to high and low amplitudes to the log spectrum even though it is known that high energy amplitudes dominate the perception of speech. This equal weight reduces the robustness of cepstral coefficients because the noise fills the valleys between formants (/speech/composition-of-speech#formant]) and harmonics and deteriorates the performance of MFCCs. [1]

4. Variations

To improve the MFCCs many variations and extensions have been proposed. This section will give a brief overview of some proposed variations and extensions.

One example of an extension of MFCC is to include cepstral mean normalization (/preprocessing/cepstral-mean-normalization]). This extensions tries to reduce channel effects such as different microphones or different locations by subtracting the cepstral mean from the MFCC feature vector. Another extensions method normalizes the spectrum of the speech signal with the hearing threshold to prohibit features, which could not be identified with the human auditory system (/speech/sense-of-hearing) [3].

Other methods substitute some parts of MFCC to improve the error rate of ASR. For example the Bark Frequency Cepstral Coefficients (BFCC) use N_d equidistant band-pass filters on the bark scale instead of the mel scale [5]. Another example is the root-Cepstrum coefficients and the μ-Law coefficients, which substitute