Implementation of MFCC vector generation in classification context

Adam Pelikant
Lodz University of Technology

JOURNAL OF APPLIED
COMPUTER SCIENCE
Vol. xx. No xx (200x), pp. xx-xx

Implementation of MFCC vector generation in classification context

Dominik Niewiadomy, Adam Pelikant
Institute of Mechatronics and Information Systems
Technical University of Lodz
ul Stefanowskiego 18/22Lodz, Poland
e-mail: dominik.niewiadomy@gmail.com

Abstract. *The article presents the implementation of each step of Mel Frequency Cepstral Coefficients (MFCCs) generation from signal acquisition to delta MFCCs computation. Additionally the article presents potential application of MFCCs in context of database multimedia queries (Query By Voice/Query By Humming).*

1. Introduction

Automatic speech classification and recognition was topic of research in many laboratories in the world. The variety of possible applications is so big that sound classification can be used in many branches of science. Speech recognition and classification systems could be applied for many commonly used devices and software. This article will point the problem of multimedia – voice data acquisition, parameterisation and some possible methods of obtaining useful sound information. This work is mainly devoted to the problem of effective acquisition of signal discriminating coefficients, but on the other hand this article will present the possible use of Mel Frequency Coefficients for data querying in new Query By Humming/Query By Voice [1][2][3][4] algorithm. The algorithm designed by article authors in future will by implemented and deployed on Oracle Relational Database Management System [5]. The main development in research conducted in speech classification was connected with algorithms and hardware progress. The development of faster computers and signal processors enabled previously impossible computations to be done in relatively short time so it enabled new useful application to be created.

2. Input signal

To understand the article you need to understand a digitalised sound signal nature [6]. To acquire the raw signal you need to have an analog input such as microphone. The gathered signal is converted by A/D (analog-digital) converter into a sample set. Measurement of the signal intensity is made at regular intervals – such operation is called sampling. Each sample must have measured intensity/amplitude in a strict range of values. If the range is wider the signal representation is better. Digital signal sample must be stored as a number of specified precision therefore samples are saved as 8, 16 or 32 bits. With rise of sample bits number, the sample intensity is more precise, but the processing is much more resource and time consuming.

Second important feature in digital signal representation is sampling frequency. The choice of sampling rate is based on Nyquist-Shannon theorem which says that "any sampled signal can be reproduced exactly as long as it is sampled at a frequency greater than twice the bandwidth of the signal". Therefore a sampling rate of 44.1 kHz would be enough to capture all the information contained in a signal having frequency bandwidth up to 22 kHz. The human frequency range of hearing is 16Hz – 22 kHz but in most cases below 200 Hz and above 8 kHz the level of hearing falls dramatically (Fig.1).

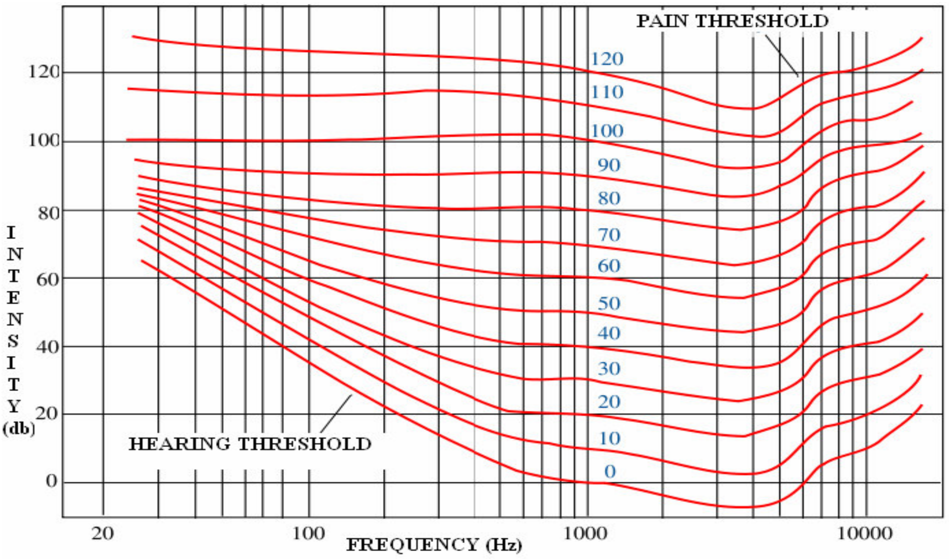

Fig.1 Sound perception

This provides us with information that the most valuable frequency range is above 200Hz and below 8 kHz so the sampling rate should be at least 16 kHz. The telecommunication companies mostly use 8 kHz sampling in their devices because such frequency is good enough to understand the speech and it is much cheaper than 16 kHz signal in computations and transfer.

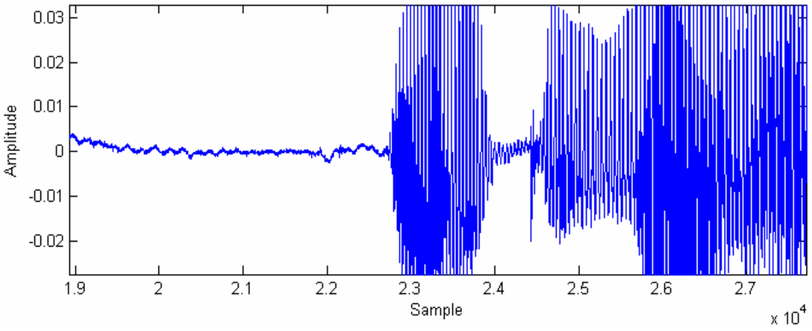

Fig.2 Human speech digitalized signal

As it can be seen on Fig. 2 sound signals are very variable. The figure presents the short signal representation which stands for a part of human three words speech sentence.

3. Mel Frequency Cepstral Coefficients

MFCCs – Mel Frequency Cepstral Coefficients (MFCCs) are audio representation coefficients. At first MFCCs were used in speech signal processing. MFCCs are a representation of the spectral power envelope that allows data compression. The MFCCs [7] analysis is derived from cepstral representation of audio data. The cepstrum is defined as the inverse Fourier transform of the log – spectrum. The difference between Mel frequency cepstrum and standard cepstrum is that the MFCC frequency bands are positioned logarithmically. The coefficients generated by algorithm are fine representation of signal spectra, with great data compression. The main use of the coefficients is audio compression, automatic speech recognition and signal classification.

4.1. Samples preparation

The MFCCs generation algorithm input is sampled acoustic signal of human speech gathered from a microphone. In order to acquire signal samples, the sound is stored in PCM WAVE format file or it is read directly from microphone. Such prepared sound stream is read in order to obtain 16bit precision values collection. Next all values are normalized to range <-1, 1>. After that in order to improve Signal to Noise Ratio factor the algorithm computes a preemphasis. The preemphasis is a preprocessing phase which increases, within a band of frequencies, the magnitude of higher frequencies with respect to the magnitude of lower frequencies.

$$x'(n) = x(n) - \alpha \cdot x(n-1) \tag{1}$$

In order to boost magnitude of higher frequencies, the algorithm uses FIR filter described as (1). In automatic speech recognition and classification systems most commonly the α factor is set to 0.95.

4.2. Framing and windowing

As the MFCCs algorithm is based on spectral analysis, it requires transformation from time domain to frequency domain. The sound signal is a quasi stationary signal. For sound sample that lasts over 200ms the signal is assumed to be not periodical. For sample that lasts from 30ms to 200ms it cannot be determined as periodical or not periodical. For samples shorter than 30 ms the sound can be assumed as a periodical one. This feature creates the need for signal framing (2). Each frame should be approximately from 20 ms to 30 ms (t).

$$N = t \cdot f_s \tag{2}$$

On the other hand in order to apply Fast Fourier Transform in further processing the frame length should be the power of 2 (3).

$$N = 2^k \tag{3}$$

Those requirements can be done by reducing the frame length N of result of (2) to the nearest power of 2 computed from (3) or increasing the frame length to the next power of 2 by inserting zeros to the frame.

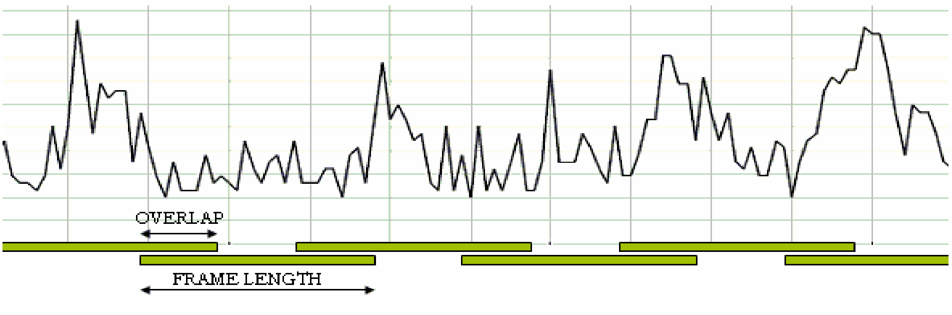

Fig.3 Human speech digitalized signal

Additionally to avoid loss of information of signal variability frames are overlapped each by another. The average overlap level in the algorithm was set to 20%. Additionally to reduce the effect of overlapping frames and aliasing the frame is modified by specially precalculated windows such as: Hamming (4), Bartlett (5) or Blackman (6).

$$w(n) = 0.53836 - 0.46164 \cdot \cos\left(\frac{2\pi n}{N-1}\right) \tag{4}$$

$$w(n) = \frac{N-1}{2} - \left| n - \frac{N-1}{2} \right| \tag{5}$$

$$w(n) = 0.42323 - 0.49755\left[\cos\left(\frac{2\pi \mathrm{n}}{\mathrm{N-1}}\right)\right] + 0.07922\cos\left[\cos\left(\frac{4\pi \mathrm{n}}{\mathrm{N-1}}\right)\right] \tag{6}$$

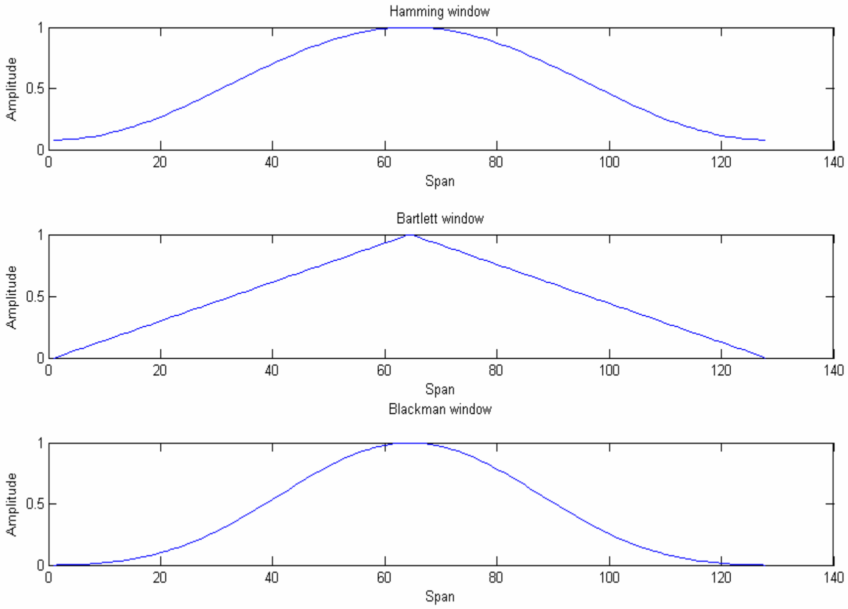

Fig.4 Windows – Hamming, Bartlett, Blackman

Such gathered, filtered, framed and windowed data is input for next part of system processing that is separating into channels.

4.3. Perceptual scales and Mel Filterbank

The human perception resolves frequencies in non linear way across the audio spectrum. Empirical evidences such as Absolute Threshold of Hearing that is “the minimum sound level of a pure tone that an average ear with normal hearing can hear in a noiseless environment” suggests, that operating in a non linear scale, improves accuracy of recognition. As it can be seen on ATF chart, for higher frequencies, the ear is less sensitive, than for the lower with the same sound intensity. In order to reconstruct and model human perception the MFCCs generation algorithm uses perceptual scale.

Perceptual scales are based on empirical experiments with groups of listeners. Most perceptual scales are linear for low frequencies and logarithmical for higher frequencies. The most commonly used perceptual scales are Bark (7) and Mel (8, 9) scales. The MFCCs algorithm uses Mel scale transformation where the equation (8) present transformation from Hertz scale to Mel scale and the equation (9) presents reverse transformation.

$$f_{bark} = 13 \cdot arcus \tan\left(\frac{0{,}76f}{1000}\right) + 3.5 \cdot arcus \tan\left(\frac{f}{7500}\right)^{2} \quad (7)$$

$$f_{mel} = 1127.01048 \log_e(1 + \frac{f_{Hz}}{700}) \quad (8)$$

$$f_{Hz} = 700\left(e^{\frac{f_{mel}}{1127.01048}} - 1\right) \quad (9)$$

Figure 5 presents relation between Mels and Hertz values. It can be assumed that Mel values beneath 1127Hz raise linearly while above 1127Hz the Mel values raise logarithmically.

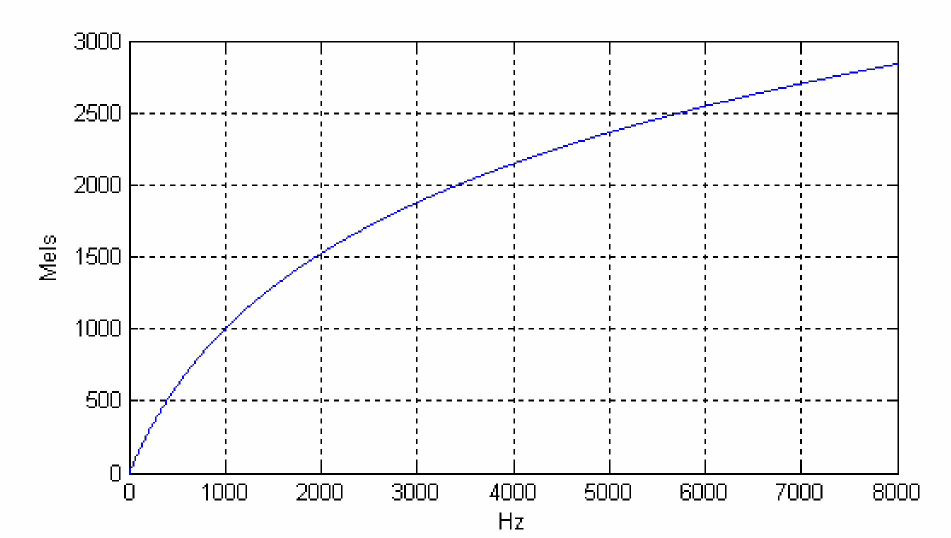

Fig.5 Relation between Hz and Mels

In order to split spectrum of signal into channels that are based on perceptual scales Mel filterbank needs to be generated. The Mel filterbank is a bank of L bandpass filters, with center frequencies linearly spaced in perceptual scale. The most common shape for filters is triangular, but in some implementations others shapes such as rectangular, trapezoidal or Gaussian can be found. Each filter in perceptual scale has the same width and is overlapped for half of its width with next filter. Because the human perception is limited in some range the filterbank defines boundary frequencies f_{min} and f_{max}. The knowledge of boundaries and number of filters enable algorithm to compute constant filter width (10) in perceptual scale.

$$f_w = \frac{f_{\max} - f_{\min}}{L} \tag{10}$$

Such prepared filterbank in Mel scale is transformed to Hertz scale by (9). The result of the transformation is presented at Figure 6 and Figure 7. Presented filterbank is generated for 2048 point FFT transform, where number of filters is 25, minimum frequency is 0 Hz, maximum frequency is 4000 Hz and sampling frequency is 8 kHz.

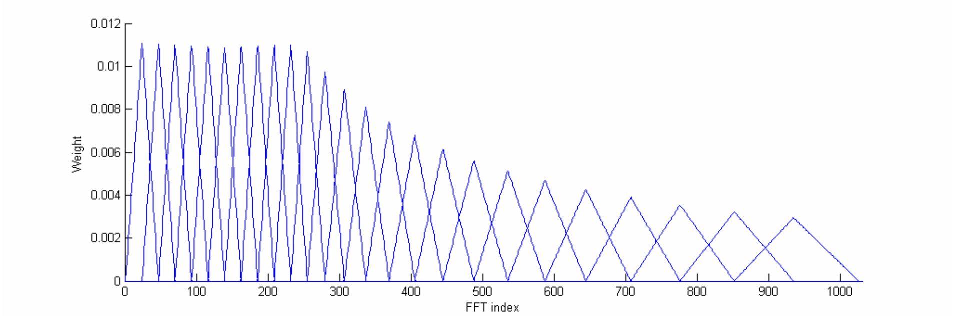

Fig.6 Filterbank – 25 triangular filters

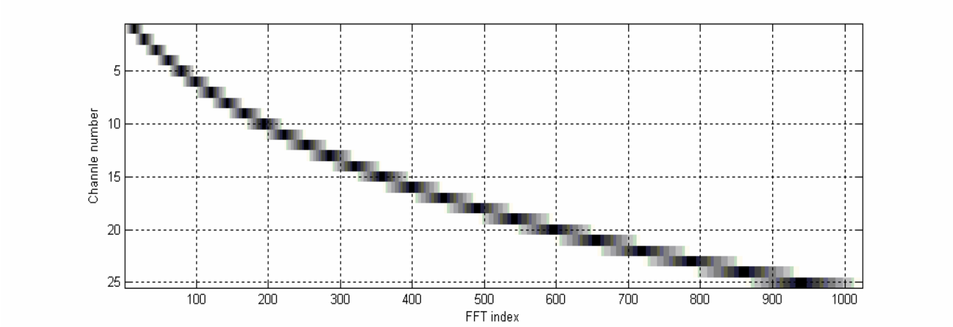

Fig.7 Filterbank – 25 triangular filters – matrix of coefficients

The algorithm of MFCCs generation creates the filterbank before all processing is done. Once created filterbank is reused in each frame channeling phase because filterbank parameters are constant.

4.4. MFCCs generation

Most processing in MFCCs generation algorithm is done on frequency domain data. In order to transform samples form time domain to frequency domain Fast Fourier Transform is computed for each frame. After that on the FFT result of each m-th frame power spectrum is calculated (11).

$$\bar{x}_m(i) = [real(X_m(i))]^2 + [imag(X_m(i))]^2 \tag{11}$$

Secondly the result of power spectrum is filtered with each k-th filter from L filters created in Mel filterbank and the result for each filter is summed as it is defined in (12). The M_k (i) is i-th FFT index weight of k-th filter in filterbank.

$$S_m(k)=\sum_{i=0}^{N/2}\left(x_m(i)\cdot M_k(i)\right) \tag{12}$$

The result of channeling operation is stored in $S_m(k)$ vector. Next the logarithm (12) out of S_m is calculated.

$$SL_m(k)=\log_e\left(S_m(k)\right) \tag{13}$$

After that the $SL_m(k)$ is transformed by Discrete Cosine Transform (14):

$$C_m(n)=\sum_{i=0}^{L-1}\left(SL_m(i)\cdot\cos\left(\frac{\pi n}{2L}(2i+1)\right)\right) \tag{14}$$

The result of this operation is C_m which is the Mel Frequency Cepstral Coefficients vector for m-th frame and n is the coefficient number out of M coefficients. In most cases [8] number of MFCCs is 12 coefficients and 1 additional logarithm of frame energy:

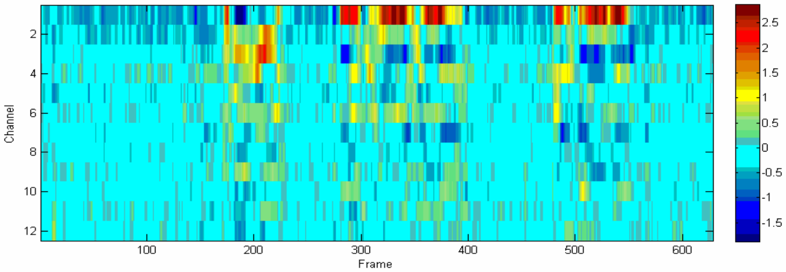

Fig.8 Mel Frequency Cepstral coefficients (1 – 12)

As it was mentioned previously additional 0th coefficient is energy logarithm defined as (15) and presented on Figure 9:

$$C_m(0)=E_m=\log_e\left(\sum_{n=1}^{N}x(n)\right) \tag{15}$$

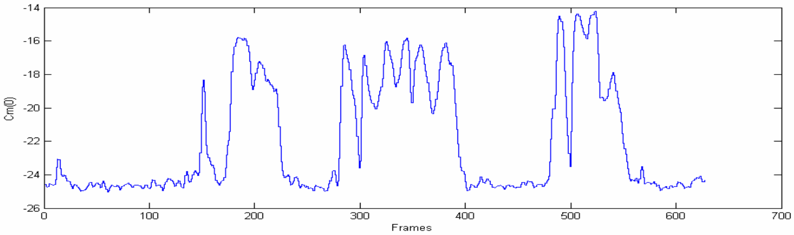

Fig.9 Mel Frequency Cepstral coefficient (0)

Previously defined coefficients were static coefficients. On the other hand in addition to static coefficients in order to increase number of information carried by parameters vector the deltas MFCCs are generated. Delta MFCCs are defined in equation (16):

$$\Delta C_m(n) = \frac{2(C_m(n+2) - C_m(n-2)) - C_m(n-1) + C_m(n+1)}{10} \tag{16}$$

Such prepared 26 element set of coefficients is compressed output of the algorithm for m-th frame.

4.5. MFCCs application in Query By Voice

In near future databases audio data will be a standard data type as numeric data type is now. Such innovation requires advanced query engine. All previously mentioned information was used to design an implement query by voice classification algorithm. Mel Frequency Cepstral Coefficients generation algorithm by use of transformation from time domain to frequency domain and use of perceptual filterbank is able to compress 1024 sample frame to 26 MFCCs. Those coefficients carry information of signal based on human perception. Additionally those features indicate that MFCCs can be used as very good and useful data source for smart sound classifier used in voice samples database search engine. The comprehensive and widely used Query By Humming/Voice algorithm is still not implemented so it is very interesting topic for researchers. The MFCCs generation algorithm and similarity classifier based on MFCCs was implemented by article authors in order to apply it as Query By Voice search mechanism for Oracle Relational Database Management Platform. There were conducted tests based on checking of similarity of MFCCs matrixes for querying single digit in multi-digit sentences. After 2800 tests the accuracy of combined algorithms was estimated approximately on level 82.5% of successful cases.

4. Conclusions

Nowadays Automatic Speech Recognition and database search engines develop very fast. The use of MFCCs in multimedia database search engines gives opportunity to improve accuracy of Query By Humming results. The phase of parameters acquisition is critical stage of speech recognition and classification. Additionally MFCCs and delta MFCCs can be used as parameters stored in database. This gives great possibility to classify them by data mining algorithms or text mining algorithms. This possibility creates a great need for further research to be conducted on improving the accuracy of database querying.

References

[1] Weinstein E.: *Query By Humming: A Survey*, NYU and Google

[2] Ghias A.,Logan J.,Chamberlin D.: *Query by humming - musical information retrieval in an audio database*, ACM Multimedia 95, 1995

[3] Zhu Y., Shasha D.:Warping *Indexes with Envelope Transforms for Query by Humming*, ACM SIGMOD 2003 International Conference on Management of Data, June 9-12, 2003, San Diego, California

[4] Pelikant A., Niewiadomy D.: *Klasyfikator podobieństwa w zapytaniach QBH oparty o współczynniki MFCC*, BDAS 2008

[5] Oracle Corporation: www.oracle.com

[6] Richard G. Lyons, *Wprowadzenie do cyfrowego przetwarzania sygnałów*, 2000

[7] Todor Ganchev, Nikos Fakotakis, George Kokkinakis, *Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task*, Wire Communications Laboratory

[8] Michael Seltzer*, SPHINX III Signal Processing Front End Specification*, 1999