ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 5, March 2016

MFCC Extraction Algorithm for Power Limited Speech Recognition System

R.M.Sneha[1], K.L.Hemalatha[2], S.Sudha[3]

PG Scholar, Dept. of ECE, Easwari Engineering College, Chennai, Tamilnadu, India[1]

Assistant Professor, Dept. of ECE, Easwari Engineering College, Chennai, Tamilnadu, India [2]

Professor, Dept. of ECE, Easwari Engineering College, Chennai, Tamilnadu, India [3]

ABSTRACT*:* Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format.The front end analysis in speech recognition is the spectral analysis which parameterizes the speech signal into feature vectors. The Mel Frequency Cepstral Coefficients (MFCC) is used to extract the features of speech signal. They are based on standard power spectrum estimate which is given to a mel- frequency scale where the signal becomes linear, and then decorrelated by using discrete cosine transform. As the speech recognition system is very sensitive to background noises which degrade the speech signal, it is difficult to process and identify the signal. The main goal is to reduce the noise and improve the signal quality and intelligibility using mel filter making the process easy. The time required to extract the features of speech is minimized by analyzing the MFCC extraction system. As a result, the power consumption is also considerably reduced compared with previous MFCC extraction systems.

KEYWORDS: Speech Recognition, MFCC extraction, Cepstral Coefficients, Power Consumption.

I. INTRODUCTION

Speaker recognition is a process that enables machines to understand and interpret the human speech by making use of certain algorithms and verifies the authenticity of a speaker with the help of a database. First, the human speech is converted to machine readable format after which the machine processes the data. The data processing deals with feature extraction and feature matching. Then, based on the processed data, suitable action is taken by the machine. The action taken depends on the application. Every speaker is identified with the help of unique numerical values of certain signal parameters called “template” or “code book” pertaining to the speech produced by his or her vocal tract. Normally the speech parameters of a vocal tract that are considered for analysis are (i) formant frequencies, (ii) pitch, and (iii) loudness. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC is perhaps the best known, robust, accurate and most popular. In Mel frequency scale the frequency filters are spaced linearly at low frequencies and are logarithmically at high frequencies which have been used to capture the important characteristics of speech. This is an important property of a human ear. Hence the MFCC processor mimics the human ear of perception. This is the process of feature extraction. Pattern recognition does the job of feature extraction which is to classify objects of interest into one of a number of categories or classes. The objects are sequences of acoustic vectors that are extracted from an input.

II.MFCC ALGORITHM

Mel frequency Cepstral coefficients algorithm is a technique which takes voice sample as inputs. After processing, it calculates coefficients unique to a particular sample. In this project, a simulation software called MATLAB R2013a is used to perform MFCC. The simplicity of the procedure for implementation of MFCC makes it most preferred technique for voice recognition.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 5, March 2016

A. GENERATION OF COEFFICIENTS USING MFCC

MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore are best for speech/speaker recognition. The step-by-step computation of MFCC is explained

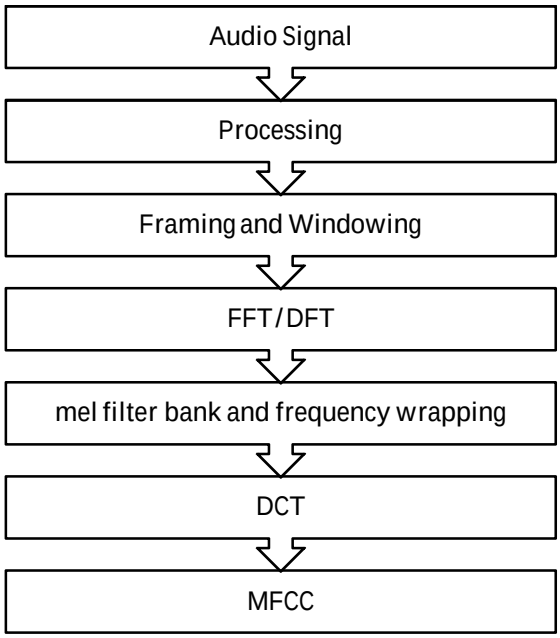


Fig 1. Step by Step Computation of MFCC

B. PRE EMPHASIS

The speech signal x(n) is sent to a high-pass filter :

$y(n) = x(n) - a * x(n - 1)$

where y(n) is the output signal and the value of a is usually between 0.9 and 1.0.

The Z transform of this equation is given by

$H(z) = 1 - a * z^{-1}$

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants.

C. FRAME BLOCKING

The input speech signal is segmented into frames of 15~20 ms with overlap of 50% of the frame size. Usually the frame size (in terms of sample points) is equal to power of two in order to facilitate the use of FFT. If this is not the case, zero padding is done to the nearest length of power of two. If the sample rate is 16 kHz and the frame size is 256 sample points, then the frame duration is 256/16000 = 0.016 sec = 16 ms. Additional, for 50% overlap meaning 128 points, then the frame rate is 16000/(256-128) = 125 frames per second. Overlapping is used to produce continuity within frames.

D.HAMMING WINDOW

Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame. If the signal in a frame is denoted by

x(n), n = 0,…N-1, then the signal after Hamming windowing is,

$x(n) * w(n)$ (3)

where w(n) is the Hamming window defined by

$w(n) = 0.54 - 0.46 * \cos(2\pi n/(N-1))$ where $0 \leq n \leq N-1$

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 5, March 2016

D.FAST FOURIER TRANSFORM

Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore FFT is performed to obtain the magnitude frequency response of each frame. When FFT is performed on a frame, it is assumed that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, FFT can still be performed but the discontinuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem, we multiply each frame by a hamming window to increase its continuity at the first and last points.

E.TRIANGULAR BANDPASS FILTERS

The magnitude frequency response is multiplied by a set of 40 triangular band pass filters to get the log energy of each triangular band pass filter. The positions of these filters are equally spaced along the Mel frequency. From centre frequencies from 133.33 Hz to 1 kHz, there are 13 overlapping (50%) linear filters, while for centre frequencies from 1 kHz to 8 kHz there are 27 overlapping filters spaced logarithmically

F.DISCRETE FOURIER TRANSFORM

In this step, DCT is applied to the output of the N triangular bandpass filters to obtain L mel-scale cepstral coefficients. The formula for DCT is,

$C(n) = \sum Ek * \cos(n * (k - 0.5) * \pi/40))$

where n = 0,1,..to N

where N is the number of triangular bandpass filters, L is the number of mel-scale cepstral coefficients. In this project, there are N = 40 and L = 13. Since we have performed FFT, DCT transforms the frequency domain into a time-like domain called quefrency domain. The obtained features are similar to cepstrum, thus it is referred to as the mel-scale cepstral coefficients, or MFCC. MFCC alone can be used as the feature for speech recognition.

III. MFCC IMPLEMENTATION

The following are the major steps involved in the implementation of the MFCC algorithm:

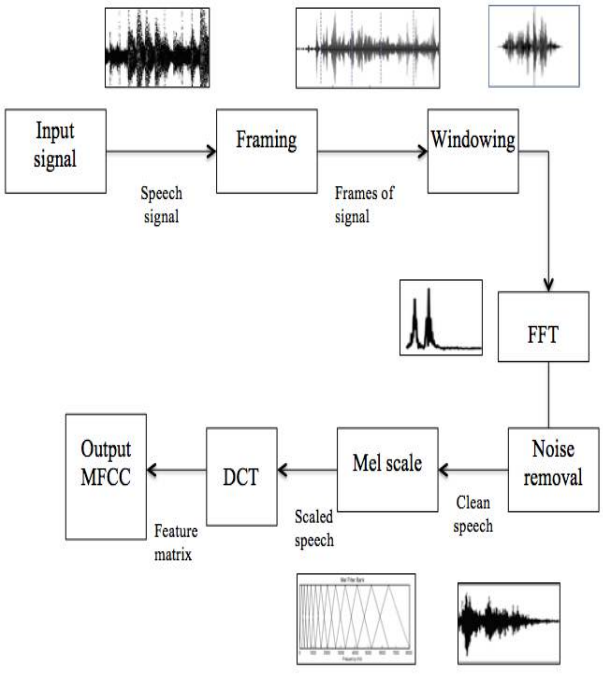


Fig 2. Implementation of MFCC Algorithm

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 5, March 2016

A. RECORDING AND SAMPLING

The recorded speech signals are sampled and stored using Audacity. The sampling is done at a rate of 16000 samples per second. Each speech signal is divided into windows of 16 ms each and hence, 256 samples each. MFCC is implemented for each of these windows and a set of parameters is extracted per window. The first window consists of first 256 samples. The second window overlaps half of the first window and consists of 128 samples of the first window and 128 samples after it. Hence a 50% overlap is used. It is observed that the same speaker saying the same word at two different instants have many variations. So it is important to calculate of the coefficients which almost remain same for a speaker at different instants becomes important.

B.MEL FILTER BANK

There are 40 Mel filters that form Mel filter Bank. Each filter passes a particular set of frequencies corresponding to samples from a frame. For a 256 sample frame, the filter bank spreads over 128 samples only because the FFT is symmetric.

C. MEL FREQUENCY CEPSTRAL COEFFICIENTS

Voice samples of a speaker saying the same word at two different instants are passed through the MFCC algorithm and their respective MFCC Coefficients are extracted.

IV. PROPOSED ARCHITECTURE FOR MFCC EXTRACTION

This section presents a new floating-point MFCC extraction architecture derived to realize the MFCC extraction with a smallhardware cost. This approach is completely different from that have utilized a separate hardware unit for each operation. The proposed architecture is described with setting N to 256, M to 13, and L to 32. For sound signals sampled with 16 bits at 16 kHz, in addition, the bit-widths of F and E in the floating-point representation are determined to 6 and 7 bits, respectively.

One of the buffers stores a half of a sound frame and the other buffer is used to save the remaining data of the frame. Since subsequent frames share a half frame, only one buffer is updated for the next sound frame. The MFCC extraction system can generate MFCC feature vectors continuously by alternatively accessing thedouble buffers. By analyzing the dataflow of the modified MFCC algorithm,we propose a new MFCC extraction system implementable with a small hardware cost. The overall architecture of the proposed system which consists of a multiply-andaccumulate (MAC) unit, an address generation unit, a controller, memories, and counters. Though the proposed architecture has one MAC unit, it is sufficient to process the entire MFCC extraction in real time. The constraint for real-time processing is that the MFCC vectors of a frame should be computed in a time limit corresponding to a half frame. Accordingly, a frame should be processed in 8 ms. The total number of MAC operations in the modified MFCC extraction is ∼15k, meaning that the modified MFCC algorithm can be processed in real time if∼2-M MAC operations are supported in 1 s. This constraint is not hard for a modern embedded system to meet.

We now explain major blocks in detail. For floating-point operations, the fixed-point-to-floating-point unit is included to convert the sound data loaded from the double buffers to the floating-point representation. The MAC unit processes floating-point multiplication and accumulation in serial. Each operator consists of small fixed-point adders and multipliers, and the resultant fraction F is normalized to ensure that

$1 \leq F < 2$ if $F \_ = 0$.

The results of the MAC unit are saved into one of four memories:

1) general purpose registers (GPRs);
2) register files (RFs);
3) C buffers (CBs); and
4) C_ buffers (CBFs).

The GPRs are used to store intermediate values such as the interim sound energy of a frame.The RFs are included to effectively compute such processes storing many values as FFT, mel filtering, DCT, and derivative computations. Grounded on the dataflow analysis of the modified MFCC algorithm, an efficient memory structure consisting of four separate RFs is derived. In terms of memory size, the proposed memory structure is more efficient than those of previous works, since it is shared with several processes.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

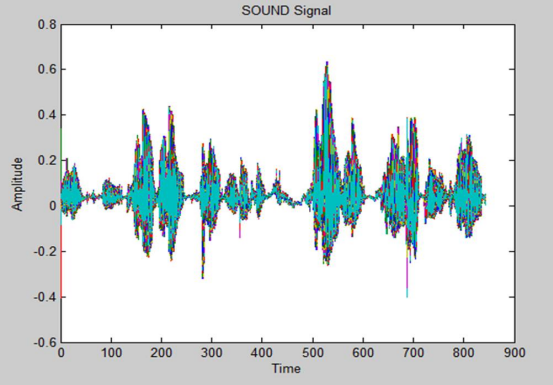
Vol. 5, Special Issue 5, March 2016

To access an entity of a memory, the corresponding address is computed by employing counters. To fetch data for aMAC operation, each counter is increased by a certain amount. The proposed architecture utilizes two counters to generate two addresses needed to access two memories simultaneously.
The controller manipulates all the blocks to process the MFCC extraction algorithm, and its main role is to decide the input signals to be fed to the MAC unit and the storage to be used to store the results.

V. RESULTS

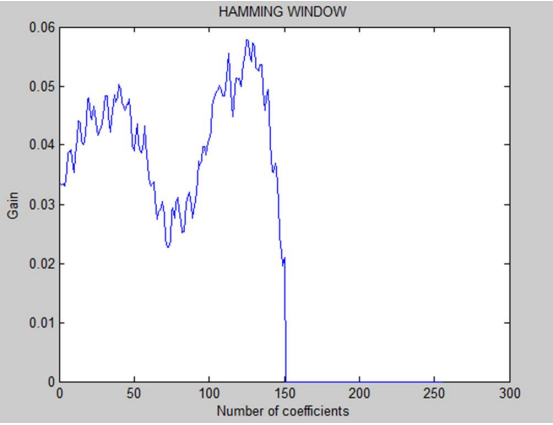
A. INPUT SIGNAL

A real time signal is given as an input through the microphone. The input signal is mixed with other external noises such as whispering, sound of the fan, system noise etc.



B. HAMMING WINDOW

The speech signal is divided into a sequence of frames where each frame can be analyzed independently and represented by a single feature vector. Since each frame is supposed to have stationary behaviour, a compromise, in order to make the frame blocking, is to use a 20-25 ms window applied at 10 ms intervals (frame rate of 100 frames/s and overlap between adjacent windows of about 50%).In order to reduce the discontinuities of the speech signal at the edges of each frame, a tapered window is applied to each one. The most common used window is Hamming window



C. MINIMISED NOISE OUTPUT

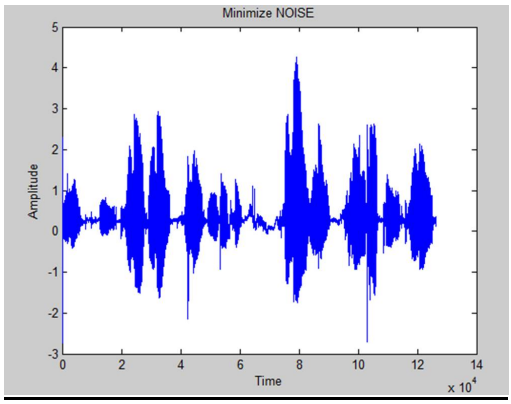
The speech recognition system is very sensitive to noise such as whispering, sound of the running fan etc. This output tells us that the noise in the given input signal is reduced considerably by using the mel filters which separates single frame and overlapped frame. Then each frame is filtered by removing the background noise or unwanted noise

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

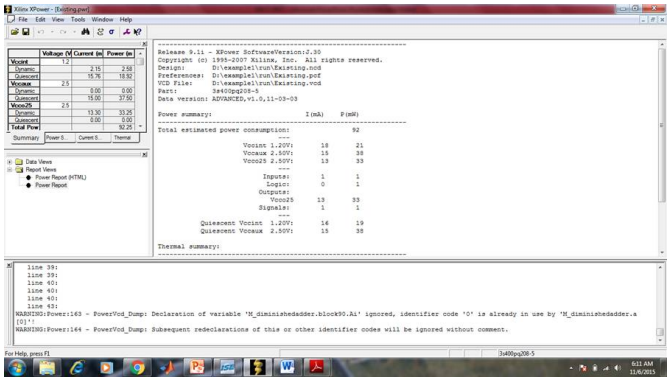
(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 5, March 2016

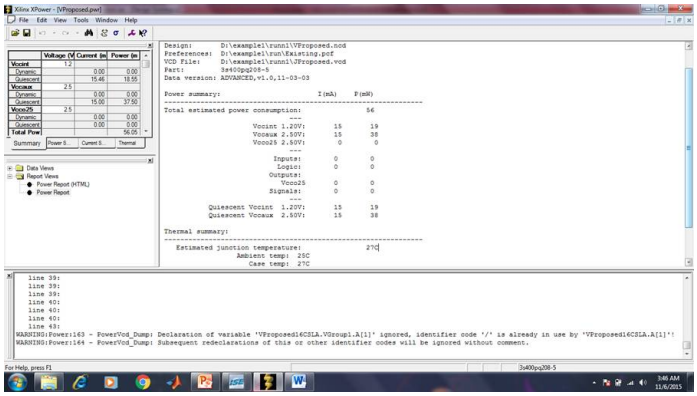


D. POWER ANALYSIS

The power consumed by the MFCC extraction method is high. Multipliers which are used in the filters is a strong operator. It consumes more power therefore replace the multipliers with adders which are a weak operator. Below is the power analysis of filter with multipliers where the power consumed in 92mW.



The power analysis of the filter with adders where the power consumed is 56mW. So the power is reduced considerably.



VI. CONCLUSION

In speech recognition system, the features of the speech are extracted through mel frequency cepstral coefficient. The background noises in the speech signal are reduced with the help of mel filter. Power consumption which is another important factor in speech recognition is reduced by decreasing the computation time of the MFCC extraction system.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 5, March 2016

REFERENCES

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet ofThings (IoT): A vision, architectural elements, and future directions," *Future Generat. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660,Sep. 2013.
[2] H.-W. Hon, "A survey of hardware architectures designed for speechrecognition," Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-91-169, Aug. 1991.
[3] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*.Englewood Cliffs, NJ, USA: Prentice-Hall, 1993, pp. 1–9.
[4] D. R. Reddy, "Speech recognition by machine: A review," *Proc. IEEE*,vol. 64, no. 4, pp. 501–531, Apr. 1976.
[5] S. Davis and P. Mermelstein, "Comparison of parametric representationsfor monosyllabic word recognition in continuously spoken sentences,"*IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4,pp. 357–366, Aug. 1980.
[6] S. Nedevschi, R. K. Patra, and E. A. Brewer, "Hardware speechrecognition for user interfaces in low cost, low power devices," in *Proc.42nd DAC*, Jun. 2005, pp. 684–689.
[7] N.-V. Vu, J. Whittington, H. Ye, and J. Devlin, "Implementation ofthe MFCC front-end for low-cost speech recognition systems," in *Proc. ISCAS*, May/Jun. 2010, pp. 2334–2337.
[8] P. EhKan, T. Allen, and S. F. Quigley, "FPGA implementation forGMM-based speaker identification," *Int. J. Reconfig. Comput.*, vol. 2011, no. 3, pp. 1–8, Jan. 2011, Art. ID 420369.
[9] R. Ramos-Lara, M. López-García, E. Cantó-Navarro, andL. Puente-Rodriguez, "Real-time speaker verification systemimplemented on reconfigurable hardware," *J. Signal Process. Syst.*,vol. 71, no. 2, pp. 89–103, May 2013.
[10] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum:A guide to processing," *Proc. IEEE*, vol. 65, no. 10, pp. 1428–1443, Oct. 1977.
[11] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient MFCCextraction method in speech recognition," in *Proc. IEEE ISCAS*, May 2006, pp. 145–148.