IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 3, Issue 1, January 2015 ISSN 2321-5992

# Comparison of conventional MFCC with new Efficient MFCC Extraction Method in Speech Recognition

**Mahaveer Chougala[1], A.H.Unnibhavi[1]**

[1]Department of Electronics and Communication Engineering, Basaveshwar Engineering College, Bagalkot, Karnataka, INDIA

**ABSTRACT**

***this paper introduces a new method of extracting MFCC for speech recognition and it is compared with the conventional MFCC method. The new algorithm reduces the calculation steps by 53% compared to conventional method. Simulation result indicates the new method has a recognition accuracy of 92.93% only 1.5% less than the conventional MFCC method which is has accuracy of 94.43%. However, the number of logic gates required to implement the new method is about half of the conventional MFCC method, which makes a new method very efficient in speech recognition.***

**Key words:**-Automatic Speech Recognition (ASR), Mel-Frequency Cepstral Coefficients (MFCC).

## 1. INTRODUCTION

Speech recognition is more commonly known as automatic speech recognition (ASR), is the process of interpreting human speech in a computer. In technical definition ASR is system which maps acoustic signals into the strings of words. There are several kinds of parametric representations for the acoustic signals. Among them the Mel-Frequency Cepstrum Coefficients (MFCC) is the most widely used. There are many reported works on MFCC, especially on the improvement of the recognition accuracy. However, all these algorithms require large amount of calculations, which will increase the cost and reduce the performance of the hardware speech recognizer. In this paper we propose a novel and an efficient way to calculate MFCC. Section II introduces the conventional MFCC extraction algorithm. The new approach is introduced in Section III and the comparison between the two methods presented in Section IV. Section V presents the simulation results for the various feature sets and recognition accuracy for each feature set. The speech signal has a 10 dB signal-to-noise ratio and a spectrum between 0.3 kHz to 3.4 kHz at a sampling frequency of 8 kHz.

### 1.1 Challenges

The general problem of automatic transcription of speech by any speaker in any environment is still far from solved. But recent years have seen ASR technology mature to the point where it is viable in certain limited domains.

### 1.2 Difficulties

One dimension of variation in speech recognition tasks is the vocabulary size.

A second dimension of variation is how fluent, natural or conversational the speech is isolated word recognition, in which each word is surrounded by some sort of pause, is much easier than recognizing continuous speech.

A third dimension of variation is channel and noise. Commercial dictation systems, and much laboratory research in speech recognition, is done with high quality, head mounted microphones.

## 2. METHODOLOGIS

### 2.1 Conventional MFCC extraction method

Figure 1 shows the block diagram of the conventional MFCC extraction method.

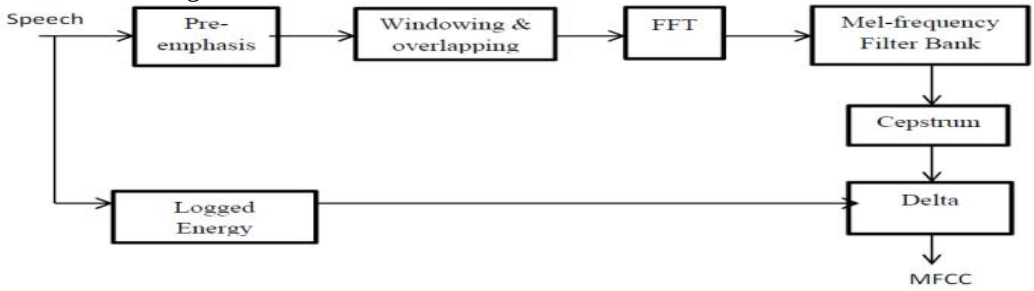


**Figure1.** Block diagram of conventional MFCC extraction method.

IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 3, Issue 1, January 2015 ISSN 2321-5992

The speech is first pre-emphasized with pre-emphasis filter to flatten the spectrum of the speech signal and to increase the amplitude of the high frequencies in the speech signal. In the time domain the relation between input and output of the pre-emphasis filter is shown in (1).

$$y_n = x_n - ax_{n-1} \qquad (1)$$

Where "a" is the value between 0 and 1. The default value of a is 0.97.
Then pre-emphasized speech is separated into short segments called frame. The frame duration is set to 20ms (160 speech samples) and the 50% overlapping between each adjacent frame is considered to ensure stationary in the frames. To reduce the discontinuities in the frames 160 length hamming window is multiplied with each of the frames which is shown in the figure 2.

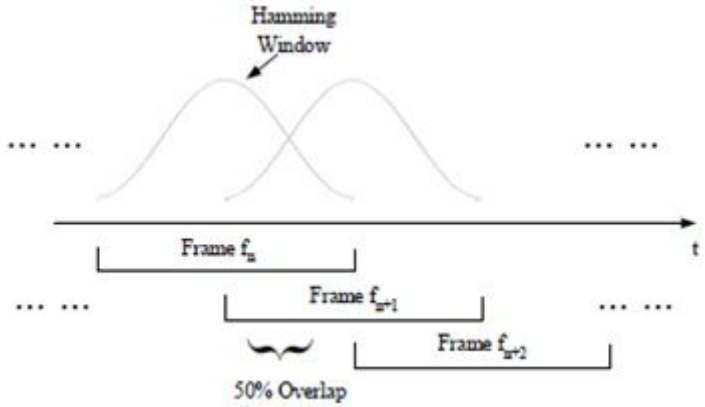


**Figure 2.** Frame blocking and overlapping in conventional MFCC method.

The hamming window is generated by the below equation (2),

$$w(n) = 0.53836 - 0.46164\cos(\frac{2\Pi n}{N-1}) \qquad (2)$$

Where N is equal to 160, the number of points in one frame, and n is from 1 to N.

After windowing the speech signal 256 point FFT is calculated from each frame to convert speech from time domain to frequency domain and to obtain good frequency resolution. After the FFT block, the spectrum of each frame is filtered by a set of filters, and the power of each band is calculated. Because of the symmetry property of FFT, we only need to calculate the first 128 coefficients. The filter bank consists of 33 triangular shaped band-pass filters, which are centred on equally spaced frequencies in the Mel domain between 0 Hz and 4 kHz, as shown in Fig. 3. The mapping from linear frequency to Mel-Frequency is shown in equation (3).

$$Mel(f) = 2595 * \log_{10}(1 + \frac{f}{700}) \qquad (3)$$

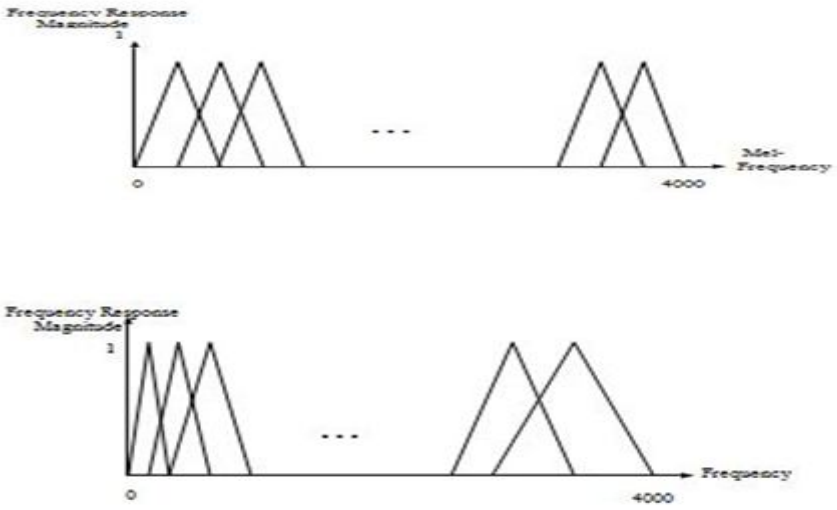


**Figure 3.** Mel-Frequency filter bank in Mel scale and normal scale.

In the final step we calculate the mel frequency cepstrum from the output power of the filter bank so as to transform data from mel scale to time scale and the result itself is called mel frequency cepstral coefficients and which can be calculated using equation (4),

$$Cn = \sum_{k=1}^{K} (\log Xk) \cos[n(k-0.5)\frac{\Pi}{K}] \quad (4)$$

Where is the output power of the filter of the filter bank, and n is from 1 to 12. We can also calculate the logged energy of each frame as one of the coefficients using equation (5),

$$E = \log \sum_{n=1}^{160} X^2 n \quad (5)$$

This is calculated without any windowing and pre-emphasis. Up to now we have got 13 cepstrum coefficients. To enhance the performance of the speech recognition system, time derivatives are added to the basic static parameters. The delta coefficients are obtained from the equation (6),

$$dc_t = \frac{2(c_{t+2} - c_{t+2}) + (c_{t+1} - c_{t-1})}{10} \quad (6)$$

After all the calculations, the total number of MFCC for one frame is 26.

**2.2 New efficient MFCC method for feature extraction**

In conventional MFCC method each frame requires 160 multiplications for the window operations, 128xlog2(256) multiplications for the FFT calculation, 128 multiplications for the filter power calculation and 33x12 multiplications for the DCT calculation. A total of 1708 multiplications are required for each frame, which requires a huge amount of computational power. The proposed a new MFCC method that only requires half of the multiplication steps as that of conventional MFCC method. The block diagram of new MFCC method is shown in Fig. 4.

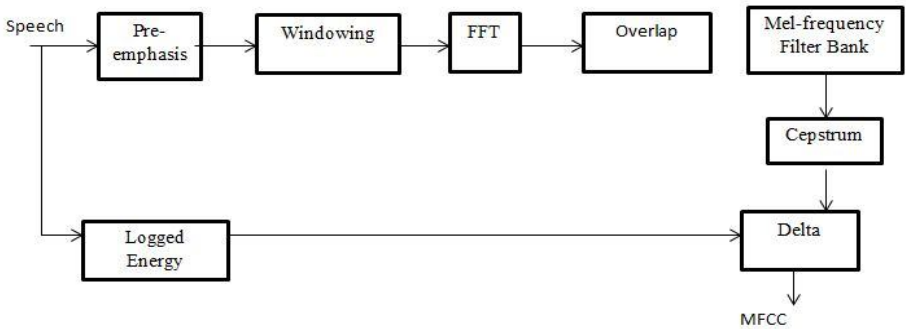


**Figure 4**. Block diagram of the proposed MFCC extraction method.

In this method there is no change to the pre-emphasis block. However, we make modification in the step (1) to eliminate the multiplication step. Approximate the value of "a" by 31/32 instead of considering it as 0.97, then equation (1) becomes,

$$y_n = x_n - ax_{n-1} = x_n - \frac{31}{32}x_{n-1}$$

$$y_n = x_n - (x_{n-1} - \frac{x_{n-1}}{32}) \quad (7)$$

We have replaced the multiplication n−1 as in (1) with simple addition and shift operations. The divide by 32 operations in (7) is simply shifting the binary number 5 bits to the right. The complex multiplication operation is replaced with simple shift operation without affecting the recognition accuracy. The original approach shown in Fig. 1 combines the window and overlap functions. In the new design, we move the overlap function after the filter bank as shown in Fig. 4. The speech is separated into segments called sub-frame here. One sub-frame is composed of 80 points and no overlap between them. Thus, one can picture a conventional frame fn consisting of two adjacent sub-frames sfn and sfn+1, as shown in Fig. 5. As stated in Section II, the Hamming window is used mainly to reduce the edge effect,

IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 3, Issue 1, January 2015 ISSN 2321-5992

so the length of the Hamming window can be reduced from 160 points to 80 points consequently. As if the window size becomes smaller, the short-time spectrum will give a poorer frequency resolution but a better estimate of the overall spectral envelope, this modification will affect the recognition accuracy slightly as shown in Section IV.

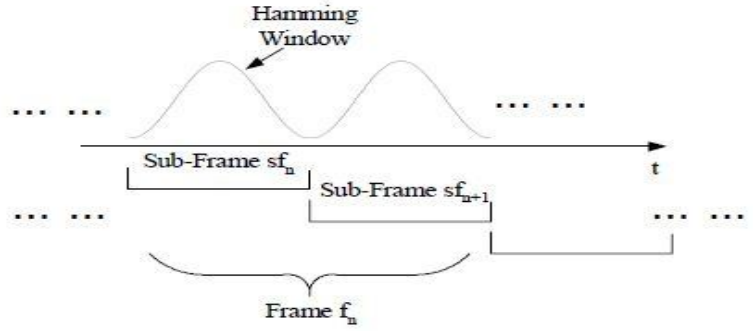


**Figure 5.** Frame blocking and overlapping in the proposed MFCC extraction method.

The next block is the FFT, which is the same as the original. However, the calculation is reduced by half to 128 points because of the new window size. We only need to calculate the first 64 coefficients because of the FFT symmetry. We modify the filter bank from equally spaced triangular filters as shown in Fig. 3 to equally spaced rectangular filters as shown in Fig. 6. A filter bank is acceptable for speech recognition so far as its composite frequency response is flat over the entire frequency range of interest. Thus, a rectangle filter bank satisfies this requirement. In the conventional approach, the FFT outputs are multiplied by the characteristic of the triangular filter to generate the filter outputs and then these filter outputs are summed to generate the power of each filter. However, if we use a rectangular filter, the output characteristic of a rectangular filter is either a "1" or a "0". For a 128-point FFT, the rectangular filter bank is reduced to 23 equally spaced rectangular filters which indicates 23 filters produces the highest recognition accuracy. The original triangular filters require 128 multiplications per frame. However, the rectangular filters only require 120 additions per frame.


**Figure 6.** The new Mel-Frequency filter bank. (The Magnitude of all filters is 1.

The new overlap method is illustrated in Fig. 7, where fn and fn+1 represent the original 160-point frames with 50% overlap, and sfn and sfn+1 represent the new 80-point non overlapped sub-frames. We add the filter bank outputs SFn,k and SFn+1,k to generate the power coefficient Sn,k. The next power coefficient Sn+1, k is equal to the sum of the filter outputs of sub-frame sfn+1 and sfn+2. Thus, Sn,k and Sn+1,k are identical to the kth power coefficient of the original frame fn and fn+1. We have reduced almost half of the computation by moving the overlap operation to the end of the spectrum calculation.

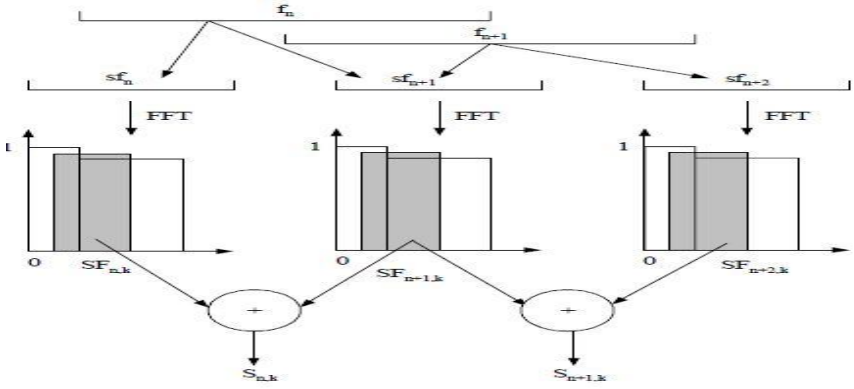


**Figure 7.** Frame blocking and overlapping in the new method.

IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 3, Issue 1, January 2015 ISSN 2321-5992

The following DCT and delta calculations are the same as the conventional algorithm. There are also 26 features in each frame as the original algorithm. The new feature extraction algorithm reduces the total number of multiplications from 1708 to 804 per frame. Each frame of the new algorithm requires 80 multiplications for the window operation, 64xlog2(128) multiplications for the FFT calculation, and 23x12 multiplications for the DCT calculation.

## 3. COMPARISON OF CONVENTIONAL MFCC WITH THE NEW MFCC METHOD

The table I gives the comparison between the original method and the new method

**Table 1**. Comparison of conventional MFCC with the new MFCC method

| Parameters | Conventional MFCC method | New efficient MFCC method |
|---|---|---|
| Window length | 160 points | 80 points |
| FFT points | 256 points | 128 points |
| Type of Filter bank | Triangular | Rectangular |
| Number of filters | 33 | 23 |
| Number of filter coefficients | 128 | 64 |
| Number of multiplications required | 1708 | 804 |
| Computation power | High | Exactly half of the Conventional method |
| Recognition accuracy | 94.13% | 92.93% |

## 4. RESULTS

The recognition results of the new and old methods with different window lengths and FFT points are summarized in Table II. F1 is the original MFCC features extraction method. F5 is the proposed new method. As we have pointed out in Section III, if we change the “a” value of equation (1) from 0.97 to 31/32 then the multiplication step in the pre-emphasis block can be replaced with shift operation without affecting the recognition accuracy. This approximation is verified by the fact that there is no change of recognition accuracy of F1 and F2. F3 segments the speech into 80-point sub-frames, using the proposed architecture in Fig. 4 but keeping the other settings as same as F1. There is a slight drop in recognition accuracy. F3 and F4 compare the recognition accuracy between triangular and rectangular filters. F4 and F5 compare the new method at different FFT points. From Table II, we can find that F5 produces relatively high recognition accuracy with the minimum requirement of calculation power.

**Table 2.** Recognition accuracy of different feature set

| Feature set | a | Window Length | FFT Point | Filter Shape | No of Filters | Recognition Accuracy |
|---|---|---|---|---|---|---|
| F1 | 0.97 | 160 | 256 | Triangle | 33 | 94.43% |
| F2 | 31/32 | 160 | 256 | Triangle | 33 | 94.43% |
| F3 | 31/32 | 80 | 256 | Triangle | 33 | 92.29% |
| F4 | 31/32 | 80 | 256 | Rectangle | 33 | 92.08% |
| F5 | 31/32 | 80 | 128 | Rectangle | 23 | 92.93% |

## 5. CONCLUSION

We have demonstrated that the new extraction method that reduces the number of multiplications from 1708 to 804 with only 1.5% drop in recognition accuracy. The new method is more efficient for hardware implementation than the original method. We expect the new method will have significant improvements on the hardware performance such as power consumption, speed, and cost.

## REFERENCES

[1] Wei HAN, Cheong-Fat CHAN et. Al "An Efficient MFCC Extraction Method in Speech Recognition" The Chinese University of Hong Kong.
[2] Steven B. Davis and Paul Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, No. 4, August 1980.
[3] L. Rabiner and Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall PTR, c1993.
[4] B. A. Dautrich, L. R. Rabiner, and T. B. Martin. "On the effects of varying filter bank parameters on isolated word recognition." IEEE Trans. Acoust., Speech, Signal Processing, ASSP-31(4):793-807, 1983.
[5] L. Muda, et al., "Voice recognition algorithm Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," Journal of computing vol. 2, 2010.
[6] B. A. Dautrich, L. R. Rabiner, and T. B. Martin. "On the effects of varying filter bank parameters on isolated word recognition." IEEE Trans. Acoust., Speech, Signal Processing, ASSP-31(4):793-807, 1983.
[7] Fang Zheng, Guoliang Zhang and Zhanjiang Song, "Comparison of Different Implementations of MFCC", J. Computer Science & Technology, 16(6): 582-589, Sept. 2001.