Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# MODIFIED MFCC METHODS BASED ON KL- TRANSFORM AND POWER LAW FOR ROBUST SPEECH RECOGNITION

[1]JOHN SAHAYA RANI ALEX, [2]NITHYA VENKATESAN
[1]School of Electronics Engineering, VIT, Chennai, INDIA
[2] School of Electrical Engineering, VIT, Chennai, INDIA

E-mail: [1]jsranialex@vit.ac.in , [2]nithya.v@vit.ac.in

**ABSTRACT**

This paper presents robust feature extraction techniques, called Mel Power Karhunen Loeve Transform Coefficients (MPKC), Mel Power Coefficients (MPC) for an isolated digit recognition. This hybrid method involves Stevens' Power Law of Hearing and Karhunen Loeve(KL) Transform to improve noise robustness. We have evaluated the proposed methods on a Hidden Markov Model (HMM) based isolated digit recognition system with TIDIGITS data for clean speech and also with noisy speech data. An increase in the recognition accuracy rate is observed with the proposed methods compared to conventional Mel Frequency Cepstral Coefficients (MFCC) technique.

**Keywords:** *Feature extraction; Stevens' power law; MPKC; KL Transform; CMN; MFCC; HMM*

## 1. INTRODUCTION

Automatic Speech Recognition (ASR) can be defined as an independent, computer-driven transcription of spoken language into readable text in real time. Most of the speech recognition applications are used in noisy environment. The main stages in ASR are feature extraction, training and decoding [1] as shown in Figure 1.

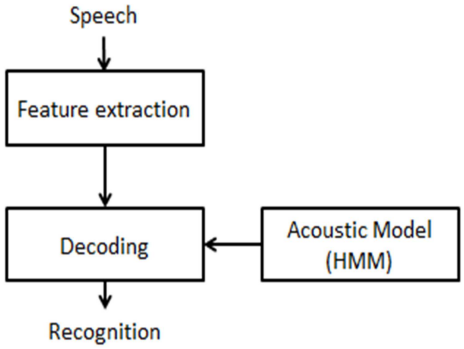


*Figure 1: Block Diagram Of The HMM-Based Speech Recognition System.*

Speech is considered as a non-stationary signal hence, to make it quasi stationary, short speech segments of 25 ms are considered for feature extraction. The speech signal is blocked into frames with adjacent frames overlapped. Each individual frame is passed through a window so as to minimize the signal discontinuities at the beginning and end of each frame. The feature extraction method is then applied to this frame to obtain a vector that represent useful characteristics of that particular frame which are used for training and recognition. Feature extraction method plays a vital role in speech recognition system. The most popular methods are Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), and Linear Predictive Coding (LPC).

Much study has been done to improve the performance of the conventional MFCC method [2]. Shannon and Paliwal [3] studied the effect of Mel, Bark and Uniform filter banks on the recognition accuracy. Zhao and Wang [4] performed a detailed analysis on each stage of the conventional MFCC method and compared its performance with that of Gammatone Frequency Cepstral Coefficients (GFCC). Hossan, Memon and Gregory [5] proposed a speaker verification test where they have used on Discrete Cosine Transform (DCT) - II in the conventional MFCC procedure.

Our aim is to introduce modifications in the MFCC feature extraction method so that the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

modified MFCC method should give better recognition accuracy in the noisy environment. The modification suggested in this paper is the usage of exponent 0.33 for intensity-loudness conversion according to Stevens' power law [6] and application of KL Transform for de-correlation and energy compaction [7]. By applying these modifications in MFCC method, we expect a better performance in recognition accuracy in the noisy environment.

This paper is organized as follows: Section II briefly describes the conventional MFCC method, Stevens' power law, Karhunen-Loève Transform and Cepstral Mean Normalisation. The proposed changes are incorporated in the conventional method in Section III. Section IV analyses the performance of discussed methods. Finally, a brief observation is given in Section V.

## 2.FEATURE EXTRACTION

### 2.1 Mel Frequency Cepstral Coefficients Method

This analysis technique uses cepstrum with a nonlinear frequency axis called Mel scale. The signal is first segmented into short overlapping frames and passed through a Hamming window. Next the power spectrum of each frame is obtained and applied to the Mel filter bank which contains 26 filters to obtain energies in various frequency regions. The Mel scale relates perceived frequency of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Frequency in Mel scale is obtained by the formula :

$$M(f) = 1125 * ln(1 + f/700) \quad (1)$$

where f is frequency in Hertz scale and M(f) is frequency in Mel scale.

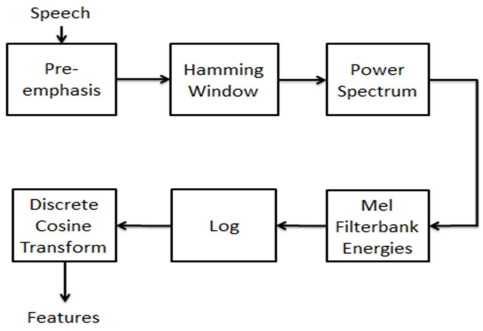


*Figure 2: Block Diagram Of MFCC Method*

After obtaining the filter-bank energies, logarithmic compression is performed and energy compaction is done using DCT. Of the 26 DCT coefficients obtained, only the lower 13 are retained and the rest are ignored. This is because the lower DCT coefficients represent the smooth spectral variations of the speech signal. A block diagram of the MFCC method is shown in Figure 2.

### 2.2 Stevens' Power Law

Stevens' power law [6] describes a relationship between the magnitude of a physical stimulus and its perceived intensity or strength.

The general form of the law is

$$\psi(I) = k\, I^{a} \quad (2)$$

where I is the magnitude of the physical stimulus, $\psi(I)$ is the subjective magnitude of the sensation evoked by the stimulus, $a$ is an exponent that depends on the type of stimulation and k is a proportionality constant. For loudness, the exponent considered here for our study is to be 0.33.

### 2.3 Karhunen-Loève Transform

Karhunen-Loève (KL) Transform takes a given collection of data i.e. an input collection which creates an orthogonal basis for the data. The basis vectors of the KL Transform are the eigenvectors of the covariance matrix. Its effect is to diagonalize the covariance matrix, removing the correlation of neighbouring elements. After KL Transform is performed, most of the energy of the transformed coefficients is concentrated within the first few components. This is the energy compaction property of KL Transform.

The KL Transform [8] is calculated using the following steps performed by finding the eigenvectors of the covariance matrix.

1. Divide the signal into sub-vectors.
2. Obtain the covariance matrix of the subvectors using the formula

$$\text{cov}(x) = E[x.x^{T}] - x_m\, x_m^{T} \quad (3)$$

where x is the sub-vector formed from the original linear vector, $x_m$ is the mean of the input vectors.

3. Determine the eigenvalues of the covariance matrix using the formula,

$$|\text{cov}(x) - \lambda I| = 0 \quad (4)$$

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. Determine the eigenvectors of the covariance matrix using ,

$$(\text{cov}(x) - \lambda_n I)\, \varphi_n = 0 \quad (5)$$

where I is the identity matrix and $\varphi_n$ is the eigenvector corresponding to the eigenvalue $\lambda_n$.

5. Form a transformation matrix which contains all the eigenvectors T.
6. Retain only those eigenvectors whose corresponding eigenvalue magnitudes are high, to achieve dimensionality reduction. Thus KL Transform of input matrix x is given by

$$Y = T^T x \quad (6)$$

### 2.4 Cepstral Mean Normalisation

Cepstral Mean Normalisation (CMN) [9] is one of the simplest feature normalization techniques available. Consider a sequence of cepstral vectors $\{s_1, s_2, \ldots, s_T\}$ obtained from a speech sample. CMN involves subtracting the mean feature vector $\mu_s$ from each vector $s_t$ and normalizing by variance $\sigma_s$ to obtain the normalized vector $\hat{s}_t$ .

$$\hat{s}_t = \frac{x_t - \mu_s}{\sigma_s} \quad (7)$$

Where

$$\mu_s = \frac{1}{T}\sum_t s_t \quad \text{and} \quad \sigma_s^2 = \frac{1}{T}\sum_{t=1}^{T}(s_t^2 - \mu_s^2)$$

## 3 PROPOSED METHODOLOGY

### 3.1 MPKC Method

The proposed method Mel-Power-KLT-Coefficients (MPKC) is obtained by making the following changes to the conventional MFCC method.

The MFCC procedure uses logarithmic compression for intensity-loudness conversion because human ears perceive loudness in decibel scale. The method proposed here replaces this logarithmic compression with exponential compression with the exponent being 0.33 in accordance with Stevens' Power Law. The coefficients obtained are squared to desensitize the MFCC coefficients to spurious low-energy spectral perturbations [10], hence making the effective exponent as 0.66.

The MFCC uses DCT for de-correlation and energy compaction of Mel filter bank energies. Its performance is very close to the KL Transform. It is optimum in energy compaction and it also has applications in noise reduction [11]. Using KL transform 19 feature vectors are extracted per frame. Figure 3 shows the block diagram of MPKC method.

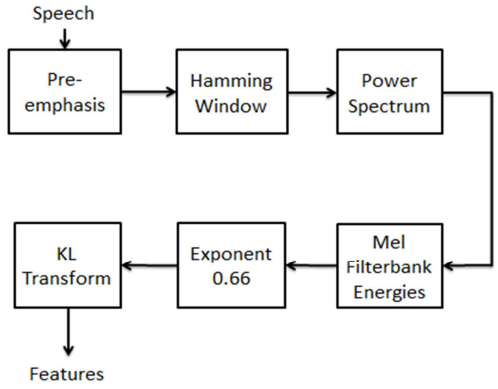


*Figure 3: Block Diagram Of MPKC Method*

### 3.2 MPC Method

The proposed Mel-Power-Coefficients (MPC) method is obtained by making the following changes to the conventional MFCC method.

The MFCC procedure uses logarithmic compression for intensity-loudness conversion because human ears perceive loudness in decibel scale. The method proposed replaces logarithmic compression with exponential compression with the exponent being 0.33 in accordance with Stevens' Power Law. Figure 4 shows the block diagram of MPC method. The dimension of feature vector per frame is 13 same as that of MFCC.

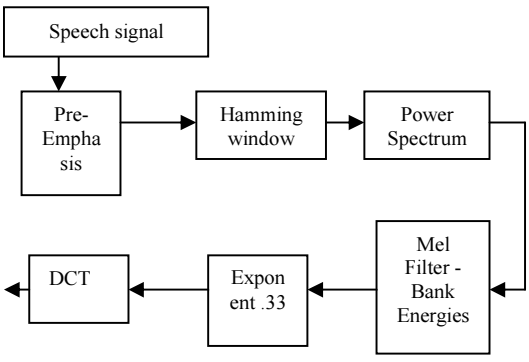


*Figure 4: Block Diagram Of MPC Method*

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

## 4.EXPERIMENTAL SETUP AND RESULTS

The isolated digit recognition system [12] is implemented with Single Gaussian HMM [13] using MATLAB. We are using samples from the TIDIGITS [14] database for training and testing of speech features. Each digit (zero, one, two, three, four, five, six, seven, eight, nine and oh) is spoken twice by each speaker and total 4576 utterances are used for training and testing. Each digit is modeled as one HMM. Using the training utterances, the HMMs are trained. Viterbi decoder is used as the decoder in HMM based isolated digit recognition system. In all the experiments, the speech signal is divided into frames of duration 25 ms, with 10 ms overlap between adjacent frames. 13 MFCC feature vectors per frame are considered for this research as baseline. Recognition rates are obtained for baseline MFCC, proposed MPKC,MPC method with the clean speech. Next, Additive White Gaussian Noise(AWGN) of various Signal to Noise Ratio(SNR) (20 dB, 10 dB, 5 dB, 0 dB) are added to the clean speech, feature vectors are extracted from this noisy speech and used to check the recognition rate of noisy speech for the proposed and baseline methods. Finally, feature vector normalization method namely Cepstral Mean Normalization (CMN) is applied to study the effect of normalization in the noisy feature vectors and the results are analyzed.

For the conventional MFCC method, a recognition rate of 96.98 % is achieved for the clean TIDIGITS whereas for the MPKC method, a rate of 94.81 % is achieved. Table 1 compares the recognition rates of both methods in the presence of AWGN of various SNR.

*Table 1: Recognition Rate Of MFCC And The Proposed Methods*

| SNR | Conventional MFCC | MPC | MPKC |
|---|---|---|---|
| Clean | 96.98 | 96.02 | 94.81 |
| 20 | 90.27 | 89.62 | 87.49 |
| 10 | 73.42 | 73.25 | 69.67 |
| 5 | 50.20 | 50.24 | 45.53 |
| 0 | 25.02 | 24.26 | 22.45 |

According to our research hypothesis, MPKC/MPC method should have outperformed than the conventional MFCC, but it did not. New set of feature vectors are formed by concatenating MFCC with the MPC, MFCC with the MPKC and the same experiment is performed for the same set of noisy data. It is observed that the concatenated vectors resulted in better performance than the conventional MFCC.

*Table 2: Recognition Rates Of MFCC, [MFCC,MPC] And [MFCC,MPKC]*

| SNR | MFCC | [MFCC, MPC] | [MFCC, MPKC] |
|---|---|---|---|
| Clean | 96.98 | 97.79 | 97.06 |
| 20 | 90.27 | 90.91 | 90.47 |
| 10 | 73.42 | 74.62 | 73.81 |
| 5 | 50.20 | 53.34 | 52.37 |
| 0 | 25.02 | 27.35 | 26.47 |

Table 2 compares the recognition rates of the conventional MFCC and the new method obtained by combination of MFCC , MPKC and MPC in the presence of AWGN of various SNR. Figure 5 compares the recognition rate of all the three methods.

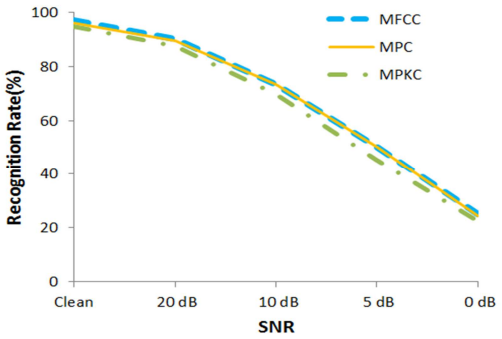


*Figure 5: Comparison Of MFCC, MPC And MPKC Feature Extraction Methods*

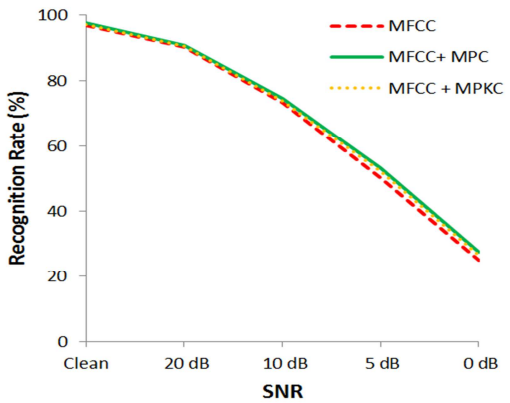


*Figure 6: Comparison Of MFCC, MFCC+MPC And MFCC+MPKC Feature Extraction Methods*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3 compares the recognition rates of the conventional MFCC and the new method [MFCC,MPKC],[MFCC,MPC] after the application of CMN.

*Table 3: Recognition Rates Of MFCC, [MFCC,MPC] And [MFCC,MPKC] With CMN*

| SNR | [MFCC] | [MFCC, MPC] | [MFCC, MPKC] |
|---|---|---|---|
| 20 | 92.24 | 92.96 | 92.28 |
| 10 | 75.22 | 78.00 | 76.51 |
| 5 | 47.51 | 52.65 | 51.17 |
| 0 | 20.51 | 22.73 | 22.65 |

Figure 7 compares the recognition rates of MFCC and [MFCC,MPKC],[MFCC,MPC] method after the application of CMN.

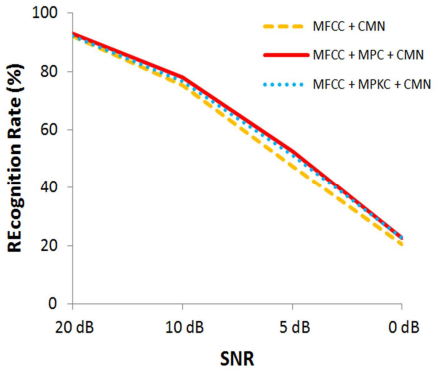


*Figure 7 Recognition Rates Of MFCC And Proposed Methods With CMN*

The proposed methods perform better than the conventional MFCC when applied on clean and noisy speech which is as shown in Figure 6 and 7.

Second set of experiments were done using the samples recorded from a classroom. These samples consist of isolated digits from 1 to 10, spoken in Tamil. The samples are recorded from 21 different speakers, in the presence of varying degree of babble noise. The proposed feature extraction methods are used to obtain the recognition accuracy of the isolated digits. Table 4 shows the recognition rates of the conventional MFCC and the proposed methods in the presence of babble noise before and after the application of CMN.

*Table 4: Recognition Rates Of MFCC And Proposed Methods With And Without CMN*

| | MFCC | [MFCC MPC] | [MFCC, MPKC] |
|---|---|---|---|
| Without CMN | 96.67 | 95.24 | 91.9 |
| With CMN | 98.57 | 99.52 | 94.29 |

## 5.CONCLUSION

In this paper, two feature extraction methods are proposed to perform better than the conventional MFCC method both in clean and noisy environment. The first method, called MPC, involves the use of Stevens' Power Law for intensity-loudness conversion of Mel-filter bank energies instead of logarithm that is used in the conventional MFCC method. The second method, called MPKC uses Stevens' Power Law for intensity-loudness conversion and KL Transform instead of the DCT for de-correlation and energy compaction of coefficients. Both methods when used individually had almost the same performance in the clean and noisy environment as that of MFCC. But when the feature vectors of MPC, MPKC were concatenated with MFCC, the recognition accuracy exceeds the baseline MFCC method accuracy. Concatenation of feature vectors results in increase in the dimensions of feature vectors per frame. It is observed that these methods perform much better in the lower SNR regions giving an improved accuracy by 2-5 %. Another set of experiments, which are conducted on the samples recorded from a classroom, shows that the proposed methods outperforms the conventional MFCC by 1%. It is also observed that the proposed configuration along with the feature vector normalization method such as CMN resulted increase in recognition accuracy in the higher SNR such as 20dB,10dB not in the lower SNR 5dB,0dB .

## REFERENCES:

[1] Iosif Mporas, Todor Ganchev, Mihalis Siafarikas, Nikos Fakotakis, "Comparison of Speech Features on the Speech Recognition Task", *Journal of Computer Science* 3 (8): 608-616, 2007.

[2] Clarence Goh Kok Leon, "Robust computer voice recognition using improved MFCC algorithm", *International Conference on New Trends in Information and Service Science*, IEEE 2009.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[3] Ben J. Shannon, Kuldip K. Paliwal, "A comparative study of filter bank spacing for speech recognition", *Microelectronic Engineering Research Conference* 2003.

[4] Xiaojia Zhao, DeLiang Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification", *IEEE International Conference on Acoustics, Speech, and Signal Processing* 2013.

[5] Md. Afzal Hossan, Sheeraz Memon, Mark A Gregory,, "A novel approach for MFCC feature extraction", 978-1-4244-7907-8/10 , *IEEE* 2010.

[6] Stevens, S.S. (1957). "On the psychophysical law," Psychol.Rev. 64, 153-181.

[7] Ing Yann Soon, Soo Ngee Koh, Chai Kiat Yeo, "Noisy Speech Enhancement Using Discrete Cosine Transform", *ELSEVIER Speech Communication* 24 (1998) 249-257.

[8] Jayaraman, "Digital Image Processing", *Tata McGraw-Hill Education*, 2011.

[9] Navnath S Nehe and Raghunath S Holambe, "DWT and LPC based feature extraction methods for isolated word recognition", *EURASIP Journal on Audio, Speech, and Music Processing* 2012.

[10] Vivek Tyagi and Christian Wellekens, "On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition" *IEEE ICASSP* 2005.

[11] Jingdong Chen, Jacob Benesty and Yiteng Huang, "Study of the noise-reduction problem in the Karhunen–Loève expansion domain", *IEEE Transactions On Audio, Speech, And Language Processing*, Vol. 17, No. 4, May 2009.

[12] Santosh V. Chapaneri, Deepak J. Jayaswal, "Efficient speech recognition system for isolated digits", *International Journal of Computer Science & Engineering Technology* (IJCSET), ISSN : 2229-3345 Vol. 4 No. 03 Mar 2013.

[13] Lawrence R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", *Proceedings of IEEE*, Vol. 77, No. 2, Feb 1989.

[14] TIDIGITS - Rutgers University http://cronos.rutgers.edu/~lrr/