ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Comparison Of Cepstral And Mel Frequency Cepstral Coefficients For Various Clean And Noisy Speech Signals

M.Kalamani[1], Dr.S.Valarmathy[2], C.Poonkuzhali[3], R.Karthiprakash[4]

ECE Department, Bannari Amman Institute of Technology, Sathyamangalam, India[1, 2, 3, 4]

ABSTRACT: Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech recognition can be roughly divided into two stages: feature extraction and classification. Although significant advances have been made in speech recognition technology, it is still a difficult problem to design a speech recognition system for speaker-independent, continuous speech. One of the fundamental questions is whether all of the information necessary to distinguish words is preserved during the feature extraction stage. If vital information is lost during this stage, the performance of the following classification stage is inherently crippled and can never measure up to human capability. Thus, this work finds out an improved feature extraction algorithm based on Mel frequency cepstral coefficient analysis. The results show the comparative analysis of various noise signals and their performance measure using SNR and peak power signal.

KEYWORDS: Speech recognition, Cepstral, MFCC

I. INTRODUCTION

Overview of ASR

A. Components of ASR

Speech Recognition (also known as Automatic Speech Recognition (ASR) or computer speech recognition) is the process of converting a speech signal to a sequence of words which is shown in Fig.1 and it is implemented as algorithm in computer. In the first step, the Feature Extraction, the sampled speech signal is parameterized. The goal is to extract a number of parameters ('features') from the signal that has a maximum of information relevant for the following classification.

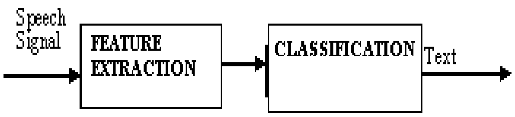


Fig 1:.Block diagram of Basic components of ASR system

That means features are extracted that are robust to acoustic variation but sensitive to linguistic content [1]. Put in other words, features that are discriminate and allow distinguishing between different linguistic units (e.g., phones) are required. On the other hand the features should also be robust against noise and factors that are irrelevant for the recognition process (e.g., the fundamental frequency of the speech signal).

In the classification module the feature vectors are matched with reference patterns, which are called acoustic models. The reference patterns are usually Hidden Markov Models (HMMs) trained for whole words or, more often, for phones as linguistic units. HMMs cope with temporal variation, which is important since the duration of individual phones may differ between the reference speech signal and the speech signal to be recognized [2]. A linear normalization of the time axis is not sufficient here, since not all phones are expanded or compressed over time in the same way. In section 2, feature extraction methods are given and section 3 is for simulation and results discussion.

II. FEATURE EXTRACTION METHODS

Feature extraction can be understood as a step to reduce the dimensionality of the input data, a reduction which inevitably leads to some information loss [2]. Typically, in speech recognition, speech signals are divided into frames and extract features from each frame. During feature extraction, speech signals are changed into a sequence of feature vectors. Then these vectors are transferred to the classification stage.

A. Cepstral Coefficient Analysis

The objective of cepstral analysis is to separate the speech into its source and system components without any a priori knowledge about source and/or system. According to the source filter theory of speech production, voiced sounds are produced by exciting the time varying system characteristics with periodic impulse sequence and unvoiced sounds are produced by exciting the time varying system with a random noise sequence [5, 6]. The resulting speech can be considered as the convolution of respective excitation sequence and vocal tract filter characteristics. If e (n) is the excitation sequence and h (n) is the vocal tract filter sequence, then the speech sequence s (n) can be expressed as follows:

$$S(n) = e(n) * h(n) \qquad (1)$$

this can be represented in frequency domain as,

$$S(\omega) = E(\omega).H(\omega) \qquad (2)$$

The Eqn (2) indicates that the multiplication of excitation and system components in the frequency domain for the convolved sequence of the same in the time domain. The speech sequence has to be deconvolved into the excitation and vocal tract components in the time domain. For this, multiplication of the two components in the frequency domain has to be converted to a linear combination of the two components [9]. For this purpose cepstral analysis is used for transforming the multiplied source and system components in the frequency domain to linear combination of the two components in the cepstral domain.

B. Mel Frequency Cepstral Coefficient Analysis

To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. Speech signal pre-processing covers digital filtering and speech signal detection. Filtering includes pre-emphasis filter and filtering out any surrounding noise using several algorithms of digital filtering. Finally 36 coefficients are extracted from the Mel Frequency Cepstral Coefficient Method. The block diagram representing MFCC is shown in Fig 2.

Pre Emphasis Filter:

In general, the digitized speech waveform has a high dynamic range and suffers from additive noise. In order to reduce this range, pre-emphasis is applied. This pre-emphasis is done by using a first-order FIR high-pass filter.

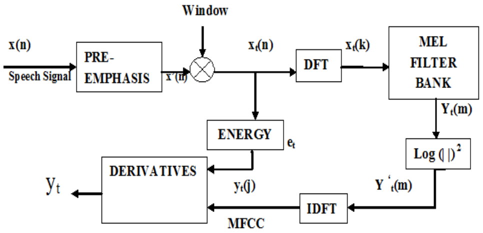


Fig 2: Block diagram for representing Mel Frequency Cepstral Coefficient Analysis

In the time domain, with input x[n] and $0.9 \le a \le 1.0$, the filter equation,

$$y[n]=x[n]-a.x[n-1] \qquad (3)$$

And the transfer function of the FIR filter in z-domain is:

$$H(Z) = 1-\alpha.z-1 ,0.9\le\alpha\le1.0 \qquad (4)$$

where α is the pre-emphasis parameter.

The pre-emphasizer is implemented as a fixed coefficient filter or as an adaptive one, where the coefficient a is adjusted with time according to the auto-correlation values of the speech. The aim of this stage is to boost the amount of energy in the high frequencies [9]. The drop in energy across frequencies (which is called spectral tilt) is caused by the nature of the glottal pulse. Boosting the high frequency energy makes information from these higher formants available to the acoustic model. The pre-emphasis filter is applied on the input signal before windowing.

B. Framing and Windowing

First we split the signal up into several frames such that we are analyzing each frame in the short time instead of analyzing the entire signal at once, at the range 10-30 ms the speech signal is for the most part stationary [4]. Also an overlapping is applied to frames. Here we will have something called the Hop Size. In most cases half of the frame size is used for the hop size. The reason for this is because on each individual frame, we will also be applying a Hanning window which will get rid of some of the information at the beginning and end of each frame. Overlapping will then reincorporate this information back into our extracted features.

C. Windowing

Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum. In speech recognition, the most commonly used window shape is the Hanning window [5]. We used hanning window the most common one that being used in speech recognition system. The hanning window
W (n), defined as [8],

$$w(n) = 0.5\left(1 - cos\left(2\pi\frac{n}{N}\right)\right),0\leq n \leq N \qquad (5)$$

The use for Hanning windows is due to the fact that MFCC will be used which involves the frequency domain (Hanning windows will decrease the possibility of high frequency components in each frame due to such abrupt slicing of the signal).

D. Mel-Scaled Filter Bank

The filter-bank analysis consists of a set of Bandpass filter whose bandwidths and spacing's are roughly equal to those of critical bands and whose range of the centre frequencies covers the most important frequencies for speech perception. The speech signal consists of tones with different frequencies. For each tone with an actual Frequency, f, measured in Hz, a subjective pitch is measured on the 'Mel' scale. We can use the following formula to compute the mels for a given frequency f in Hz:

$$mel(f) = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \qquad (6)$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

E. Cepstrum

In the final step, the log mel spectrum has to be converted back to time. The result is called the mel frequency cepstrum coefficients (MFCCs) [3]. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis.

F. Discrete Cosine Transform

The signal is real (we took the magnitude) with mirror symmetry. The IFFT needs complex arithmetic, the DCT does not [7]. The DCT implements the same function as the FFT more efficiently by taking advantage of the redundancy in a real signal. The DCT is more efficient computationally.

G. Calculation of Log Energy and Derivatives

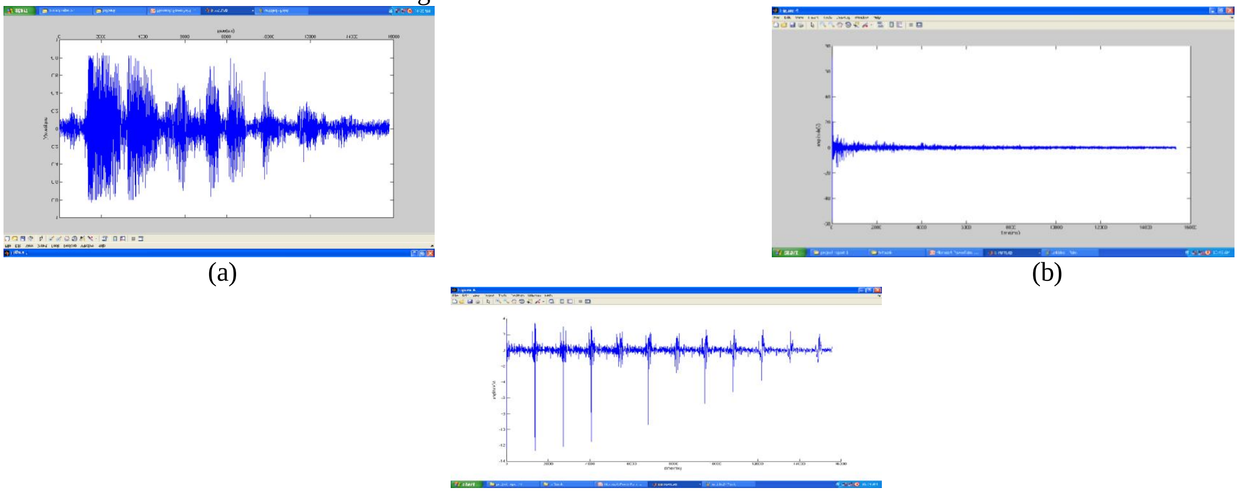
The derivative of coefficient x (n) can be calculated as

$$x(n) \equiv \frac{d}{dx} x(n) \approx \sum_{m=-M}^{M} m x(n+m) \quad (7)$$

where 2M + 1 is the number of frames considered in the evaluation.

III. SIMULATION AND RESULTS

A. Results for Cellular Phone Clean Signal



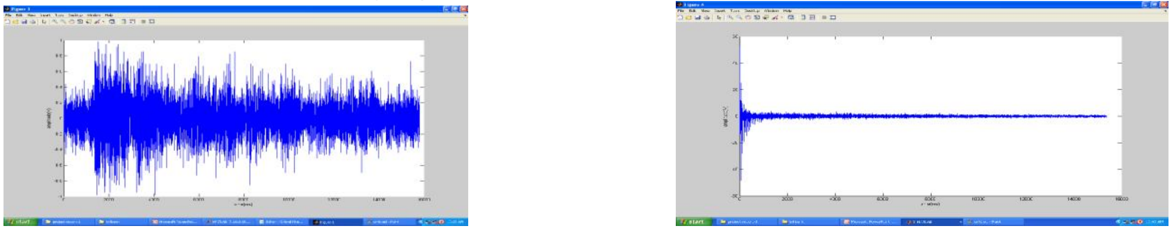
(a)

(b)

(c)

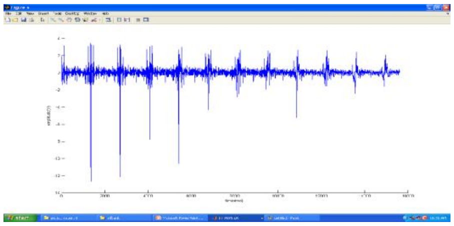
Fig 3: Results for Cellular Phone clean signal (a) Input signal (b)cepstral output (c) MFCC output

B. Results for Cellular Phone Noisy Signal



(a)

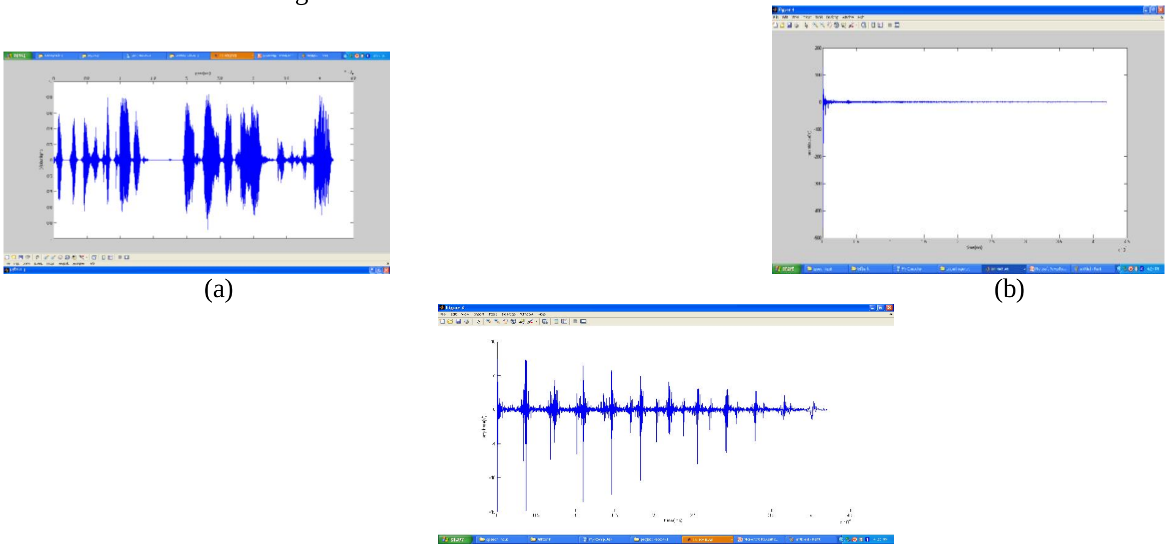
(b)



(c)

Fig 4: Results for Cellular Phone noise signal (a) Input signal (b)ceptral output (c) MFCC output

C. Results for Car Clean Signal



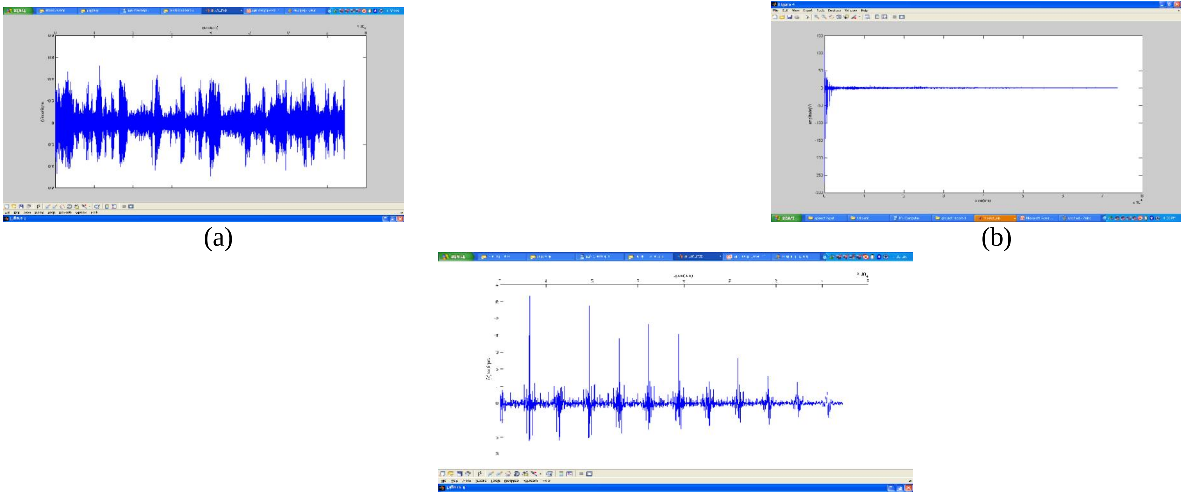
(a)

(b)

(c)

Fig 5: Results for Car clean signal (a) Input signal (b)cepstral output (c) MFCC output

D. Results or Car Noisy Signal



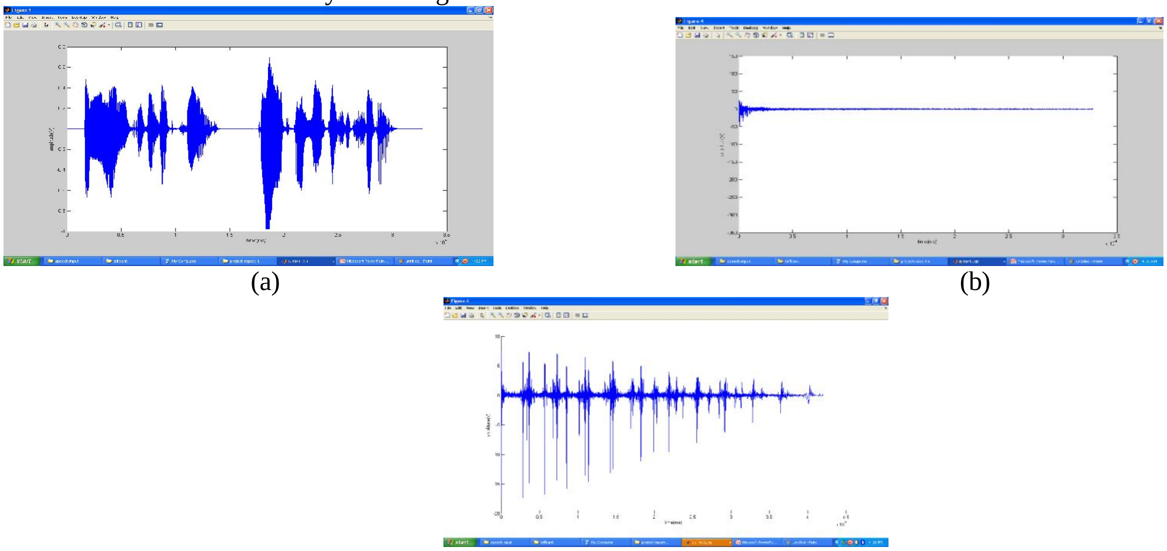
(a)

(b)

(c)

Fig 6: Results for Car noisy signal (a) Input signal (b)ceptral output (c) MFCC output

E. Results for White Stationary Clean Signal

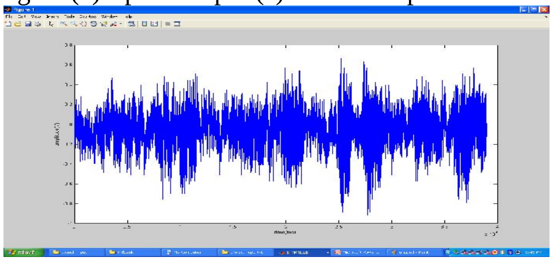


(a)

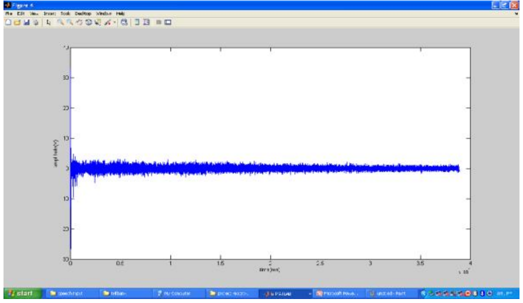
(b)

(c)

Fig 7: Results for white stationary clean signal (a) Input signal (b)ceptral output (c) MFCC output

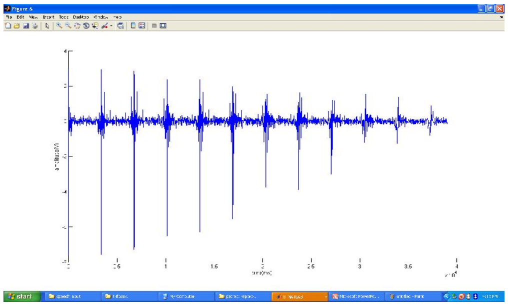


F. Results for White Stationary Noisy Signal



(a)

(b)



(c)

Fig 8: Results for white stationary noisy signal (a) Input signal (b)ceptral output (c) MFCC output

From the results it is observed that the peak signal power of MFCC algorithm is higher than the cepstral method for the feature extraction.

IV. PERFOMANCE ANALYSIS

A. Signal to Noise Ratio

Signal-to-noise ratio (SNR or S/N) is a measure used in science and engineering that compares the level of a desired signal to the level of background noise. It is defined as the ratio of signal power to the noise power in decibels.

$$SNR = \frac{P_{signal}}{P_{noise}} \qquad (8)$$

B. Peak Power Signal

The value at which the signal reaches its maximum is called Peak Power Signal. The highest signal information is obtained at this maximum value.

TABLE 1: SNR comparison results for various noise signals

Input signal	SNR(cepstral)	SNR(MFCC)
Car phone noisy	1.5045	4.8431
Cellular noisy	0.9679	7.5926
White non stationary noisy	1.1187	9.4644

TABLE 2: Peak power signal comparison results for various clean signals

Input signal	Cepstral	MFCC
Car phone clean	4.85	6.68
Cellular clean	0.716	5.854
White non stationary clean	3.44	11.553

V. CONCLUSION

Even though many speech recognition systems have obtained satisfactory performance in clean environments, recognition accuracy significantly degrades if the test environment is different from the training environment. These environmental differences might be due to additive noise, channel distortion, acoustical differences between different speakers, and so on. Hence, MFCC method has been developed to enhance the Accuracy and reduce the computational time for environmental robustness of speech recognition systems. Future work is to implement the system modeling and matching phase.

REFERENCES

[1]. A. Biem and S. Katagiri, "Cepstrum-based filter-bank design using discriminative feature extraction training at various levels," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 1997, pp. 1503–1506.

[2]. B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 1, pp. 24–33, Jan. 2007.

[3]. Chulhee Lee and Chungyong Lee, "Optimizing Feature Extraction for Speech Recognition", IEEE Trans. Audio, Speech, Lang. Process., Vol. 11, No. 1, pp.80, January 2003.

[4]. D. R. Sanand and S. Umesh, "VTLN Using Analytically Determined Linear-Transformation on Conventional MFCC", IEEE Trans. on Speech and Audio Processing, VOL. 20, NO. 5, pp.1573, JULY 2012

[5]. Daniele Giacobello, Mads Græsbøll Christensen, Manohar N. Murthi, Søren Holdt Jensen and Marc Moonen, " Sparse Linear Prediction and Its Applications to Speech Processing", IEEE Transactions on Speech and Audio Processing, Vol. 20, No. 5, pp.1644, July 2012

[6]. Dimitrios Dimitriadis, Petros Maragos and Alexandros Potamianos, "On the Effects of Filter bank Design and Energy Computation on Robust Speech Recognition",IEEE Transactions on Audio, Speech and Language Processing, Vol. 19, No. 6, August 2011.

[7]. D. Giacobello, M. G. Christensen, M. N. Murthi, S. H.Jensen, and M. Moonen, "Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2010, pp. 4650–4653.