Variation in Noise Parameter Estimates for Background Noise Classification

[1] Md. Danish Nadeem
[1]Greater Noida Institute of Technology,
Gr. Noida

[2] Mr. B. P. Mishra
[2]Greater Noida Institute of Technology,
Gr. Noida,

Abstract— In current paper, authors try to investigate regarding variation in speech parameter estimates which can be used to classify environmental noise for grouping a large range of environmental noise into a reduced set of classes of noise with similar type of speech characteristic parameters. One hundred original noises from environment were recorded with the help of a microphone connected to personal computer & stored as a noise database in memory of the computer. Built-in programs for Linear predictive coding (LPC) and Real cepstral parameter (RCEP) have been used while user defined program was written in MATLAB for Mel Frequency Cepstral coefficient (MFCC) in MATLAB to estimate variation in speech parameters which may be utilized for speech analysis through any one of the soft computing techniques viz. neural networks, fuzzy logic, genetic algorithms or a combination of these. Twenty five samples each of four commonly encountered environmental noises (o2car1-o2car25, o3office1-o3office25, o4market1-o4marke25 & o5train1-o5train25) i.e. 100 noises in total have been considered in our study for estimation of three coefficients viz. Mel Frequency Cepstral coefficient, Linear predictive coding and real cepstral parameter. Our experimental results show that Mel Frequency Cepstral Frequencies are robust features for finding out variation in noise parameter estimates. Twenty seven filter banks were used and filter bank output along with power spectrum was obtained in MATLAB. By experimentation through trial & error method, it was found that while considering average of second highest & third highest MFCC coefficients, the noise parameter estimates varied by at most 1% only when internet noise samples were compared to those of original noise samples.

Index Terms- Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC), Real Cepstral Parameter (RCEP).

I. INTRODUCTION

Since over two decades, several algorithms and techniques have been proposed by many researchers regarding classification of environmental noise using parameters such as power spectral density (PSD), zero crossing rate (ZCR), line spectral frequency (LSF) and log area ratio (LAR) coefficients but none of the techniques have proven to be highly effective because of their own inherent limitations associated with each technique so far. Recently, different research groups have carried out studies on new methods and algorithms for environmental noise classification but in current paper, authors have tried to explore noise parameter estimation variants for speech analysis. In our day-to-day life, we encounter different types and levels of environmental acoustical noises like train noise, office noise, market noise etc. In various speech analysis and processing systems such as speech recognition, speaker verification and speech coding, the unwanted noise signals are picked up along with the speech signals which often cause degradation in the performance of communication systems [1]. After modification of processing according to the type of background noise, the performance can be enhanced which requires noise classification based on speech parameter estimation and characterization. Background noise classifier can be used in various fields as, speech recognition and coding being the main ones. Acoustic features can be made adaptable to the type of environmental noise by choosing the most appropriate set to ensure separability between phonetic classes. Since low cost DSP's are increasingly becoming popular, therefore, the next generation of speech coders and intelligent volume controllers are likely to include classification modules in order to improve robustness to environmental/ background noise [2] .

II. II. ENVIRONMENTAL NOISE CLASSIFICATION METHODOLY

The type of methodology that can be adopted for environmental noise classification through parameter estimation variants is based on exploring any one or a few of the environmental noise parameters viz .Linear Predictive Coding, Mel-cepstral based parameters, Real Cepstrum based parameters, line spectral frequencies coefficients, log area ratio coefficients, zero crossing rate and power spectral density [3]. From these noise parameters, we have explored and analyzed two main parameters Linear predictive coding, Mel frequency cepstral coefficients and one allied parameter i.e. real cepstrum parameter for internet noise samples as well as original recorded samples in this paper. Noise database created can be explored on basis of noise classes as follows:

- Automobiles noise class (ANC): Cars, trucks, buses, trains, ambulance, police cars etc

- Babble noise class (BNC**)**: Cafeteria, sports, stadium, office etc
- Factory noise class (FNC): Tools such as drilling machines, power hammer etc.
- Street noise class (SNC): Shopping mall, market, busy street, bus station, gas station etc.
- Miscellaneous noise class (MNC): Aircraft noise, thunder storm etc

Out of these noise classes, only three noise classes have been considered viz. car & train noise from automobile noise class (ANC), office noise from babble noise class (BNC) and market noise from street noise class (SNC).

III. SPEECH PARAMETER ANALYSIS

The variants of speech parameters have been analyzed by acoustic-phonetic approach after spectral analysis. The first step in speech processing is feature measurement which provides an appropriate spectral representation of the characteristics of the time-varying speech signal by filter bank method implemented in MATLAB. Signal representation of internet downloaded and original car noise is as follows:

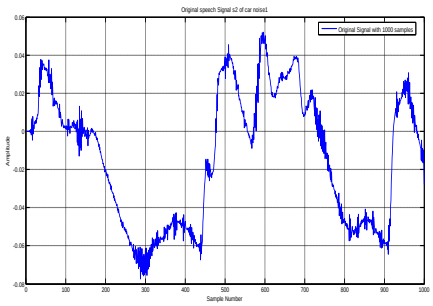

Fig 1. Internet Car noise signal (s2car1) representation in MATLAB

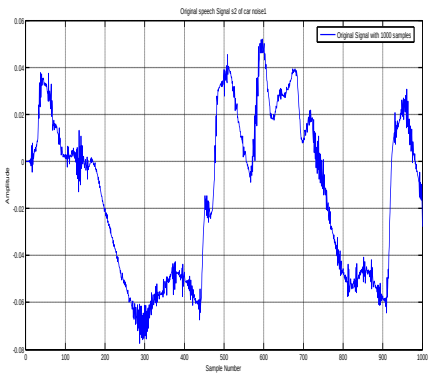

Fig 2.Original Car noise signal (o2car1) representation in MATLAB

The most common type of filter bank used for speech analysis is the uniform filter bank for which the center frequency, fi, of the ith band pass filter is defined as

$$Fi = \frac{Fs}{N} i, \qquad 1 \leq i < Q,$$

where Fs is the sampling rate of the speech signal, and N is the number of uniformly spaced filters required to span the frequency range of the speech [4]. The actual number of filters used in the filter bank, Q, of our work satisfies the relation

$$Q < N / 2 \leq 54/2 < 27$$

with equality meaning that there is no frequency overlap between adjacent filter channels, and with inequality meaning that adjacent filter channels overlap..The digital speech signal, s(n), was passed through a bank of 27 band pass filters whose coverage spans the frequency range of interest in the signal (e.g., 100-3000 Hz for telephone-quality signals, 100-8000 Hz for broadband signals) & output in MATLAB is as follows [5]-

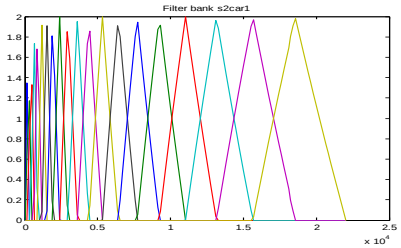

Fig.3 Filter-bank output of Internet Car noise signal (s2car1) in MATLAB (s2car1) in MATLAB

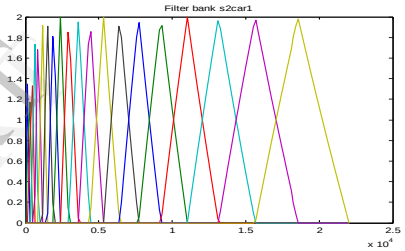

Fig.4 Filter-bank output of Original Car noise signal (o2car1) in MATLAB

Similarly, filter bank outputs were obtained for other noises.

Power spectrum output of all noises were obtained in MATLAB and that of car noise obtained is as follows-

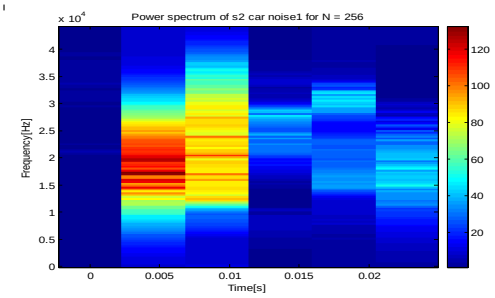

Fig.5 Power spectrum output of Internet Car noise signal (s2car1) in MATLAB

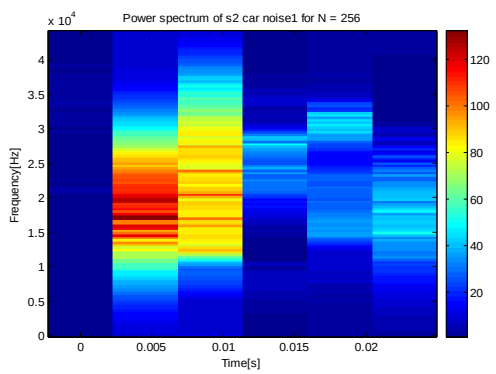

Fig.6 Power spectrum output of Original Car noise signal (o2car1) in MATLAB

IV. SPECTRAL MODELS USED FOR ENVIRONMENTAL NOISE CLASSIFICATION

Following models are widely used for environmental noise classification:

A. LPC Model

Speech synthesis based on LPC model in vocal tract of human throat may be assumed as follows in figure 7

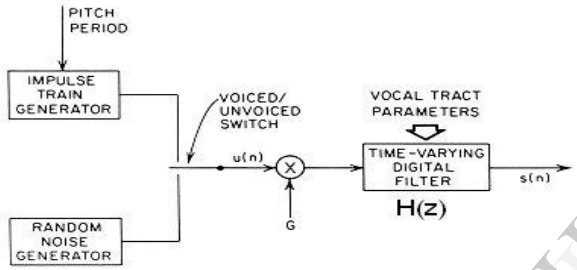

Fig. 7 Speech synthesis based on LPC model in human throat

The object of linear prediction is to form a model of a Linear Time Invariant (LTI) digital system through observation of input and output sequences [6]. The basic idea behind linear prediction is that a speech sample can be approximated as a linear combination of past speech samples. By minimizing the sum of the squared differences (over a finite interval) between the actual speech samples and the linearly predicted ones, a unique set of predictor coefficients can be determined.

If u(n) is a normalized excitation source and being scaled by 'G' , the gain of the excitation source, then LPC model is the most common form of spectral analysis models on blocks of speech (speech frames) and is constrained to be of the following form, where H (z) is a pth order polynomial with z-transform and the coefficients a1, a2, …, ap are assumed to be constant over the speech analysis frame

$$H(z) = 1 + a1z^{-1} + a2\, z^{-2} + a3\, z^{-3} \ldots + ap\, z^{-p}$$

Here the order 'p' is called the LPC order. Thus the output of the LPC spectral analysis block is a vector of coefficients (LPC parameters) that specify (parametrically) the spectrum that best matches the signal spectrum over the period of time in which the frame of speech sample was accumulated [7].

If 'N' is the number of samples per frame and 'M ' is the distance between the beginnings of two frame, then for a given speech sample at time 'n'; S(n), can be approximated as a linear combination of the past 'p' speech samples, such that

$$s(n) \approx a1s(n-1) + a2s(n-2) + aps(n-p), \quad (1)$$

where the coefficients a1, a2… ap are assumed constant over the speech analysis frame. We convert eq. (1) to an equality by including an excitation, G u (n), giving:

$$s(n) = \sum_{i=1}^{P} ais(n-i) + G\,u(n), \quad (2)$$

where u(n) is a normalized excitation and G is the gain of the excitation. By expressing eq (2) in the z-domain we get the relation

$$S(z) = \sum_{i=1} ai\, z^{-i}\, S(z) + G\, U(z), \quad (3)$$

leading to the transfer function

$$H(z) = \frac{S(z)}{G\,U(z)} = \frac{1}{p\quad H(z)} = \frac{1}{} \quad (4)$$

Because speech signals vary with time, this process is done on short chunks of the speech signal, which are called *frames*. Usually 30 to 50 frames per second give intelligible speech with good compression. When applying LPC to audio at high sampling rates, it is important to carry out some kind of auditory frequency warping, such as according to mel or Bank frequency scales.

B. MFCC MODEL

The perception of human frequency content of sounds, either for pure tones or for speech signals, does not follow a linear scale. This research has led to the idea of defining subjective pitch of pure tones [8]. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the "mel" scale. As a reference point, the pitch of a 1 KHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels .Other subjective pitch values are obtained by adjusting the frequency of a tone such that it is half or twice the perceived pitch of a reference tone (with a known mel frequency).A filter bank, in which each filter has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. (The spacing is approximately 150 mels and the width of the triangle is 300 mels). Mel scale cepstral analysis uses cepstral smoothing to smooth the modified power spectrum. This is done by direct transformation of the log power spectrum to the cepstral domain using an inverse Discrete Fourier Transform (DFT).

The modified spectrum of S(w) thus consists of the output power of these filters when S(w) is the input. Denoting these power coefficients by Sk, k = 1, 2. . . K, we can calculate what is called the mel-frequency cepstrum, Cn,

$$Cn = \sum_{k=1}^{k} (\log Sk) \cos [n\,(k - 1/2)\,\pi/K],$$

n = 1, 2. . . L,

w here L is the desired length of the cepstrum. The first 12 coefficients (1st frame) can be discarded since they are the mean of the signal and hold little information. Hence 13th coefficient (1st frame) is usually considered.
The difference between the cepstrum and the mel-frequency cepstrum is that in the Mel frequency cepstrum, the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system's response more closely than the linearly-spaced frequency bands obtained directly from the FFT or DCT. This can allow for better processing of data, for example, in audio compression. However, unlike the sonogram, MFCCs lack an outer ear model and, hence, cannot represent perceived loudness accurately.

1) Thus, in the sound processing, the mel-frequency cepstrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Steps in MFCC extraction are as follows:
Framing- Human speech is a non stationary signal, but when segmented into parts ranging from 10-40 msec, these divisions are quasi-stationary [9]. For this reason the human speech input is to be divided into frames before feature extraction takes place. The selected properties for the speech signals are a sampling frequency of 16 kHz, 8-bit monophonic PCM format in WAV audio. The chosen frame size is of 256 samples, resulting in each frame containing 16 msec portions of the audio signal..It seems that a value of 256 for N is an acceptable compromise. Furthermore the number of frames is relatively small, which will reduce computing time [10].

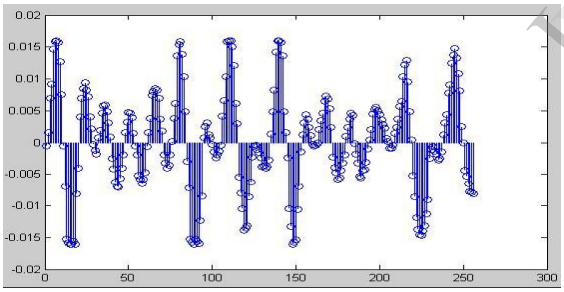

Fig.8 Frame of Internet Car noise signal (s2car1) in MATLAB

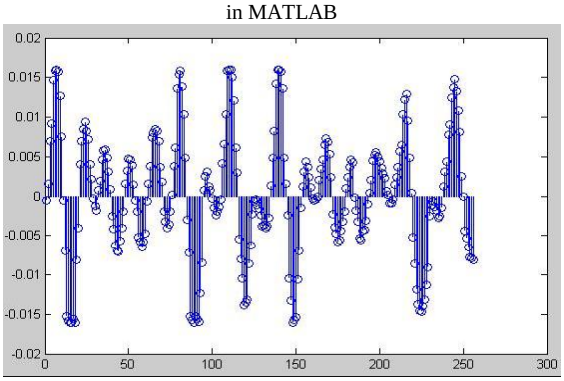

Fig.9 Frame of Original Car noise signal (o2car1) in MATLAB

Windowing-. The use of the window function reduces the frequency resolution by 40%, so the frames must overlap to permit tracing and continuity of the signal. The motive for utilizing the windowing function is to smooth the edges of each frame to reduce discontinuities or abrupt changes at the endpoints. The windowing serves a second purpose and that is the reduction of the spectral distortion that arises from the windowing itself.

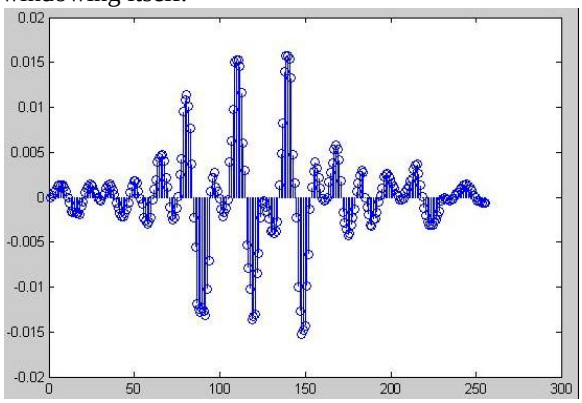

Fig.10 Internet Car noise signal (s2car1) windowed data after Hamming in MATLAB

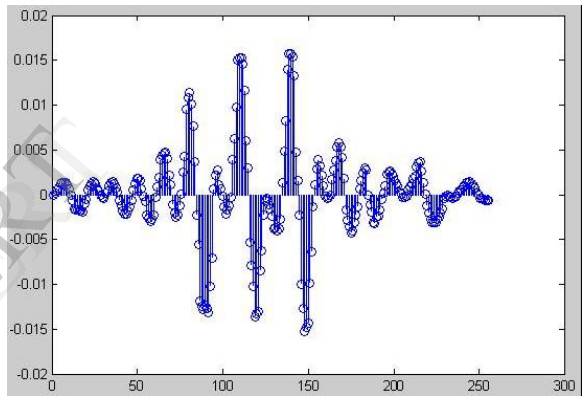

Fig.11 Original Car noise signal (o2car1) windowed data after Hamming in MATLAB

Fast Fourier Transform- The frame size is not a fixed quantity and therefore can vary depending on the resulting time portion of the audio signal. The reason that the authors selected number of samples as 256 is that it is a power of 2, which enables the use of the Fast-Fourier Transform [11]. The FFT is a powerful tool since it calculates the DFT of an input in a computationally efficient manner, saving processing power and reducing computation time. The operation results in the spectral coefficients of the windowed frames.

Mel-scale Filter bank Frequency Transformation- Mel-cepstral coefficients are the features that will be extracted from speech during our work. The key difference between MFCCs and cepstral coefficients lies in the processing involved when extracting each of these characteristics of a speech signal[12]. The process of obtaining Mel-cepstral coefficients involves the use of a Mel-scale filter bank. The spectral coefficients of each frame are then converted to Mel scale after applying a filter bank. The Mel-scale is a logarithmic scale resembling the way that the human ear perceives sound. The filter bank is composed of triangular filters that are equally spaced on a logarithmic scale. The Mel-scale warping is approximated and represent by the following

Mel (f) = 2595 $\log_{10}$ (1 + f / 700), where f is frequency.

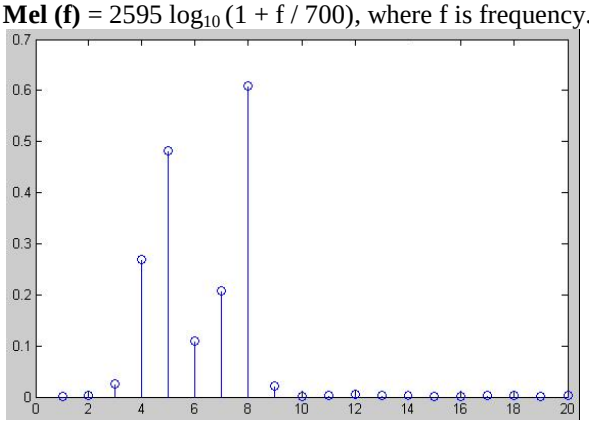

Fig.12 Mel Spectral Coefficients of Internet Car noise signal (s2car1) in MATLAB

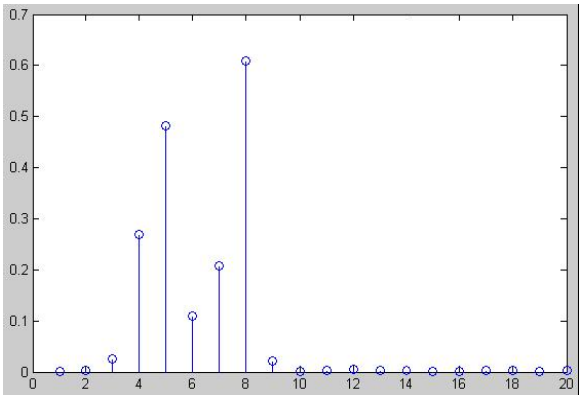

Fig.13 Mel Spectral Coefficients of Original Car noise signal (o2car1) in MATLAB

Discrete Cosine Transform- The Discrete Cosine Transform is applied to the log of the Mel-spectral coefficients to obtain the Mel-Frequency Cepstral Coefficients.

Only the first 12 coefficients of each frame are kept, since most of the relevant information is kept amongst those at the beginning[13] . The first 12 coefficients (1st frame) can be discarded since they are the mean of the signal and hold little information. Hence 13th coefficient (1st frame) is usually considered and the use of the DCT minimizes the distortion in the frequency domain.

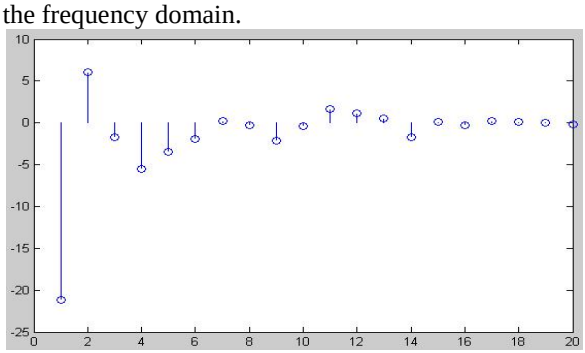

Fig.14 Mel-frequency cepstral coefficients of Internet Car noise signal (s2car1) in MATLAB

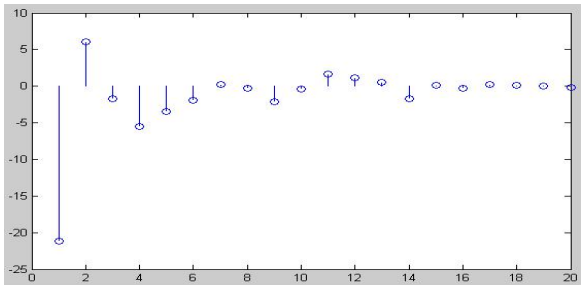

Fig.14 Mel-frequency cepstral coefficients of Original Car noise signal (o2car1) in MATLAB

C. RCEP MODEL

As per theoretical point of view, the Cepstrum is defined as the inverse Fourier transform of the real logarithm of the magnitude of Fourier transform [14]. Therefore, by keeping only the first few cepstral coefficients and setting the remaining coefficients to zero, it is possible to smooth the harmonic structure of the spectrum. Cepstral coefficients are therefore very convenient coefficients to represent the speech spectral envelope.

Hence, the following function calculates the real Cepstrum of the signal x.

$$y = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(e^{i\omega t})|\, e^{i\omega t}\, d\omega,$$

This denotes the Fourier Transform of x and hence real Cepstrum as a real-valued function can be used for the separation of two signals convolved with each other [15]. Thus, RCEP is a Cepstrum-based technique for determining a Harmonics-to-Noise Ratio (HNR) in Speech Signals and is a valid technique for determining the amount of spectral noise, because it is almost linearly sensitive to both noise and jitter for a large part of the noise or jitter continuum.

Thus real Cepstrum block gives the real Cepstrum output of the input frame and is also a popular way to define the prediction filter. Last, the line spectrum frequencies (a.k.a. line spectrum pairs) are also frequently used in speech coding [16]. Line spectrum frequencies are another representation derived from linear predictive analysis which is very popular in speech coding.

V. RESULTS OBTAINED IN MATLAB (UPTO TENTH ORDER FOR FIVE SAMPLES OF FOUR INTERNET NOISES)

(a)MFCC

Car s1	**Car s2**	**Car s3**	**Car s4**	**Car s5**
0.7606	0.8497	-0.1915	0.5952	-0.6787
Office s1	Office s2	Office s3	Office s4	Office s5
1.1829	0.2051	0.6141	0.7134	0.2297
Market s1	Market s2	Market s3	Market s4	Market s5
0.8646	0.4135	0.8211	0.0903	0.1616
Train s1	Train s2	Train s3	Train s4	Train s5
0.0271	-0.5599	-0.1922	0.9966	0.9129

(b)LPC

Car s1	Car s2	Car s3	Car s4	Car s5
0.2164	0.1270	0.2298	0.0988	0.1835
Office s1	Office s2	Office s3	Office s4	Office s5
0.5474	0.5195	0.2179	0.1775	0.2018
Market s1	Market s2	Market s3	Market s4	Market s5
0.1504	0.1527	0.1558	0.1181	0.1645
Train s1	Train s2	Train s3	Train s4	Train s5
0.6579	0.6629	0.7030	0.6006	0.6627

(c)RCEP

Car s1	Car s2	Car s3	Car s4	Car s5
0.0011	0.0009	0.0003	-0.0004	0.0000
Office s1	Office s2	Office s3	Office s4	Office s5
0.0007	0.0010	-0.0009	-0.0001	-0.0000
Market s1	Market s2	Market s3	Market s4	Market s5
0.0006	-0.0003	0.0001	-0.0007	-0.0002
Train s1	Train s2	Train s3	Train s4	Train s5
-0.0012	0.0005	0.0013	0.0008	-0.0017

(d) AVERAGES OF COEFFICIENTS

MFCC

M F C C Coefficients	C1	C2	C3	C4	C5
Car Noise (S1-S5)	16.613	0.978	2.040	1.572	2.101
Office Noise (S1-S5)	19.651	1.074	1.787	1.397	1.331
Market Noise (S1-S5)	19.284	1.302	1.748	1.377	1.718
Train Noise (S1-S5)	18.976	1.154	1.687	1.437	1.594

LPC

LPC Coefficients	C1	C2	C3	C4	C5
Car Noise (S1-S5)	1.000	-0.497	-0.546	-0.281	0.543
Office Noise (S1-S5)	1.000	-0.590	-0.430	-0.172	0.099
Market Noise (S1-S5)	1.000	-0.298	-0.553	-0.619	0.209
Train Noise (S1-S5)	1.000	-1.432	0.005	0.769	-0.731

RCEP

RCEP Coefficients	C1	C2	C3	C4	C5
Car Noise (O1-O5)	7.998	0.114	0.013	0.104	0.013
Office Noise (O1-O5)	8.632	0.005	0.185	-0.016	-0.146
Market Noise (O1-O5)	8.814	0.829	0.017	-0.182	-0.011
Train Noise (O1-O5)	8.513	0.018	0.143	0.016	0.121

VII. CONCLUSION

On experimentation, our MATLAB results show that out of three noise parameters under consideration, Mel Frequency Cepstral Frequencies are robust features in variants of noise parameter estimation and its characterization. By trial & error method, it was found that the best result of MFCC was obtained at maximum difference of 0.182 when average of second highest & third highest MFCC coefficients was taken since scaling becomes easier at maximum difference while undergoing defining membership in fuzzy logic operation for noise classification. Also, the noise parameter estimates varied by at most 1% only when internet noise samples were compared to those of original noise samples. In future, these results can be explored for finding out classification accuracy during implementation of a practical background/ environmental noise classifier.

VIII. REFERENCES

[1] Schafer, R. and Rabiner, L.Digital Representation of Speech Signals.. Proceedings of the IEEE 63 (1975): 662-677.
[2] Gray, R.M.Vector Quantization.. IEEE ASSP Magazine 1 (1984): 4-29.
[3] Schafer, R. and Rabiner, L Systems for Automatic Formant Analysis of Voiced Speech..Journal of the Acoustical Society of America 47 (1970): 634-648.
[4] Tokhura, Y. A weighted cepstral distance measure for speech recognition..IEEE Transactions on acoustics, speech and signal processing 35 (1987): 1414-1422.
[5] Fujimura, O.Analysis of nasal consonants.. Journal of the Acoustical Society of America 34 (1962): 1865- 1875.
[6] Hughes, G. and Halle, M. .Acoustic Properties of Stop Consonants..Journal of the Acoustical Society of America 30 (1957): 07-116.
[7] Atal, B.S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Journal of the Acoustical Society of America 55 (1974): 1304-1312.
[8] Furui, Sadaoki. Digital Speech Processing, Synthesis, and Recognition. New York: Marcel Dekker, 2001
[9] F. Beritelli, S. Casale, and P.Usai, "Background Noise classification in Mobile Environments Using Fuzzy Logic," contrib.. ITU-T (WP 3/12), Geneva, Switzerland, Apr. 1997.
[10] Blumstein, S. and Stevens, K. .Perceptual invariance and onset spectra for stop consonants in different vowel environments.. Journal of the Acoustical Society of America 67 (1980): 648-662.
[11] Blumstein, S. and Stevens, K. .Invariant cues for place of articulation in stop consonants Journal of the Acoustical Society of America 64 (1978): 1358-1368.
[12] Itakura, F. and Saito, S.Speech information compression based on the maximum likelihood spectrum estimation. Journal of the Acoustical Society of Japan 27 (1971):463-470.
[13] F. Beritelli, S. Casale, G. Ruggeri, "New Results in Fuzzy Pattern Classification ofr Background Noise", Proceedings of ICSP 2000.
[14] W.C. Treurniet and Y. Gong, "Noise independent speech recognition for a variety of noise types", Proc. IEEE ICASSP 94 Adelaide, pp. 437-440, April 1994.
[15] F. beritelli, S. Casale, "Background Noise Classification in Advanced VBR Speech Coding for Wireless Communications", Proc. 6th IEEE International Workshop on Intelligent Signal Processing And Comunication systems (ISPACS98), Melbourne, Australia, 4-6 Nov. 1998,pp. 451-455.
[16] Khaled El-Maleh, Ara Samouelian, Peter Kabal, "Frame-Level Noise Classification in Mobile Environments" ICASSP 99, Phoenix, Arizona, May 15-19, 1999. (3)