Non-intrusive Objective Speech Quality Evaluation using Multiple Time-scale Estimates of Multi-resolution Auditory Model (MRAM) Features

Rajesh Kumar Dubey
Department of Electronics and Communication Engg.
Jaypee Institute of Information Technology
Noida, India
rajeshk_dubey@yahoo.com

Arun Kumar
Center for Applied Research in Electronics,
Indian Institute of Technology, Delhi
New Delhi, India.
arunkm@care.iitd.ac.in

***Abstract*—The effects of short time-transients of additive noise present over some specific active regions in speech utterances cannot be captured in features computed over an entire speech utterance. Thus the uses of multiple time-scale estimates of auditory features have been sought in this work for non-intrusive speech quality evaluation. It is capable in capturing the time localized information of short-time transient distortions and their distinction from plosive sounds of speech. The features are computed from the combination of different active speech regions of a speech utterance using multi-resolution auditory model (MRAM) on frame-by-frame basis. The voice activity detection (VAD) algorithm has been used for the selection of active speech regions and rejection of silence region from the speech utterance. The multiple time-scale MRAM features are probabilistically modelled to map into mean opinion score (MOS) value using Gaussian Mixture Model (GMM) for each combination of active speech regions. The average value of these multiple time-scale estimates MOS values of the different combinations of active speech regions give the overall objective MOS value of a degraded speech utterance. The results are given in terms of correlation coefficient between the subjective MOS and the overall objective MOS. The results are also compared with the ITU-T Recommendation P.563, the standard for non-intrusive speech quality assessment for telephone band speech.**

Keywords—Listening test; Speech Quality; Mean Opinion Score; Speech degradations, Auditory features.

I. INTRODUCTION

The speech quality measurement is important at different nodes of communication networks or systems using speech processing algorithms. Thus, the requirement of automatic evaluation of speech quality objectively and continuously is becoming increasingly popular for speech processing algorithms. The subjective evaluation of speech quality according to absolute category rating (ACR) method as given in ITU-T Recommendation P.800 [1] is considered to be an ideal method. It uses generally 16-25 listeners to rate the quality of speech signal played for them on a scale of one to five: 5-excellent, 4-good, 3-fair, 2-poor and 1-bad in subjective listening test. The average value of their opinions about the speech quality is expressed in terms of subjective mean opinion score (MOS) value. This approach is impractical, time consuming and expensive for system automation. Thus, the subjective listening test for speech quality evaluation is supplemented by different objective methods using different computational techniques using time-domain information and frequency-domain information of speech waveforms such as different temporal, spectral and auditory perception features. The objective methods are thus becoming increasingly popular. The objective speech quality measurement is expressed in terms of objective MOS value. The objective methods of speech quality measurement are of two types: intrusive and non-intrusive. If the original clean speech signal is used as reference for comparison, it is intrusive method and if not then it is non-intrusive method. Non-intrusive method do not require the original clean speech signal for any comparison and evaluate the quality of a degraded speech utterance using received signal only as shown in figure 1. The standard for non-intrusive speech quality evaluation is published in May 2004 [2] as the ITU-T Recommendation P.563.

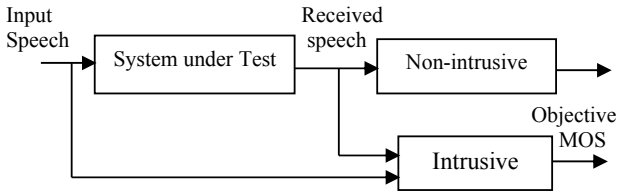

Figure 1: Intrusive and non-intrusive objective speech quality evaluation.

In [3], the mapping of local and global speech features has been done using GMM to find the objective MOS values of degraded speech signal. These features are obtained directly from speech coders without considering any degradation or channel model. The features are computed for entire speech utterances, here it is called as single time-scale features. The parameters of GMM are obtained by training it from the subjective MOS enabled speech database using Expectation Maximization (EM) algorithm [4]. The parameters of the trained GMM and the speech features are used for the computation of objective MOS value of a degraded speech utterance. The human auditory system is modelled explicitly or implicitly as filtering, detection and compression in cochlea as computational model and perception leading to the human brain to obtain an opinion about the quality of a speech signal given in [5]. In [6], the combination of several speech features obtained from different auditory models and speech production model are used for non-intrusive speech quality estimation. In [7], a comparison of performance in terms of correlation between the subjective MOS and the objective MOS of speech features obtained from mel-frequency cepstral coefficients (MFCC) and reconstructed phase space (RPS) used for non-intrusive speech quality estimation is presented. The effectiveness of algorithms in non-intrusive speech quality measurement problem using the combination of speech features obtained from perceptual linear prediction coefficients (PLPC), MFCC and LSF is given in [8]. In [9], the speech quality evaluation for narrowband telephonic speech has been done using multi-resolution auditory model (MRAM) features computed for entire speech on single-time scale and improvement is reported for the combination of MRAM, MFCC and LSF features. The performance of two

978-1-4673-9080-4/16/$31.00 ©2016 IEEE

different auditory features, Lyon's auditory features and MRAM features, both computed at single time-scale are compared in terms of correlation in [10] for narrowband speech. The concept of instantaneous speech quality measurement at different instants of time in a speech utterance using E-model has been given in [11].

These algorithms in the literature use single time-scale features for evaluation of objective MOS in non-intrusive speech quality evaluation problem. These features are not accurate in capturing the time-localized information of short-time transient distortions and their distinction from plosive sounds of speech. Hence, the importance of estimating speech features at multiple time-scales. In this work, multiple time-scale estimates of MRAM features are used for the mapping of objective MOS in non-intrusive speech quality evaluation objectively. The multiple time-scale estimates of MRAM features can capture the simultaneous as well as temporal masking effect of the human auditory system in the listening process and cater for the effect of short-time transient additive noise present in some specific active regions in a degraded speech utterance. The MRAM features computed for multiple time-scales can capture both the local features at different time-scales within the speech utterance and the global features at a single time-scale corresponding to entire speech utterance.

II. MULTIPLE TIME-SCALE MRAM FEATURES

The MRAM features are computed on multiple time-scales to capture more detailed statistical information of localized features. If non-stationary short-time transients of additive noise are present in the active regions of a speech utterance particularly for contiguous active speech segments, then it is expected that multiple time-scale auditory features used for the speech quality mapping problem may improve the correlation. The different active regions of a speech utterance are selected using voice activity detection (VAD) algorithm. The different segregated (SEG) combinations are formed by concatenating one active speech region to the next contiguous active speech region till all the active speech regions are accounted for as shown in figure 2.

The computation of MRAM features are done for each SEG on per frame basis. The duration of frame size are taken to be 16 ms and windowed with a Hamming window of same length as shown in figure 3. The windowed speech frame is used for discrete wavelet packet decomposition of level-3. In the next step, squaring of DWT coefficients is done to compute energy and to incorporate the effect of absolute hearing threshold (AHT) outer and middle ear (OME) weighting has been done. In the multi-resolution spectral spreading, energy for low frequency to larger time duration and for high frequency to shorter time duration is applied. To capture the effect of temporal masking there is temporal smearing and finally, the subjective loudness adjustment of the speech intensity has been done to obtain the MRAM features. The detailed explanation of each blocks and related mathematical model are given in [12]. In this work, third level packet decomposition of discrete wavelet transform (DWT) has been used as shown in figure 4. A total of 17 critical bands are considered and corresponding to them 17 MRAM feature values are obtained for a narrowband telephonic speech sampled at 8 kHz. The mean, variance, skewness, and kurtosis of 17 MRAM features over the frames for each SEG are computed and concatenated to obtain a 68-dimensional MRAM feature vector. The dimensionality of MRAM feature vector is reduced from 68 to 30 that retain 98% the energy using principal component analysis (PCA) [9]. The process is repeated to compute 30 dimensional feature vectors for all the SEGs i.e. $i=1,\ldots, K$ for each speech utterance.

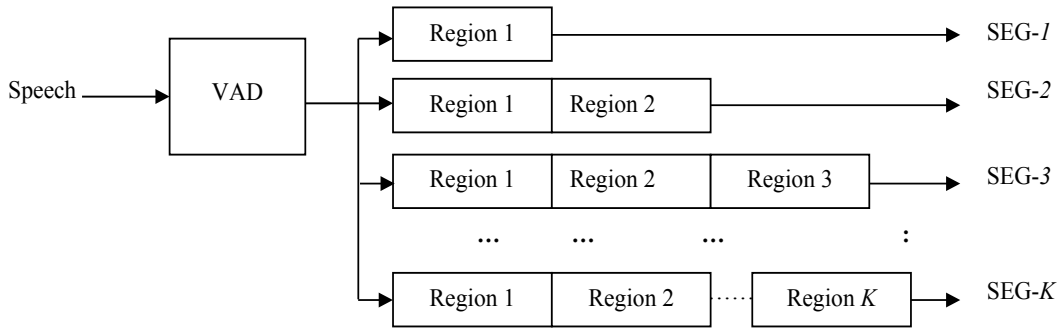

Figure 2: Concatenations of active speech regions to form different segregated (SEG) combinations for multiple time-scale estimates of features.

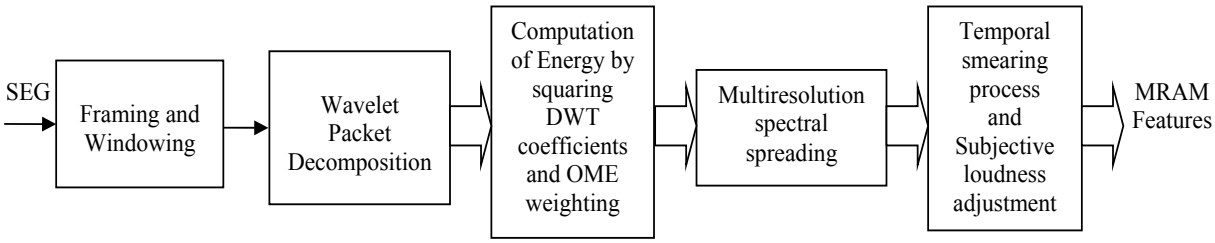

Figure 3: The Lyon's auditory features computation.

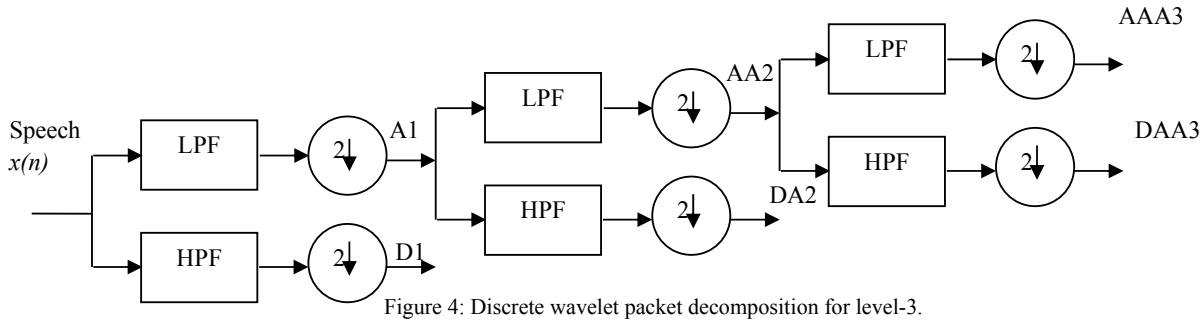

Figure 4: Discrete wavelet packet decomposition for level-3.

III. GMM TRAINING AND SPEECH QUALTY MAPPING USING AUDITORY FEATURES

The 30-dimensional MRAM feature vector Ψ for each SEG is appended with the subjective MOS value θ_j of the corresponding speech from MOS labelled speech databases to train the joint Gaussian Mixture Model (GMM) using Expectation Maximization algorithm [4] as shown in figure 5. The 30-dimensional reduced size MRAM feature vectors and GMM parameters $\Pi(\mu^{(k)},\omega^{(k)},\sum^{(k)})$ with $k=1,2,3\ldots, M$ are then used to compute the objective MOS value of the test speech SEG using probabilistic approach as shown in figure 6, where $\mu^{(k)},\omega^{(k)}$, and $\sum^{(k)}$are the mean, mixture weight and covariance matrix respectively of the k-th mixture component.

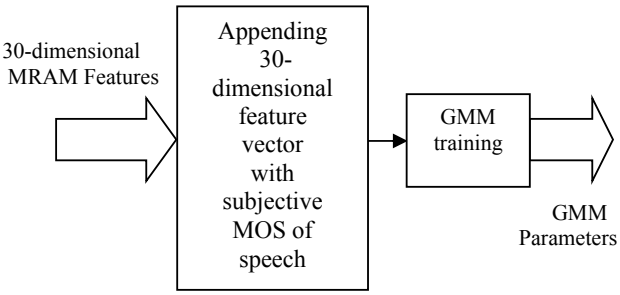

Figure 5: Appending 30-dimensinal feature vector of SEG-1 with the subjective MOS value of the corresponding speech signal for the training of joint GMM.

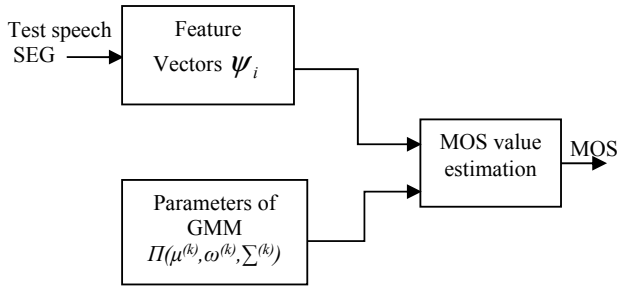

Figure 6: The computation of objective MOS value

Now, the objective MOS value $\hat{\theta}$ is computed using MMSE criterion [2] given by,

$$\hat{\theta} = \hat{\theta}(\psi) = \arg\min_{\hat{\theta}(\psi)} E\left\{(\theta - \hat{\theta}(\psi))^2\right\} = E\{\theta / \psi\} \qquad (1)$$

In this work, GMM of $M=12$ mixture components, with "leave one out" procedure and ten-fold cross-validation process is used [6]. The overall objective MOS value of a speech utterance is computed by averaging of the objective MOS values of the different SEGs formed using multiple time scale estimates as shown in figure 7.

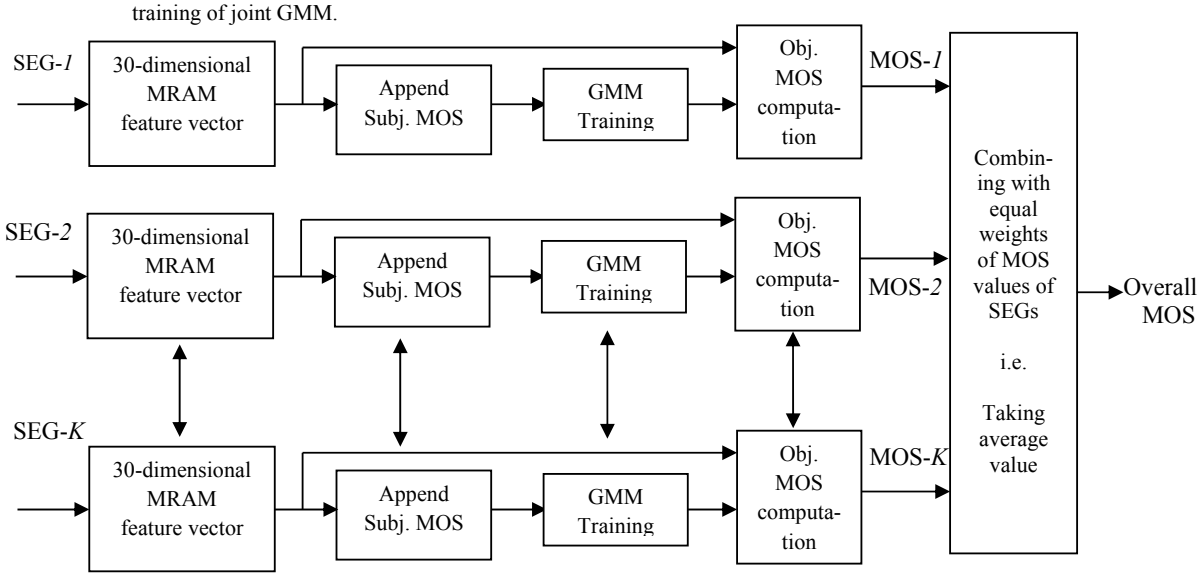

Figure 7: The overall objective MOS of the speech utterance computed as average of MOS values of different SEGs.

The overall objective MOS value of a speech utterance $\hat{\theta}$ is the average of objective MOS values of K SEGs. Thus, $\hat{\theta}_i$ given by,

$$\hat{\theta} = \frac{1}{K}\sum_{i=1}^{K}\hat{\theta}_i \tag{2}$$

where, K is the number of SEGs of multiple time scale estimates i.e. number of active speech regions present in the speech utterance.

IV. RESULTS AND ANALYSIS

The segregation of SEG-1 and SEG-2 using first two active speech regions of a sample speech utterance is shown in figure 8 for illustration. The SEG-1 is the first active region and SEG-2 is formed by concatenating the first and second active speech regions obtained from VAD algorithm. The first active speech region i.e. SEG-1 is shown in figure 8 (a), the second active speech region obtained from VAD algorithm is shown in figure 8 (b) and the concatenation of first and second active speech regions to obtain SEG-2 is shown in figure 8 (c). The segregation and formation of all SEGs in this way will be continued till all active speech regions are accounted for.

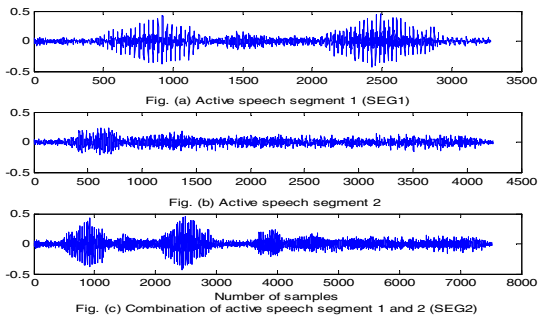

Figure 8: A speech utterance and SEGs formation (a) First active speech region (SEG-1), (b) Second active speech region, and (c) The concatenation of first and second active speech regions to obtain (SEG-2).

In this work, three speech databases have been used: (i) ITU-T Supplement-23 database [13] having 1328 subjective MOS labelled degraded speech utterances according to ACR subjective listening test, (ii) NOIZEUS-960 database having 960 degraded speech utterances without subjective MOS and (iii) NOIZEUS-2240 databases [14] having 2240 degraded speech utterances without subjective MOS. The subjective MOS values for these 960 and 2240 speech utterances are determined in our laboratory according to ACR subjective listening test.

The Pearson's correlation coefficient between the subjective MOS value and the overall objective MOS value is used to represent the result of algorithm as a figure of merit in speech quality evaluation problem. The condition averaged MOS and the unconditioned MOS values are used for the computation of Pearson's correlation coefficients. In Table 1, the results are given for condition averaged MOS values and in Table 2 the results are given for unconditioned MOS values computed for both single time-scale and multiple time-scale estimates of MRAM features for speech utterances. The results are given for single time-scale estimates and multiple time-scale MRAM features along with ITU-T Rec. P.563. The comparison has been also given in form of visual plots in figure 9 for unconditioned MOS values.

Table 1: Comparison of Pearson's correlation coefficients between the condition averaged subjective MOS and the condition averaged overall objective MOS using single time-scale and multiple time-scale estimates of MRAM features.

Data of Different Experiments	**ITU-T Rec. P.563**	**Single time-scale MRAM features**	**Multiple time-scale MRAM Features**
Exp.1(A)-French	0.885	0.860	0.944
Exp.1(D)-Japanese	0.842	0.875	0.973
Exp.1(O)-Am. English	0.902	0.876	0.973
Exp.3(A)-French	0.867	0.815	0.862
Exp.3(C)-Italian	0.854	0.803	0.944
Exp.3(D)-Japanese	0.929	0.792	0.899
Exp.3(O)-Am. English	0.918	0.857	0.891
NOIZEUS-960	0.951	0.993	0.994
NOIZEUS-2240	0.955	0.987	0.989
Weighted Average	**0.934**	**0.945**	**0.971**
Std. Dev.	**0.041**	**0.073**	**0.047**

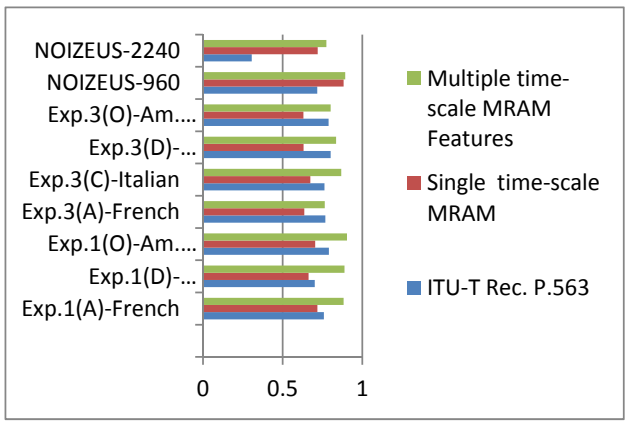

Figure 9: A visual performance comparison in terms of Pearson's correlation coefficients.

Table 2: Comparison of Pearson's correlation coefficients between the unconditioned subjective MOS and the unconditioned overall objective MOS using single time-scale and multiple time-scale estimates of MRAM features.

Data of Different Experiments	ITU-T Rec. P.563	Single time-scale MRAM features	Multiple time-scale MRAM Features
Exp.1(A)-French	0.759	0.718	0.883
Exp.1(D)-Japanese	0.701	0.663	0.888
Exp.1(O)-Am. English	0.790	0.704	0.904
Exp.3(A)-French	0.768	0.636	0.764
Exp.3(C)-Italian	0.762	0.674	0.868
Exp.3(D)-Japanese	0.801	0.631	0.835
Exp.3(O)-Am. English	0.788	0.630	0.801
NOIZEUS-960	0.717	0.883	0.893
NOIZEUS-2240	0.306	0.720	0.774
Weighted Average	**0.529**	**0.738**	**0.821**
Std. Dev.	**0.155**	**0.079**	**0.054**

IV. CONCLUSIONS

The MRAM features are computed on multiple time-scales by forming different combinations of active speech regions called SEGs and corresponding objective MOS values of different SEGs are mapped using GMM probabilistic approach. The overall objective MOS value of a speech signal is computed by assigning equal weights to different multiple time-scale estimates for non-intrusive speech quality evaluation problem. The Pearson's correlation coefficient between the subjective and the estimated overall objective MOS for different types of noisy speech databases are obtained as performance measure and compared with the performance of single time-scale MRAM features and ITU-T Rec. P.563. The improved performance results are obtained for multiple time-scale estimates of MRAM features that are better than the single time-scale estimates of MRAM features for non-intrusive speech quality evaluation.

ACKNOWLEDGEMENTS

The authors are thankful to Mr. Yi Hu and Dr. Philipos C. Loizou for providing the NOIZEUS database of 2240 speech utterances degraded at different degradations.

REFERENCES

[1] ITU-T Rec. p. 800, "Methods for Subjective Determination of Transmission Quality", 1996.

[2] ITU-T Rec. p. 563, "Single Ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications", 2004.

[3] Grancharov, V., Zhao, D.Y., Lindblom, J. and Kleijn, W.B. (2006), "Low Complexity Non-intrusive Speech Quality Assessment", *IEEE Trans. on Audio, Speech and Lang. Process*., Vol. 14, No. 6, pp. 1948–1956.

[4] Dempster, A.P., Laird, N. and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society,* Series B (Methodological), Vol. 39, No. 1, pp. 1–38.

[5] Lyon, R.F. (1982), "A Computational Model of Filtering, Detection and Compression in the Cochlea", in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Proces., Palo Alto*, CA, pp. 1282–1285.

[6] Dubey, R.K. and Kumar, A. (2013), "Non-intrusive Speech Quality Assessment using Several Combinations of Auditory Features", *Int. Journal of Speech Technology, Springer*, Vol. 16, No. 1, pp. 89–101.

[7] Parmar, N. and Dubey, R.K. (2015), "Comparison of Performance of the Features of Speech Signal for Non-intrusive Speech Quality Assessment", in *Proc. of Int. Conf. on Signal Process. and Communication*, Noida, India, pp. 243–248.

[8] Dubey, R.K. and Kumar, A. (2013), "Non-intrusive Objective Speech Quality Assessment using a Combination of MFCC, PLP and LSF Features", in *Proc. of Int. Conf. on Signal Process and Communication*, Noida, India, pp. 297–302.

[9] Dubey, R.K. and Kumar, A. (2015), "Non-intrusive Speech Quality Estimation using Multi-resolution Auditory Model Features", *IET Signal Process*, Vol. 9, Ino. 9, pp. 638–346.

[10] Dubey, R.K. and Kumar, A. (2016), "Lyon's Auditory Features and MRAM Features Comparison for Non-intrusive Speech Quality Assessment in Narrowband Speech", in *Proc. of 3rd Int. Conf. on Signal Process. & Integrated Networks*, Noida, India, Feb. 2016.

[11] Singh, M. and Dubey, R.K. (2012), "Non-intrusive Speech Quality with Different Time Scale", *IOSR Journal of Computer Engineering,* Vol. 2, No. 5, pp. 49–53.

[12] Karmakar, A., Kumar, A. and Patney, R.K. (2016), "A Multiresolution Model of Auditory Excitation Pattern and its Application to Objective Evaluation of Perceived Speech Quality", *IEEE Trans. on Audio, Speech and Lang. Process*., Vol. 14, No. 6, pp. 1912–1923, Nov. 2006.

[13] ITU-T Rec. p. Supplement-23 "ITU-T coded-speech Database", 1998.

[14] http://www.utdallas.edu/~loizou/speech/noizeus last accessed Feb.2009.