A NOVEL REFERENCE-FREE OBJECTIVE SPEECH QUALITY PERCEPTION MEASUREMENT USING MULTI-INSTANCES FEATURES OF DEGRADED SPEECH SIGNAL

Rajesh Kumar Dubey* and Arun Kumar

Centre for Applied Research in Electronics,
Indian Institute of Technology Delhi,
New Delhi 110016, India.
E-mail: rajeshk_dubey@yahoo.com

Abstract: In modern telecommunication networks, it is an important requirement to measure the quality of speech objectively and continuously at different nodes of the network. Reference-free (non-intrusive) speech quality estimation algorithms measure the quality of speech signals without using the original clean speech signal as a reference. In this work, reference-free speech quality assessment is done for telephony band speech signal using multi-instance features which are probabilistically modelled using Gaussian Mixture Model (GMM). The use of single-instance features, as in existing algorithms, is not accurate in capturing the time localized information of short-time transient distortions and their distinction from plosive sounds of a speech signal. Hence, the importance of estimating features at multi-instances that are relevant for objective speech quality measurements. The silence segments are removed from the speech signal and only active speech segments are considered for features computation using frame by Lyon's auditory model. The features thus computed are combined by taking mean, variance, skewness and kurtosis over the frames to obtain the features of the active speech segment. A principal component analysis is done to reduce the dimensionality of features. In a similar manner, mel-frequency cepstral coefficients (MFCC) and line spectral frequencies (LSF) are also computed on a per-frame basis and combined by taking mean over the frames to obtain the features. Then, the active speech segments are combined across the segments across an increasing number of active segments till all the segments of complete speech utterance are accounted for. The features of the combination of active speech segments are computed in a similar manner to obtain the resultant features of the combination of active segments. For training of the algorithm, the subjective Mean Opinion Score (MOS) of the speech signal that is available from a suitably large and varied training database is taken as the subjective Mean Opinion Score (MOS) for each active speech segment or the combination of active speech segments. These features along with the subjective MOS are used for the training of a joint GMM probability density function and then used to measure the objective MOS of each active speech segment or the combination of active speech segments. The overall objective MOS of the speech utterance is obtained by taking average of the objective MOS of the segments. A results in terms of correlation coefficient of subjective MOS and objective MOS and their comparison with the ITU-T Recommendation P.563 has been presented here.

Keywords: *Speech quality, Degraded signal, Gaussian mixture model (GMM), Mel-frequency cepstral coefficients (MFCC), Line spectral frequencies (LSF).*

*Also, working as Assistant Professor in the Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida (India), *Corresponding author

1. Introduction

The speech processing algorithms and codecs are used in modern telecommunication systems and thus for monitoring and maintaining the quality of service the speech quality assessment is essential at different nodes of the telecommunication system. If the quality of speech is not up to the level for customer satisfaction, then feedback can be given to the base station for the proper bandwidth allocation to improve the quality of speech and thus the quality of service. There are two methods for signal-based speech quality assessment: Intrusive and Non-intrusive or reference-free. Original clean speech is required for comparison in intrusive speech quality assessment method as a reference but in the non-intrusive (reference-free) method of speech quality assessment, the original clean speech is not required as a reference and thus it is suitable for speech quality assessment at any node of the telecommunication network and system automation.

Reference-free (non-intrusive) speech quality assessment algorithms depend only on the received (degraded) speech utterance to estimate its quality called mean opinion score objective listening quality (MOS-LQO) [1]. The ITU-T has standardized Recommendation P.563 in 2004 for non-intrusive speech quality assessment [2]. Ideally, the speech quality should be assessed by subjective listening test using the Absolute Category Rating (ACR) method as given in ITU-T Recommendation P.800-Aug.1996 [3]. In which speech is played to the human listeners and the average of their opinions about the quality of speech is considered as speech quality for a particular speech utterance and called the mean opinion score-subjective listening quality (MOS-LQS). The low complexity approach for non-intrusive or reference-free speech quality assessment by GMM training and speech quality evaluation with different local and global features obtained from speech coders and without considering any degradation model is explained in [4]. The human auditory system is modelled explicitly or implicitly leading to the brain to get an opinion score MOS-LQO of the speech. The auditory models used in this work are Lyon’s cochlear model [5], which takes into account the critical band and masking effect of the human auditory system.

In this work, the features are computed at multiple time scales called multi-instance features, which capture the features of speech at different time scales. Each speech utterance is passed through a voice activity detection (VAD) algorithm to get the active speech segments. Now, each active speech segment or the combination of multiple contiguous active speech segments of speech utterance is divided into frames and features are computed on a per-frame basis using Lyon’s auditory model. These per-frame features are combined over the frames to give features of an active speech segment or the combination of multiple contiguous active speech segments. In a similar manner, MFCC [6], [7] and LSF features [8] are computed at multiple time scales and combined to Lyon’s features at multiple time scales to estimate the quality of speech utterance. The subjective MOS of the speech utterance is taken as the subjective MOS for each active speech segment or the combination of multiple contiguous active segments. These features along with the subjective MOS are used for the training of joint GMM and the objective MOS of each active speech segment or the combination of multiple contiguous active speech segments are computed using GMM parameters. The objective MOS of the speech utterance is computed by taking equal weights of the segments. The results in terms of correlation coefficient of subjective MOS and objective MOS are obtained and compared with single time-scale features approach of non-intrusive speech quality estimation as well as ITU-T Recommendation P. 563.

2. Multiple time-scale or multi-instances auditory features

The more detailed statistical information of local features particularly for contiguous speech segments can be captured on multiple time-scales, if non-stationary noise is present in the speech utterance. Thus, this approach of multiple time-scale features may improve the correlation of subjective and objective MOS in speech quality estimation. The degraded speech is input to the multiple time-scale auditory feature computation modules. Each speech utterance is passed through the voice activity detection (VAD) algorithm to obtain the active speech segments. For a speech utterance having three active speech segments, the output of the VAD algorithm is schematically shown in Fig. 1.

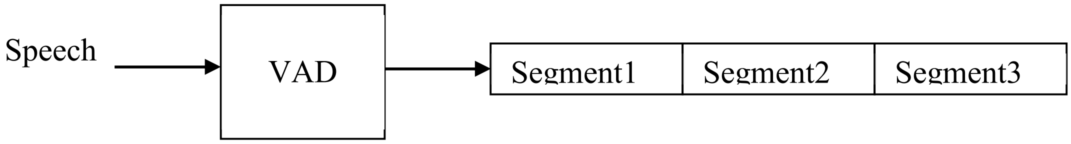

Figure 1: Concatenation of active speech segments.

The active speech segments at the output of the VAD algorithm are used to make the different combinations of multiple time duration active speech segments till all the active speech segments are accounted for. The method of making the combinations of active speech segments for a speech utterance having three active speech segments is shown in Fig. 2.

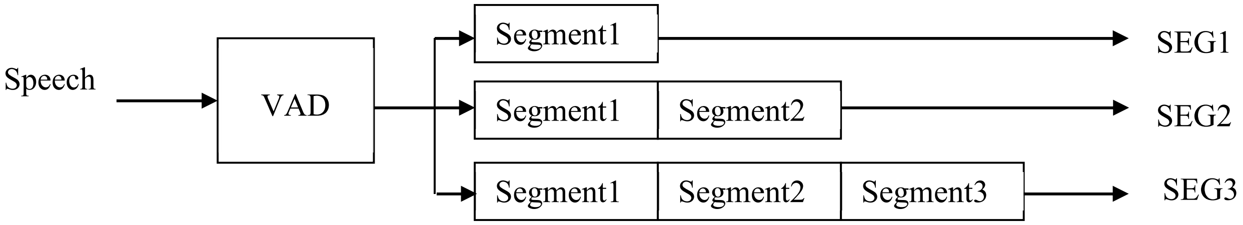

Figure 2: Combinations of three active speech segments for different multi-instance estimates.

The first active segment is, say SEG1. Next, the combinations of the first and second active speech segments is, say, SEG2. Finally, the combination of the first, second and third active speech segments is, say, SEG3. In a similar manner, for *S* active speech segments in an utterance, there shall be *K* combinations of segments, on the lines of SEG1, SEG2 …..up to SEG*K.* The active speech segment, SEG1 or the combinations of active speech segments such as SEG2 SEG3,…..up to SEG*K* are divided into frames of fixed duration of 16 ms and 64-channel Lyon's auditory features are computed on a frame-by-frame basis after windowing with a Hamming window of 16 ms duration. These computed features are combined by taking the mean, variance, skewness and kurtosis over the frames by concatenating them. Thus, 256-dimension Lyon's feature vector set is generated for a 64-channel Lyon's auditory model which is reduced to a 30-dimensional feature vector set for the first active speech segment, SEG1, by principal component analysis. In the multiple time-scale features approach, the duration of active speech segments is varying over time.

To preserve more than 99% of the energy, 30 principal components of Lyon's auditory features are used in the multiple time-scale features approach.

In a similar manner, other features such as 13-dimensional MFCC and 10-dimensional LSF are also computed on a frame-by-frame basis by dividing active speech into frames for the first active speech segment. The computed 30-dimensional Lyon's features, 13-dimensional MFCC and 10-dimensional LSF features are now concatenated to give a 53- dimensional feature vector and appended with the subjective MOS of the active speech segment SEG1, which is the subjective MOS of the corresponding speech utterance as shown in Fig. 3.

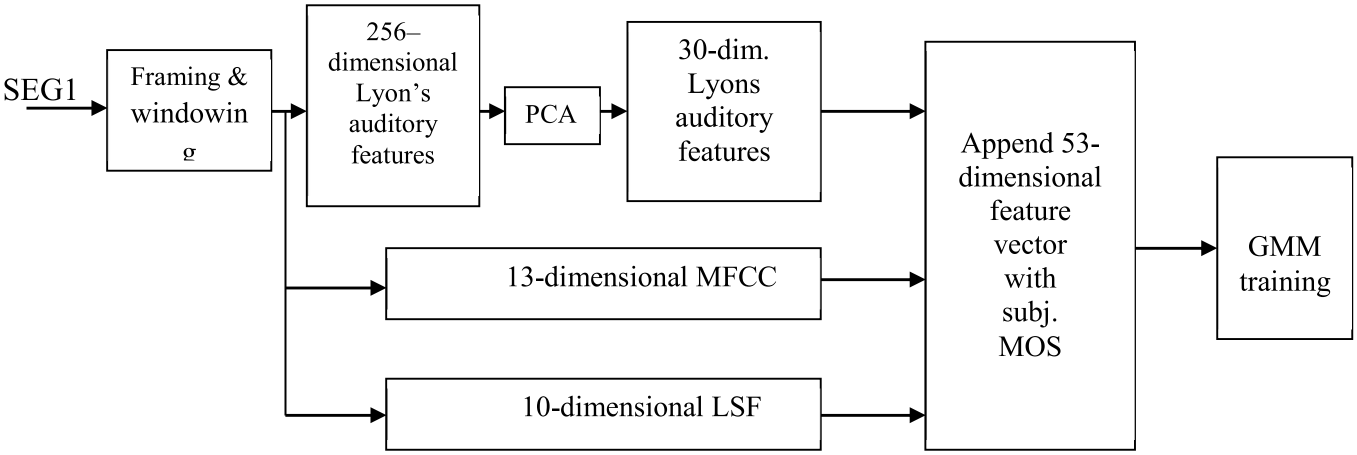
SEG1
Framing & windowin g
256–dimensional Lyon's auditory features
PCA
30-dim. Lyons auditory features
13-dimensional MFCC
10-dimensional LSF
Append 53-dimensional feature vector with subj. MOS
GMM training

Figure 3: Computation of 53-dimensinal feature vector, and appending with the subjective MOS for GMM training.

In a similar manner, 53-dimensional feature vectors are computed for the combinations the active speech segments such as SEG2, SEG3 and so on up to SEG*K* and appended with the subjective MOS of the speech utterance. The subjective Mean Opinion Score (MOS) for each active speech segment or for the combinations of the active segments is taken as the subjective MOS of the speech utterance during the training of the joint GMM. The objective MOS of each active speech segment or the combinations of the active speech segments are computed using GMM parameters and the 53-dimensional feature vectors. The averaging of the objective MOS of the single or multiple combinations of the active speech segments is done i.e. equal weights are given to the objective MOS of the single or multiple combinations of the active speech segments to compute the objective MOS of the corresponding speech utterance. If $\hat{\theta}$ is the objective MOS of speech utterance, then it is computed by taking the average of the objective MOS of *K* SEGs, $\hat{\theta}_i$ given by,

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

where, *K* is the number of the active speech segments for the speech utterance.

3. GMM training and speech quality estimation

The subjective MOS score θ_j from MOS labelled speech databases is appended to the 53-dimensional feature vector Ψ, (which is a combination of 30-dimensional Lyon's feature vector, 13-dimensional MFCC, and 10-dimensional LSF features) and used for the training of a joint Gaussian Mixture Model (GMM) using Expectation-Maximization algorithm [9] to obtain the parameters of the joint GMM $\Pi(\mu^{(k)},\omega^{(k)},\sum^{(k)})$ with $k=1,2,3\ldots, M$ mixture components, where $\mu^{(k)},\omega^{(k)}$, and $\sum^{(k)}$ are the mean, mixture weight, and covariance matrix respectively of the k-th mixture component. Thus, $[\Psi_j,\ \theta_j]$ is the 54-dimensional feature vector for the j-th training utterance, where j=1, 2, 3,….., J is the number of speech utterances used for the training of the joint GMM.

Now, the aim is to get an objective estimator $\hat{\theta}$ for the quality of a speech utterance as a function of the feature vector i.e., $\hat{\theta} = \hat{\theta}(\psi)$ and given the trained joint GMM parameters $\Pi(\mu^{(k)},\omega^{(k)},\sum^{(k)})$, the objective MOS estimate $\hat{\theta}$ is obtained using the MMSE criterion:

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

The modelling of the joint density function of the feature vector variables along with the subjective MOS scores as a GMM facilitates the estimation:

$$f(\psi/\Pi) = \sum_{k=1}^{K} \omega^{(k)} N\left(\psi/\mu_{\psi}{}^{(k)}, \Sigma_{\psi\psi}{}^{(k)}\right) \tag{3}$$

where, $N(\psi/\mu^{(k)},\Sigma^{(k)})$ are the multivariate Gaussian densities, with $\mu^{(k)}$ being the mean vectors and $\Sigma^{(k)}$ the covariance matrices of the k-th mixture components of Gaussian density.

$$E\{\theta/\psi\} = \sum_{k=1}^{K} X^{(k)}(\psi)\, \mu_{\theta/\psi}^{(k)} \tag{4}$$

where,

$$X^{(k)}(\psi) = \frac{\omega^{(k)} N\left(\psi/\mu_{\psi}{}^{(k)}, \Sigma_{\psi\psi}{}^{(k)}\right)}{\sum_{k=1}^{K} \omega^{(k)} N\left(\psi/\mu_{\psi}{}^{(k)}, \Sigma_{\psi\psi}{}^{(k)}\right)} \tag{5}$$

and

$$\mu_{\theta/\psi}^{(k)} = \mu_0^{(k)} + \Sigma_{\psi\theta}^{(k)} (\Sigma_{\psi\psi}^{(k)})^{-1} (\psi - \mu_{\psi}{}^{(k)}) \tag{6}$$

where, $\mu_{\Psi}{}^{(k)}$, is the mean of feature vector Ψ, $\mu_{\theta}{}^{(k)}$, is the mean of subjective MOS θ, $\sum_{\Psi\Psi}{}^{(k)}$, is the covariance matrix of Ψ, *and* $\sum_{\Psi\theta}{}^{(k)}$ is the cross-covariance matrix of Ψ and θ. In this investigation, $M=12$ mixture components are used in the GMM for the modelling of the probability density function for the combination of the feature vector.

For GMM training and objective MOS computation using different feature vectors and GMM parameter, "leave one out" procedure is used. Out of 10 subsets of the database, 9 subsets are used for training of the GMM and the remaining one subset is used for objective MOS computation.

This procedure is repeated (randomly always) to get an objective MOS score list for all speech utterances.

4. Description of Databases

Three databases are used in this work. First one is ITU-T P. Supplement 23 database [10] of 1328 speech utterances each of duration 8 seconds, at 332 different degradation conditions and sampled at 8 kHz of sampling rate are available in this database along with the ACR labelled subjective MOS. The second one is the NOIZEUS-2240 database, obtained from the University of Texas; Dallas, USA is a noisy speech database of 2240 speech sentences which contains 20 clean speech utterances all sampled at an 8 kHz sampling rate and 3-second duration. The speech utterances are degraded by passing through 4 different types of noise namely babble, car, street and train noise at 5 dB and 10 dB SNR levels each. To process each speech utterance there are 14 different speech enhancement and noise suppression schemes namely MMSE-STSTA (6 algorithms), spectral subtraction (3 algorithms), subspace-approach (2 algorithms) and Weiner filtering (3 algorithms) are used. A total of 2240 degraded speech sentences with 112 different conditions were used for conducting the subjective listening tests in our laboratory. The third database is the NOIZEUS-960 database which is taken from the NOIZEUS database of noisy speech corpus of 960 speech sentences [11]. There are 30 clean speech sentences each of duration 3 seconds and sampled at an 8 kHz sampling rate. Each speech utterance is degraded with 8 different types of noise namely airport, babble, car, exhibition, restaurant, station, street and suburban train at 4 different SNR levels (0 dB, 5 dB, 10 dB and 15 dB). Thus, resulting in 960 degraded speech sentences at 32 different degradation conditions of noise. These speech utterances are used for conducting the subjective listening test to get the subjective MOS to make it suitable for the training of GMM.

5. Computation of results and analysis

The Karl-Pearson's correlation coefficient and root mean square error (RMSE) between estimated objective speech quality MOS score $\hat{\theta}$ and the subjective MOS score θ are used for the performance evaluation of different reference-free (non-intrusive) speech quality assessment techniques. Results in terms of the correlation coefficient are computed and compared in **Table 1** for condition averaged MOS and **Table 2** for unconditioned MOS using three databases. Although, in most of the literature condition averaged case is considered but the unconditioned case is more realistic and seems to be the true measure of performance as it shows one-to-one speech utterances correlation. Results in terms of correlation coefficient by the proposed model (the combination of Lyon's auditory model features, MFCC and LSF feature at multiple time scales or multi-instance features) are also compared with ITU-T Rec. P.563, standard for non-intrusive speech quality estimation. In Fig. 4 and Fig. 5 results are shown in form of a bar chart for better visualization of comparison. Fig. 4 (i) is for correlation comparison and Fig. 4 (ii) for RMSE comparison for unconditioned estimated objective MOS. In a similar manner, Fig. 5 (i) is presented for correlation comparison and Fig. 5 (ii) for RMSE comparison for condition averaged estimated objective MOS. It is observed that there is significant increase in correlation coefficient and reduction in RMSE for both conditioned average case of MOS as well as unconditioned MOS for the chosen datasets of speech utterances.

Table 1: Correlation and RMSE between the unconditioned subjective and the unconditioned estimated objective MOS.

Database	No. of speech utterances	ITU-T Rec. P. 563		Proposed model	
		Correlation	RMSE	Correlation	RMSE
ITU-T Supp. 23	1328	0.7168	0.5801	0.9233	0.3356
NOIZEUS-960	960	0.7169	0.8567	0.9180	0.2770
NOIZEUS-2240	2240	0.3057	0.9988	0.7007	0.3791

Table 2: Correlation and RMSE between the subjective and the estimated objective MOS for the condition averaged case.

Database	No. of speech utterances	ITU-T Rec. P. 563		Proposed model	
		Correlation	RMSE	Correlation	RMSE
ITU-T Supp. 23	1328	0.8159	0.4502	0.9667	0.1686
NOIZEUS-960	960	0.9512	0.2505	0.9950	0.0398
NOIZEUS-2240	2240	0.9548	0.4225	0.9862	0.0699

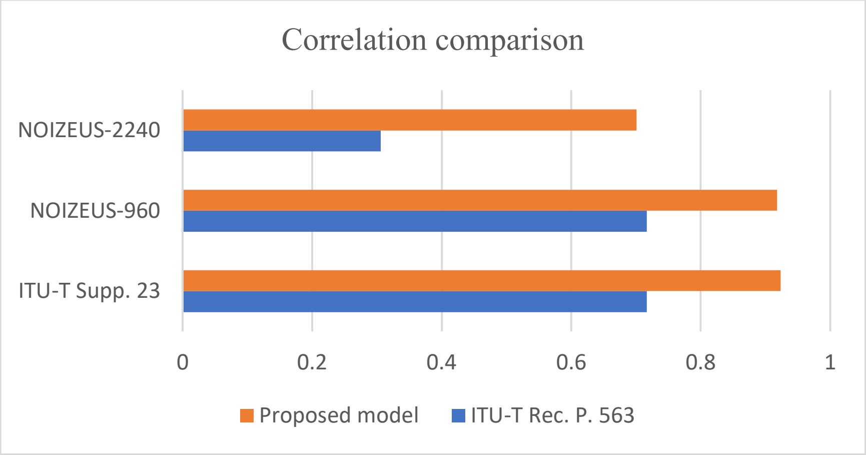

Fig.4 (i) Correlation comparison for unconditioned case of estimated objective MOS

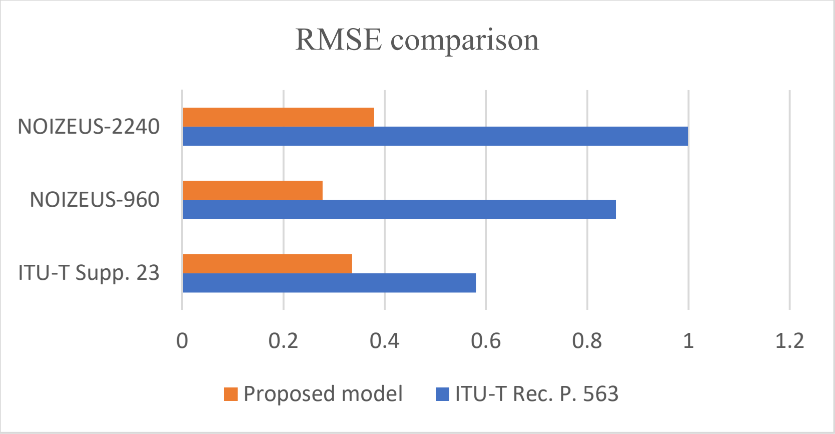

Fig.4 (ii) RMSE comparison for unconditioned case of estimated objective MOS

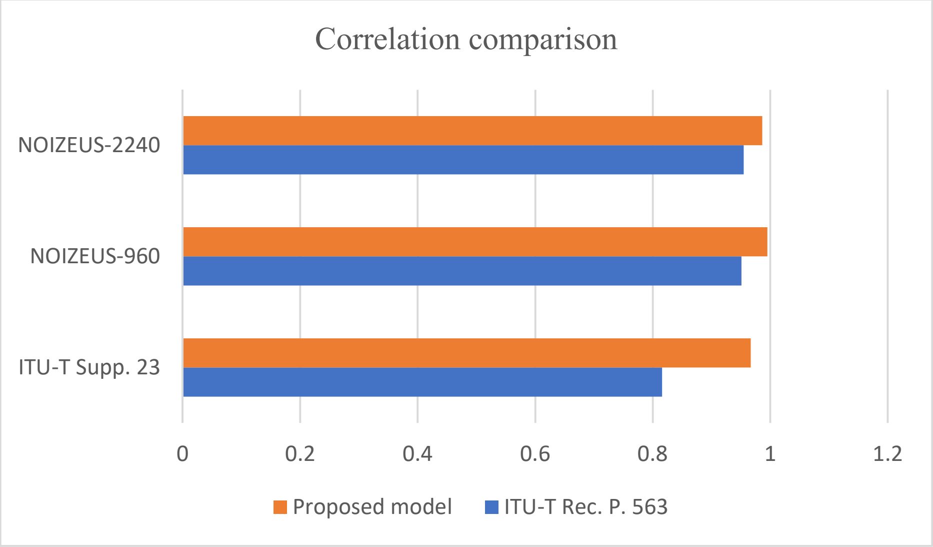

Fig. 5 (i) Correlation comparison for condition averaged case of estimated objective MOS

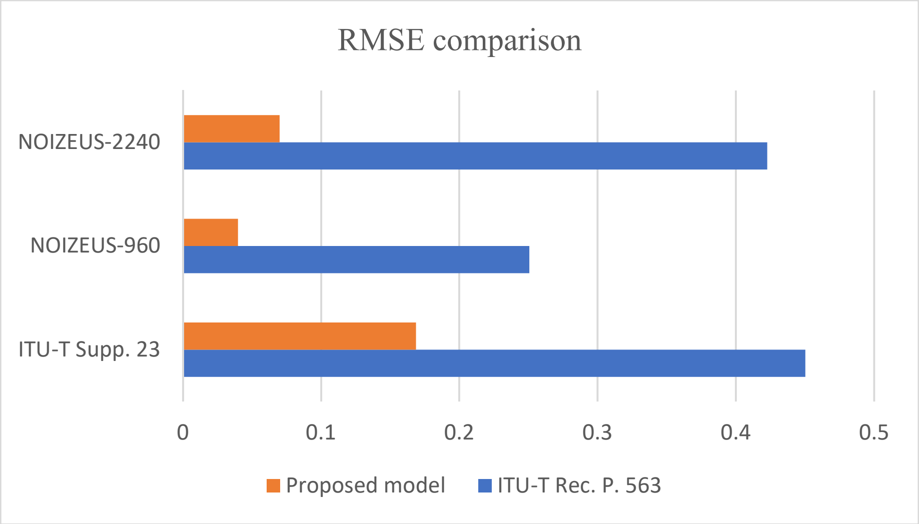

Fig. 5 (ii) RMSE comparison for condition averaged case of estimated objective MOS

6. Conclusions

Lyon’s auditory features, MFCC and LSF features are computed for multiple time scales or multi-instances for an active speech segment or for the combination of active segments. These multiple time-scale features are combined for a speech utterance for a non-intrusive (reference-free) speech quality assessment. The overall objective MOS of the speech utterance is computed by taking equal weights (averaging) of the MOS of the multiple time-scales estimates. The results in terms of correlation of the subjective and the estimated objective MOS for different types of noisy speech database are obtained and compared with the single time-scale results and ITU-T Rec. P.563 and found that the multiple time scales or multi-instance features approach outperforms as compared to the ITU-T Rec. P.563, the prevailing standard for non-intrusive speech quality evaluation.

ACKNOWLEDGEMENTS

The authors would like to thank Mr. Yi Hu and Dr. Philipos C. Loizou of Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX75083-0688, USA for providing the NOIZEUS-2240 database of 2240 speech utterances passed through different speech processing algorithms at different noisy conditions to create the different degradations.

REFERENCES

[1] L. Malfait, J. Berger, and M. Kastner, 2006. "P.563-The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. on Audio, Speech and Language Processing*, 14 (6), 1924-1934.

[2] 2004. "Single-ended method for objective speech quality assessment in narrow-band telephony applications," *ITU-T Rec. P.563*.

[3] 1996. "Methods for subjective determination of transmission quality," *ITU-T Rec.P. 800*.

[4] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, 2006. "Low-complexity, non-intrusive speech quality assessment," *IEEE Trans. on Audio, Speech and Language Processing*, 14 (6), 1948-1956.

[5] R. F. Lyon, 1989. "A computational model of filtering, detection, and compression in the cochlea," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Palo Alto, CA, 1282-1285, May (1982).

[6] M. R. Hasan, M. Jamil, M. G. Rabbani, and M. S. Rahman,2004. "Speaker identification using Mel-frequency cepstral coefficient," in *3rd Int. Conf. on Electrical & Computer engineering*, Dhaka, Bangladesh, 565-568, Dec.(2004).

[7] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel and L. T. Chia, 2012. "Nonintrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression," *IEEE Trans. on Audio, Speech and Language Processing*, 20 (4), 1217-1232.

[8] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem,2010. "Use of line spectral frequencies for emotion recognition from speech," in *IEEE Int. Conf. on Pattern Recognition*, Istanbul, 3708-3711, Aug. (2010).

[9] A. P. Dempster, N. Laird, and D. B. Rubin, 1977. "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1), 1-38.

[10] "ITU-T coded-speech database," *ITU-T Rec. P. Supplement 23*, 1998.

[11] URL: http://www.utdallas.edu/~loizou/speech/noizeus.