# Multiple Time-Instances Features of Degraded Speech for Single Ended Quality Measurement

*Rajesh Kumar DUBEY*[1], *Arun KUMAR*[2]

[1]Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Sector-62, 201309 Noida, Uttar Pradesh, India
[2]Centre for Applied Research in Electronics, Indian Institute of Technology Delhi, Huaz Khas, 110016 New Delhi, India

rajesh.dubey@jiit.ac.in, arunkm@care.iitd.ac.in

DOI: 10.15598/aeee.v15i3.2330

***Abstract.*** *The use of single time-instance features, where entire speech utterance is used for feature computation, is not accurate and adequate in capturing the time localized information of short-time transient distortions and their distinction from plosive sounds of speech, particularly degraded by impulsive noise. Hence, the importance of estimating features at multiple time-instances is sought. In this, only active speech segments of degraded speech are used for features computation at multiple time-instances on per frame basis. Here, active speech means both voiced and unvoiced frames except silence. The features of different combinations of multiple contiguous active speech segments are computed and called multiple time-instances features. The joint GMM training has been done using these features along with the subjective MOS of the corresponding speech utterance to obtain the parameters of GMM. These parameters of GMM and multiple time-instances features of test speech are used to compute the objective MOS values of different combinations of multiple contiguous active speech segments. The overall objective MOS of the test speech utterance is obtained by assigning equal weight to the objective MOS values of the different combinations of multiple contiguous active speech segments. This algorithm outperforms the Recommendation ITU-T P.563 and recently published algorithms.*

## Keywords

***Auditory feature, degraded speech, speech quality.***

## 1. Introduction

The speech processing algorithms and codecs are used in modern telecommunication systems and thus the monitoring and maintaining the quality of speech is important from customer satisfaction point of view to maintain and improve the quality of service. One aspect of this requirement for the automated system is to evaluate the speech quality objectively and continuously. If the quality of speech is not up to the mark, the proper bandwidth allocation or other speech enhancement techniques can be utilized to improve the quality of speech and thus the quality of service. There are two methods for signal based speech quality measurement: Double ended (Intrusive technique) and single ended (Non-intrusive technique). Double ended (Intrusive technique) requires original clean speech signal along with the received degraded speech signal to compute the quality rating called objective MOS, while single ended (Non-intrusive technique) uses only received degraded speech signal to compute the quality rating [1]. The non-intrusive method of speech quality measurement is suitable for system automation and real-time applications where the original clean speech signal is practically impossible to obtain such as mobile communications, telephonic communication, Direct-to-Home (DTH) signal of television (TV), Voice over Internet Protocol (VoIP) signal, etc. The Recommendation ITU-T P.563 (May 2004) is the standard for single ended (non-intrusive) speech quality measurement [2]. The subjective measurement is the ideal way to obtain the speech quality rating of degraded speech signal where the speech signal is played and average value of opinions of about 16–20 listeners is treated as quality rating for a particular speech utterance and called the subjective MOS as per the Recommendation ITU-T P.800-Aug.1996 [3]. The measurement of

speech quality has been done using different types of features obtained from speech encoder and GMM mapping in [4], without considering any degradation model. The human auditory system modelled explicitly or implicitly as Lyon's cochlear model is used in this work. Reference [5], which takes into account for the critical band and different auditory phenomenon such as masking the effect of human auditory system. The functional role of the human auditory system and the articulator system characteristics in the form of temporal envelope representation of speech have been utilized in the Auditory Non-Intrusive Quality Estimation (ANIQUE) model [6]. The Lyon's auditory features computed for entire speech let us call as "single time-instance features" and their mapping to the speech quality score by GMM has been given in [7]. The combination of different single time-instance speech features including auditory features and features related to vocal-tract resonances are used for GMM mapping and speech quality evaluation in [8]. The method given in [9] is assessing dimensions of perceptual quality space using linear regression and the dimension used is the loudness of speech which describes a non-optimal sound level. Estimating the quality and intelligibility of speech degraded by additive noise and distortions associated with telecommunication networks, based on a data driven framework of feature extraction and tree based regression, is given in [10].

The limitations of current research in the literature are that the features used for speech quality measurement are single time-instance, where the entire speech utterance is used for the computation of features, and these features are mapped to the objective quality rating score. In this work, the features are computed at multiple time-instances which capture the presence of noise at different locations of the speech utterance instead of averaging the effect over the entire speech utterance. Thus, the use of single time-instance features is not accurate and adequate in capturing the time localized information of short-time transient distortions and their distinction from plosive sounds of speech, particularly degraded by impulsive noise. The Voice Activity Detection (VAD) algorithm is employed to get the active speech segments of different speech utterances [11]. Here, active speech means both voiced and unvoiced frames except silence. Now, the combinations of multiple contiguous active speech segments of speech utterance are made in increasing order till all the active speech segments are accounted for. These combinations of active segments are divided into frames and features are computed on per frame basis using Lyon's auditory model. These per frame features are combined over the frames to give features of the different combinations of multiple contiguous active speech segments. In similar manner, Mel-Frequency Cepstral Coefficients (MFCC) [12] and [13] and Line Spectral Frequencies (LSF) features [14] are computed at multiple time-instances and concatenated to obtain the feature vector. The subjective MOS of the speech utterance is taken as the subjective MOS for each of the different multiple time scale estimates (the combination of multiple contiguous active speech segments) during GMM training. The objective MOS values for each of the multiple time scale estimates are computed using the GMM parameters and different multiple time-scale features of test speech utterance. The overall objective MOS of the test speech utterance is computed by assigning equal weights to the objective MOS values of different multiple time scale estimates. The results are compared with Recommendation ITU-T P.563, the standard for non-intrusive technique of speech quality measurement, and different state-of-art recently published works [13], [15], [16], [17] and [18], which are using single time-instance features approach in terms of Pearson's correlation coefficient and RMSE between the subjective MOS and the overall objective MOS of speech utterances. The proposed algorithm using the combination of Lyon's auditory features, MFCC and LSF features, all computed at multiple time-instances, outperforms the state-of-art recent works.

## 2. Multiple Time-Instances Auditory Features

The more detailed statistical information of local features, particularly for contiguous speech segments, can be captured in multiple time-instances estimates, if non-stationary noise is present in the speech utterance. Thus, it is expected that the correlation between the subjective and the objective MOS in speech quality measurement problem will improve in multiple time-instances features approach. The degraded speech is input to the multiple time-instance auditory feature computation modules. At the very first stage, it will have to pass through VAD algorithm to remove silence region and find out the different active speech regions present in the speech utterances. For a speech utterance having three active speech segments, the output of VAD algorithm is schematically shown in Fig. 1.

The active speech segments at the output of the VAD algorithm are used in increasing order 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 concatenation to obtain different multiple time-instances estimates as the combinations of active speech segments for a speech utterance having three active speech segments is shown in Fig. 2. It will be continued till all the active segments are accounted for.

The first active segment is, say SEG1. Next, the combinations of the first and second active speech seg-

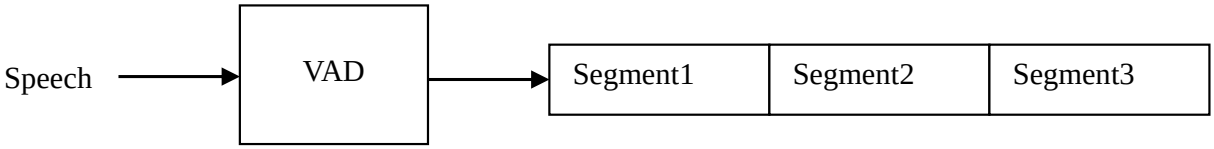


**Fig. 1:** Active speech segments and their concatenation.

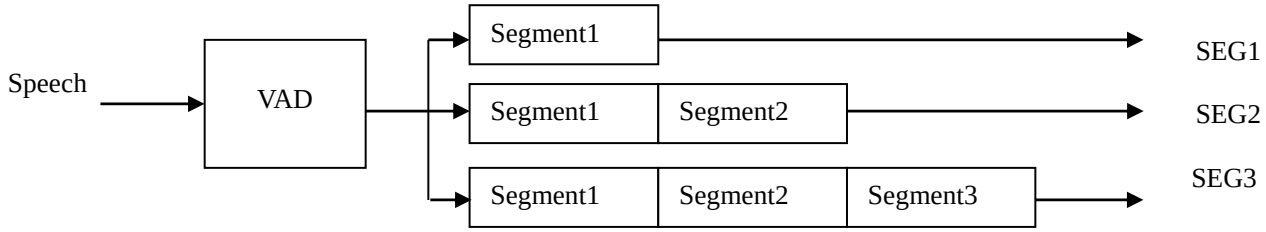


**Fig. 2:** Combinations of three active speech segments for different time-instances estimates for illustration.

ments is, say, SEG2. The combinations of the first, second and third active speech segments is, say, SEG3 and so on. In a similar manner, for $K$ number of active speech segments in a speech utterance, there will be K different combinations of multiple contiguous active speech segments, on the lines of SEG1, SEG2, ... up to SEGK. These combinations of multiple contiguous active speech segments such as SEG2, SEG3, ... up to SEGK are divided into frames of fixed duration of 16 ms and passed through 64-channel Lyon's auditory model to compute 64 auditory features on frame-by-frame basis after windowing with a Hamming window of 16 ms duration with 50 % overlap. The mean, variance, skewness and kurtosis over the frames of 64 auditory features are computed and concatenated to obtain 256-dimensional Lyon's feature vector. The dimensionality of the feature vector is reduced from 256 to 30 by using Principal Component Analysis (PCA) to preserve more than 98 % of the energy. In the multiple time-instances features approach, the duration of active speech segments is varying over time.

In a similar manner, 13-dimensional multiple time-instances MFCC and 10-dimensional multiple time-instances LSF feature vectors are also computed on per frame basis. All these feature vectors are now concatenated to obtain a 53-dimensional feature vector. In a similar manner, 53-dimensional feature vectors are computed for all multiple time-instances estimates such as SEG2, SEG3 and so on up to SEGK. For the training of joint GMM according to Expectation Maximization (EM) algorithm [19], the 53-dimensional feature vectors are appended with the subjective MOS values of the corresponding speech utterance. The subjective MOS for each of the multiple time-instances estimates is taken as the subjective MOS of the speech utterance, as shown in Fig. 3, because no separate subjective MOS will be available for the multiple time-instances estimates in any database. The objective MOS of each of the multiple time-instances estimates is computed using GMM parameters namely mean, mixture weight, and covariance matrix and 53-dimensional feature vectors of the corresponding multiple time-instances estimates. The objective MOS value of $i^{th}$ multiple-time scale estimate $\hat{\theta}_i$ as a function of 53-dimensional multiple time-instances feature vector $\psi$ is obtained using the Minimum Mean Square Error (MMSE) criterion [4]:

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

where $\theta$ is the subjective MOS of corresponding speech utterance. The three databases are randomized to use leave-one-out10-fold cross validation process for training and testing. That is, 90% data are used for training and 10 % data are used for testing. The process is repeated 10-times to obtain the objective MOS values for all the multiple time-instances estimates. In this work, GMM with 12 mixture components are used and all the GMM training parameters are computed offline and stored in a library. In real-time monitoring, only test speech will be used but there will be an algorithmic buffering delay corresponding to one sentence speech utterance before the multiple time-instances speech quality evaluation algorithms are applied.

The averaging of the objective MOS values of the multiple time-instances estimates is done i.e. equal weights are assigned to the objective MOS values of the different multiple time-instances estimates to compute the overall 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 values of $K$ SEGs, $\hat{\theta}_i$ is given by:

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

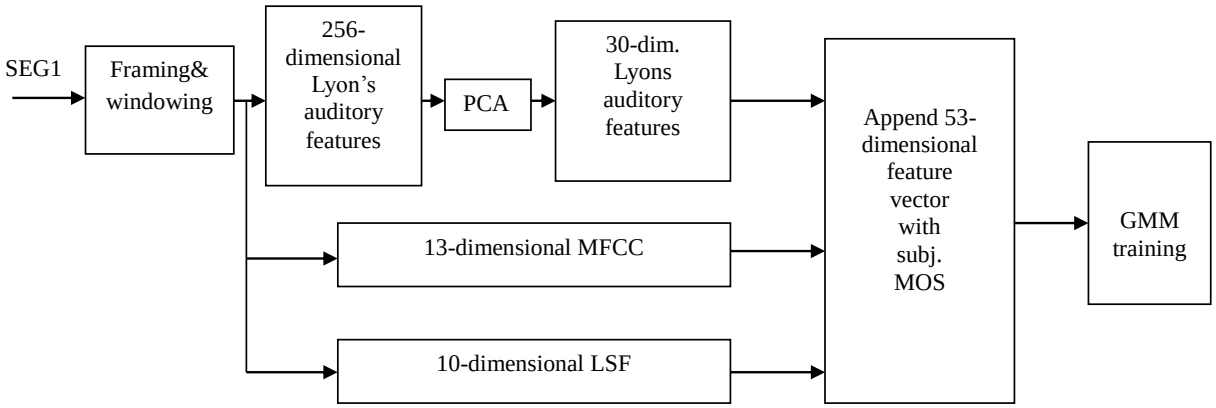


**Fig. 3:** Computation of 53-dimensional feature vector, and appending with the subjective MOS for GMM training.

where $K$ is the number of active speech segments which will be equal to the number of combinations of multiple contiguous active speech segments.

## 3. Description of Databases

In this work, three databases are used namely ITU-T Supplement-23 database [20], NOIZEUS-2240 and NOIZEUS-960 [21]. The first database of 1328 speech utterances constitute Expt.-1 (A, D, O) having a total of 528 speech utterances with 8 kbps ITU & ETSI standard CODECS interworking and Expt.-3 (A, C, D, O) is having a total of 800 speech utterances with channel errors and background noises. All these 1328 speech utterances degraded at 332 different degradation conditions and are of 8 second duration each, sampled at 8 kHz and subjective MOS labelled according to Absolute Category Rating (ACR). The second database is having a total of 2240 degraded speech utterances of 3 second duration each, sampled at 8kHz and degraded by 4 different types of noise namely babble, car, street and train noise at two different SNR levels, 5 dB and 10 dB. A total of 20 clean speech utterances are degraded at 112 different conditions of degradation. The third one, NOIZEUS-960 database, which is taken from NOIZEUS database of noisy speech corpus of 960 speech sentences of 30 clean speech signals are sampled at 8 kHz and of 3 seconds duration each. The clean signals are degraded by 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). The NOIZEUS-2240 and NOIZEUS-960 speech utterances are not having subjective MOS associated with them and thus subjective listening test was conducted to obtain the subjective MOS in our laboratory. The statistical analysis of the subjective MOS rating is presented in [22] to ensure the high degree inter-and-intra-rater reliability.

## 4. Results and Analysis

The Pearson's correlation coefficient and RMSE between the subjective MOS score $\theta$ and estimated overall objective MOS score $\hat{\theta}$, both computed as condition averaged value, are used as figure of merit in most of the literatures of single ended speech quality measurement algorithms. In this work, unconditioned values of the subjective and objective MOS are also used for the computation of Pearson's correlation coefficient and RMSE [8], where MOS values of speech sentence-by-sentence are used, because it will be more realistic. Results are given and compared in Tab. 1 for condition averaged MOS values and Tab. 3 for unconditioned MOS values using three databases. The comparison of results between single time-instance [8] and multiple time-instances approaches is presented along with Recommendation ITU-T P.563. The overall weighted average of the correlation using multiple time-instances estimates is 0.980 as against single time-instance features approach which is 0.960 [8], whereas the correlation is 0.934 using the ITU-T Rec. P.563 algorithm over the three databases for condition averaged MOS case as given in Tab. 1.

In [8], on same databases 37-dimensional feature vectors formed by combining 14-dimensional reduced size

**Tab. 1:** Correlation coefficients and RMSE between the subjective and the estimated overall objective MOS for the condition averaged MOS case.

| Database | No. of speech utterances | ITU-T Rec. P.563 | | Proposed model | |
|---|---|---|---|---|---|
| | | Correlation | RMSE | Correlation | RMSE |
| ITU-T Supp. 23 | 1328 | 0.815 | 0.450 | 0.966 | 0.168 |
| NOIZEUS -960 | 960 | 0.951 | 0.250 | 0.995 | 0.039 |
| NOIZEUS -2240 | 2240 | 0.954 | 0.422 | 0.986 | 0.068 |

Lyon's auditory model features, 13-dimensional MFCC features, and 10-dimensional LSF features, all computed at single time-instance for entire speech utterances are used. In this work, 53-dimensional feature vectors formed by combining 30-dimensional reduced size Lyon's auditory features, 13-dimensional MFCC features, and 10-dimensional LSF features, all computed at multiple time-instances are used. The basis for dimensionality reduction of Lyon's auditory features using PCA from 256 to 14 in the case of single time-instance is preservation of 98 % energy. According to this criterion, the dimensionality of multiple time-instances Lyon's auditory features is reduced from 256 to 30 using PCA. Moreover, the MFCC features are 13-dimensional and LSF features are 10-dimensional. Thus, single time-instance feature vectors are 37-dimensional and multiple time-instances feature vectors are 53-dimensional.

**Tab. 2:** Comparison of single time-instance and multiple time-instances features approach using equal weights with ITU-T Rec. P.563 taking condition averaged subjective MOS and estimated objective MOS.

| Data of Different Expts. | No. of Speech Utterances | ITU-T Rec. P.563 | Single time-instance features Lyon's, MFCC & LSF | Multiple time-instances features Lyon's MFCC, & LSF with equal weight |
|---|---|---|---|---|
| Exp.1(A) -French | 176 | 0.885 | 0.912 | 0.967 |
| Exp.1(D) -Japanese | 176 | 0.842 | 0.933 | 0.975 |
| Exp.1(O) -A.English | 176 | 0.902 | 0.946 | 0.988 |
| Exp.3(A) -French | 200 | 0.867 | 0.887 | 0.949 |
| Exp.3(C) -Italian | 200 | 0.854 | 0.851 | 0.954 |
| Exp.3(D) -Japanese | 200 | 0.929 | 0.908 | 0.948 |
| Exp.3(O) -A.English | 200 | 0.918 | 0.891 | 0.961 |
| NOIZEUS -960 | 960 | 0.951 | 0.993 | 0.995 |
| NOIZEUS -2240 | 2240 | 0.955 | 0.980 | 0.985 |
| **Weighted Average** | | **934** | **0.960** | **0.980** |
| Std. Dev. | | 0.041 | 0.046 | 0.018 |
| Confidence Interval (95 %) | | 0.027 | 0.030 | 0.012 |

The results in terms of Pearson's correlation coefficient and RMSE for condition averaged MOS are also compared in Tab. 5 with the published results of recent works in [13], [15] and [16] which were using a database of 1792 speech utterances that was a subset of NOIZEUS-2240 database of 2240 speech utterances used in this work. The comparison is also shown by

**Tab. 3:** Correlation coefficients and RMSE between the unconditioned subjective MOS and the unconditioned estimated overall 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 | 0.580 | 0.9233 | 0.335 |
| NOIZEUS -960 | 960 | 0.7169 | 0.856 | 0.9180 | 0.277 |
| NOIZEUS -2240 | 2240 | 0.3057 | 0.998 | 0.7007 | 0.379 |

**Tab. 4:** Comparison of single time-instances and multiple time-instances features approach using equal weights with ITU-T Rec. P.563 taking unconditioned subjective and estimated objective MOS.

| Data of Different Expts. | No. of Speech Utterances | ITU-T Rec. P.563 | Single time-instance features Lyon's, MFCC & LSF | Multiple time-instances features Lyon's MFCC, & LSF with equal weight |
|---|---|---|---|---|
| Exp.1(A) -French | 176 | 0.759 | 0.837 | 0.921 |
| Exp.1(D) -Japanese | 176 | 0.701 | 0.828 | 0.934 |
| Exp.1(O) -A.English | 176 | 0.790 | 0.828 | 0.956 |
| Exp.3(A) -French | 200 | 0.768 | 0.773 | 0.889 |
| Exp.3(C) -Italian | 200 | 0.762 | 0.753 | 0.903 |
| Exp.3(D) -Japanese | 200 | 0.801 | 0.806 | 0.901 |
| Exp.3(O) -A.English | 200 | 0.788 | 0.745 | 0.91 |
| NOIZEUS -960 | 960 | 0.717 | 0.859 | 0.918 |
| NOIZEUS -2240 | 2240 | 0.306 | 0.690 | 0.695 |
| **Weighted Average** | | **0.529** | **0.756** | **0.807** |
| Std. Dev. | | 0.155 | 0.054 | 0.076 |
| Confidence Interval (95 %) | | 0.101 | 0.035 | 0.050 |

bar graph in Fig. 4. Here, we have conducted subjective listening tests to obtain the subjective MOS for 2240 speech utterances, while in [13], [15] and [16] they have used their own respective subjective scores. The value of correlation reported in [13] for the condition averaged case is 0.9002 and the RMSE to be 0.33, whereas in this proposed work the correlation obtained is 0.986 and the RMSE to be 0.068 respectively for the NOIZEUS-2240 database. In [15], the maximum value of Pearson's correlation coefficients obtained is 0.910 in test-1 which uses 8-fold cross validation process, whereas 10-fold cross-validation process has been used

**Tab. 5:** Comparison of results in terms of Pearson's correlation coefficient and RMSE with recently published works [11], [13] and [14] on NOIZEUS-2240 database.

| Methods | | Correlation | RMSE |
|---|---|---|---|
| Ref [11] | | 0.9002 | 0.33 |
| Ref [13] | Test-1 | 0.910 (Estimated from given bar chart) | 0.190 (Estimated from given bar chart) |
| | Test-2 | 0.886 | 194 |
| | Test-3 | 0.842 | 248 |
| Ref [14] | Mean | 0.77 | 0.29 |
| | Variance | 0.83 | 0.25 |
| | Mean + Variance | 0.90 | 0.20 |
| Proposed Work | | 0.986 | 0.068 |

in this proposed work. In [16], the mean and variance statistics of Gabor PCA features gives the best performance for speech quality assessment on NOIZEUS-2240 database. It uses 80 % of data for training and 20 % for testing. The maximum Pearson's correlation coefficients obtained to be 0.90 and RMSE 0.20 in [16].

**Tab. 6:** Comparison of results in terms of Pearson's correlation coefficient with recently published works [16] on ITU-T P.Supplement-23 database.

| **Database** | **ITU-T Rec. P.563** | **Bag-of-Words Representation Algorithm** | **Multiple time-instances features Lyon's, MFCC & LSF** |
|---|---|---|---|
| Exp.1(A) -French | 0.885 | 0.933 | 0.967 |
| Exp.1(D) -Japanese | 0.842 | 0.902 | 0.975 |
| Exp.1(O) -A.English | 0.902 | 0.949 | 0.988 |
| Exp.3(A) -French | 0.867 | 0.925 | 0.949 |
| Exp.3(C) -Italian | 0.854 | 0.849 | 0.954 |
| Exp.3(D) -Japanese | 0.929 | 0.888 | 0.948 |
| Exp.3(O) -A.English | 0.918 | 0.902 | 0.961 |
| **Average** | **0.885** | **0.893** | **0.963** |

The comparison of results in terms of Pearson's correlation coefficient for NOIZEUS-960 database has also been done with [17] for condition averaged MOS, which is the same speech database used in this work. Here, we have conducted subjective listening tests to obtain the subjective MOS for 960 speech utterances, while in [17] they have used their own respective subjective scores. The Pearson's correlation coefficients obtained in [17] was 0.933 as against 0.995 in this proposed work. In [17], 70 % of data has been used for training while 30 % for testing. The comparison of results in terms of Pearson's correlation coefficient for ITU-T P. Supplement-23 database has also been done with recent work [18] in Tab. 6 for seven sub-databases for condition averaged MOS values. In these comparisons, it is observed that the proposed work performs better than these recently published works.

## 5. Inferences Drawn from Results

From the overall results expressed in tabular form and different comparisons, the following inferences are made:

- The multiple time-instances estimates to compute the objective MOS score of the overall speech utterance gives higher correlation as compared to the single time-instances features approach.
- For both, the condition averaged MOS case or unconditioned MOS case, correlation coefficients and RMSE are significantly better for multiple time-instances estimates as compare to single time-instances estimates over the different databases.
- In this algorithm, the combination of reduced size Lyon's auditory features with MFCC and LSF features are used as feature vectors in the study. In this, even there will be some duplicity of information in the features, but the combination of features gives better result in terms of correlation and RMSE between the subjective MOS and the estimated overall objective MOS for speech on sentence-by-sentence basis. By combining these feature vectors, the correlation coefficient, in both the cases of unconditioned and condition averaged MOS increases significantly.

## 6. Conclusion

Lyon's auditory features, MFCC and LSF features are computed for multiple time-instances for the different combinations of multiple contiguous active speech segments. These multiple time-instances features are combined for a speech utterance for single ended speech quality measurement. The overall objective MOS of the speech utterance is computed by assigning equal weights (averaging) of the MOS values of the multiple time-instances estimates. The results in terms of correlation coefficients between the subjective and the estimated overall objective MOS for different types of noisy speech databases are obtained and compared with the different single time-instances approaches recently published and Rec. ITU-T P.563 and found that multiple time-instances approach outperforms.

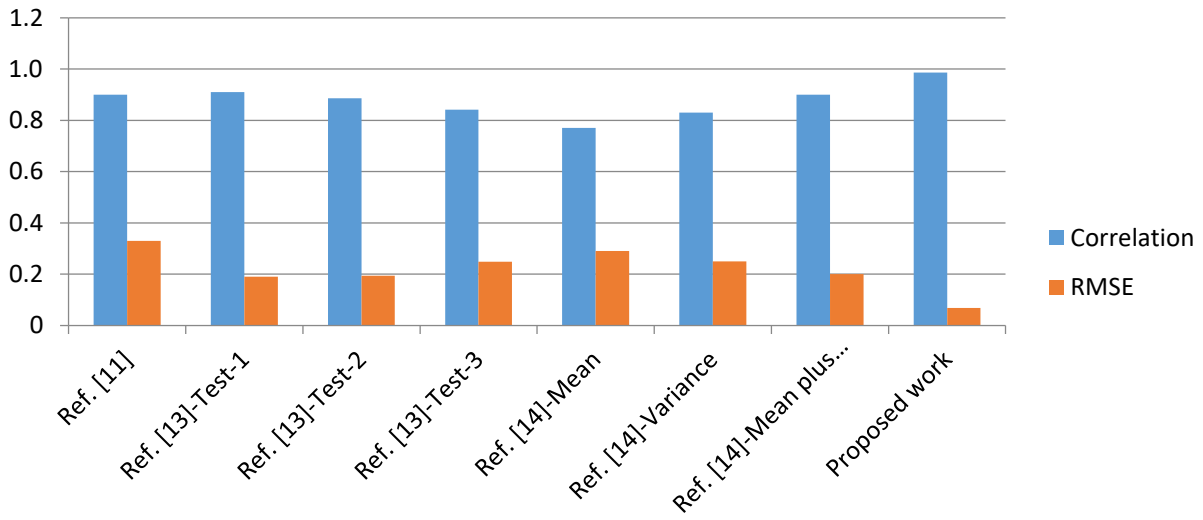


**Fig. 4:** Bar graph comparison with recent published work [11], [13] and [14].

## Acknowledgment

The authors would like to express their gratitude to 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.

## References

[1] MALFAIT, L., J. BERGER and M. KASTNER. P.563-The ITU-T standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech and Language Processing*. 2006, vol. 14, iss 6, pp. 1924–1934. ISSN 1558-7924. DOI: 10.1109/TASL.2006.883177.

[2] ITU-T RECOMMENDATION P. 563. *Single ended method for objective speech quality assessment in narrow-band telephony applications*. Geneva: ITU-T, 2005.

[3] ITU-T RECOMMENDATION P. 800. *Methods for subjective determination of transmission quality*. Geneva: ITU-T, 1996.

[4] GRANCHAROV, V., D. Y. ZHAO, J. LINDBLOM and W. B. KLEIJN. Low-complexity, non-intrusive speech quality assessment. *IEEE Transactions on Audio, Speech and Language Processing*. 2006, vol. 14, iss. 6, pp. 1948–1956. ISSN 1558-7924. DOI: 10.1109/TASL.2006.883250.

[5] LYON, R. F. A computational model of filtering, detection, and compression in the cochlea. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Paris: IEEE, 1982, pp. 1282–1285. ISBN 978-82-80125-127. DOI: 10.1109/ICASSP.1982.1171644.

[6] KIM, D. S. ANIQUE: An auditory model for single ended speech quality estimation. *IEEE Transactions on Audio, Speech and Language Processing*. 2005, vol. 13, iss. 5, pp. 821–831. ISSN 1063-6676 DOI: 10.1109/TSA.2005.851924.

[7] AUDHKHASI, K. and A. KUMAR. Two scale auditory features based non-intrusive speech quality evaluation. *IETE Journal of Research*. 2010, vol. 56, iss. 2, pp. 111–118. ISSN 0377-2063.

[8] DUBEY, R. K. and A. KUMAR. Non-intrusive speech quality assessment using several combinations of auditory features. *International Journal of Speech Technology*. 2013, vol. 16, iss. 1, pp. 89–101. ISSN 1381-2416. DOI: 10.1007/s10772-012-9162-4.

[9] KOSTER, F., V. CERCOS-LLOMBART, G. Gabriel MITTAG and S. MOLLER. Non-intrusive estimation model for the speech-quality dimension loudness. In: *Proceedings of Speech Communication, 12. ITG Symposium*. Berlin: IEEE, 2016, pp. 175–179. ISBN 978-3-8007-4275-2.

[10] SHARMA, D., Y. WANG, P. A. NAYLOR and M. BROOKES. A data-driven non-intrusive measure of speech quality and intelligibility. *Speech Communication*. 2016, vol. 80, iss. C, pp. 84–94. ISSN 0167-6393. DOI: 10.1016/j.specom.2016.03.005.

[11] RABINER, L. R. and M. R. SAMBUR. Voiced-unvoiced-silence detection using the Itakura LPC distance measure. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. New Jersey: IEEE, 1977, pp. 323–326. DOI: 10.1109/ICASSP.1977.1170330.

[12] HASAN, M. R., M. JAMIL, M. G. RABBANI and M. S. RAHMAN. Speaker identification using mel-frequency cepstral coefficient. In: *3rd International Conference on Electrical & Computer engineering*. Bangladesh: ICECE, 2004, pp. 565–568. ISBN 984-32-1804-4.

[13] NARWARIA, M., W. LIN, I. V. MCLOUGHLIN, S. EMMANUEL and L. T. CHIA. Non-intrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression. *IEEE Transactions on Audio, Speech and Language Processing*. 2012, vol. 20, iss. 4, pp. 1217–1232. ISSN 1558-7924. DOI: 10.1109/TASL.2011.2174223.

[14] BOZKURT, E., E. ERZIN, C. E. ERDEM and A. T. ERDEM. Use of line spectral frequencies for emotion recognition from speech. In: *IEEE International Conference on Pattern Recognition*. Istanbul: IEEE, 2010, pp. 3708–3711. ISBN 978-1-4244-7541-4. DOI: 10.1109/ICPR.2010.903.

[15] SONI, M. H. and H. A. PATIL. Novel deep autoencoder features for non-intrusive speech quality assessment. In: *Proceedings of the 24th European Signal Processing Conference (EUSIPCO)*. Budapest: IEEE, 2016, pp. 2315–2319. ISBN 978-1-5090-1891-8. DOI: 10.1109/EUSIPCO.2016.7760662.

[16] LI, Q., Y. FANG, W. LIN and D. THALMANN. Non-intrusive quality assessment for enhanced speech signals based on spectro-temporal features. In: *International Conference on Multimedia and Expo Workshops (ICMEW)*. Chengdu: IEEE, 2014, pp. 1–6. ISBN 978-1-4799-4716-4. DOI: 10.1109/ICMEW.2014.6890561.

[17] ISLAM, M. R., M. A. RAHMAN, M. N. HASAN, A. S. HOSSAIN, A. N. UDDIN and M. A. HAQUE. Non-intrusive objective evaluation of speech quality in noisy condition. In: *9th International Conference on Electrical and Computer Engineering (ICECE)*. Dhaka: IEEE, 2016, pp. 586–589. ISBN 978-1-5090-2964-8. DOI: 10.1109/ICECE.2016.7853988.

[18] LI, Q., Y. FANG, W. LIN and D. THALMANN. Bag-of-words representation for non-intrusive speech quality assessment. In: *China Summit and International Conference on Signal and Information Processing (ChinaSIP)*. Chengdu: IEEE, 2015, pp. 616–619. ISBN 978-1-4799-1947-5 DOI: 10.1109/ChinaSIP.2015.7230477.

[19] DEMPSTER, A. P., N. LAIRD and D. B. RUBIN. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1977, vol. 39, iss. 1, pp. 1–38. ISSN 1369-7412.

[20] ITU-T RECOMMENDATION SUPPLEMENT 23. *Methods for subjective determination of transmission quality*. Geneva: ITU-T, 1998.

[21] NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms. In: *University of Texas at Dallas* [online]. 2009. Available at: `http://ecs.utdallas.edu/loizou/speech/noizeus/`.

[22] DUBEY, R. K. and A. KUMAR. Comparison of subjective and objective speech quality ass assessment for different degradations/noise conditions. In: *International Conference on Signal Processing and Communication (ICSC)*. Noida: IEEE, 2015, pp. 261–266. ISBN 978-1-4799-6762-9. DOI: 10.1109/ICSPCom.2015.7150659.

## About Authors

**Rajesh Kumar DUBEY** received the B.Tech. degree in Electrical Engineering from the National Institute of Technology Hamirpur (HP), India in 1999, the M.Tech. degree in Electrical Engineering in 2002, and the Ph.D. degree from the Center for Applied Research in Electronics, in 2014, both from the Indian Institute of Technology Delhi, India. Since 2006, he is working as an Assistant Professor (Sr. Grade) in the department of Electronics and Communication Engineering, Jaypee Institute of Information Technology (JIIT), Noida, India. His research interests span the areas of Digital Signal and Speech Processing.

**Arun KUMAR** received the B.Tech, M.Tech, and Ph.D. degrees in Electrical Engineering from the Indian Institute of Technology Kanpur, India, in 1988, 1990, and 1995, respectively. He was a visiting researcher at the University of California, Santa Barbara, from 1994 to 1996. Since 1997, he is working as professor at the Center for Applied Research in Electronics, Indian Institute of Technology Delhi, India. His research interests span the areas of Digital Signal Processing, Speech Processing, Underwater Acoustics, Air Acoustics and Communications. In these areas, he has introduced new courses at the Masters level, supervised more than 40 funded research projects, and several Masters and Ph.D theses at IIT Delhi. Dr. Kumar is a recipient of the Young Scientist Award of the International Union of Radio Science (URSI).