Non-intrusive speech quality assessment using several combinations of auditory features

Article *in* International Journal of Speech Technology · March 2013
DOI: 10.1007/s10772-012-9162-4

2 authors:

Rajesh Kumar Dubey
Jaypee Institute of Information Technology
10 PUBLICATIONS **32** CITATIONS
SEE PROFILE

Arun Kumar
Veltech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering Coll...
24 PUBLICATIONS **45** CITATIONS
SEE PROFILE

Int J Speech Technol (2013) 16:89–101
DOI 10.1007/s10772-012-9162-4

Non-intrusive speech quality assessment using several combinations of auditory features

Rajesh Kumar Dubey · Arun Kumar

Received: 21 April 2012 / Accepted: 11 June 2012 / Published online: 27 June 2012

Abstract Quality estimation of speech is essential for monitoring and maintenance of the quality of service at different nodes of modern telecommunication networks. It is also required in the selection of codecs in speech communication systems. There is no requirement of the original clean speech signal as a reference in non-intrusive speech quality evaluation, and thus it is of importance in evaluating the quality of speech at any node of the communication network. In this paper, non-intrusive speech quality assessment of narrowband speech is done by Gaussian Mixture Model (GMM) training using several combinations of auditory perception and speech production features, which include principal components of Lyon's auditory model features, MFCC, LSF and their first and second differences. Results are obtained and compared for several combinations of auditory features for three sets of databases. The results are also compared with ITU-T Recommendation P.563 for non-intrusive speech quality assessment. It is found that many combinations of these feature sets outperform the ITU-T P.563 Recommendation under the test conditions.

Keywords Non-intrusive speech quality evaluation · Two-sided speech quality evaluation · Lyon's auditory model · Speech quality assessment

R.K. Dubey (✉)
Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology (JIIT), Noida, India
e-mail: rajeshk_dubey@yahoo.com

A. Kumar
CARE, Indian Institute of Technology (IIT), Delhi, India
e-mail: arunkm@care.iitd.ac.in

1 Introduction

The estimation of speech quality is essential at different nodes of the modern telecommunication networks for several applications such as network system design, development, monitoring and maintenance of the Quality of Service (QoS), for example in mobile communications, voice over internet protocol (VoIP) communications, conventional telecommunication networks, and multimedia applications. It is also required in performance evaluation of speech coders, automatic speech and speaker recognition systems, and text-to-speech synthesis systems.

Direct estimation of speech quality can be done by subjective listening tests using the Absolute Category Rating (ACR) method-as given in ITU-T Recommendation P.800-Aug. (1996). The mean opinion score (MOS-LSQ) in subjective listening tests is the average of opinions given by human listeners for a particular speech utterance. But, these subjective tests are time consuming, expensive and also not suitable for automation. Thus, objective estimation methods of perceived speech quality that can supplement the subjective tests are becoming increasingly important. The goal of objective estimation of speech quality is to mimic a human subject's performance in subjective quality tests. It is not known very accurately that how the human auditory system leading to the brain processes speech utterances to award a quality score to it. Thus, it is a challenging task to get a human-like objective model that well correlates with the subjective quality assessment scores. As is well known, the subjective scores do not correlate well with simple and intuitive objective measures such as mean square error, segmental SNR, spectral distortion, Itakura-Saito distance etc. A more meaningful objective model for the estimation of speech quality attempts to uses speech production and auditory perception principles to estimate the mean opinion

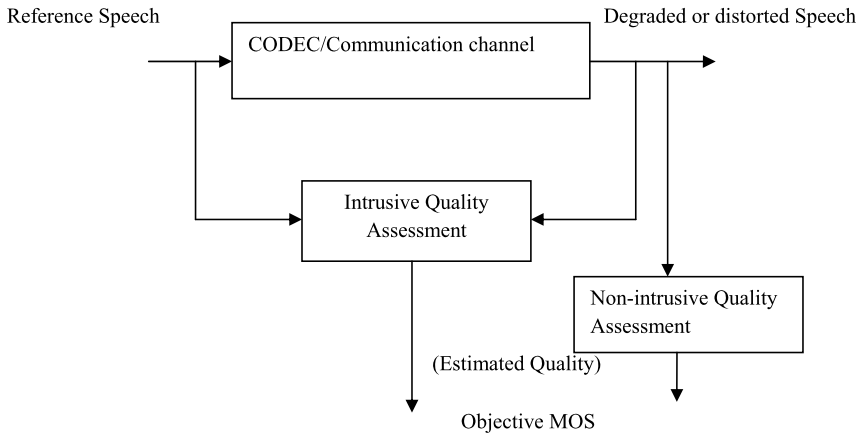

Fig. 1 Block-level comparison of intrusive and non-intrusive speech quality assessment

score (MOS). A psychoacoustic front-end using FFT based Mel filter bank, forward, and backward masking component followed by a cognitive modeling is utilized for objective speech quality assessment (Fegyo et al. 2000). A number of algorithms have been proposed in the research literature to objectively estimate the mean opinion score (MOS) of a speech utterance. These algorithms are broadly classified into two categories: Intrusive (two-sided) and Non-Intrusive (one-sided) speech quality evaluation algorithms whose principle difference is depicted in Fig. 1.

In intrusive speech quality evaluation algorithms the original clean speech signal is required as reference and it is compared with the degraded one to estimate the speech quality (Fegyo et al. 2000). The International Telecommunications Union (ITU-T) standardized intrusive speech quality assessment through its Recommendation P.862, perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks in 2001 (ITU-T Recommendation P.862 1996) replacing P.861, objective quality measurement of telephone band (300–3400 Hz) speech codecs using perceptual speech quality measurement (PSQM) of 1996 (ITU-T Recommendation P.861 1996). The intrusive approach requires the clean speech signal as a reference, which is impractical across telecommunication networks. Thus, this approach is not suitable for monitoring and maintaining the quality of service at different nodes of the telecommunication network. On contrary, only the received (degraded) speech signal is used to estimate the speech quality in non-intrusive algorithms. The ITU-T standardized its Recommendation for non-intrusive speech quality assessment by P.563, single ended method for objective speech quality assessment in narrow-band telephony applications in May 2004 (ITU-T Recommendation P.563 2004).

In the present work, we explore the effect of different combinations of meaningful feature sets to study which combinations lead to better objective scores for non-intrusive speech quality evaluation in terms of increased correlation with the subjective scores. The organization of the rest of the paper is as follows: Sect. 2 is concerned with the review of the prior art in non-intrusive speech quality evaluation. Section 3 describes the feature sets that have been combined variously to study the efficacy of objective speech quality evaluation. These feature sets include principal components of Lyon's auditory model, mel-frequency cepstral coefficients (MFCC), line spectral frequencies (LSF), and their first and second differences. In Sect. 4, Gaussian Mixture Model (GMM) training and speech quality estimation steps are explained along with the details of the speech databases that have been used in the training and evaluation of the algorithms. Section 5 gives the performance evaluation results of the speech quality estimation algorithms using different feature set combinations, and makes observations. Section 6 concludes the paper.

2 Review of prior art in non-intrusive speech quality evaluation

One of the first non-intrusive signal based model was proposed in 1994, by Liang and Kubichek (1994). Perception based speaker independent speech parameters such as perceptual linear prediction (PLP) coefficients and perceptually weighted Bark spectrum are utilized in this algorithm. The artificial reference parameters corresponding to high speech quality are derived from a variety of clean source speech material. They compared the degraded speech parameters to an artificial reference clean speech signal parameters that is appropriately selected from an optimally clustered codebook. The speech degradation was estimated by finding the average distance between the parameters of the test and reference parameter sets. Au and Lam proposed the output based objective speech quality measure, which is based on visual features of the spectrogram of distorted speech (Au and Lam 1998). In this method, average features of block- wise variance and dynamic range are calculated from the spectrogram

of typically 10 to 30 ms and 50 to 500 Hz time-frequency resolution. It is presumed that a "good" quality sentence will have discrete and dominant features while a "bad" quality sentence will typically have uniform distribution of energy in the spectrogram. Gray attempted to predict the quality of network degraded speech by the use of vocal-tract modeling technique (Gray et al. 2000). The parameterized data are used to generate physiology based rules for error assessment utilizing cavity tracking techniques and context based error spotting.

A high level description of no-reference (non-intrusive) speech quality assessment model was standardized in May 2004 as the ITU-T, Recommendation P.563 (2004). The algorithm aims at predicting the subjective quality MOS of 3.4 kHz band (narrowband) speech signals transmitted through telephone networks that may introduce background noise, filtering and variable delay, as well as distortions due to channel errors and speech coders. In principle, it uses a full reference perceptual model by reconstructing a quasi-clean intermediate reference speech signal from the degraded signal by modeling the vocal tract as a series of tubes. A total of 51 speech features are extracted from the frame of the signal. Eight key features are used to determine a dominant distortion class, and in each distortion class a linear combination of features is used to predict the intermediate speech quality. The overall speech quality is estimated by taking the linear combination of the intermediate quality and 11 additional features. The implementation of the ITU-T Recommendation P.563 for single-ended speech quality assessment in general speech communication applications is described in Malfait et al. (2006). In another approach (Kim 2005), an auditory model for single-ended (non-intrusive) speech quality estimation is presented. It utilizes the temporal envelope representation of speech for quality estimation. In this method, auditory non-intrusive quality estimation (ANIQUE) is based on the functional roles of human auditory systems and the characteristics of human articulation systems. A low complexity algorithm for monitoring the non-intrusive speech quality over a network is presented in Grancharov et al. (2006). The 11 per frame local features (e.g. spectral flatness, spectral centroid, excitation variance, speech variance, pith-period, their time derivatives, and spectral dynamics) are computed and statistical properties like mean, variance, skewness and kurtosis of these per frame features are used in the algorithm after dimensionality reduction to 14 per utterance called global features set. These features can be computed from commonly used speech-coding parameters without any perceptual transformation of the speech signal or reconstruction of any intermediate reference signal. The algorithm estimates speech quality by GMM training using 14-dimensional global generic features of speech utterances, commonly used in speech coding, without any assumption on explicit distortion modeling and their subjective MOS.

Recent algorithms based on Gaussian-mixture models (GMM) of features derived from perceptually motivated spectral envelope representations can be found in Falk et al. (2005), and for Bayesian model in Chen and Parsa (2006). The accuracy and robustness of GMM model for speech quality estimation is improved with the addition of a reference model of the behavior of speech degraded by different transmission and/or coding schemes (Falk and Chan 2006a). The design of single-ended speech quality estimation algorithm using models of speech signals, including clean and degraded speech and speech corrupted by multiplicative noise and temporal discontinuities, and machine learning methods is given in Falk and Chan (2006b). Chen and Parsa (2005) have introduced the adaptive neuro-fuzzy inference system for single-ended quality estimation of speech signal. Lyon has introduced an auditory model for filtering, detection, and compression in the cochlea of human auditory system (Slaney 1988; Lyon 1982). The models are computational expressions of the important functions of the cochlea. The main parts of the models concern mechanical filtering effects and the mapping of mechanical vibrations into neural representations. The techniques for the extraction of robust feature vector in an auditory model are given in Jing and Johnson (2002).

Non-intrusive speech quality estimation has been primarily done by exploiting either the time domain signal based parameters obtained from the speech coders and vocal tract parameters or by using frequency domain features in the Bark or Mel scale or other human auditory system based parameters. The mean opinion score (MOS) is estimated either by a linear combination of features of the distorted speech or by cognitive mapping. It is of great interest if several of these features representing different aspects of the speech production or auditory system aspects are combined to improve the quality score's correlation with a subjective test based MOS score. Such an objective measure may be gainfully used in speech quality estimation applications.

3 Description of feature sets

A human listener who gives an opinion score about a speech sentence processes the speech signal waveform in the auditory system and the brain. In the present work, Lyon's cochlear model is used for obtaining auditory features that are used for non-intrusive objective speech quality estimation. This model is based on the knowledge of how the cochlea works. The main parts of the model concern the mechanical filtering effects and the mapping of mechanical vibrations into neural representation. This approach implicitly models the fluid-dynamic wave medium of the cochlea

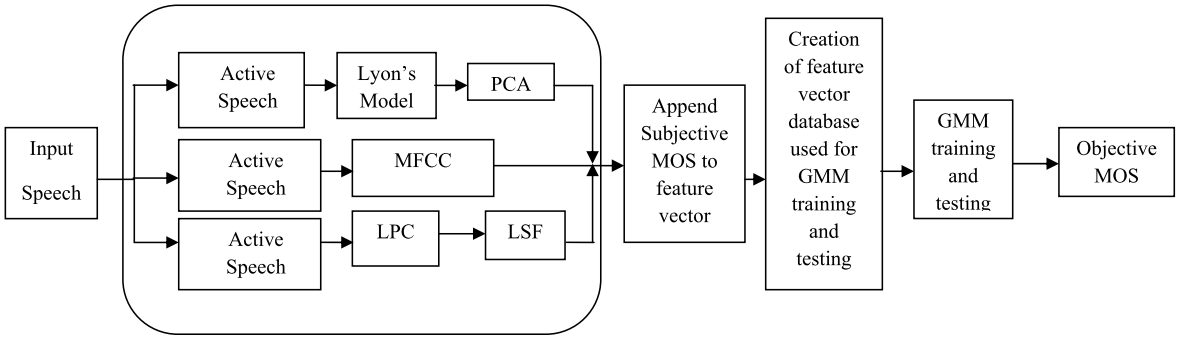

Fig. 2 Combined Feature Sets Model for GMM training and testing sequence in the ratio of 9:1

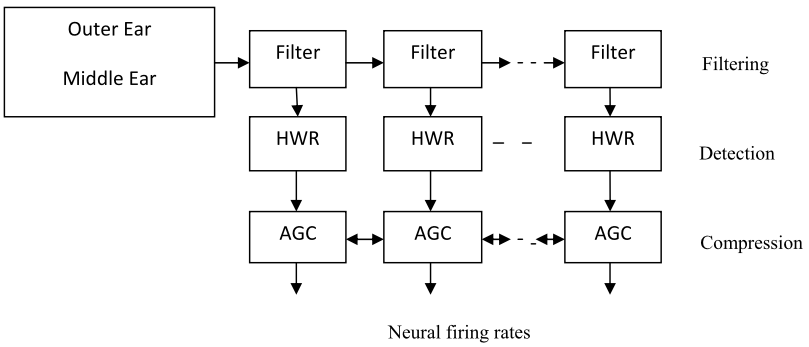

Fig. 3 Block-level signal flow in Lyon's cochlear model

by a cascade of filters based on the observed properties of the medium. The action of the active outer hair cells is modeled by a set of automatic gain controls (AGC), which simulates the dynamic compression of the intensity on the basilar membrane. A neural spike at the base of auditory nerve is only generated when the stereo cilia of the inner hair cell are bent one way and no spikes are generated when the cilia are bent in the other way. So the inner hair cells act like half-wave-rectifiers. The outputs of the model are the probability of firing of the neurons of the auditory nerve along time, called the cochleagram, a two-dimensional representation of time and frequency. To resemble the cognitive mapping function of the brain, Gaussian Mixture Model (GMM) based probability density function modeling is done using these auditory features to estimate the MOS for the speech sentence. The narrowband input speech signal, sampled at 8 kHz, is passed through the Voice Activity Detection algorithm to remove the silence region and get the active speech regions. The active speech is segmented into frames of length 16 ms with 50 % frame overlapping and the features are calculated for each frame.

This section describes the component feature sets of the Combined Feature Sets Model shown in Fig. 2. These feature sets are used in the proposed model for non-intrusive speech quality assessment.

3.1 Lyon's auditory model features

The model is based on the functional roles of human auditory systems. That is, how the human auditory system processes the speech sound that finally forms the basis of the opinion score. The human speech quality assessment process can be divided into two parts: Conversion of the received speech signal into auditory nerve excitations for the brain, and cognitive processing in the brain.

After selecting the speech part and rejecting the silence part from the speech utterance, voiced speech frames are passed through Lyon's auditory model, which has three main functional blocks, namely-(i) Filtering Stage, (ii) Detection Stage or Half Wave Rectification, and (iii) Compression Stage. The first filtering stage models, by a broadly tuned cascade of low-pass filters, the propagation of energy as waves on the basilar membrane (BM). The second detection stage non-linearly (through half wave rectification) converts membrane (BM) velocity into a representation of Inner Hair Cells (IHC) receptor potential or auditory nerve (AN) firing rate. Finally, the third compression stage continuously adapts, by an automatic gain control (AGC), the operating point of the system in response to its level of activity. Lyon's cochlear model aims to globally model the neural firing rate as a function of cochlear place or best frequency (BF) versus time. Lyon's cochlear model is shown in Fig. 3.

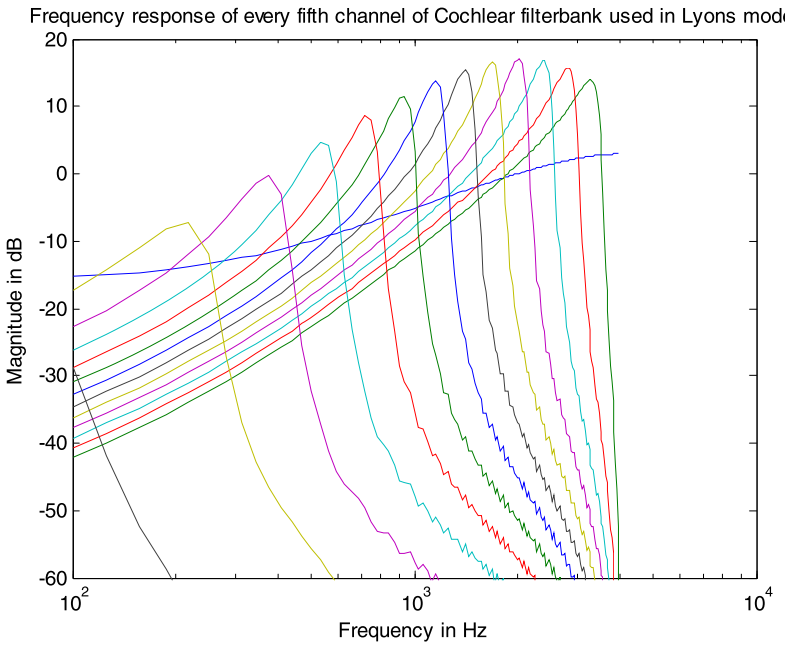

Fig. 4 Frequency response of every fifth channel of the cochlear filter bank used in Lyon's model restricted to 4 kHz

The filtering stage consists of a broadly tuned cascade of low-pass filters. Each filter bank uses a differentiator to convert pressure waves into basilar membrane motion. Each differentiator is adjusted so that at the centre frequency of the stage the differentiator has unity gain. There is another stage simulating the effects of the outer and middle ear before the filters. As the number of filters increases, accuracy of the model increases. The number of filters used depends on the sampling rate of signals, overlapping factor of the band of filters, and quality factor of the resonant part of the filters.

The cascade-parallel model of the broadly tuned low pass filter bank of 64-second order filters is implemented with freely available auditory toolbox. The resulting response for every fifth channel of the filter bank used in the Lyon's cochlear model is shown in Fig. 4. The BM is simulated by combining the stages into a cascade. The Detection stage is also called as half wave rectification (HWR). It is composed of a bank of HWRs which have the function to drop the negative portions of the waveform, modeling the directional behavior of the inner hair cells, thus cutting the energy measure of the signal by a factor of approximately 2. It nonlinearly converts basilar membrane (BM) velocity into a representation of Inner Hair Cells (IHC) reception potential or Auditory Nerve (AN) firing rate.

The Compression stage describes the adaptive features which work in our auditory system. It consists of 4 automatic gain control (AGC) stages that are cascaded. Each AGC reduces the level of the channel, but with shorter time constants. The AGC attenuate the input signal into a limited dynamic range of basilar membrane (BM) motion Inner Hair Cells (IHC) receptor potential or Auditory Nerve (AN) firing rate. The value of the gain of each stage depends on the time constant, the value of preceding output sample of the channel, and the value of preceding output sample of the adjacent channels. Thus masking effects can be reproduced in this model. The outputs of these stages approximately represent the neural firing rates.

The results from auditory model are used to form cochleagram. It is a 2-D representation of time and frequency. The frequency discrimination depends on the number of channels. The outputs of all the channels are autocorrelated. The short time autocorrelation (STA) of each output of the auditory model for an input speech signal $x_i(n)$ is calculated by:

$$r_{xi}(n,\tau) = \sum_{m=0}^{L-|\tau|-1} x_i(n+m).\hat{w}(n).x_i\big(n+m+|\tau|\big).\hat{w}\big(n+|\tau|\big) \tag{1}$$

where, $w(m)$ is a window function

$$\hat{w}(m) = w(-m) \tag{2}$$

and r_{xi} is the short time autocorrelation function, $i = 1, 2, \ldots, N$ is the channel index, L is the window length in samples, and τ is the autocorrelation lag. For windowing, the Hamming window is used.

Each of the outputs of Lyon's cochlear model is short time windowed and auto-correlated thus obtaining as many STAs as the number of channels. Each high energy frame is windowed by Hamming window and is passed through the 64 channel Lyon's model producing a 64-dimensional feature vector. For each channel output, the mean, variance,

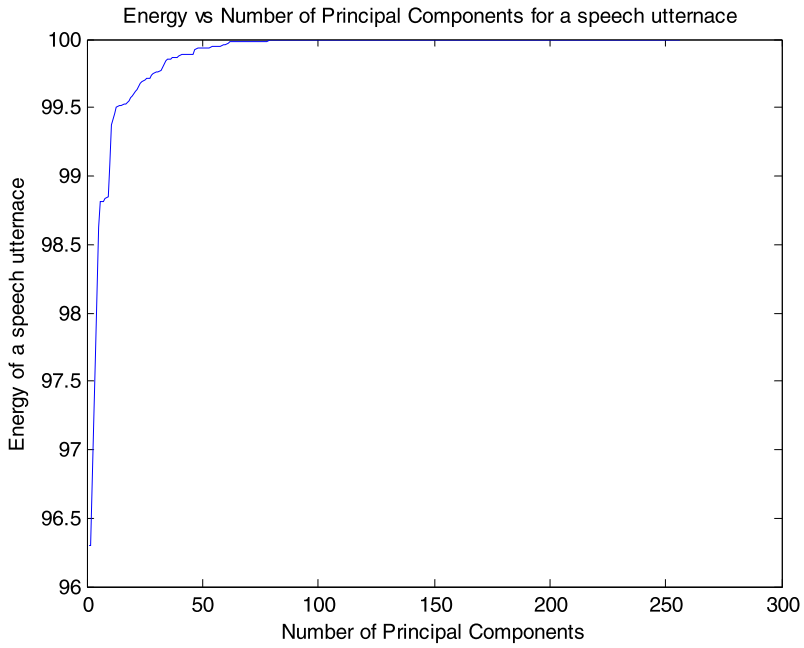

Fig. 5 Energy versus number of principal components for a speech utterance

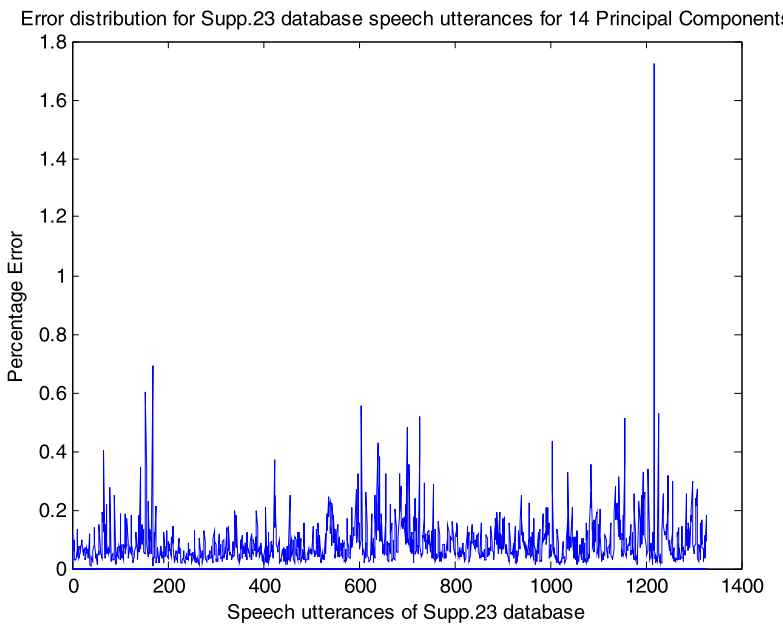

Fig. 6 Error distribution for Supp. 23 database speech utterances for 14 Principal Components

skewness and kurtosis are calculated over the frame and appended in a single vector producing a 256-dimensional feature vector. Principal component analysis is done to reduce the feature vector dimensionality from 256 to 14. The reduction of dimensionality of feature vector helps in reading the data required for training the GMM and also to reduce the computational complexity while insignificantly defeating the performance. The 14 principal components were observed to retain more than 98 % of the total energy of speech from ITU-T P.Supplement-23 database as shown in Fig. 5. Figure 6 shows the error distribution for ITU-T P.Supplement-23 database speech utterances, if 14 principal components are taken for each speech utterance. Thus, each speech frame is reduced to a 14-dimensional Lyon's auditory feature vector.

3.2 MFCC features

Mel Frequency Cepstral Coefficients (MFCC) are one of the most widely used parametric representation of the speech signal frame that captures the variations of basilar membrane (BM) of the human ear's critical bandwidth with frequency. The MFCC technique uses two types of filters. It

uses linear frequency spaced filters below 1000 Hz and logarithmic frequency spaced filters above 1000 Hz frequency. MFCC are less susceptible to the variations in the speech waveform due to the varying physical conditions of speaker's vocal cord (Hasan et al. 2004). MFCC are an effective representation of the perceptual quality variations of speech signal. Also, MFCC has a de-correlating effect by the use of Discrete Cosine Transform (DCT) on the log mel spectrum coefficients. It is therefore suitable for learning algorithms. The active speech is segmented into frames of length 16 ms and each individual speech frame of 16 ms duration is windowed using a Hamming window. Thus, 13-dimensional MFCC for each frame are obtained by performing the DCT of the log mel spectrum coefficients calculated for the frames.

Assuming there are total of K frames in a speech utterance, and the MFCC vector X_j for the jth frame is represented as

$$X_j = [m_{j1}, m_{j2}, \ldots, m_{j13}]^T \tag{3}$$

where, each m_{ji} denotes a mel-frequency cepstral coefficient (MFCC), $i = 1$ to 13 and $j = 1$ to K. The global feature vector of dimension 13 for each speech utterance is computed by taking the average of the MFCC vector over the frames $j = 1$ to K for each speech utterance (Narwaria et al. 2010).

3.3 Perceptual linear prediction (PLP), linear prediction coefficients (LPC) and line spectral frequencies (LSF) based feature vectors

The PLP coefficients proposed by Hermansky are used to represent the speech spectrum by a compact set of linear prediction coefficients after warping the frequency axis into the Bark frequency scale. It uses three psycho-acoustic concepts to estimate the auditory spectrum namely critical band spectral analysis, the equal loudness curve and the intensity power law. Here, eighth order perceptual linear prediction coefficients are computed to suppress the speaker dependent information such as pitch using PLP and RASTA Matlab functions.[1] Mean variance, skewness and kurtosis over the frames for an eighth order PLP are taken and a 36-dimensional feature vector is obtained for each speech utterance.

The line spectral frequency (LSF) features also offer an alternative efficient spectral envelope representation form for speech as borne out by its extensive use in speech coding algorithms. They carry intrinsic information of the formant structure which is related to the resonance frequencies of the vocal tract of the speaker during articulation. The active speech is segmented into frames of length 16 ms and each individual speech frame of 16 ms duration is windowed using a Hamming window. A 10th order LPC analysis over the frame is done to get 10 LSF for each frame. To get a 10-dimensional LSF feature vector the mean, variance, skewness, and kurtosis over the frames are computed. But only mean values over the frames as 10-dimensional LSF feature vector is used which are sufficient for spectral envelope representation form for speech. The experiment shows that the combination of the mean, variance, skewness, and kurtosis over the frames as a feature vector does not give any further improvement in the correlation of subjective MOS and estimated MOS.

[1] http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat.

4 GMM training and speech quality estimation

We first consider the Gaussian Mixture Model (GMM) based MOS training using the Expectation Maximization algorithm (Dempster et al. 1977). If Q is the subjective quality MOS score of a speech utterance from the MOS labeled training databases, then our aim is to get an objective estimator $\hat{Q}$ of the subjective quality as a function of the feature vector set i.e., $\hat{Q} = \hat{Q}(\psi)$, where ψ represents the reduced size feature vector according to the criterion,

$$\hat{Q}(\psi) = E(Q/\psi) = \underset{Q^*(\psi)}{\arg\,\min} E\left\{\left(Q - Q^*(\psi)\right)^2\right\} \tag{4}$$

where $E\{\}$ is the expectation operator. Thus, it is a problem of the estimation of conditional probability. It is assumed that the joint density of the feature vector variables along with the subjective MOS scores is probabilistically modeled as a GMM probability density function:

$$f(\varphi/\lambda) = \sum_{m=1}^{M} \omega^{(m)} N\left(\varphi/\mu^{(m)}, \sigma(m)\right) \tag{5}$$

where $\varphi = [Q, \hat{\psi}]$, m is the mixture component index, ω^m are the mixture weights, and $N(\varphi/\mu^{(m)})$, $\sigma(m)$ are the multivariate Gaussian densities, with $\mu^{(m)}$ being the mean vectors and $\sigma(m)$ the covariance matrices of the Gaussian densities. We experimented with 8, 12, and 16 Gaussian mixture components with diagonal and full covariance matrices and observed that it is sufficient to use 12 full covariance matrices for 14-dimensional feature vectors. The GMM is completely specified by a set of M mean vectors, covariance matrices, and mixture weights, which are estimated from a large training dataset having speech utterances and their subjective MOS score using the expectation maximization (EM) algorithm.

In the following, we describe the speech databases used for training and testing of the different feature vector based non-intrusive speech quality evaluation algorithms.

4.1 ITU-T P.Supplement-23 database

Supplement-23 to the P series of ITU-T Recommendations (1998) is a database of coded and source speech material used in the ITU-T 8 kbps codec (Recommendation G.729) characterization tests. For the purpose of the development of new and revised ITU-T Recommendations, relating to objective speech quality measures, this database provides source, pre-processed, and processed speech material with related subjective test plans and MOS scores. The database consists of absolute category rating (ACR) and degradation category rating (DCR) labeled database. Experiments-1 and experiment-3 are ACR-labeled whereas experiment-2 is DCR-labeled. Experiment-2 has not been used in this work. All speech utterances of the database are recorded in 16-bit linear raw PCM with a low-byte first format at a sampling rate of 22 kHz. Three versions of each utterance are available, namely original (clean speech), pre-processed (clean speech with some pre-processing such as sampling and quantization) and coded (speech subjected to various classes of coder/channel degradations). In this work, the coded version of speech utterances down sampled at 8 kHz sampling rate was used for which the ACR labeled MOS were available. Experiment-1 is divided into three sub-experiments (A, D and O), each having 176 speech utterances, excluding the trial/ test speech utterances, at 44 different conditions of transcoding (speech subjected to various classes of coder/channel degradations) and different MNRU levels. Experiment-3 is divided into four sub-experiments (A, C, D, and O), each having 200 speech utterances, excluding the trial/ test speech utterances, at 50 different conditions of random/bursty frame erasures type noise at 20 dB SNR level which include vehicle, street, and hoth noise. Sub-experiments A, C, D, and O correspond to the various laboratories where the recording and subjective listening quality evaluations of speech utterances were done. Thus, the total number of speech utterances in the two experiments of the database is 1328 at 8 kHz with ACR labeled subjective MOS, excluding the trial/ test speech utterances.

4.2 Subjective listening test on NOIZEUS-2240 database

The noisy speech corpus of 2240 speech utterances at 8 kHz from University of Texas, Dallas, USA, was also used in this work. The database contains 20 clean speech utterances each passed through 4 different types of noise namely babble, car, street and train at 2 SNR levels of 5 dB and 10 dB each. Resulting each speech utterance is passed through 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). Thus, there are 2240 speech utterances at 112 different conditions.

This speech database was used for conducting listening tests in our laboratory with 16 listeners and an average of their opinion score was computed to obtain the ACR labeled subjective MOS score to make it suitable for the training of GMM in the speech quality evaluation. These 2240 speech utterances are randomly divided into three groups of 840 speech utterances, 800 speech utterances, and 600 speech utterances for conducting the listening test to be completed in three sessions by each listener (subject). Each listener took three sessions to complete the listening and scoring of these three groups of speech utterances. Thus, this database contains 2240 ACR labeled subjective MOS annotated speech utterances with 112 different conditions.

4.3 Subjective listening test on NOIZEUS-960 database[2]

The noisy speech corpus of 960 speech utterances were obtained from NOIZEUS database, which are different from the 2240 speech utterances of the NOIZEUS-2240 database described previously. These speech utterances were used for the subjective listening test, which was conducted in our laboratory for the evaluation of subjective MOS score. The database contains 30 clean speech utterances each passed through 8 different types of noise namely airport, babble, car, exhibition, restaurant, station, street and suburban train at 4 SNR levels (0 dB, 5 dB, 10 dB and 15 dB) each. Noise signals for degradations of the speech utterances of the database were taken from the AURORA database. Thus, 960 speech utterances are available for 32 different degradation conditions of noise. This database was used to conduct subjective listening tests with 16 listeners and an average of their opinion score was computed to obtain the ACR labeled subjective MOS score for each speech utterance.

In the following, we consider the methodology for MOS estimation using objective algorithms. For training and testing the GMM's the following set of features are used in experiments:

(i) Lyon's auditory model feature vector of dimension 14,
(ii) MFCC feature vector of dimension 13,
(iii) LSF feature vector of dimension 10,
(iv) Lyon's auditory model feature vector of dimension 14, and MFCC feature vector of dimension 13 is appended for each speech utterance to give a 27-dimensional feature vector.
(v) Lyon's auditory model feature vector of dimension 14, and LSF feature vector of dimension 10 is appended for each speech utterance to give a 24-dimensional feature vector.

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

(vi) Lyon's auditory model feature vector of dimension 14, MFCC feature vector of dimension 13, and LSF feature vector of dimension 10, is appended for each speech utterance to give a 37-dimensional feature vector.

The subjective MOS score for each speech utterance is appended to each feature vector set increasing the dimensionality by one. The database of 1328 utterances of ITU-T Supplement-23 databases is first randomized and then leave one out procedure is used for GMM training and testing. Firstly, the database of 1328 utterances of ITU-T Supplement-23 databases is divided into 10 subsets, of which 9 subsets are used for training of the GMM while remaining one subset is used for testing. This procedure is repeated wherein all the 10 subsets are considered for testing one-by-one while the remaining 9 subsets in each case are used for training, thus generating an objective MOS score list for all 1328 speech utterances. The same approach has been followed for NOIZEUS-2240 database and NOIZEUS-960 databases for objective MOS evaluation.

5 Results and performance evaluation

The results have been obtained in terms of the correlation coefficients between the estimated MOS and the subjective MOS scores, considering two cases viz. condition averaged MOS and non-condition averaged MOS. In condition average MOS, the average of all the MOS scores are calculated over the same "condition" of the speech utterances irrespective of the sentence of the speech utterance. Here, "condition" refers to speech utterances at the same degradation or passed from the same speech processing algorithms. In non-condition averaged (or unconditioned) MOS, the estimated objective MOS score and the listeners' subjective MOS for individual sentence utterances are used to calculate the correlation coefficient. In ITU-T supplement-23 database, out of 1328 speech utterances, four different speech utterances are at the same condition thus making a total of 332 different conditions. The correlation coefficients are calculated using 332 condition averaged values of subjective MOS and 332 conditioned averaged values of the estimated objective MOS. The non-condition averaged MOS is obtained individually from all the 1328 speech utterances from which the objective estimated MOS and the listener's subjective MOS to calculate the correlation coefficient. Similarly, in NOIZEUS-2240 database, there are 2240 speech utterances at 112 different conditions, as 20 clean speech utterances are subjected to the same conditions. Thus, the correlation coefficients are calculated using 112 conditioned average values of subjective MOSs and 112 conditioned average values of objective MOS. In the non-condition averaged MOS, individual objective MOS of all 2240 speech utterances and the listener's averaged subjective MOS for individual speech utterance are used to calculate the correlation coefficient. In NOIZEUS-960, there are 960 speech utterances at 32 different conditions as 30 clean speech utterances are subjected to the same conditions. Thus, the correlation coefficients are calculated using 32 condition averaged values of subjective MOS and 32 condition averaged values of objective MOS. In the non-condition averaged MOS, individual objective MOS of all 960 speech utterances and the listener's averaged subjective MOS for individual utterances are used to calculate the correlation coefficient.

The performance of the different feature vectors is assessed using Karl-Pearson's correlation coefficient R between condition average of the predicted objective speech quality MOS score $\hat{Q}$ after 3rd order monotonic polynomial fitting and the subjective speech quality MOS score Q. Here, R is given by,

$$R = \frac{\sum_{i=1}^{N}(\hat{Q}_i - \mu_{\hat{Q}}).(Q_i - \mu_Q)}{\sqrt{\sum_{i=1}^{N}(\hat{Q}_i - \mu_{\hat{Q}})^2 . \sum_i (Q_i - \mu_Q)^2}} \tag{6}$$

where μ_Q and $\mu_{\hat{Q}}$ are the mean values of the introduced variables and the summation is over the conditions N.

Karl-Pearson's correlation coefficient R is also obtained for the predicted objective speech quality MOS score $\hat{Q}$ after 3rd order monotonic polynomial fitting and the subjective speech quality MOS score Q without taking the conditioned average. Most of the previous work has compared the results for conditioned averages of MOS only. However, in practice the MOS score is estimated sentence by sentence rather than as a condition average, and must therefore be the natural basis for correlation. It is observed that the condition averaged MOS when correlated with the condition averaged objective MOS estimate invariably produces higher correlation coefficient value compared than the non-condition averaged case.

The performance evaluations have been done for the different sets of features. The following observations can be made:

(i) It is observed that out of 36 dimensional feature vectors of PLP and 10 dimensional feature vectors of LSF, when used independently for speech quality evaluation, the LSF feature vectors perform slightly better than the PLP vector as given in Table 1.
(ii) The correlation coefficient for condition averaged MOS is better than the unconditioned case, as can be seen from Table 1 and comparing Tables 2 and 3. However, the unconditioned case gives a more definitive evaluation of the quality which is required on a sentence-by-sentence basis as in any practical use of the non-intrusive speech quality evaluation algorithm.
(iii) When considering individual feature vector types, Lyon's auditory model feature vector gives the best performance if condition averaged MOS is considered for

Table 1 Comparison of correlation coefficient with condition average of MOS and unconditioned MOS on Supplement-23 database using LSF and PLP as feature vectors

Data of different experiments	No. of speech utterances	Correlation coefficient with condition average of MOS on Supp.23 database		Correlation coefficient without condition average of MOS on Supp.23 database	
		LSF10	PLP36	LSF10	PLP36
Exp.1(A)-French	176	0.919	0.914	0.826	0.738
Exp.1(D)-Japanese	176	0.821	0.846	0.747	0.710
Exp.1(O)-A. English	176	0.917	0.871	0.791	0.625
Exp.3(A)-French	200	0.588	0.673	0.498	0.503
Exp.3(C)-Italian	200	0.738	0.698	0.637	0.558
Exp.3(D)-Japanese	200	0.783	0.802	0.682	0.680
Exp.3(O)-A. English	200	0.767	0.663	0.612	0.472
Average		**0.790**	**0.781**	**0.685**	**0.612**
Weighted average		**0.785**	**0.776**	**0.679**	**0.608**

Table 2 Correlation coefficients between subjective MOS (conditioned) and estimated MOS (conditioned) derived with different sets of feature vectors. Feature vector sets are: (1) 13-dimensional MFCC feature vector, (2) 14-dimensional feature vector of Lyon's model, (3) 10-dimensional LSF feature vector, (4) Sets 2 and 1, (5) Sets 2 and 3, (6) Sets 2, 1 and 3, (7) Set 6, the magnitude of 1st differences of 13-dimensional MFCC feature vector, and the magnitude of 1st difference of 10-dimensional LSF feature vector. (8) Set 7, the magnitude of 2nd difference of 13-dimensional MFCC feature vector, and the magnitude of 2nd difference of 10-dimensional LSF feature vector

Data of different expts.	No. of speech files	P.563 rec.	Feature vector set1	Feature vector set2	Feature vector set3	Feature vector set4	Feature vector set5	Feature vector set6	Feature vector set7	Feature vector set8
8 kbps ITU & ETSI standard CODECS interworking										
Exp.1(A) -French	176	0.885	0.885	0.849	0.913	0.896	0.887	0.912	0.937	0.933
Exp.1(D) -Japanese	176	0.842	0.869	0.885	0.827	0.930	0.918	0.933	0.936	0.940
Exp.1(O) -A. Eng.	176	0.902	0.911	0.910	0.901	0.901	0.937	0.946	0.938	0.955
Channel errors and background noise										
Exp.3(A) -French	200	0.867	0.754	0.783	0.592	0.907	0.820	0.887	0.900	0.900
Exp.3(C) -Italian	200	0.854	0.817	0.814	0.741	0.848	0.841	0.851	0.872	0.894
Exp.3(D) -Japanese	200	0.929	0.855	0.849	0.768	0.908	0.883	0.908	0.922	0.917
Exp.3(O) -A. Eng.	200	0.918	0.711	0.756	0.764	0.869	0.852	0.891	0.909	0.891
Degradations by different types of noise (Babble, Car, Exhibition hall, Restaurant, Street, Airport, and Train station) at different SNR (0 dB, 5 dB, 10 dB, and 15 dB)										
NOIZEUS -960	960	0.951	0.991	0.987	0.989	0.992	0.987	0.993	0.992	0.991
Different suppression schemes (MMSE-STSTA, Spectral subtraction, Subspace approach and Weiner filtering) at different SNR (5 dB and 10 dB)										
NOIZEUS -2240	2240	0.955	0.987	0.986	0.985	0.989	0.988	0.990	0.989	0.988
Average		**0.900**	**0.849**	**0.869**	**0.831**	**0.915**	**0.901**	**0.923**	**0.933**	**0.934**
95 % Conf. Interval		**0.027**	**0.062**	**0.053**	**0.085**	**0.032**	**0.040**	**0.031**	**0.025**	**0.025**

Table 3 Correlation coefficients between subjective MOS (unconditioned) and estimated MOS (unconditioned) derived with same sets of features vectors as in Table 2

Data of different expts.	No. of speech files	P.563 Rec.	Feature vector set1	Feature vector set2	Feature vector set3	Feature vector set4	Feature vector set5	Feature vector set6	Feature vector set7	Feature vector set8
8 kbps ITU & ETSI standard CODECS interworking										
Exp.1(A) -French	176	0.759	0.780	0.652	0.815	0.807	0.775	0.834	0.837	0.825
Exp.1(D) -Japanese	176	0.701	0.748	0.739	0.755	0.810	0.807	0.828	0.843	0.851
Exp.1(O) -A. Eng.	176	0.790	0.769	0.717	0.772	0.787	0.785	0.828	0.835	0.847
Channel errors and background noise										
Exp.3(A) -French	200	0.768	0.622	0.550	0.510	0.787	0.681	0.773	0.786	0.775
Exp.3(C) -Italian	200	0.762	0.665	0.692	0.639	0.742	0.747	0.753	0.781	0.802
Exp.3(D) -Japanese	200	0.801	0.730	0.700	0.680	0.803	0.759	0.806	0.822	0.810
Exp.3(O) -A. Eng.	200	0.788	0.564	0.562	0.602	0.711	0.710	0.745	0.774	0.764
Degradations by different types of noise (Babble, Car, Exhibition hall, Restaurant, Street, Airport, and Train station) at different SNR (0 dB, 5 dB, 10 dB, and 15 dB)										
NOIZEUS -960	960	0.717	0.869	0.778	0.698	0.852	0.807	0.859	0.858	0.854
Different suppression schemes (MMSE-STSTA, Spectral subtraction, Subspace approach and Weiner filtering) at different SNR (5 dB and 10 dB)										
NOIZEUS -2240	2240	0.306	0.676	0.674	0.654	0.684	0.691	0.700	0.732	0.733
Average		**0.710**	**0.714**	**0.674**	**0.681**	**0.776**	**0.751**	**0.792**	**0.807**	**0.807**
95 % Conf. Interval		**0.101**	**0.060**	**0.050**	**0.061**	**0.035**	**0.031**	**0.034**	**0.027**	**0.027**

obtaining the correlation coefficient (Table 2) whereas for unconditioned MOS the MFCC feature vector set gives the best performance (Table 3).

(iv) By combining two and more feature vectors, the correlation coefficient, in both the cases, of unconditioned and conditioned average increases significantly, as can be seen from Tables 2 and 3. Scattered plot of condition averaged subjective MOS versus condition averaged estimated MOS by proposed Combined Feature Sets Model is shown in Fig. 7.

(v) When combined feature vectors are used, Lyon's auditory model feature vector, MFCC, LSF, magnitude of first difference of MFCC and magnitude of first difference of LSF together form the best combination feature vector set, whether conditioned or unconditioned MOS are used for evaluating the correlation coefficient. The size of the combined feature vectors is 60. The combined features set gives a correlation coefficient of 0.934 compared to 0.900 of the ITU-T P.563 Recommendation when condition averaged MOS is used. Similarly, the combined features set gives a correlation coefficient of 0.807 compared to 0.710 of the ITU-T P.563 Recommendation when unconditioned MOS is used. The results are also compared with 95 % confidence interval and indicated against each result in Table 2 and Table 3. This combined feature vector set outperforms as compared to the P.563 Recommendation because, it utilizes the features in perceptual domain, spectral domain, and in time domain.

(vi) When the magnitude of second difference of MFCC are also included in the feature vector set, it does not lead to a further improvement of the correlation coefficient and is therefore not proposed for inclusion in the combined feature vectors set.

(vii) Absolute errors in objective MOS estimation for different databases are given in Table 4 and diagrammatically shown in Fig. 8.

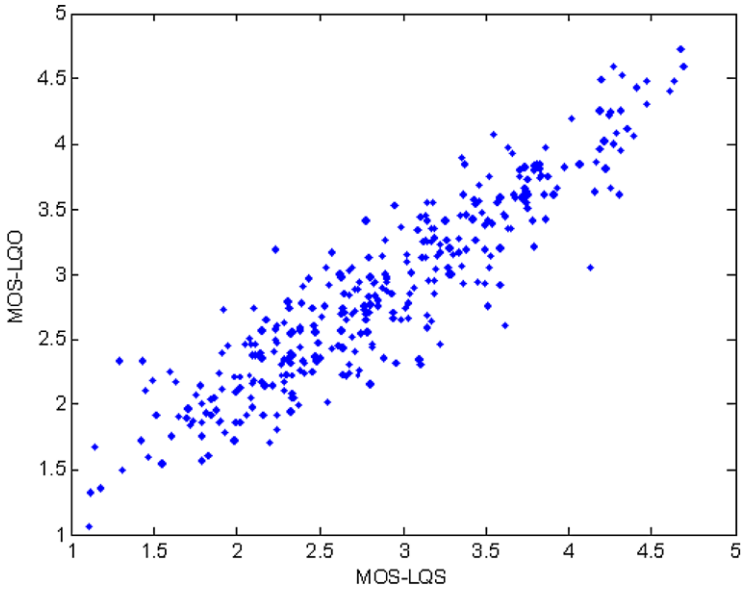

Fig. 7 Scattered plot of condition averaged subjective MOS versus condition averaged estimated MOS by proposed Combined Feature Sets Model

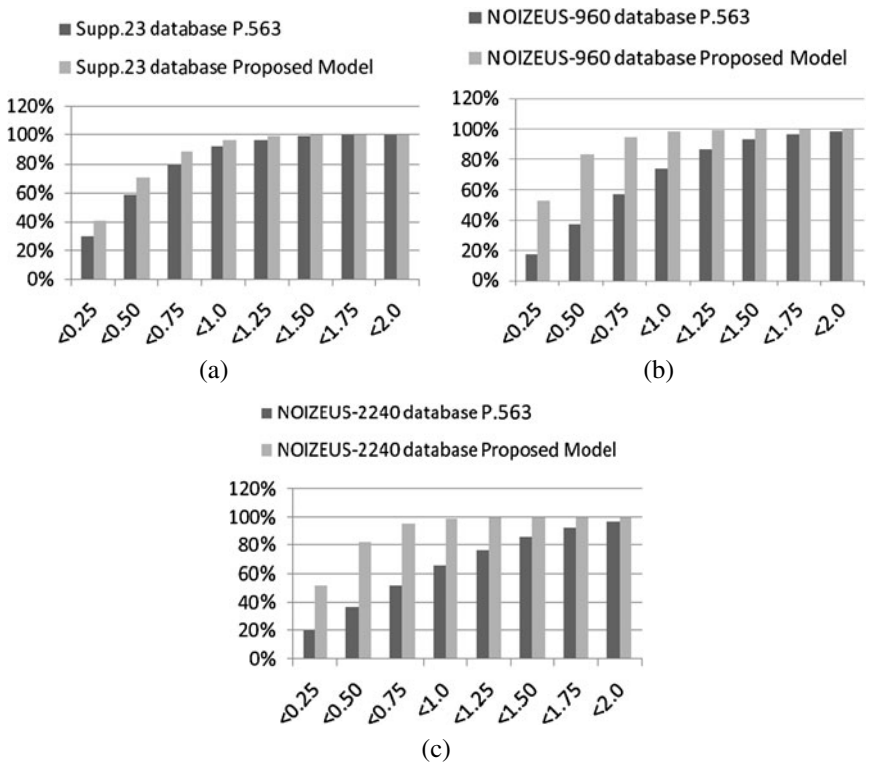

Fig. 8 Distribution of absolute errors for (**a**) Supp.23 database, (**b**) NOIZEUS-960 database and, (**c**) NOIZEUS-2240 database

Table 4 Distribution of absolute errors for supplement-23 database, NOIZEUS-960 database, and NOIZEUS-2240 database

Different database	Absolute error	<0.25	<0.50	<0.75	<1.0	<1.25	<1.50	<1.75	<2.0
Supp.23 database	P.563	30.57 %	59.26 %	80 %	92.30 %	97 %	99.70 %	100 %	100 %
	Proposed Model	40.80 %	70.80 %	88.90 %	96.70 %	99.40 %	99.90 %	100 %	100 %
NOIZEUS-960 database	P.563	17.18 %	37.40 %	56.97 %	74.17 %	86.56 %	93.44 %	96.67 %	98.95 %
	Proposed Model	53.12 %	83.22 %	94.80 %	98.95 %	99.90 %	100 %	100 %	100 %
NOIZEUS-2240 database	P.563	20.00 %	37.14 %	52.67 %	65.40 %	76.42 %	86.11 %	92.18 %	96.78 %
	Proposed Model	51.90 %	82.27 %	95.84 %	99.24 %	99.82 %	99.95 %	100 %	100 %

6 Conclusion

A method of combining the different relevant auditory features set and GMM training is used for non-intrusive speech quality evaluation. Different auditory feature sets are obtained, combined and along with their subjective MOS score, used for GMM training. The parameters of GMM thus obtained are used for the evaluation of objective MOS for speech utterance as a conditional probability. Simulations show that certain combination of feature sets perform better than the ITU-T P.563 Recommendation.

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. The authors would also like to thank Prof. S.C. Saxena, Vice Chancellor (Acting), Jaypee Institute of Information Technology, Noida, India for providing the suitable environment to the corresponding author to complete the work.

References

ITU-T Recommendation P.800 (1996). *Methods for subjective determination of transmission quality*. International Telecommunication Union-Telecommunication Standardization Sector, Geneva.

Fegyo, T., Szarvas, M., Tatai, P., & Gordos, G. (2000). Objective speech quality estimation for analog mobile channels: problems and solutions. *International Journal of Speech Technology*, *3*, 277–287.

ITU-T Recommendation P.862 (1996). *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks*. International Telecommunication Union-Telecommunication Standardization Sector, Geneva.

ITU-T Recommendation P.861 (1996). *Objective quality measurement of telephone band (300–3400 Hz) speech codecs*. International Telecommunication Union-Telecommunication Standardization Sector, Geneva.

ITU-T Recommendation P.563 (2004). *Single ended method for objective speech quality assessment in narrow-band telephony applications*. International Telecommunication Union- Telecommunication Standardization Sector, Geneva.

Liang, J., & Kubichek, R. (1994). Output based objective speech quality. In *IEEE 44th vehicular technology conf.* (Vol. 3(8–10), pp. 1719–1723).

Au, O. C., & Lam, K. (1998). A novel output based objective speech quality measure for wireless communication. In *Proc. 4th international conf. on signal processing* (Vol. 1, pp. 666–669).

Gray, P., Hollier, M., & Massara, R. (2000). Non-intrusive speech quality assessment using vocal-tract models. *IEE Proceedings. Vision, Image and Signal Processing*, *147*(6), 493–501.

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

Kim, D. S. (2005). ANIQUE: an auditory model for single ended speech quality estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, *13*(5), 821–831.

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

Falk, T., Xu, Q., & Chan, W. Y. (2005). Non-intrusive GMM based speech quality measurement. In *Proc. IEEE international conf. on acoustics, speech and signal processing* (Vol. 1, pp. 125–128).

Chen, G., & Parsa, V. (2006). Bayesian model based non-intrusive speech quality evaluation. In *Proc. IEEE international conf. on acoustics, speech and signal processing* (Vol. 1, pp. 385–388).

Falk, T., & Chan, W. Y. (2006a). Enhanced non-intrusive speech quality measurement using degradation models. In *Proc. IEEE international conf. on acoustics, speech and signal processing* (Vol. 1, pp. 837–840).

Falk, T., & Chan, W. Y. (2006b). Single-ended speech quality measurement using machine learning methods. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 1935–1947.

Chen, G., & Parsa, V. (2005). Non-intrusive speech quality evaluation using an adaptive neuro-fuzzy inference system. *IEEE Signal Processing Letters*, *12*(5), 403–406.

Slaney, M. (1988). *Lyon's cochlear model*. Advanced technology group, apple technical report no. 13, Apple Computer Inc.

Lyon, R. F. (1982). A computational model of filtering, detection, and compression in the cochlea. In *Proc. IEEE international conf. on acoustics, speech and signal processing* (pp. 1282–1285).

Jing, Z., & Johnson, M. H. (2002). Auditory modeling inspired methods of feature extraction for robust automatic speech recognition. In *Proc. IEEE international conf. on acoustics, speech and signal processing* (Vol. 4, pp. 4176–4179).

ITU-T Recommendation P. Supplement-23 (1998). *ITU-T Coded-Speech database (1998)*. International telecommunication union-telecommunication standardization sector, Geneva.

Hu, Y., & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 229–238.

Hu, Y., & Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communications*, *49*, 588–601.

Narwaria, M., Lin, W., McLoughlin, I. V., Emmanuel, S., & Tien, C. L. (2010). Non-intrusive speech quality assessment with support vector regression. In *Advances in multimedia modeling*, 16th International multimedia modeling conf. (Vol. 5916, pp. 325–335). Berlin: Springer.

Hasan, M. R., Jamil, M., Rabbani, M. G., & Rahman, M. S. (2004). Speaker identification using mel frequency cepstral coefficients. In *3rd international conf. on electrical & computer engineering* (pp. 565–568). Dhaka: ICECE.

Bozkurt, E., Erzin, E., Erdem, C. E., & Erdem, A. T. (2010). Use of line spectral frequencies for emotion recognition from speech. In *IEEE international conf. on pattern recognition*, Turkey (pp. 3708–3711).

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, *87*, 1738–1752.

Audhkhasi, K., & Kumar, A. (2010). Two scale auditory features based non-intrusive speech quality evaluation. *IETE Journal of Research*, *56*(2), 111–118.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, *39*(1), 1–38.

Karmakar, A., Kumar, A., & Patney, R. K. (2006). A multiresolution model of auditory excitation pattern and its application to objective evaluation of perceived speech quality. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 1912–1923.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals*. Englewood: Prentice-Hall.