Statistical Measurement Approach for On-line Audio Quality Assessment

Lopamudra Roychoudhuri, Ehab Al-Shaer and Raffaella Settimi

School of Computer Science, Telecommunications and Information Systems,
Depaul University, 243 S. Wabash Ave., Chicago, IL-60604, U.S.A

Abstract. The quality of audio in IP telephony is significantly influenced by the impact of packet loss rate, burstiness and distribution on audio compression techniques. On-line audio quality assessment is important to provide real-time feedback to end-to-end Internet audio transport protocols to increase the reliability and quality of the audio session. In this paper we propose a novel passive statistical measurement framework, Audio Genome, that can deduce the audio quality of an on-going Internet audio for many different codecs under any network loss condition at real-time. We use multiple polynomial regression as the modelling technique to accurately characterize the audio quality for codecs under any loss scenario. Our approach is easy to deploy and guarantees high computational speed. Our extensive evaluation experiments, that include large simulation scenarios, show that our framework is accurate and viable for adaptive real-time audio mechanisms.

1 Introduction

A reliable online speech quality assessment framework can be highly beneficial for quality monitoring of an ongoing VoIP communication, and can be part of many applications, such as multi-codec audio mechanisms, proactive error control for audio, QoS provisioning for multimedia, SLA monitoring, to name a few. Audio codecs have a diverse range of compression degrees and underlying technologies. The main factors that significantly influence the audio quality in IP telephony thus include codec type, loss rate, loss burst, inter-loss gap, delay, and recency [1] [3] [5]. However, it is a real challenge to establish a framework that derives audio quality on-line considering all of these factors.

Audio quality of any speech processing system is generally described in terms of MOS (Mean Opinion Score) [10], the formal subjective measure of received speech quality, which is a real number between 1 and 5, where 1 is bad and 5 is excellent. ITU-specified E-model [7][8][11] provides a computational model to derive relative impairments to voice quality and to estimate subjective MOS. But ITU provides no analytic methods that can directly measure the impairment due to *random loss* conditions of bursts and inter-loss gaps.

The objective of our Audio Genome is to provide a statistical framework that first quantifies the effects of packet loss on various codecs by considering loss bursts, inter-loss gaps and various loss rates. The validity of our measurement techniques is twofold. We establish the audio quality as a set of functions that are derived from sufficient data generated from a large set of simulation experiments considering various codecs and a wide range of loss scenarios. Secondly, we use multiple polynomial regression as the modelling technique to accurately characterize the curves representing the audio quality for codecs under any loss scenario. The resulting repository of quality information can be used real-time to effectively assess the expected audio quality of an ongoing communication.

In our approach we use a wide range of fixed inter-loss gaps and loss burst lengths causing increasing degrees of packet loss for a large number of short audio clips with a set of codecs. The resulting MOS, measured using PESQ, is curve-fitted to show the relation of MOS with inter-loss gap and burst size. This repository is then used to derive the audio quality of the ongoing transmission by (i) estimating MOS for a codec from a passive measurement of inter-loss gap and loss burst data using regression functions, and (ii) combining the MOS data for observed loss bursts and inter-loss gaps into a partial MOS using aggregation schemes for the session so far.

Although many researchers have attempted to establish audio quality prediction models based on packet loss [3][4][13], they do not provide frameworks as comprehensive and complete as Audio Genome. Other researchers have used neural networks [14] that require extensive training. Objective testing schemes, such as PESQ (Perceptual Evaluation of Speech Quality [12]), are automated and repeatable speech testing schemes, but they can only be used off-line. Compared to these methods, Audio Genome, being an on-line statistical approach, guarantees speed, accuracy and less overhead in terms of computation and data storage. The framework is flexible, as new codecs can be added to the system fairly easily. Our work can be directly applied to adaptive multimedia control mechanisms, and is also designed simple enough for easy deployment in handheld devices.

Subsequent sections are organized as follows. Section 2 contains the related work. In section 3 we present the Audio Genome Approach. We describe the evaluation and experiment results in section 4, and conclusion and future work in section 5.

2 Related Work

ITU-specified E-model provides a computational model to derive relative impairments to voice quality and to estimate subjective MOS. ITU provides the equipment impairment measure I_e for many codecs under no loss condition [11] and a limited number of codecs under very limited loss condition scenarios [8]. ITU framework does not directly consider random loss conditions of bursts and inter-loss gaps in measuring the impairment I_e. However, at this point not enough subjective measurements and their specifics are available expressing I_e values for various codecs in fully analytic form as a function of packet loss and burstiness.

Many authors have presented extensions to E-model. Cole and Rosenblath [4] described a method for monitoring VoIP applications based upon E-model, where they used curve fitting of ITU-published I_e values for selected codecs for various loss percentages. However, since they pointed out that ITU does not show a complete description of algorithms to generate loss data, they were unable to provide a complete framework of codec quality assessment, as in the case of our Audio Genome approach. In [13], the authors addressed the problem of predicting the quality of telephone speech and classified quality prediction models based on E-Model. But they did not provide a comprehensive measurement study of impairments due to packet loss. VQMon [3] is a non-intrusive passive monitoring system for VoIP using a Markov model incorporating packet loss and recency effect. However, VQMon uses a limited "burstiness" model that, for example, does not distinguish between "burst" situation when 3 packets are lost consecutively vs. 3 packets are lost with a gap of 1 or more packets in between each loss pair, these two scenarios producing completely different quality results.

A different approach has been taken by training a neural network with MOS for a set of codecs under various loss rates and distribution [14], that is less attractive to us because of complexity in training and computational delay. Audio Genome attempts to bridge this gap by providing a comprehensive measurement framework for on-line audio assessment, that is easy for practical deployment and can be extended to many codecs. Watson and Sasse [17] have conducted extensive subjective evaluation of audio quality under packet loss compensation in multimedia conference systems. We use PESQ instead, since subjective testing is time-consuming, cumbersome, error-prone and non-repeatable. In our earlier work [15] we used interpolation to determine the codec quality functions, which was deterministic and simple in computation. In this paper we present a regression model, a preferred statistical method that provides a measure of the accuracy of the model predictions and their deviation from the actual data.

3 The Audio Genome Approach

The Audio Genome approach can be described as follows.

o *Generation of audio clips with packet loss scenarios:* We drop packets from audio clips using a periodic drop framework for a set of chosen codecs. It is worth noting that the framework is not limited to these codecs, as other codecs can easily be added by following these steps.

o *MOS evaluation and observations:* Using PESQ, we compare 'pure' and 'poisoned' audio clips to deduce MOS scores under loss. We also observe the characteristics and behavior of each codec under various packet loss conditions.

o *Codec Quality Function Derivation:* We deduce codec quality functions for the collected data under loss using multiple polynomial regression analysis techniques.

o *Online prediction of audio quality:* We use regression functions determined in the previous step to deduce the MOS for the ongoing transmission. We use

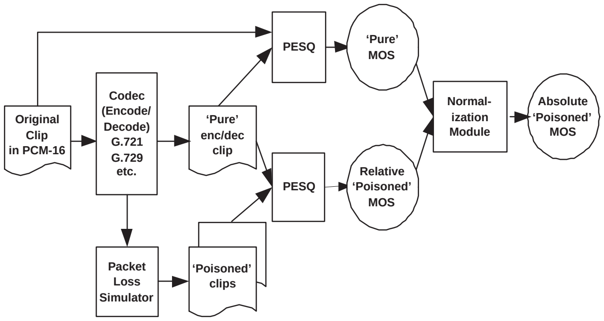

Fig. 1. Packet Loss Simulation Framework

weighted aggregation schemes that calculate MOS values for observed inter-loss gaps and burst sizes, and produce a combined MOS for the session so far.

3.1 Audio Clip Generation with Loss Scenarios

Experiment Codec Set We choose G.711, as the standard audio compression technique or ’codec’, a waveform PCM-16 (16 bit Pulse Code Modulation) coder of bitrate 128kbs with the best quality. In addition, we choose G.721, G.729, G.723.1 and GSM FR 6.10 as representative codecs of varied bitrates and underlying technology that use complex coding methods, such as Analysis By Synthesis (ABS) and Codebook Excited Linear Prediction (CELP). Apart from these codecs, G.722.2 (AMR-WB) is a fairly new Adaptive Multi-Rate Wideband codec with multiple bitrates [9]. Not much testing results on audio quality for this codec are available from any source. We choose 6 (modes 0, 1, 3, 5, 6 and 8, bitrates 6.6, 8.85, 14.25, 18.25, 19.85 and 23.85 respectively) out of 9 bitrates of G.722.2 to evaluate how the different bitrates of the codec behave under degrees of packet loss in relation to each other.

Experiment Methodology Fig. 1 depicts the packet loss simulation framework. Each original PCM-16 audio clip is encoded and decoded with every codec to create a ’pure’ image with no loss, and is compared with the original clip to deduce the ’Pure’ MOS. To create the ’Poisoned’ clips, we drop packets during encoding with each combination of gap and burst, and decode back to PCM-16. In the Packet Loss Simulator we use a wide combination of fixed inter-loss gaps from 300 down to 2 packets and a set of fixed loss burst lengths of 1, 2, 3 and 4 packets (the most occurring burst sizes as observed in the Internet [1][2]). We achieve a wide range of loss rates from 0.3% to 66.7% for the sake of completeness, though loss rates greater than 30% are too high for any meaningful result. We choose fixed inter-loss gaps and burst sizes in order to measure the effect of packet loss on each codec in a precise and controlled manner. We compare the ’Pure’ and the ’Poisoned’ images using PESQ to deduce the ’Poisoned’ MOS.

Since the 'Pure' image is an encoded-decoded clip, the comparison produces a quality score relative to the codec score under no loss. We refer to quality as the measured relative 'Poisoned' MOS for the rest of the paper. The absolute quality is obtained by normalizing the relative quality with the ratio of measured codec 'Pure' MOS and 4.5, the PCM MOS under no loss, as a scaling factor in the Normalization process.

We choose a total of 24 short PCM-16 audio clips, created by sequentially truncating 4 larger clips, about 1 minute each, spoken by 2 males and 2 females, into 6 short segments of 9-12 seconds, as prescribed by PESQ [12]. The periodic drop experiment is run using every clip for each codec, and the MOS for each gap-burst combination is taken as the average of the MOS for all clips under the same degree and distribution of loss. For each codec, the Data Collection step produces four tables of data, one for each burst size, containing the MOS scores for all possible gaps of 2 to 300 packets.

3.2 Codec Quality Function Derivation

We observe that the quality patterns vary from codec to codec under similar loss distribution scenarios. The codecs exhibit inherent differences in their quality degradation patterns [15]. This motivates us to derive a model of the codec behaviors that will capture these differences effectively. We use multiple polynomial regression modelling to determine the codec quality functions, as opposed to interpolation analysis in our earlier work [15]. The regression model is a preferred statistical method that provides a measure of the accuracy of the model predictions and their deviation from the actual data.

Multiple Regression Analysis The purpose of regression analysis is to derive a model that best represents the observed relationships between MOS, inter-loss gap and burst size. We found a logarithmic transformation of the gap length to be the most appropriate to represent the relationship between MOS and inter-loss gap. Fig. 2(a) shows the polynomial (almost linear) relationship between y (MOS) and $ln(1 + x)$, x being the gap length, for G.721 burst size 1. We thus defined the new predictor variable as $g(x) = \ln(1 + \alpha x)$, where $\alpha = 0.07$ for GSM, and $\alpha = 1$ for the others. The data were fitted using a polynomial model where the MOS score was the response variable y, and the regressors were the transformed variables $g(x)$. We noticed differences from codec to codec while modelling, some codecs (G.729, G.723.1 and GSM) requiring a quadratic model and the others requiring a cubic model.

In order to analyze the differences between the four burst sizes, we included three dummy variables D_1, D_2 and D_3 in the model. The statistical analysis was computed using the statistical package SAS. The regression models were selected using the backward selection algorithm. Standard statistical diagnostic methods, such as tests for normality and the residual analysis showed that the model assumptions of normality and constant variance were satisfied [6]. The following is the estimated regression equation for the data for G.729 after substituting

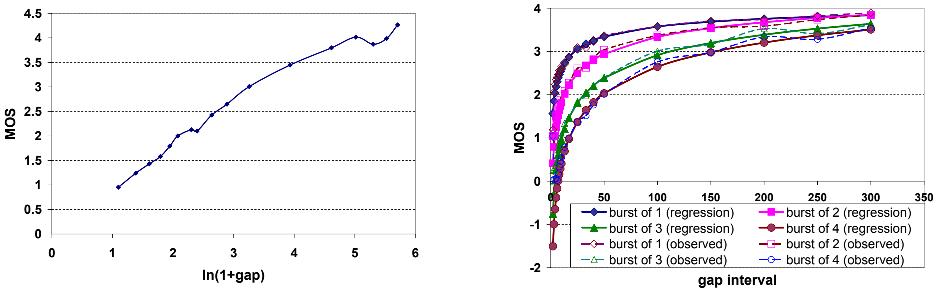

Fig. 2. (a) MOS vs. ln(1+gap): G.721, (b) Regression and Observed MOS: G.729

$(1 + x)$ by x'.

$$
\begin{aligned}
y = &-2.41603 + 1.35213 * ln(x') - 0.05501 * (ln(x'))^2 \\
&+ 3.44836 * D_1 - 0.54690 * ln(x') * D_1 \\
&+ 2.13615 * D_2 - 0.31430 * ln(x') * D_2 \\
&+ 0.83908 * D_3 - 0.12394 * ln(x') * D_3 \qquad (1)
\end{aligned}
$$

The regression equations for bursts of 1, 2 and 3 are derived by setting D_1, D_2, D_3 to 1 respectively, with the rest of D's equal to 0. The equation for burst of 4 is derived by setting all D's to 0. In Fig. 2(b) the regression equations match the observed MOS for G.729 well. Regression equations for the other codecs have been derived in the same manner [16].

3.3 Prediction of Audio Quality using Passive Measurement of Loss

The process of deducing a quality MOS using regression equations for an ongoing transmission can be described in the following steps:

(i) *Loss Pattern Identification*: First, we passively measure the loss for the session to keep track of the loss distribution and degree of burstiness observed so far, in terms of single, burst of 2, burst of 3, burst of ≥ 4, and inter-loss gaps preceding loss bursts, in a series of (gap,burst) pairs. This will be input to determine the MOS so far.

(ii) *Real-time deduction of MOS*: We use the observed loss pattern to derive the individual MOS values for gaps and bursts using regression functions, and combine them to deduce an aggregate MOS for the ongoing session so far. To find the MOS for each individual (gap,burst) pair, we simply fit the gap value as x in the codec regression equation for the particular burst size, and derive the MOS as a function of gap. We deduce the aggregate MOS by combining the individual MOS using weighted aggregation schemes described next.

Weighted Aggregation Schemes - In order to get the estimated MOS as close as possible to PESQ-observed value, we need to take into account the factors considered by PESQ in MOS deduction. PESQ dampens the effects of individual segment disturbances, after treating each occurrence of packet loss as a

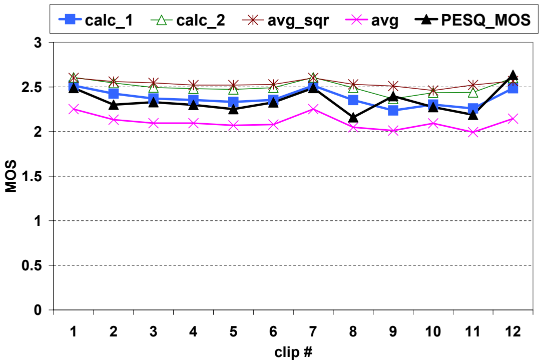

Fig. 3. Example comparing computed MOS and PESQ MOS : G.729

disturbance [12]. We mimic the dampening by accentuating the effects of larger inter-loss gaps, as they are analogous to lack of disturbance. We increase the accentuation in weighted average schemes, where we give larger weights to better individual MOS values due to bigger gaps, by using a factor of the gap value as the weight. In another weighted aggregation scheme, we use squares of individual MOS scores to accentuate the effect of better MOS values on the overall sum. The weighted average schemes *avg*, *calc_1*, *calc_2* and *avg_sqr* are computed as follows:

$$avg = (\sum_{i=1}^{P} MOS_i)/P \qquad calc\_1 = \frac{\sum_{i=1}^{P}(gap_i/10) * MOS_i}{\sum_{i=1}^{P}(gap_i/10)}$$

$$calc\_2 = \frac{\sum_{i=1}^{P}(gap_i/5) * MOS_i}{\sum_{i=1}^{P}(gap_i/5)} \qquad avg\_sqr = \frac{\sum_{i=1}^{P}(MOS_i)^2}{\sum_{i=1}^{P} MOS_i}$$

where P is the number of (gap,burst) pairs observed so far.

In order to evaluate the 4 weighted aggregation schemes and to derive a unified MOS deduction scheme for each codec, we conducted random packet drop experiments ranging in a wide degree of packet loss ratio (2% to 30%) and burst degree (single burst to burst of 4). In an example randomized test on G.729 for 12 clips with 9-12% loss per clip (Fig. 3), PESQ MOS matches *calc_1* better than the others. Since the goal is to assess the quality degradation during transmission, our aim is to deduce the aggregate scheme(s) for each codec that will predict MOS closest to the PESQ MOS under all loss circumstances. We define the MOS deviation as the deviation between the PESQ MOS and the predicted MOS to determine the measure of the prediction accuracy, as follows:

$$MOSdev = (\sum_{i=1}^{M}(|MOS\_PESQ_i - MOS\_pred_i|))/M \qquad (2)$$

where M is the set of clips in an experiment set. We calculate the error percentage as $MOSdev/4.5$, since 4.5 is the best score possible.

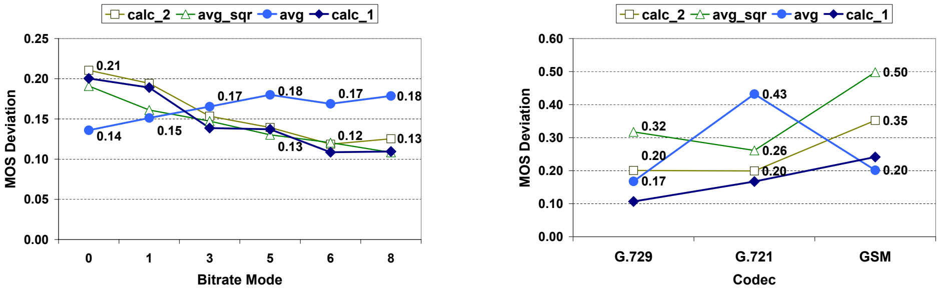

Fig. 4. Accuracy of Aggregate schemes (a) G.722.2 all modes, (b) G.729, G.721 & GSM

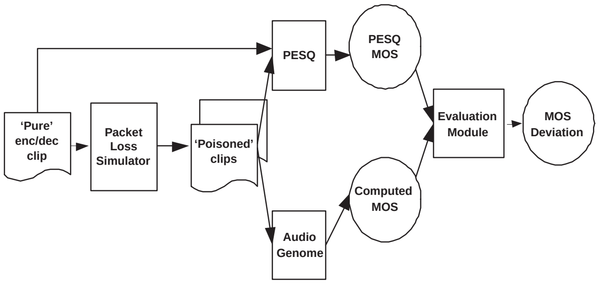

Fig. 5. Evaluation Experiment Framework

Fig. 4 depicts the similarities and the differences of the behavior of the aggregation schemes for different codecs. For most codecs *calc_1* stands out to be the best, except for GSM and G.722.2 modes 0 and 1, that have *avg* as the preferred scheme. A complete report on the behavior of aggregate schemes for all codecs under test can be found in [16]. Table 1 column 2 depicts the aggregate MOS deduction schemes of highest accuracy for each codec under test. We use this set of schemes to evaluate the Audio Genome framework.

4 Evaluation Results

The purpose of the evaluation experiments is to assess the accuracy of the Audio Genome framework under various ranges of packet loss scenarios. Fig. 5 depicts the experimental framework we use to evaluate the accuracy of Audio Genome. In the Packet Loss Simulator we conducted sets of random packet drop experiments ranging in a wide degree of packet loss ratio (2% to 40%) and burst degree and distribution most likely to be observed in the Internet [1][2]. More details of the loss simulation scenarios can be found in [16]. We appended small audio clips spoken by the same person in sequence to create incrementally larger 6 clips for each speaker and poisoned them randomly in order to consider the recency factor [5] of disturbance in PESQ scoring. For each of the 6 small clips as well

Table 1. Test Codec Set: Genome Accuracy

Codec	Overall Scheme	$MeanMOSdev$	Error %	Std Dev
G.729	$calc\_1$	0.11	2.4%	0.07
G.721	$calc\_1$	0.17	3.8%	0.09
GSM	avg	0.20	4.4%	0.11
G.722.2 0	avg	0.14	3.1%	0.12
G.722.2 1	avg	0.13	2.9%	0.12
G.722.2 3	$calc\_1$	0.13	2.9%	0.12
G.722.2 5	$calc\_1$	0.12	2.7%	0.10
G.722.2 6	$calc\_1$	0.11	2.4%	0.12
G.722.2 8	$calc\_1$	0.11	2.4%	0.09

as 6 extended clips for 4 speakers, 5 loss-degree groups, 3 testing schemes, i.e. a total of 720 clips per codec (a total of 6480 clips), we performed the following: (i) Extract the sequence of (gap, burst) pair data from each "poisoned" clip, (ii) Evaluate MOS_PESQ using PESQ, (iii) Deduce MOS_pred from Audio Genome, (iv) Evaluate the accuracy by computing the MOS deviation $MOSdev$ (eqn. 2) in Evaluation Module.

We calculated $MOSdev$ for all experiment sets and computed the accuracy of Genome under various loss degree and burstiness scenario. We observed that Audio Genome shows good accuracy for the appended clips in particular, which shows that it accommodates PESQ recency factor very well. We present the aggregate and an overall accuracy result of Genome for each codec in Table 1. A more complete description of the results can be found in [16]. Though the loss rate and degree is varied considerably over the experiments, the mean MOS deviation of Audio Genome for every codec is observed to be in a low range with low error percentage and standard deviation. Hence Audio Genome shows high accuracy under a wide range of loss scenarios.

5 Conclusion and Future Work

This paper presents a novel passive statistical measurement framework for real-time audio quality assessment, Audio Genome, that can deduce the audio quality of an on-going Internet audio for many different codecs under any network loss condition. We first provide an extensive experimental framework with 5 codecs G.721, G.729, G.723.1, GSM and 6 bitrate modes of G.722.2, where we quantify the effect of packet loss on the audio quality objectively by considering a wide range of loss bursts, inter-loss gaps and loss rates. For each codec, we model the relationship of audio quality with inter-loss gaps and loss burst sizes using multiple polynomial regression. For an ongoing communication, we estimate the partial MOS by aggregating the MOS using the inter-loss gaps and bursts seen in the session so far. We evaluate 4 aggregation schemes and derive a *unified* set of aggregation schemes for the codecs under test as the Audio Genome Model. We evaluate Audio Genome by conducting a set of extensive random loss experiments with loss degrees ranging from 2% to 40% and a wide range of packet burst distribution. For all codecs, under all loss scenarios, Audio Genome shows high accuracy of 96%-98% in average with low standard deviation of 0.07-0.12 and minimum accuracy of 91%. As future work, we would like to integrate this framework as a part of an adaptive multi-codec audio control mechanism that switches and mixes codecs according to changing bandwidth and delay conditions to maintain optimal quality.

References

1. Bolot,J., Vega-Garcia,A.: Control Mechanisms for Packet Audio in the Internet. IEEE Infocom, San Francisco (1996) 232-239
2. Borella,M.S.: Measurement and Interpretation of Internet Packet Loss. Journal of Communications and Networks (2000)
3. Clark, A.D.: "Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality." IPTel April (2001)
4. Cole,R.G., Rosenbluth,J.H.: Voice over IP Performance Monitoring. ACM SIGCOMM (2001)
5. Cox, R., Perkins, R.: Results of a Subjective Listening Test for G.711 with Frame Erasure Concealment. Committee contribution T1A1.7/99-016 (1999)
6. Draper, N.R., Smith, H.: Applied Regression Analysis, 2nd Ed., John Wiley & Sons (1981)
7. ITU-T Recommendation G.107 (12/98). "The E-Model, a computational model for use in transmission planning."
8. ITU-T Recommendation G.113 (02/96). "Transmission impairments."
9. ITU-T Recommendation G.722.2 (11/02). "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)."
10. ITU-T Recommendation P.800 (08/96). "Methods for subjective determination of transmission quality."
11. ITU-T Recommendation P.833 (02/01). "Methodology for derivation of equipment impairment factors from subjective listening-only tests."
12. ITU-T Recommendation P.862 (02/01). "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs."
13. Moeller, S., Raake, R.: Telephone speech quality prediction: towards network planning and monitoring models for modern network scenarios. Speech Communication, Vol. 38, Issue 1, pp. 47-75 (2002)
14. Mohamed,S., Cervantes-Perez,F., Afifi,H.: Integrating Network Measurements and Speech Quality Subjective Scores for Control Purposes. IEEE Infocom, Anchorage, Alaska, (2001)
15. Roychoudhuri,L., Al-Shaer,E.: Real-time Audio Quality Evaluation for Adaptive Multimedia Protocols. IEEE Management of Multimedia Networks and Services (MMNS), Barcelona, Spain (2005)
16. Roychoudhuri,L., Al-Shaer,E., Settimi, R.: Audio Genome: An On-Line Audio Quality Assessment Framework. Technical Report, May (2005). http://www.mnlab.cs.depaul.edu/ lroychou/Genome_techrep.pdf
17. Watson A., Sasse M. A.: Evaluating Audio and Video Quality in Low-Cost Multimedia Conferencing Systems. Interacting with Computers Vol. 8 No. 3 (1996), pp. 255-275