QoE Model Performance Evaluation

By Dr. Irina Cotanis

Initiated during the VQEG Multimedia QoE Models project, then extensively refined, tested, and validated during ITU-T SG12 POLQA, P.NAMS, and P.NBAMS projects, the ITU-T P.1401 recommendation uses state of the art statistics to define methods, metrics, and procedures for the statistical evaluation, qualification, and comparison of objective quality prediction models, regardless of the assessed media type—e.g., voice, video-audio/multimedia. The recommendation describes an evaluation framework, provides guidance on model selection, and discusses special use cases.

Voice and video-audio (multimedia) QoE modeling experts contributed throughout the years to the development and continuous improvement of a stable and self-sustained statistical evaluation procedure for QoE model comparison. The final work resides with the ITU-T P.1401 recommendation, released in July 2012.

Evaluation Framework

Evaluation framework assumes that subjective tests in place are taking into consideration all new types of degradations that have emerged from a rapid technology evolution, one that brings with it a large variety of multimedia services which impact users more and more in a non-traditional way (e.g., re-buffering effect for multimedia streaming). In addition, it is assumed that aspects related to objective models, such as model type (e.g. parametric, perceptual), evaluation scope (e.g. comparison between

Based on well-established aspects related to both subjective tests and objective model development, an evaluation framework covers data preparation techniques, analysis types, numeral scale predictions, statistical evaluation metrics, and evaluation metrics' associated statistical confidence and significance.

models or against pre-defined performance thresholds), and application type, are well-defined prior to the evaluation process.

Data preparation

Known to drastically impact the evaluation results, the content of the databases is recommended to cover conditions related to the main scope of the QOE models (e.g., network design/deployment, performance evaluation and/or monitoring) as well as simulated conditions specific to the network's design/deployment life phase and the real live recordings required by the evaluation/monitoring phase. In addition, each experiment should contain conditions with quality levels that uniformly cover the 1-5 MOS scale. A thorough cleansing that removes unexpected subjective outlier scores ensures the quality of the databases.

Analysis types

There are four main analysis types that are dependent on the application and model types. Analysis per individual experiment and across multiple experiments are required regardless of the application or the model type. Analysis per media sample is necessary for live recorded databases, while per condition analysis is needed in the case of simulated databases. However, for live recorded databases, a recorded sample can equate to a field condition.

Prediction on a numerical quality scale

Prediction on a numerical scale is a determining factor of the accuracy of the QoE models' evaluation and involves the following relevant topics:

- The comparison of MOS values from different experiments
- The scale calibration of a QoE model

- The compensation for variance between subjective experiments in the evaluation process

The systematically observed differences between MOS scores from different experiments, even when the experiments followed the same guidance, can be grouped into three problem categories: bias (offset), different gradient, and different quantitative rank order. Bias represented in the result of the "overall" quality experiment is generally caused by different listening gear or environmental noises. A different gradient, defined as the relative quality distance between two identical stimuli or conditions during two experiments, is usually caused by a test design that does not cover the entire quality range. A different quantitative rank order is caused by MOS scores' statistical uncertainty expressed in the confidence interval, which needs to be considered when quality ranking is required. Ranking relies only on statistically significant differences, and resolutions finer than 0.3 MOS are not expected since a MOS confidence interval is usually in the range of 0.15 MOS. A generally adopted strategy to minimize scaling effects, such as biases and differing gradients, is to introduce defined anchor and reference conditions in two experiments; this can then be used to align the scores of the two experiments. In addition, other alternatives, such as MOS score normalization across experiments and design constraints to make the distribution of distortion types and quality ranges comparable between different experiments, are under discussion.

The scale calibration of QoE models is needed due to the fact that objective models predict quality based on technical information, and often partial results of individual analysis are combined in a late aggregation step into a single value that is generally dimensionless and not tied to the numerical 1-5 MOS quality scale. The scaling involves multidimensional optimization against the statistical evaluation metrics across a large pool of media samples (e.g., voice, video, audio) carefully selected to uniformly cover all test conditions for which the algorithm has been designed. The scaling procedure is based on a large number of well-balanced subjective reference experiments, and it is calculated such that the prediction widely follows the scale interpretation of the reference experiments, e.g., by choosing a scaling function that results in a minimum root mean square error (rmse) between the subjective reference experiments and the scaled objective predictions. Therefore, the selection of reference experiments is essential to how the model uses or interprets the quality scale.

The compensation for variance between subjective experiments in the QoE model evaluation process is required due to the inevitable differences between the objective QoE model, which predicts an average MOS value across many experiments as described above, and the subjective MOS value obtained in an individual experiment. As a strategy to minimize this dependency on subjective experiments, an individual compensation is used. The basic assumption is that well-balanced and well-designed subjective experiments are reproducing the qualitative rank-order with high accuracy, while the actual scale range and the gradient, as explained above, may be subject to individual interpretation. Both can be compensated for by individual mappings, where bias and gradient become aligned towards a generalized scale as used by the objective model. Usually, a monotonous linear, or a more sophisticated monotonous part of a third order polynomial, or a logistic mapping function can be applied. The purpose of the mapping function is to minimize the rmse or another metric as well as compensate for offsets, different biases, and other shifts between scores without changing the rank-order. The function is usually applied to the predicted scores before any statistical evaluation metric is calculated.

Per Experiment Statistical Evaluation

The recommended statistical metrics for objective quality assessment need to cover three main aspects: accuracy, consistency, and linearity against subjective data.

It is recommended that the prediction error be used for accuracy; the outlier ratio (OR), or the residual error distribution, for consistency; and the Pearson correlation coefficient for linearity. In addition, confidence intervals, as well as the statistical significance tests, are required for the comparison of these metrics calculated for different QoE models. The ITU-T P.1401 recommendation provides details on how these metrics should be calculated and compared.

Statistical Evaluation in the Context of Subjective Uncertainty: Epsilon-insensitive rmse

For stricter performance evaluation, ITU-T P.1401 introduces the ***epsilon-insensitive rmse*** (***rmse*)*** statistical metric, which considers differences related to an epsilon-wide band around the target value, with ***epsilo***' defined as the 95% confidence interval of the subjective MOS value, which reflects the uncertainty of the MOS scores. The ***modified*** rmse (rmse*) uses as ***modified*** prediction error (Figure1)

$$Perror(i) = \max(0, \left|MOSLQS(i) - MOSLQO(i)\right| - ci_{95}(i)) ,$$

where ci_{95} is the 95% confidence interval of the individual MOS scores. The rmse* is calculated per database, and it describes how the prediction error exceeds the ci_{95}. As a modified rmse, the statistical significance of the difference between two rmse* values is calculated as in the traditional rmse case.

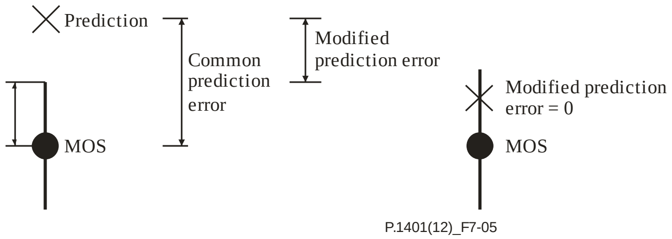

Figure 1. Rmse* calculation.

Statistical Evaluation of the Overall Performance

The overall performance of a model is defined by its performance across each experiment (i.e., test database) as well as across all experiments. Therefore, results per experiment should be aggregated in an overall figure of merit. In order to do so, three steps need to be performed:

- Weighting of databases based on their importance within the QoE model evaluation scope
- Calulation of the aggregated statistical significant distance measure (SSDM) per experiment
- Calculation of the overall performance and statistical significance testing between QoE models.

The SSDM represents the figure of merit of a model per experiment and can be calculated as follows:

$$d_{k,v} = \sum_{i=1}^{Nmetric} W(i) * \max(0, StatMetric F(0.05, N_k, N_k) Result)$$

where *StatMetricF(0.05,*N_k*,*N_k*)Result* denotes the result of the statistical significance test for each evaluated metric *i*=1...Nmetric (e.g., correlation coefficient, OR, rmse). The index *k* denotes the experiment, while index *v* denotes the objective model. F(0.05, n1, n2) is the tabulated value of the F-distribution for n1 and n2 degrees of freedom and 95% significance level. N_k describes the number of considered samples (files or conditions) in experiment *k.* The function *W(i)* represents the weight that is allocated to each statistical metric based on their importance to the evaluation process. The highest importance should be allocated to the primary metric which the QoE models have been optimized against.

The overall performance for an algorithm *v* is defined as

$$p_v = \sum_{k=1}^{M} w_k \times d_{k,v}$$

where M is the total number of databases across the sets, k is the index of the database, $d_{k,v}$ is the distance measure of the model v for the database k, and w_k represents the weight of the database k.

The statistical significance test is applied to the aggregated distance measures p_v calculated for all models. The value p_v is the aggregated distance for v model, p_{min} is lowest p_v in the evaluation and the value K describes the degree of freedom of the F distribution:

$$t_v = \max(0, \frac{p_v}{\left(p_{\min} + c\right)} - \mathrm{F}(0.05, K\ \ , K\ \ )) .$$

If $t_v = 0$, the model v is considered as statistically equivalent to the model with $p = p_{min}$. If $t_v > 0$, the model v is considered as significantly statistically worse than the lowest $p = p_{min}$. The constant c is recommended to be set to 0.0004 based on proved calculations performed for the speech QoE models.

Guidance on Models' Selection

To select the best performing model, it is recommended to consider ***per experiment*** and ***overall performance***, as well as the analysis of ***the worst performance cases***. The models with statistically equal lowest SSDM values per experiment perform the best for that particular experiment. The overall best performing models should exhibit the lowest statistically-equal overall figure of merit calculated as the aggregated SSDM across all experiments. The analysis of the worst performance experiments ensures that the best performer does not show as the worst case in any of the evaluation instances (e.g., per one experiment).

Selecting a best-performing QoE model depends on a variety of factors, such as scope of the evaluation, media and model type, approach used for the QoE model development, etc.

In addition, the evaluation process should use both test databases (e.g. databases used to train the models) as well as validation databases (e.g. databases that are completely unknown to the model). After the selection process is accomplished and a winner is selected, then a characterization phase should take place, with the scope of identifying strengths and weaknesses of the best performing model.

Special Cases

In the case of models designed to estimate the subscriber's perception of various dimensions of quality degradation (e.g., blurriness and blockiness in video, or loudness and coloration in voice), the evaluation is required for each degradation type, as well as on the overall performance.

Special evaluation cases refer either to models with multi-dimensional outputs or to scenarios when only one model is evaluated. In both cases the same framework and same statistical metrics are used.

The second special case refers to the evaluation of one single model. In this scenario, the comparison is performed against pre-defined minimum performance thresholds defined based on previous experiences, whenever available. These scenarios include the case of either a new or improved standard, or a parametric (including planning) or hybrid model when a perceptual model is already in place. In this case, the role of the "best performing model" is played by the minimum performance thresholds defined *a priori* to the evaluation process.

Dr. Irina Cotanis is principal technologist with Ascom Network Testing CTO Office. She holds a Doctorate in Electrical Engineering, and a score card of more than 25 years of experience in wireless and radio communications systems, statistical signal processing and analysis, and statistics, as well as more than 10 years as an active member in standardization organizations, and several publications in IEEE conference proceedings, standards, and text books. She has also acted as reviewer to IEEE papers as well as session chair for various IEEE conferences.