ISCA Archive
http://www.isca-speech.org/archive

First ISCA ITRW on
Auditory Quality of Systems
Akademie Mont-Cenis, Germany
April 23-25, 2003

Application of Objective Speech Quality Metrics

Isabelle, Scott K.

PCS Acoustic Technology Center
Motorola, Inc., Libertyville, IL USA
scott.isabelle@motorola.com

Guzman, Sandra J.

PCS Consumer Experience Design Human Factors
Motorola, Inc., Libertyville, IL USA

Saliba, Rindala

PCS Consumer Experience Design Human Factors Audio
Motorola, Inc., South Plainfield, NJ USA

Novorita, Robert

CGISS Global Technology Development Group, Advanced Technology
Motorola, Inc., Schaumburg, IL USA

Abstract

Results of subjective evaluations of perceived sound quality preferences for mobile telephone handsets with different acoustic design strategies are presented and compared to predictions from two objective speech quality metrics.

1. Introduction

The pursuit of the ability to understand the perceived quality of mediated speech communication has a long history, going back at least to the work of Campbell [1] and Crandall [2] at the Bell Telephone Labs in the early part of the last century. More recently, a great deal of effort has been aimed at developing predictive models for perceived speech quality. Some of these models have been drawn from other work in auditory science, based on a detailed understanding of the nature of the physiology of the auditory periphery (e.g., [3]). Others have been more broadly aimed at developing phenomenological models more specifically in the context of telephonic quality (e.g., [4]). Recent results have compared some of these metrics to a variety of problems in end-to-end speech quality assessment of telephone networks [5, 6, & 7]. Most of these studies have focused on the entire network, including impairments that arise from a multitude of sources, and report varying degrees of success in predicting subjective quality. This work is an attempt to examine the ability of objective metrics to predict the impact of acoustic effects of telephone handsets on perceived speech quality.

2. Background

2.1. Motivation and objectives

The primary motivation for this work is to determine the ability of modern objective speech quality metrics to characterize and quantify the expected user experience of quality of speech when using digital wireless telephone handsets. The major goal is to compare predictions from objective speech quality metrics to subjective (quantitative) ratings of the perceived quality of handsets that implement a variety of acoustic design approaches. The results will enable a better understanding of the relative value to the customer of those varying design approaches.

2.2. General Methods for Subjective Evaluation

All the subjective evaluations reported here follow generally accepted practices for such studies as outlined, for example, in the relevant series of Recommendations of the ITU-T [8, 9]. Each person in a proper-sized panel of listeners, drawn from an appropriately representative population, is presented, in randomized sequence and in a well-controlled acoustic environment, a set of auditory stimuli. The auditory stimuli are developed from appropriate source speech material that has been processed according to the needs of the particular evaluation. Subsequently, each listener gives responses to the stimulus or stimuli using appropriate Likert scales. Relevant supplementary details will be given below.

2.3. General Methods for Prediction by Objective Metrics

The computations required for generating predictions are performed on the same signals that are used as auditory stimuli. Values of the objective metrics are compared to subjective results as appropriate for the particular subjective evaluation. While comparisons are made for two objective metrics, PESQ [4] and TOSQA [5, 6], the majority of the comparisons with subjective results reported here are with PESQ.

3. Evaluation of Acoustic Design Preference

Because the primary motivation for this work is to determine how objective speech quality metrics predict the user's experience of perceived speech quality in mobile telephone handsets, the focus of this report is on a subjective study that was designed to directly compare three handsets that implement different acoustic design approaches.

3.1. Acoustic Design Differences

3.1.1. Traditional telephone handsets

The acoustic design of traditional landline telephones handsets often includes the assumption that there is a good acoustic seal between the user's ear and the receiver or earpiece. The quality of the acoustic seal depends on the geometry of the earcap of the handset. For purposes of evaluating the performance of these designs, a sealed ear simulator, the ITU-T Type 1 artificial ear, has been developed and standardized [10].

3.1.2. Modern mobile telephone handsets

For modern mobile telephones, the demands of the market have dictated smaller devices with shapes driven in part by stylistic concerns. One implication of these shapes is that the acoustic seal between a user's ear and the handset earpiece can become compromised or "leaky." Handsets that employ an acoustic design that is intended to provide a transfer function that is robust to acoustic sealing (i.e., roughly constant transfer function shape across acoustic sealing conditions) are said to be "leak-insensitive." For purposes of evaluating such acoustic designs, two ear simulators intended to represent a range of leakage conditions, the ITU-T Type 3.2 low-leak and high-leak artificial ears, have been developed and standardized [10].

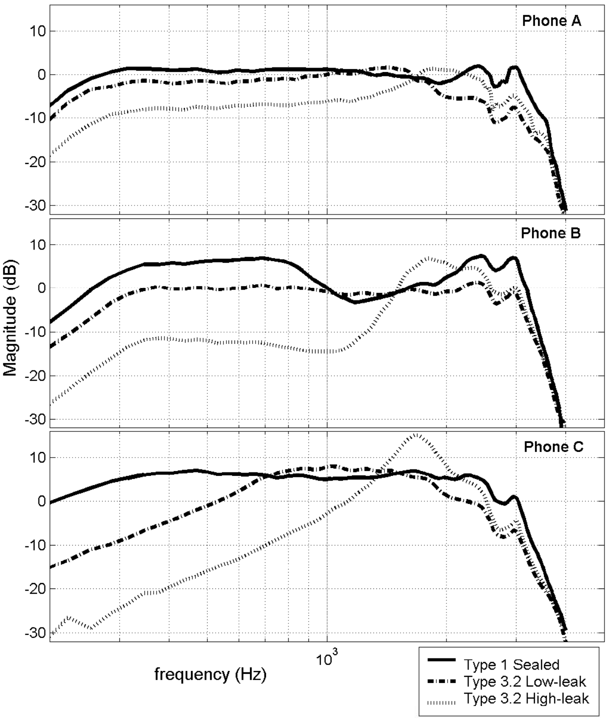

Figure 1: Transfer functions of three mobile telephone handsets using three artificial ears.

Figure 1 shows the transfer functions of three mobile telephone handsets (the three panels of the figure) measured using three artificial ears (the three line types in each panel). The solid line is for the Type 1 artificial ear, the dashed line is for the Type 3.2 Low-leak artificial ear, and the dotted line is for the Type 3.2 High-leak artificial ear). It can be seen in this Figure that the variation of the transfer functions across artificial ears (or leakage conditions) is smallest for Phone A. Thus, Phone A would be termed the most "leak-insensitive." Phone B demonstrates larger sensitivity changes with leakage than does Phone A, particularly for frequencies below about 1.5 kHz. Finally, it can be seen that the changes with leakage are largest for Phone C. Therefore, Phone C would be described as "leak-sensitive," with Phone B intermediate between A and C.

While a discussion of the governing acoustic principles is beyond the scope of this study, the relation of acoustic loading, as defined by acoustic leakage, and the characteristics of an acoustic design that produce robust transfer function dependence on leakage are well-known to experts in the field of acoustic design.

3.2. User Experience of Acoustic Design Differences

From a user perspective, it is hypothesized that a leak-insensitive handset, providing a more consistent transfer function shape with leakage, leads to an experience of higher speech quality than does a leak-sensitive handset. A subjective evaluation was conducted to test this hypothesis.

3.2.1. Stimuli and Subjects

The source speech consisted of a set of phonetically balanced sentences produced by native speakers of American English, both male and female talkers [11]. The source speech was filtered to approximate transmission over a telephone system (Modified Intermediate Reference System Send Characteristic, [9]).

The seven subjects all had normal hearing and spoke American English.

3.2.2. Telephone Handsets

It is well known in the subjective evaluation literature that extraneous factors can influence the perception of a quality being evaluated [12]. For example, differences in the visual appearance of different handsets can influence the perception of differences in their audio quality. In order to minimize the effect of visual and tactile cues on the subjective evaluation of audio quality, and to maintain an ecologically valid experience for the participants, all three handsets A, B, and C were realized using one common physical design. That is, one common physical handset type was used, but three distinct instances were created, each containing an acoustic system with the transfer functions shown in Figure 1. The common physical design was selected on the basis of addition data (not reported here) indicating that users perceive it to have high comfort on the ear, good cues for establishing a proper listening position on the ear, and robustness with respect to position on the ear.

3.2.3. User Evaluation

In order to evaluate the three acoustic designs, users were presented pairs of handsets in random order. The source speech played out of each handset simultaneously. Users indicated which handset they preferred, and the magnitude of their preference. Handset pairs were presented until all possible handset pairs and orders were judged.

3.2.4. Results of User Evaluation

Figure 2 shows the results of the subjective evaluations. In this figure, every possible pair is shown on the abscissa. Magnitude of preference is shown on the ordinate. A value of zero indicates no preference. The error bars indicate standard error of the mean. The ratings are plotted such that a larger value indicates a stronger preference for the first handset in the pair.

For handset pair B vs C (the leftmost bar), a value of 1.0 indicates that there is a moderate preference for handset B over handset C. Similarly, the middle bar indicates that handset A is slightly preferred over handset B, whereas the rightmost bar indicates a larger preference for handset A over handset C.

Overall, the results suggest that a leak-insensitive design leads to a greater perceived speech quality, as was hypothesized. Phone A was most preferred, followed by phone B. Phone C was least preferred.

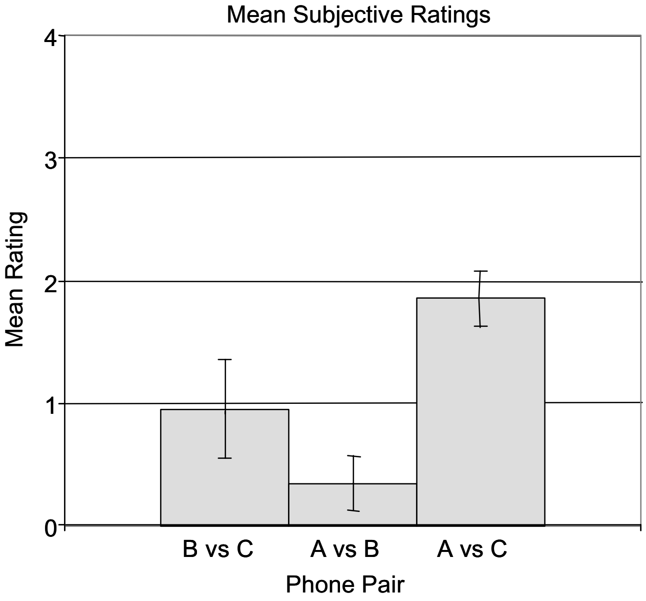

Figure 2: Subjective ratings of preference for pairs of phones.

3.3. Prediction of Users' Subjective Evaluation

In order to generate predictions of the subjective preference ratings, it is necessary to simulate the listening conditions. Because the subjects used the handsets in the typical, and ecologically-valid, manner, the effective transfer function for a given handset is likely to have been different between listeners and even between repeated placements of the same handset on the same listener. Furthermore, because measurements of the range of the transfer functions of the handsets on the listeners' ears were not possible to make, it becomes necessary to make the assumptions described in the next section.

3.3.1. *Processing of stimuli used for prediction*

We assume that the transfer functions as measured using Type 1 Sealed artificial ear (see Figure 1, solid lines) represent typical upper bounds for handset sensitivities on our subjects. We further assume that the transfer functions as measured using the Type 3.2 High-leak ear (see Figure 1, dotted lines) represent typical lower bounds for handset sensitivities on our subjects. Intermediate conditions are represented by interpolating for ten steps between the measured transfer functions for Type 1 Sealed and Type 3.2 Low-leak artificial ears (see Figure 1, dashed lines), and also interpolating for ten steps between the measured transfer functions for the Type 3.2 Low-leak and the Type 3.2 High-leak artificial ears. Figure 3 shows the interpolated transfer functions for Handset C.

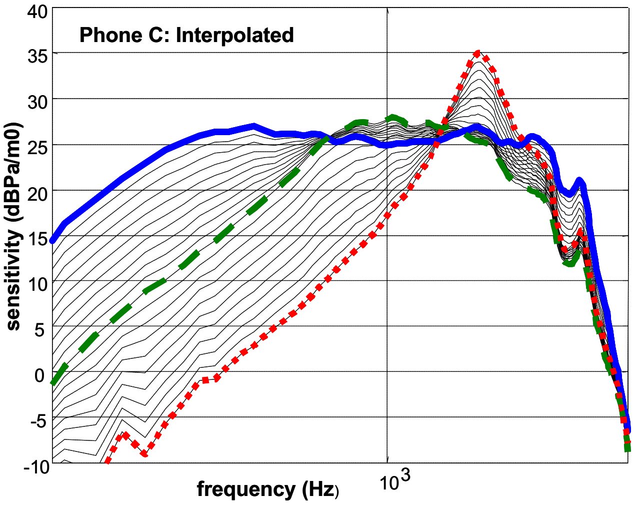

Figure 3: Interpolated transfer functions for Handset C.

A heavy solid line shows the measured transfer function for the Type 1 Sealed artificial ear. The measured transfer functions for the Type 3.2 Low-leak, and Type 3.2 High-leak artificial ears are shown by a heavy dashed line, and by a heavy dotted line, respectively. Light solid lines indicate the interpolated transfer functions. The sentences used in the subjective evaluation were each filtered by the set of 21 transfer functions for each handset to create the stimulus set used for the predictions.

3.3.2. *Procedure for predictions*

Both the PESQ and TOSQA metrics require a pair of signals for each computation, an original signal used as the reference, and that signal after processing by the system under test, called the degraded signal [3]. Because we are interested in predicting the subjective preferences between different acoustic systems, we used the source speech sentence as the reference signals, and the filtered speech sentence as the degraded signals. It is worthwhile to note that the use of these metrics to predict speech quality preferences solely due to acoustic differences is not a typical application for these metrics. However, given the demonstrated ability of these tools to predict a variety of conditions [5, 7], it is of interest to examine the ability of these metrics to predict the subjective results under these conditions

3.3.3. *Results of predictions of subjective results*

The results of the PESQ predictions are shown in Figure 4. The results for each handset are plotted along the abscissa, with the predicted PESQMOS score on the ordinate. Note that the values shown are the PESQ MOS values directly from the computation, without any rescaling or mapping. The mean value for the predicted score for each phone is given by the centerline of the diamond-shaped symbol. The upper and lower 95% confidence intervals are shown by the upper and lower points of each diamond shape. Also shown are quartile box plots, marking median and interquartile values. The mean values are ordered (largest to smallest) as A, B, C. The Tukey-Kramer Honestly Significant Differences (HSD) comparison of all means show that while the difference between means for A and C is statistically significant, the mean value for B is not statistically different from either A or C. Thus, while the differences among the means are consistent with the differences among the subjective ratings (A most preferred, followed by B, with C least preferred), the PESQMOS scores do not fully distinguish among these cases in a statistically significant manner for these data.

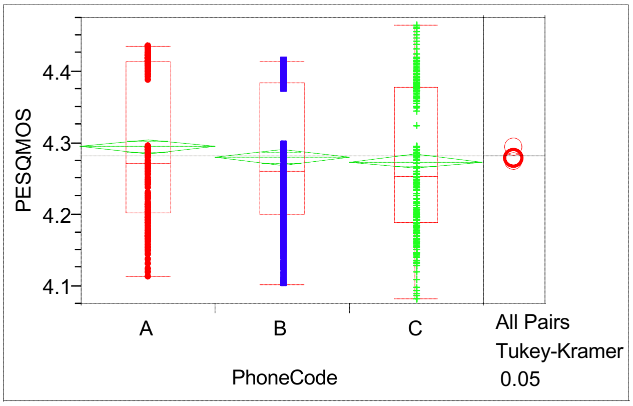

Figure 4. PESQ Predictions

Figure 5 shows the results for the TMOS prediction from TOSQA, plotted similarly as for Figure 4. It can be seen from these results that the ordering of the mean TMOS scores is also consistent with the ordering of subjective preferences. However, for TMOS, the Tukey-Kramer HSD test indicates that the mean TMOS values are significantly different for all pairs of phones.

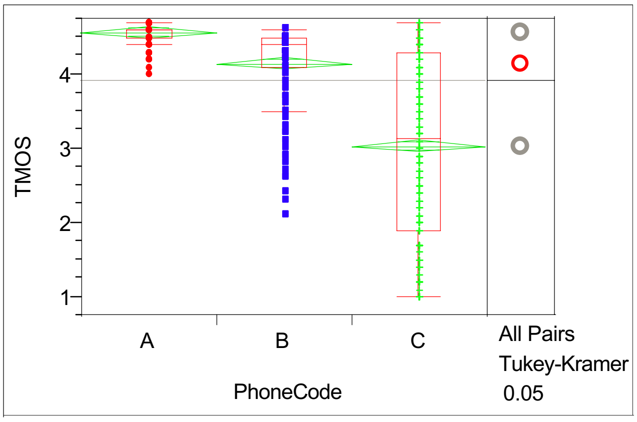

Figure 5. TOSQA TMOS predictions

Comparing Figure 5 to Figure 4, it can be seen that the range of predicted values across phones appears to be roughly constant for PESQMOS, but appears to increase with decreasing mean TMOS. The distributions of the predicted values are further explored in Figure 6.

In Figure 6, there appear to be substantial differences in distribution shape between the two metrics. The PESQMOS scores (upper panel) appear to have a bimodal distribution, unlike the TMOS scores (lower panel).

To address this issue, we consider the *paired* differences of the objective ratings. Because the subjective ratings were made for paired comparison, it is reasonable to use the same approach for the objective scores. The paired-comparison differences are computed simply by subtracting the predicted scores for each phone on a stimulus-by-stimulus basis. Figure 7 shows the distributions of the paired-comparison differences of the objective metrics.

Although there are still apparent differences in shape between the distributions of the paired-comparison differences for PESQMOS and TMOS metrics, both distributions are unimodal with modes close to zero and positive.

Figures 8 and 9 show the predictions of the paired-comparison differences for PESQMOS and TMOS respectively. For both PESQMOS (Figure 8) and TMOS (Figure 9), the pair A-C has the largest difference, indicating that A is more strongly preferred over C, consistent with the subjective results. However, there is a difference between the PESQMOS and TMOS

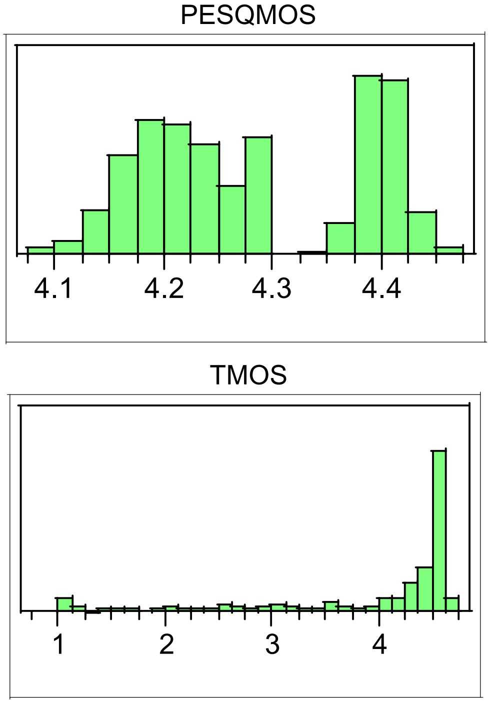

Figure 6: Distribution of PESQMOS & TMOS values

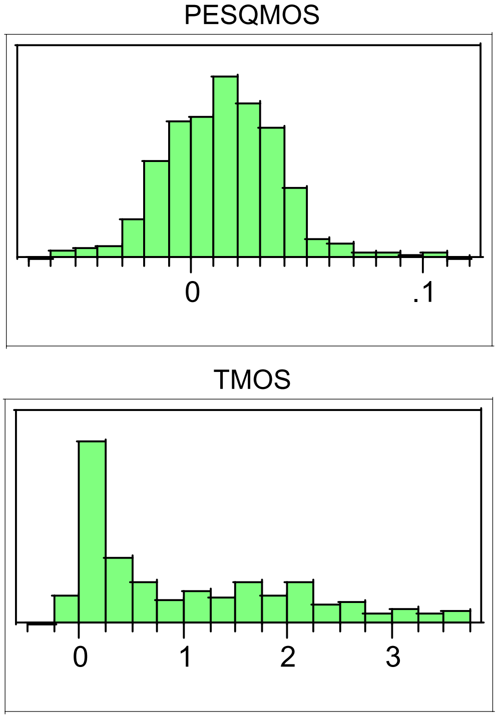

Figure 7: Distribution of paired-comparison differences of PESQMOS and TMOS values.

differences when considering the pair with the second-largest predicted difference. PESQMOS predicts that pair A-B has the second-largest differences, whereas TMOS predicts that pair B-C has the second-largest differences. The subjective results shown in Figure 2 are that pair B-C has the second-largest differences, as do the TMOS predictions. Note, however, that the PESQMOS prediction for pair B-C is still significantly larger than zero (at the 95% confidence level), predicting that phone B is preferred over phone C.

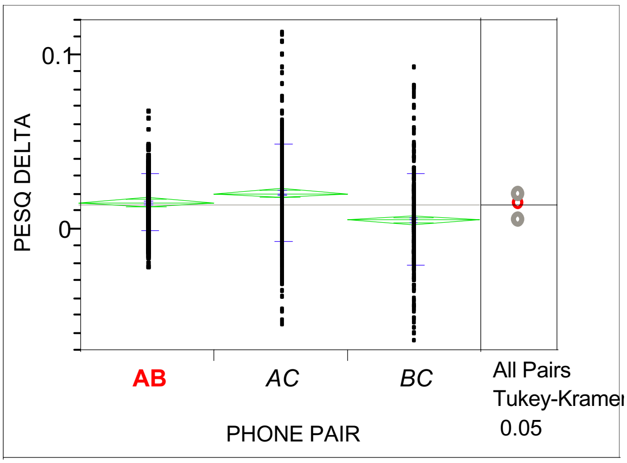

Figure 8: Predictions for PESQMOS paired-comparison differences

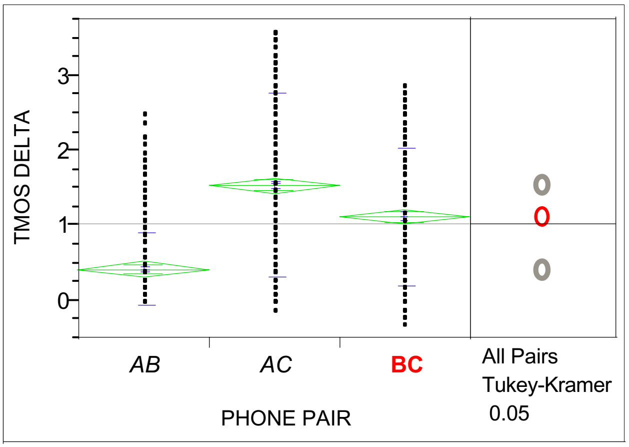

Figure 9: Predictions for TMOS paired-comparison differences

4. Summary and Conclusions

Results are presented on the subjective preferences of speech quality for mobile handsets with varying acoustic designs. The results indicate that leak-insensitive handsets are preferred over leak-sensitive handsets. While it should be noted that the data presented here are restricted to American English, we have conducted similar studies using Mandarin Chinese and have drawn similar conclusions.

Results are also presented comparing predictions of two objective speech quality metrics, PESQ and TOSQA, to the subjective ratings. Both metrics successfully predicted which handset received the highest preference ratings, and both made predictions that did indicate the relative order of preference within each pair. However, for this particular study, only TOSQA properly predicted the relative order of preference across all three handset designs. We consider these results to be preliminary and there are other factors (e.g., gender of talker) that have not yet been analyzed.

5. Acknowledgements

Acknowledgements are gratefully made to David Cranfill of the PCS Consumer Experience Design Human Factors Audio team, Ian Lewis of the PCS Audio Architecture Systems Engineering team, and other members of the CXD Human Factors and Audio teams for their valuable contributions to this research.

6. References

[1] Campbell, G (1910). Telephonic Intelligibility. Phil. Mag. 19(6):152-159.

[2] Crandall, B (1916). A quantitative model for telephonic quality. Bell Telephone Labs Journal.

[3] Hauenstein, M (1998). Applications of Meddis' hair-cell model to the prediction of subjective speech quality. Proc. ICASSP98.

[4] ITU-T Recommendation P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.

[5] ETSI EP TIPHON 22-TD031 (11/27/2001). Report of 1st ETSI VoIP Speech Quality Test Event.

[6] Pennock, S (2002). Accuracy of the PESQ algorithm, MESAQIN 2002.

[7] ETSI EP TIPHON 28-TD045 (6/282002). ETSI 2nd Speech Quality Test Event Anonymized Test Report.

[8] ITU-T Recommendation P.800, Methods for subjective determination of transmission quality.

[9] ITU-T Recommendation P.830, Subjective performance assessment of telephone-band and wideband digital codecs.

[10] ITU-T Recommendation P.57, Artificial ears.

[11] IEEE (1969). IEEE Recommended Practice for Speech Quality Measurements, IEEE Trans. Audio Electroacoust. 17(3):225-246.

[12] Toole, FJ and Olive, SE (11/1994). Hearing is Believing vs. Believing is Hearing: Blind vs. Sighted Listening Tests. 97th Convention, Audio Eng. Soc. Preprint No. 3894.