Objective Speech Quality Measures for Internet Telephony

Timothy A. Hall

National Institute of Standards and Technology
100 Bureau Drive, STOP 8920
Gaithersburg, MD 20899-8920

ABSTRACT

Measuring voice quality for telephony is not a new problem. However, packet-switched, best-effort networks such as the Internet present significant new challenges for the delivery of real-time voice traffic. Unlike the circuit-switched public switched telephone network (PSTN), Internet protocol (IP) networks guarantee neither sufficient bandwidth for the voice traffic nor a constant, acceptable delay. Dropped packets and varying delays introduce distortions not found in traditional telephony. In addition, if a low bitrate codec is used in voice over IP (VoIP) to achieve a high compression ratio, the original waveform can be significantly distorted. These new potential sources of signal distortion present significant challenges for objectively measuring speech quality. Measurement techniques designed for the PSTN may not perform well in VoIP environments.

Our objective is to find a speech quality metric that accurately predicts subjective human perception under the conditions present in VoIP systems. To do this, we compared three types of measures: perceptually weighted distortion measures such as enhanced modified Bark spectral distance (EMBSD) and measuring normalizing blocks (MNB), word-error rates of continuous speech recognizers, and the ITU E-model. We tested the performance of these measures under conditions typical of a VoIP system. We found that the E-model had the highest correlation with mean opinion scores (MOS). The E-model is well-suited for online monitoring because it does not require the original (undistorted) signal to compute its quality metric and because it is computationally simple.

Keywords: speech quality, Internet telephony, voice over IP, network metrology

1. INTRODUCTION

In recent years, there has been growing interest in using the Internet and other Internet protocol (IP) networks for telephony. Motivations such as reduced cost, simplification of infrastructure through network convergence, and the opportunity to provide new and programmable services have driven this interest. However, success of Internet telephony depends upon the reliable delivery of good voice quality, and speech quality metrics are needed for designing, building, and maintaining such VoIP systems. While the problem of measuring speech quality of telephony systems is not new, the characteristics of VoIP systems are different in many respects from those of the existing PSTN. Best-effort IP networks present significant new challenges to the delivery of real-time voice traffic. Whereas the circuit-switched PSTN guarantees that sufficient bandwidth is reserved and available for the duration of the call, IP networks, in general, do not. Delay is not guaranteed to be either minimal or constant in an IP network. In addition, dropped packets and packet delay variation, or jitter, introduce distortions not found in traditional telephony. Low bitrate (high compression ratio) codecs used to reduce required bandwidth distort the original waveform significantly before it is even transmitted. The compressed speech produced by such codecs is also more sensitive to packet loss. These and other characteristics of VoIP make delivery of toll quality speech challenging. These same characteristics make measuring the speech quality difficult as well. Most existing objective speech quality measures have been developed for high bit-rate, error-free telephony environments and do not accurately predict subjective voice quality in the presence of the significant impairments introduced by VoIP systems. In this paper, we evaluate several objective speech measures to determine their effectiveness in predicting human perception of speech quality in VoIP networks. We also discuss the suitability of the algorithms for implementation in an online monitoring environment capable of providing speech quality measures in real time.

The paper is organized as follows. We first give a general background on speech quality measurement, along with brief descriptions of the algorithms we evaluated. Second, we describe the two experiments we conducted to evaluate them, including the data sets used and the distortions introduced. Third, we present the results of the two experiments, and, finally, we discuss implications of the results.

2. MEASURING SPEECH QUALITY

There are two broad classes of speech quality metrics: subjective and objective. Subjective measures involve humans listening to a live or recorded conversation and assigning a rating to it. This rating can be either a single overall quality rating or a rating of a particular characteristic (i.e., clarity or listening effort) or a particular distortion (i.e., clipping, hum). Because they use human subjects, subjective measures are often very accurate and useful for evaluating a telephony system. The mean opinion score (MOS) is one such useful metric. Although the MOS is not the only subjective measure, it is one of the most widely used and recognized. ITU-T Recommendation P.830[10] describes in detail how to conduct a subjective test experiment, but the procedure can be summed up as follows. A panel of subjects listens to a set of speech samples, assigning to each sample an overall quality score ranging from 1 (Bad) to 5 (Excellent). The average score of the panel for a given sample is that sample's MOS.

Clearly, a metric such as MOS that uses human subjects can be a good measure of perceived speech quality. However, subjective metrics have disadvantages, too. In particular, they can be time-consuming and expensive. Some researchers or organizations may not have the resources to conduct the tests. Certainly, such metrics cannot be used in any sort of real-time or online application. These shortcomings, among other reasons, have led to the development of objective metrics. Such measures predict perceived speech quality based typically on a computation of distortion between the original (clean) signal and a received (noisy) signal. In some algorithms, something other than the difference between the received and original signals is used, such as a quantitative measure of the distortion.

Typically, the accuracy or effectiveness of an objective metric is determined by its correlation, usually the Pearson (linear) correlation, with MOS scores for a set of data. If an objective metric has a high correlation with MOS, then it is deemed to be an effective measure of perceived speech quality, at least for speech data and transmission systems with the same characteristics as those in the experiment. Indeed, metrics that work well under some conditions are not necessarily good predictors of perceived voice quality under other conditions.

Our goal is to find a speech quality metric that accurately predicts human perception under conditions typical of VoIP systems. To do this, we compare three types of measures. The first type is perceptually weighted distortion measures, which include the enhanced modified Bark spectral distance (EMBSD) [11][12][13] and measuring normalizing blocks (MNB) [9][1][2][3] algorithms. The second uses the word-error rates output by a continuous speech recognizer of the original and received signals to predict voice quality[4]. The third is the ITU E-model[6][7][8].

2.1. Perceptually Weighted Distortion Measures

Modern objective metrics use knowledge of the human auditory system to compute a perceptually weighted distortion measure. Distortions that are most significant to the human ear are weighted more heavily while those that are inaudible or nearly so are weighted lightly or not at all. A number of algorithms exist in this class of measures. We chose the two best performers, according to the literature: measuring normalizing blocks (MNB), which is found in Appendix A of ITU-T Recommendation P.861, and enhanced modified Bark spectral distance (EMBSD).

The MNB algorithm comprises two stages: a simple perceptual transformation, and a distance measure that uses hierarchies of measuring normalizing blocks. For perceptual transformation, the time-aligned, normalized signals, original and received, are divided into 50% overlapping frames of 128 samples. Each frame is multiplied by a Hamming window and transformed using the fast Fourier transform (FFT). Only the squared magnitudes of the FFT coefficients are preserved. The coefficients are transformed to the Bark scale, a psychoacoustic frequency scale where

$$b = 6 \cdot \sinh^{-1}\left(\frac{f}{600}\right)$$

defines the transformation. This is accomplished by grouping the squared FFT coefficients into bins of equal width on the Bark scale. The total energy of each frame is computed, and frames below an energy threshold in either the original or received signals are discarded. All samples in remaining frames are transformed using a logarithm to model perceived loudness.

The distance measure used is a linear combination of the distances computed in the time and frequency MNBs. There is one frequency MNB (FNMB) for each power spectrum coefficient. A frequency MNB averages the difference at that coefficient between the original and received signals across all frames that exceed the above-mentioned energy thresholds. Four measurements covering the lower and upper band edges of telephone band speech are saved in measurement vector m. There are two different time MNB (TMNB) structures using different frequency scales,

producing either eight or seven measurements saved in m, depending upon which is used. The measurement vector m is multiplied by a weight vector w to compute a single auditory distortion (AD) number, which is passed through the logistic function

$$L(z) = \frac{1}{1 + e^{a \cdot AD + b}}$$

to map it to (0,1).

The EMBSD algorithm consists of a perceptual transform followed by a distance measure that incorporates a new cognition model. As with the MNB, the original and received signals are normalized and divided into 50% overlapping frames of 320 samples. Frames that exceed an energy threshold are transformed to the Bark frequency domain. The Bark coefficients are transformed to dB to model perceived loudness, then scaled. The first 15 Bark spectral components in each frame are used to compute the loudness difference. Only distortion that exceeds a noise masking threshold and lasts longer than 200 ms (10 frames) is considered. The distortion is computed as the average distortion for all valid frames.

The MNB and EMBSD algorithms require the original and received signals to be time-aligned. This requires a measure or estimate of the time delay between the two signals. The two-state delay estimation algorithm described in ANSI Standard T1.801.04-1997[14] is an effective algorithm for this purpose. The first stage of the algorithm uses cross-correlation between the speech envelopes of the two signals to give a coarse estimate of the delay to within ±4 ms. The second stage uses the cross-correlation between the power spectral densities (PSDs) of the two signals to refine the estimate to within ±1 ms, if the PSD of the received signal has not been overly distorted by the codec used.

2.2. Word-error Rates of Continuous Speech Recognizers

This technique is very straightforward. The received speech sample is input to a continuous speech recognizer (CSR), and a transcription is output. This transcription is compared to the original and the scored accordingly. The better the transcription, the better the speech quality. Our approach builds on promising earlier work at NIST's Information Technology Lab, which used phoneme recognition rates as a predictor of speech quality[4]. We made two changes in method: (1) replacing the research-oriented *hidden Markov-model toolkit (HTK)*[15] speech recognition system with the commercial product *Dragon Systems' Naturally Speaking*[16] *, and (2) using word recognition and word error rates instead of phoneme recognition and phoneme error rates. We considered two measures: percentage of words correctly recognized (CSR 1) and percentage of errors (CSR 2), which includes word insertions and omissions as well as incorrectly identified words.

2.3. ITU E-model

Unlike the two approaches described above, the E-model does not compare the original and received signals directly. Instead, it uses the sum of equipment impairment factors I_e, each one quantifying the distortion due to a particular factor. Impairment factors include codec used, echo, average packet delay, packet delay variation, and fraction of packets dropped. As an example, in a system with distortion due to the codec, average one-way delay, packet delay variation (jitter) and packet loss, the rating R is computed as follows:

$$R = R_0 - I_{codec} - I_{delay} - I_{pdv} - I_{packetloss}$$

where R_0 is the highest possible rating for this system with no distortion. For our tests we used $R_0 = 100$. R can then be used as-is or mapped to MOS. Values for many of the impairment factors are found in ITU-T Rec. G.113, Transmission impairments, Appendix I: Provisional planning values for the equipment impairment factor, I_e, [8].

*No approval or endorsement of any commercial product is intended or implied. Certain commercial products are identified in this paper to facilitate understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

2.4. Other Algorithms

There are several objective speech quality algorithms that we did not use in our experiments. In [3], Voran compares the MNB algorithm with six well-known algorithms: signal-to-noise ratio (SNR), segmented signal-to-noise ratio, perceptually-weighted segmented signal-to-noise ratio, cepstral distance, Bark spectral distance, and the noise disturbance estimator produced by the perceptual speech quality measure (PSQM) described in the main body of ITU-T Recommendation P.861[9]. His results showed that the MNB algorithm compares favorably with PSQM, given higher correlation with subjective test results in most of the tests, with significantly higher correlation in several cases and never significantly lower correlation. The MNB outperformed the five other algorithms in all of the nine tests.

The ITU is proposing a successor to ITU-T Recommendation P.861 that is intended to account for distortions introduced by a VoIP system called Perceptual Evaluation of Speech Quality (PESQ) [20]. ITU-T Rec. P.862 is due out sometime in 2001, but was not available at the time of this writing. It requires the original and received signals in order to measure the distortion and predict perceived speech quality.

3. EXPERIMENTS

We conducted two experiments. For the first, we used the same set of speech samples, original and distorted, as in [4] in order to repeat its baseline experiment. The original (clean) speech samples were from the TIMIT speech database, frequently used in evaluating speech recognizers[18]. The samples were coded using a CELP coder, then run through a simulated channel with Gaussian-distributed bit error rates of 0%, 0.1%, 0.5% ,1%, 2%, and 5%. All speech samples were MOS scored. Using these speech samples, we computed the objective measures EMBSD, MNB 1 and MNB 2 (two different time MNB structures) and CSR 1 and CSR 2. We conducted this experiment to test our assertion that an off-the-shelf CSR recognizing words would give similar results to the *HTK* package used for phoneme recognition in [4]. We used the *sclite* [5] speech recognition scoring package developed at NIST to compute percentage of words correctly identified and percentage of errors. Errors comprise substituted (mis-identified), inserted, and deleted words; therefore, percentage of errors is not simply 100% minus percent correctly identified. We also wanted to compare CSR performance with more modern perceptually weighted distortion measures that were not evaluated in [4], namely the MNB and EMBSD algorithms. The E-model was not used in this experiment since no published values existed for bit error rates (only packet loss rates) or CELP coders. Using MOS scores from [4], we computed the linear (Pearson) correlation between each of the measures and MOS.

While the first experiment linked our work to [4], in our second experiment, we used speech samples more representative of VoIP systems. The samples were coded with the G.711 and G.729 codecs. Two manufacturers's codecs, a G.711 and G.729 codec from each one, were used in the tests. Errors introduced by transmission over an internet were emulated using the *NISTNet* network emulator [17] by varying packet loss rates (0%, 1%,and 5%) and jitter values (0, 50, and 100 ms). The jitter is defined as the standard deviation of the one-way delay. Average one-way delay was not considered. Although the E-model accounts for impairment due to delay, the other algorithms we tested require removal of any delay (i.e., the original and received signals must be aligned in time) and do not account for it. As in the first experiment, we used MOS scores and the computed measures from EMBSD, MNB 1 and MNB 2, and CSR 1 and CSR 2.

In the second experiment, we also computed the E-model rating (R-value) for each of the received samples. We considered the following equipment impairment factors: codec used, packet loss percentage, and packet delay variation (jitter). The values for impairment due to codec used and percent packet loss were taken from Tables 1.2 and 1.3 in ITU-T Recommendation G.113, Appendix I[8]. We had no available published values for packet delay variation; therefore, we calculated values for each of the codec/packet delay variation pairs, as shown in Table 1. We chose values for packet delay variation that maximized the correlation of the rating R with MOS given the values for codec and packet loss from [8].

Table 1. Calculated values for packet delay variation equipment impairment factor

PDV	G.711	G.729
0 ms	0	0
50 ms	5	10
100 ms	20	20

As in the first experiment, we compute the linear correlation between each objective measure and MOS. The higher the absolute value of the correlation, the better the measure predicts human perceived speech quality.

4. RESULTS

The results of the first experiment are shown in Table 2. The two MNB algorithms and the two CSR measures gave very similar results, both achieving fairly high correlation with MOS. The EMBSD algorithm performed significantly worse than the others. In [4], the correlation betwen phoneme recognition rate and MOS was 0.816 ± 0.064 for speech samples from speakers on which the recognizer had been trained (eleven of the nineteen speakers) and 0.745 ± 0.074 for samples from the other speakers. This shows that the off-the-shelf CSR using word recognition gives substantially the same results as HTK using phoneme recognition.

Table 2. Correlation between MOS and objective measures for first experiment

Metric	Corr.
EMBSD	0.60
MNB(1)	0.77
MNB(2)	0.77
CSR(1)	0.75
CSR(2)	0.76

The correlation of the objective measures with MOS for the second experiment are shown in Table 3. The second column in Table 3 shows the correlation between the objective measures and MOS for the entire data set. Columns three and four show the correlation between the objective measures and MOS for data coded using the G.711 and G.729 codecs, respectively, over the entire range of packet loss and packet delay variation conditions. Clearly, the E-model gives the best results for the entire data set, and the other algorithms do not predict perceived voice quality as well, although the MNB algorithm and the E-model have nearly identical performance for the G.729 codec case. In particular, the CSR 1 and CSR 2 algorithms perform poorly. This is primarily due to difficulty in training the recognizer. This illustrates the sensitivity of this approach to the training set, a weakness to which the other algorithms are not subject. Plots of the results for the entire data set for the MNB 2 algorithm and the E-model are shown in Figures 1 and 2, respectively.

Table 3. Correlation between MOS and objective measures for VoIP data

Metric	Both codecs	G.711	G.729
EMBSD	0.38	0.52	0.39
MNB(1)	0.51	0.67	0.61
MNB(2)	0.54	0.67	0.63
E-model	0.70	0.86	0.62
CSR(1)	0.15	0.13	0.14
CSR(2)	0.11	0.03	0.15

Our analysis of the results showed significant difference in the correlation of the objective measures with MOS between the two codecs used in the experiment, G.711 and G.729. The third and fourth columns of Table 3 give these results. Plots of the E-model scores vs. MOS for the G.711 and G.729 codecs are shown in Figures 3 and 4, respectively. These results are discussed more fully in the next section.

5. DISCUSSION

As noted above, the first experiment served two purposes. The first is to establish that a commercially-available CSR doing word recognition would perform similarly to a CSR doing phoneme recognition. This it did. The correlation between word-error rates and MOS we measured was substantially the same as that between phoneme-error rates and MOS as measured in [4]. The second purpose was to test the MNB and EMBSD algorithms on the data set used in [4]. As shown in the results, the MNB algorithm metric had almost exactly the same correlation with MOS as the CSR (0.77 vs. 0.76). The EMBSD performed significantly worse, with a 0.60 correlation with MOS.

However, our primary intent was to investigate how effective objective speech measures are at predicting perceived speech quality for VoIP systems, and, as noted earlier, the codec and the error conditions used in the first experiment are not representative of VoIP systems. Therefore, we are most interested in the results of the second experiment, which used the G.711 and G.729 codecs and introduced distortions typical of an IP network. For the entire data set, the E-model proved to be the best predictor of perceived speech quality, as measured by the correlation between the rating produced by the model and MOS. Its correlation with MOS of 0.70 greatly exceeded that of the next best performer, the MNB 2 algorithm, which had a 0.54 correlation with MOS. EMSD had significantly lower correlation with MOS, 0.38. The CSR techniques did not perform well in the second experiment due to the difficulty in training the recognizer adequately, as noted above. This highlights the dependence of this technique on the training set.

Analyzing the results for the two codecs (G.711 and G.729) separately provides additional insights. All algorithms except the CSR ones showed a higher correlation with MOS for the G.711 codec than for the G.729 codec. Although the E-model rating is intended to be used as a means for comparing systems with different codecs, Figures 3 and 4 show that the relationship or mapping between E-model rating and MOS is different for the two codecs. The trendlines on the two figures reveal a much steeper slope in the mapping for the G.711 codec than for the G.729 codec.

One application of an objective speech measure is monitoring the quality of a conversation at a gateway. The computational simplicity of the E-model makes it ideal for this use. Additionally, the E-model's use of statistical descriptions of types of errors (i.e., percent packet loss, packet delay variation) suits this application well. These statistics are often collected and are readily available at gateways and endpoints. For online monitoring of a VoIP conversation, it is not necessary to have a highly accurate estimate of MOS over the entire range of 1 to 5. Rather, it is more important to know generally if the quality is excellent, fair, or unacceptable. Also important is knowing the cause of the distortion, in order to make adjustments such as changing the codec or the route.

In conclusion, we investigated several objective speech quality metrics in VoIP environments and found the ITU E-model to have the highest correlation with MOS. In addition, it is well suited to online monitoring applications, because it does not require the original signal in order to compute its metric, and it is computationally simple.

REFERENCES

1. S. Voran, Objective Estimation of Perceived Speech Quality- Part I: Development of the Measuring Normalizing Block Technique, *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 4 July 1999.
2. S. Voran, Objective Estimation of Perceived Speech Quality- Part II: Evaluation of the Measuring Normalizing Block Technique, *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 4 July 1999.
3. S. Voran, Objective Estimation of Perceived Speech Quality Using Measuring Normalizing Blocks, NTIA Report 98-347, 1998.
4. C. M. Chernick, S. Leigh, K. L. Mills, and R. Toense, Testing the Ability of Speech Recognizers to Measure the Effectiveness of Encoding Algorithms for Digital Speech Transmission, *Proceedings of MILCOM '99*.
5. Speech Group Tools and APIs `http://www.nist.gov/speech/tools`
6. N. O. Johannesson, The ETSI Computation Model: A Tool for Transmission Planning of Telephone Networks, *IEEE Communications Magazine* , Januray 1997.
7. International Telecommunications Union, Application of the e-model, a planning guide, ITU-T Recommendation G.108, 1999.
8. International Telecommunications Union, Transmission impairments, Appendix I: Provisional planning values for the equipment impairment factor I_e ITU-T Recommendation G.113, Appendix I, 1999.
9. International Telecommunications Union, Objective quality measurement of telephone-band (300-3400 Hz) speech codecs, ITU-T Recommendation P.861, 1998.
10. International Telecommunications Union, Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs ITU-T Recommendation P.830, 1996.
11. W. Yang, M. Benbouchta, and R. Yantorno Performance of the Modified Bark Spectral Distortion as an Objective Speech Quality Measure, *Proceedings ICASSP*, vol. 1, Seattle, 1998.
12. Speech Processing Lab `http://nimbus.temple.edu/~ryantorn/speech`
13. W. Yang, K. R. Krishnamachari, and R. Yantorno, Improvement of the MBSD Objective Speech Quality Measure Using TDMA Data, *Proceedings ICASSP*, vol. 2, 1999.

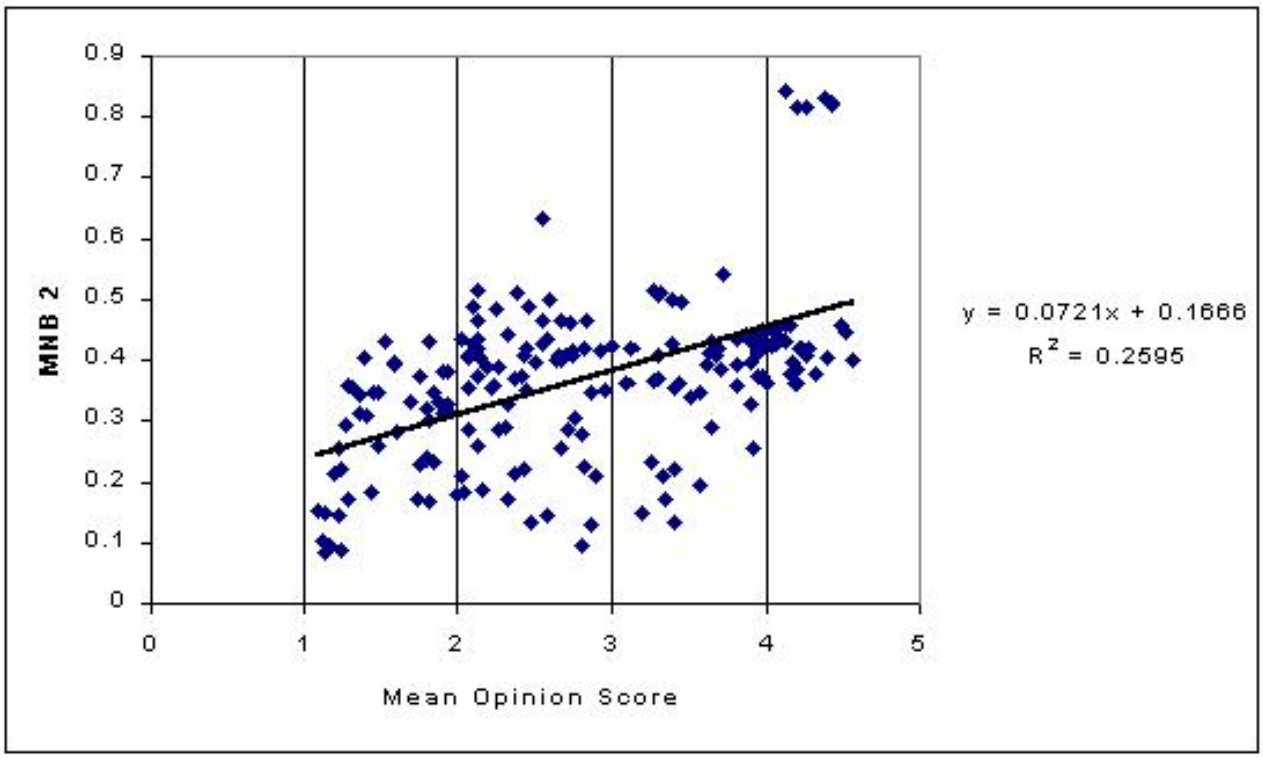

Figure 1. Plot of MNB-2 vs. MOS

14. ANSI Standard T1.801.04-1997, Multimedia Communications Delay, Synchronization, and Frame Rate Measurement, New York, 1997.
15. HTK Hidden Markov Model Toolkit – speech recognition research toolkit `http://htk.eng.cam.ac.uk`
16. Dragon Systems, Inc.`http://www.dragonsys.com`
17. Application and protocol testing through network emulation (the NIST network emulation tool) `http://www.antd.nist.gov/itg/nistnet`
18. Darpa TIMIT Acoustic-Phonetic Continuous Corpus CD-ROM, J. Garofolo, L. F. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, NISTIR 4930, February 1993.
19. J. Anderson, Methods for Measuring Perceptual Speech Quality, Agilent Technologies white paper. `http://onenetworks.comms.agilent.com/downloads/PerceptSpeech2.pdf`
20. The New Web Portal for Advanced Voice Quality Testing in Telecommunications `http://www.pesq.org`

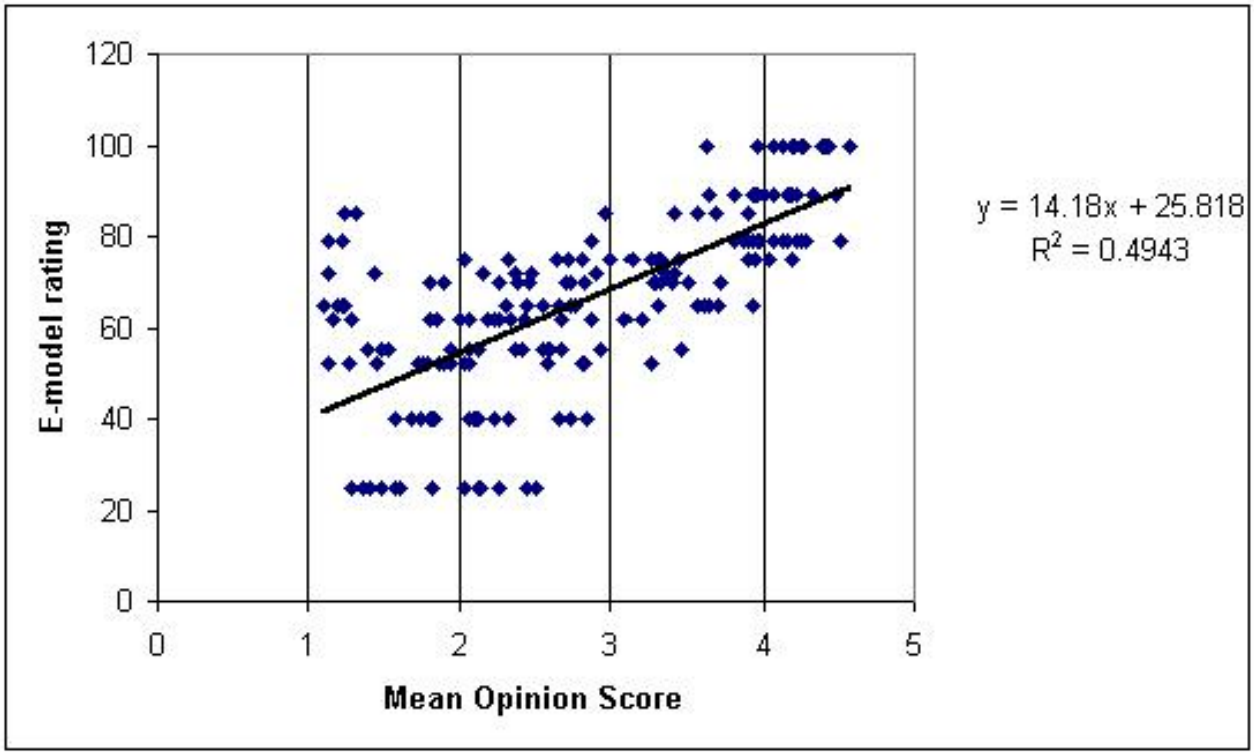

Figure 2. Plot of E-model rating vs. MOS

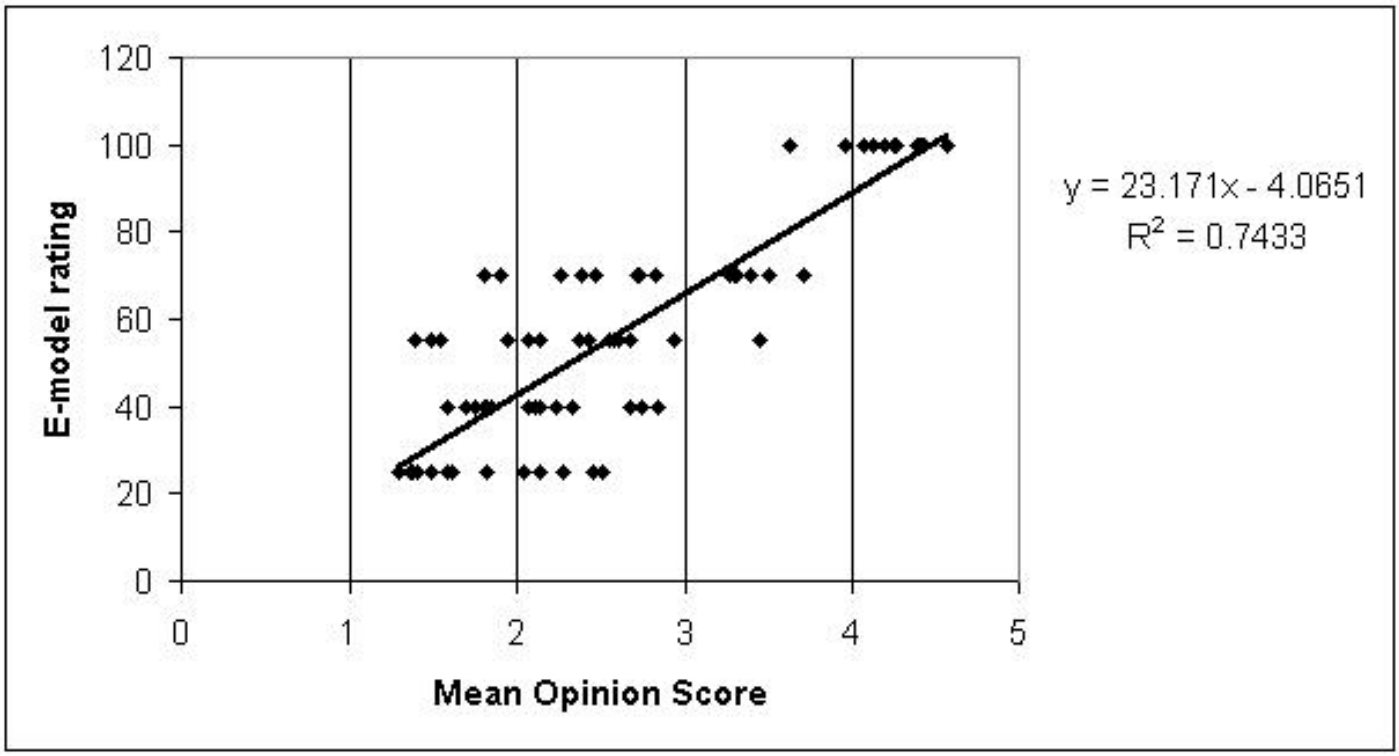

Figure 3. Plot of E-model rating vs. MOS for G.711 codec only

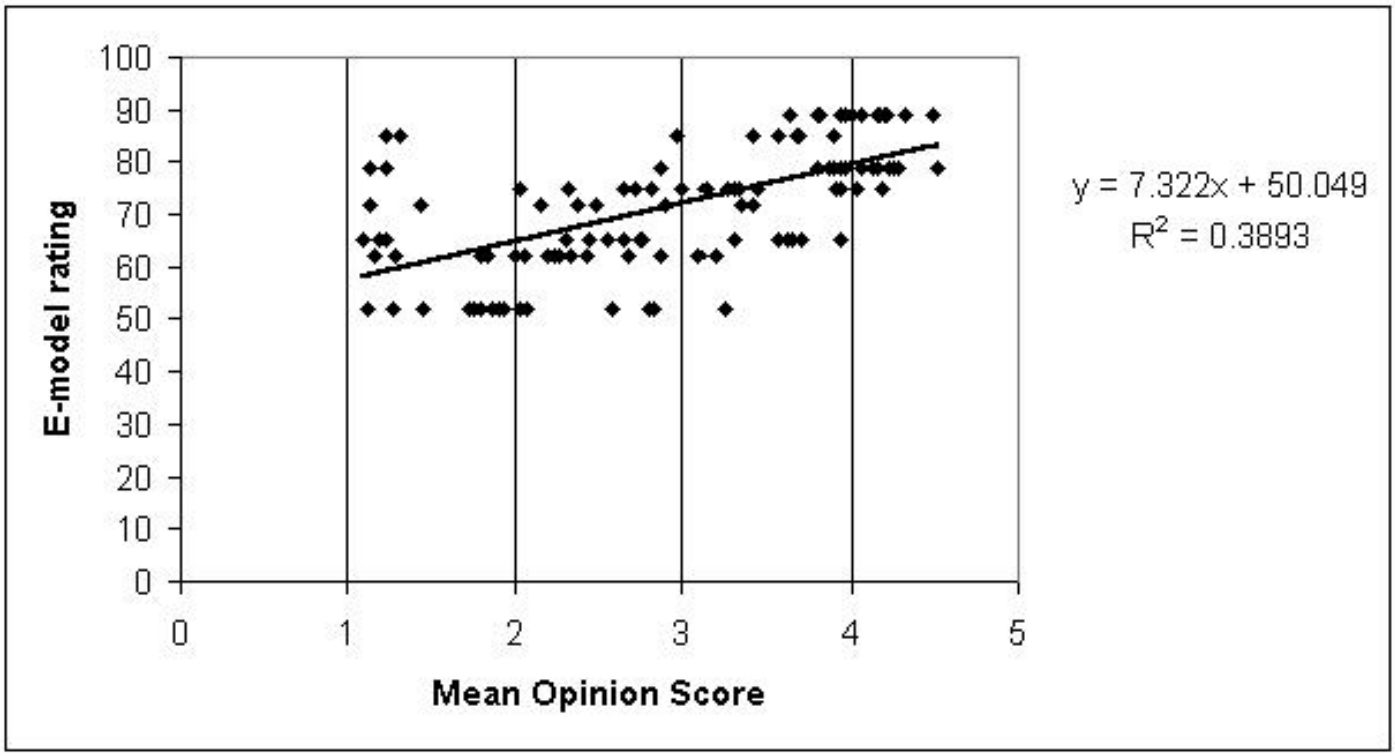

Figure 4. Plot of E-model rating vs. MOS for G.729 codec only