INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2009-2012

COM 12 – C 288 – E

October 2011

English only

Original: English

Question(s): 9/12

STUDY GROUP 12 – CONTRIBUTION 288

Source: Audience

Title: P.ONRA Contribution – Additional results from a candidate algorithm

ABSTRACT

There is a need in the industry for an accurate objective predictor of the performance of high-performance noise suppressors, standardized at ITU-T in the P.ONRA initiative. This contribution describes additional work extending an approach introduced in COM 12 – C 184 intended to predict SIG, BAK, and OVRL scores (SMOS_LQO, NMOS_LQO, and GMOS_LQO, respectively) obtained using the ITU-T P.835 methodology. These extensions include accommodation of non-stationary distracters and different noise suppressor strategies. Preliminary work on validation is presented. Further work is needed to extend the algorithm to explicitly handle voice processing apart from noise suppression such as speech codecs and time-varying dynamic range compression. Also, as this current version was developed based on narrowband data, further work is needed to collect wideband data and extend the algorithm accordingly.

1. Introduction

There is a need in the industry for an accurate objective predictor of the performance of high-performance noise suppressors, standardized at ITU-T. This contribution describes additional work on an algorithm first described in COM 12 – C 184 [1]. This work demonstrates early feasibility of predicting the SIG, BAK, and OVRL scores (SMOS_LQO, NMOS_LQO, and GMOS_LQO) obtained using the P.835 Amendment 1 Appendix III methodology [2].

2. Algorithmic Approach

The approach assumes the availability of the input signal (noisy mix) and output signal (noise-reduced speech) of the device under test, as well as the original speech signal and noise signal, as shown in Figure 1:

Contact: Scott Isabelle, Ph.D.
Audience Inc.
440 Clyde Avenue, Mountain View
CA 94043, USA

Tel: +1 650.224.2866
Fax: +1 650.254.1440
Email: sisabelle@audience.com

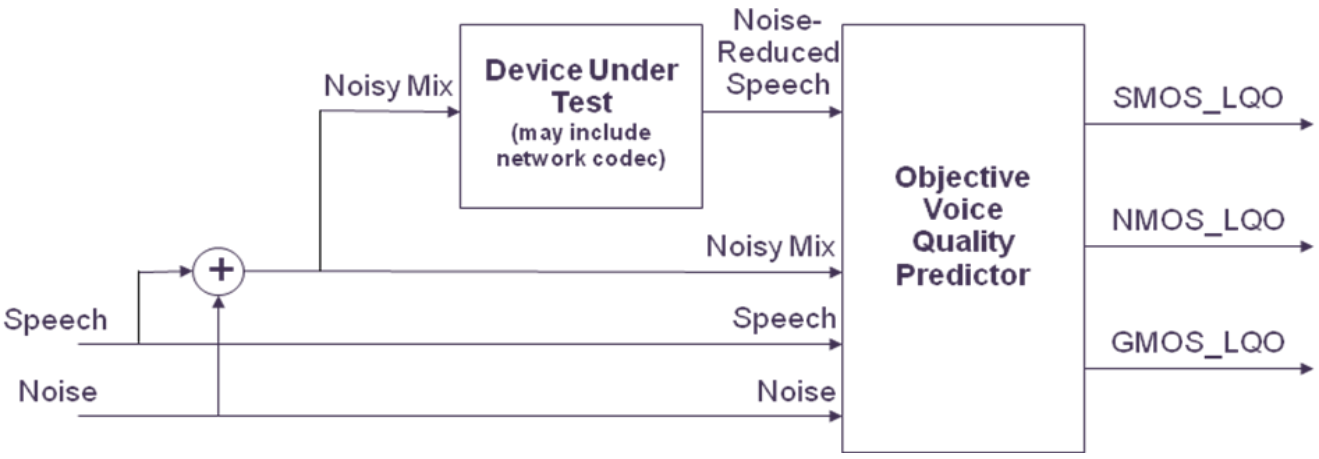

Figure 1: System Diagram

The Objective Voice Quality Predictor takes those four signals, and performs the following operations:

- Estimate the speech gain and noise attenuation from the Device Under Test,
- Construct a corresponding reference signal for an ideal noise suppressor (Estimated Idealized Noise-Reduced Reference, or EINRR),
- Compare the EINRR to the Noise-Reduced Speech to estimate the speech distortion and noise masking effects (used to predict SMOS_LQO),
- Compare the Noisy Mix to the Noise-Reduced Speech to determine the amount of noise suppression and noise distortion (used to predict NMOS_LQO).
- Combine the SMOS_LQO and NMOS_LQO and their constituent components to predict the overall score (GMOS_LQO).

3. Development Methodology

The training data previously described in COM 12 – C 184 comprised a range of input SNRs from 0 to 30dB for babble noise only, presented to one noise suppressor algorithm, operating over a range of fixed suppression levels from 0 to 35dB.

For this work, additional training data was collected for a set of eight noise types, including the six types defined in ITU-T P.835, Amendment 1 Appendix III. Five of the noise samples were taken from ETSI EG 202 396-1 [3]. Table 1 lists the names, descriptions, and filename from ETSI EG 202 396-1 if applicable. The SNR levels were 0, 6, 12, and 24 dB.

Table 1. Noise names and descriptions for training set

Noise Type Name	Description	EG 202 396-1 Filename
Mensa	Recording in a cafeteria	Mensa_binaural
Car	Recording at the driver's position	Fullsize_Car1_130kmh_binaural
Street	Recording at pavement	Outside_Traffic_Crossroads_binaural
Train	Recording at departure platform	Train_Station_binaural
School	Recording beside schoolyard	Schoolyard_Noise2_binaural
Music	Rock music, guitar and drums	n/a
Voice	Alternating male and female talker	n/a
Pink	Uncorrelated pink noise	n/a

The noise suppressor algorithm investigated here was a two-microphone hybrid system comprising a canceller followed by a fixed multiplicative suppressor. The canceller portion is implemented at two levels based on the distance between the two microphones, or Mic Spacing: 2-cm and 8-cm, where the former provides better noise reduction than the latter. The subsequent multiplicative suppressor stage is implemented at six fixed levels of Noise Suppression: 0, 6, 12, 18, 24, and 30dB.

The speech source for the P.835 tests training data was provided by Dynastat and included sixteen sentences, two from each of four male and four female talkers, all native speakers of American English. Four additional sentences were added to the beginning of the 16 test sentences to accommodate any convergence in processing. These 4 additional sentences were not used in listening tests or algorithm training.

For each noise type in Table 1, a P.835 listening test was conducted. Each test included 48 test conditions: 4 SNR x 2 Mic Spacing x 6 Noise Suppression.

The generation of conditions was simulated, in a manner similar to that described in COM 12 – C 184. Two sets of impulse responses were created, one for each level of Mic Spacing, by building two acoustic mock-up handsets, and measuring speech signal impulse responses from HATS artificial mouth to each microphone on the two devices. Impulse responses from the four loudspeakers in a test room consistent with ETSI EG 202 396-1 to each microphone on the two devices were also measured to obtain noise signal impulse responses. Input signals for the algorithm from Figure 1, clean speech, noise-alone, and noisy mix, were produced by convolution of speech and noise files with the appropriate impulse responses and mixing at the specified SNRs before processing by the noise reduction systems. No additional signal processing (e.g., speech codec) was applied in the test conditions for training data. All processing was performed at a sample rate of 8-kHz for narrowband speech.

Twelve reference conditions were included, based on the reference system proposed in AH-11-029 [4], which is intended as an improvement over the MNRU reference system for the SIG rating when used for P.835 evaluation of noise reduction systems.

In each test, 32 naïve native speakers of American English participated, listening monaurally at 79 dBSPL. A total of 128 votes were collected for each of the 60 conditions per test. The results from the School condition were a pilot test for the hybrid canceller/suppressor, covering a wider range of mic spacing, and so were not included in the final training set. Combined across the seven tests, excluding school, the new training database consists of 336 test conditions. These were added to the 72 test conditions reported in COM 12 – C 184 for a total training set size of 408 conditions.

4. Results – Training Set

The operations described in Section 2 above were performed on the four input audio signals for each of the 408 listening conditions, to determine the estimated values of speech distortion, noise distortion, noise masking, and noise suppression strength. A model fit was then performed to map those four extracted signal values to the desired outputs SMOS_LQO, NMOS_LQO, and GMOS_LQO. Figure 2 below shows the results of the model fit to the training data. Three sets of panels are shown, one for S-MOS (top), one for N-MOS (middle) and one for G-MOS (lower). In each set, there are columns for each noise type. In each set, the left-most panel is for the training results in babble, as reported in COM 12 – C 184. For each dimension (e.g., S-MOS), there are two rows of results, with the upper row for the 2-cm microphone spacing, and the lower row for the 8-cm mic spacing. Results are plotted as a function of the amount of noise suppression, with SNR coded by color: blue for 0 dB; green for 6 dB; red for 12 dB; and magenta for 24 dB. Thin lines with error bars show subjective results; thick lines with open symbols show model fits.

A simple linear remapping, derived from the reference conditions only, was applied to the subjective scores prior to fitting the model. The remapping was based on common practice as used by the Global Analysis Lab in subjective studies, and is described in the Appendix.

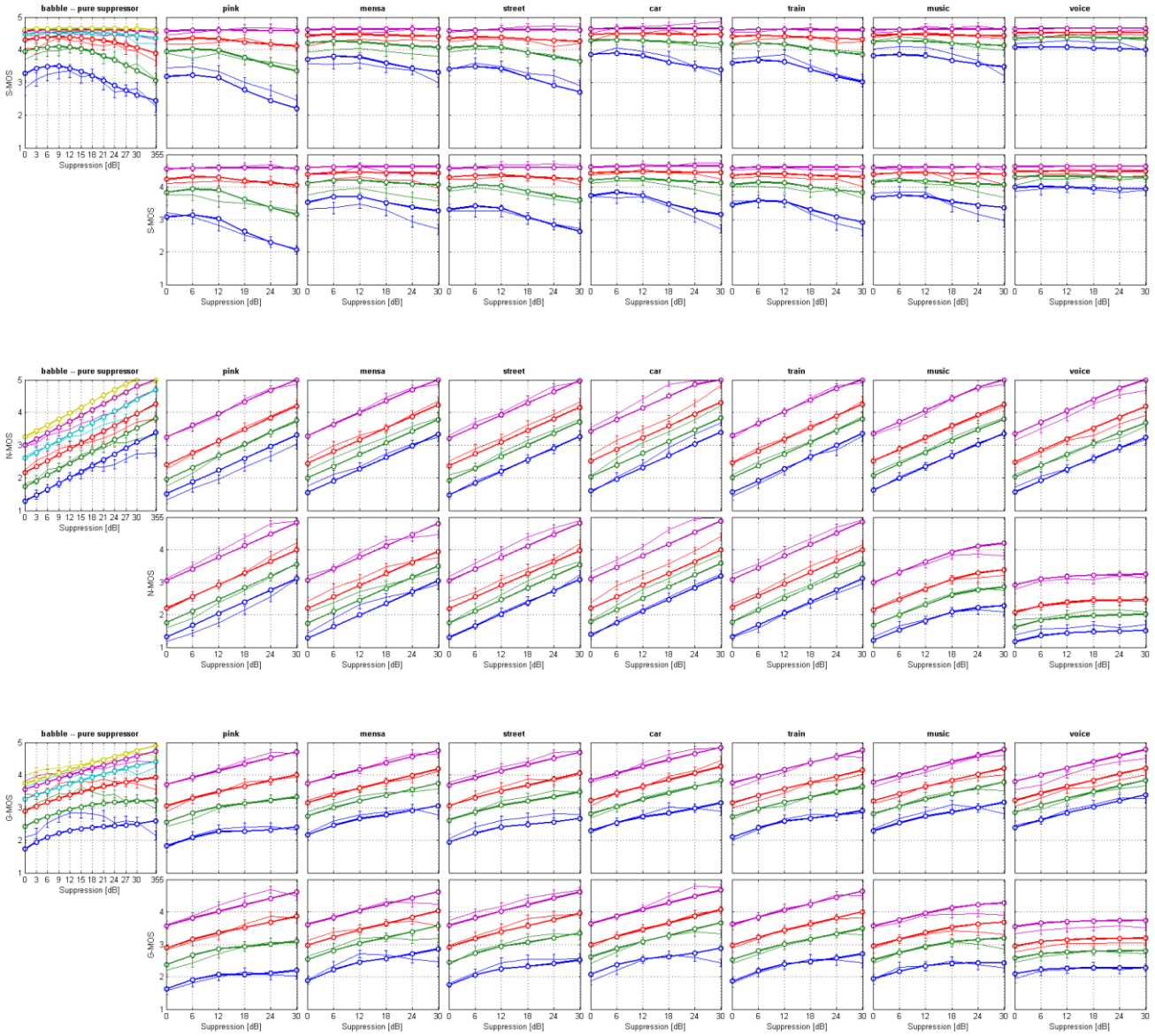

Figure 2: Model Fit on training data. Thin lines with error bars are the subjective scores. Bold lines with circle points are the predictions. Upper panel for S-MOS, middle panel for N-MOS, lower panel for G-MOS. SNR values are coded by color: Blue for 0 dB; Green for 6 dB; Red for 12 dB; and magenta for 24 dB.

The predictions above show that the extracted signals can be used to accurately predict the subjective responses to the audio samples, within approximately +/- 0.25 MOS absolute accuracy in general.

The same data can be re-plotted in the familiar scatter-plot format, as shown below in Figure 3a (S-MOS), 3b (N-MOS), and 3c (G-MOS).

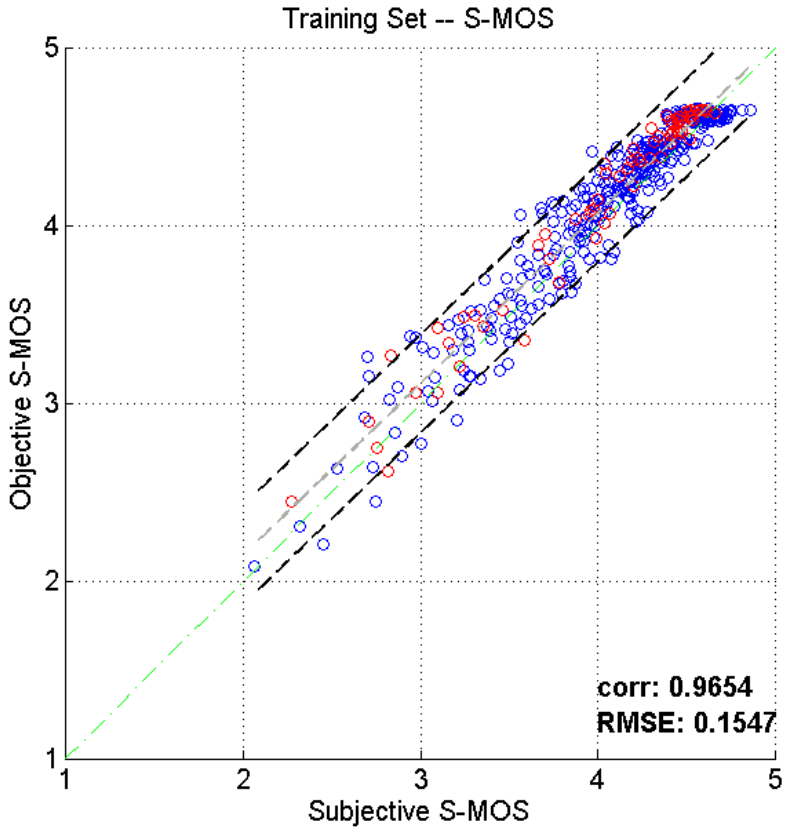

Figure 3a: Model Fit on Training Data – Scatter-plot format, S-MOS. Red symbols are for the pure suppressor. The dashed grey line shows the best linear fit.

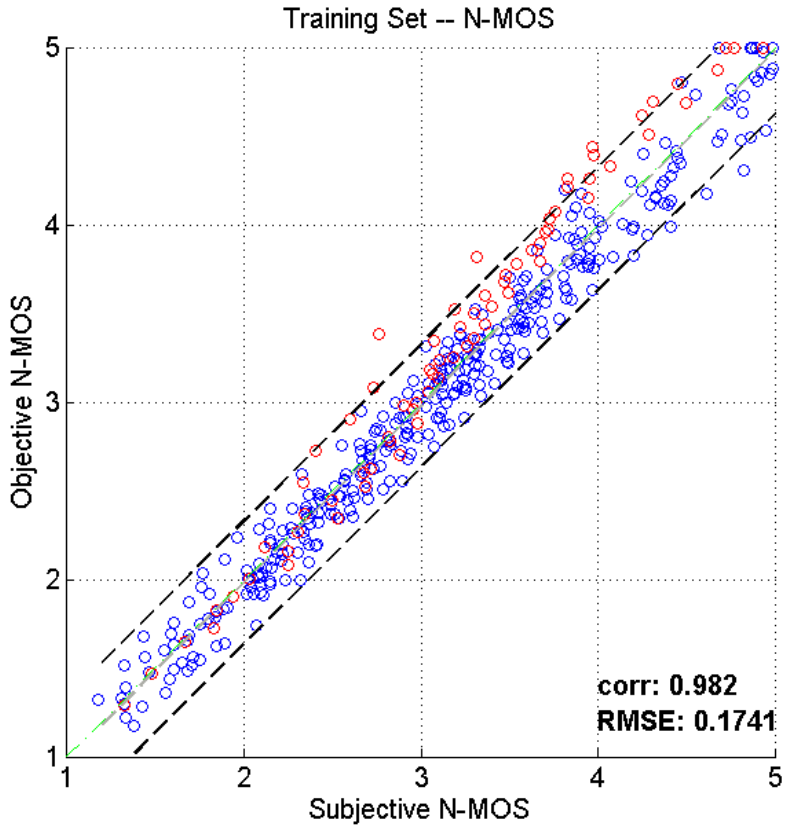

Figure 3b: Model Fit on Training Data – Scatter-plot format, N-MOS. Red symbols are for the pure suppressor. The dashed grey line shows the best linear fit.

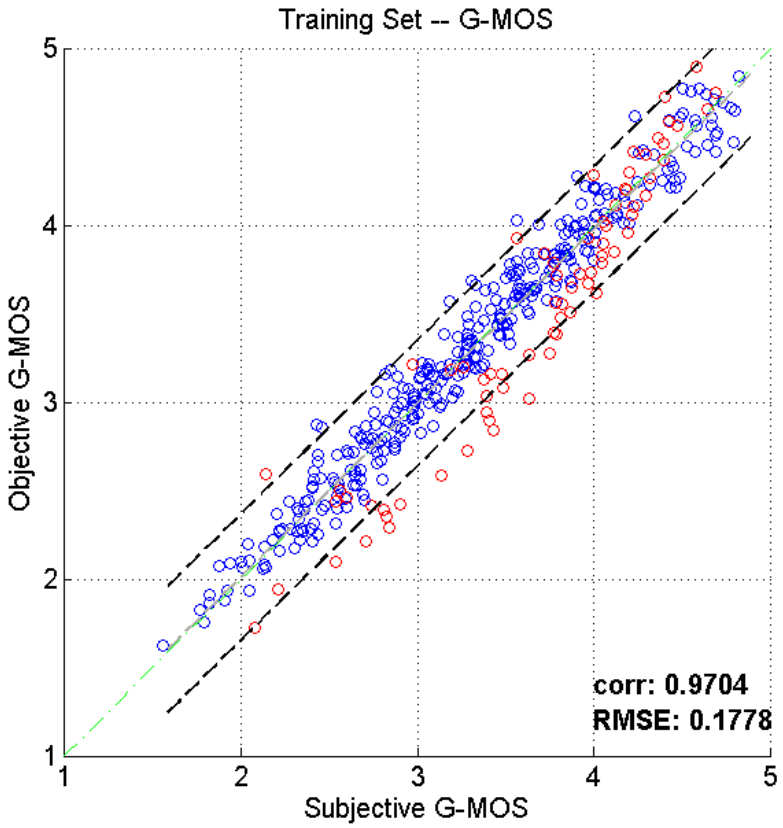

Figure 3c: Model Fit on Training Data – Scatter-plot format, G-MOS. Red symbols are for the pure suppressor. The dashed grey line shows the best linear fit.

The results for the pure suppressor, from COM 12 – C 184, are color-coded separately, as the subjective test conditions for these differed somewhat from the eight training tests described in Table 1. These results were obtained using a different speech sample. Also, the MRNU reference system was used for that subset.

The fit to the training set is generally fairly good, with correlation of 0.97 to 0.98 and RMSE of 0.15 to 0.18 across the 408 training conditions. As a subset, the fit is slightly less good on the 72 conditions reported in COM 12 – C 184. Note that because the reference conditions were different, the remapping described in the Appendix was not applied to these data.

5. Results – Preliminary Validation Set

A validation dataset was collected for seven commercially available narrowband handsets. Three noise types were tested as listed in Table 2.

Table 2 Noise names and descriptions for validation set

Noise Type Name	Description	EG 202 396-1 Filename
Babble	Recording in a pub	Pub_Noise_binaural_V2
Car	Recording at the driver's position	Fullsize_Car1_130kmh_binaural
Music	Rock music, guitar and drums	n/a

The speech source for the validation data was different from that used in training, and was also provided by Dynastat. It consists of 32 sentences, 4 from each of 4 male and 4 female talkers, all native speakers of American English. Each sentence was normalized to -26 dBov Active Speech Level. Four additional sentences were added to the beginning of the 32 test sentences to accommodate any convergence in processing; these 4 sentences were not used in listening tests or algorithm validation.

The room set up used for acoustic reproduction of noise and speech is consistent with ETSI EG 202 396-1, and as described in P.835 Amendment 1 Appendix III. The speech was played through an equalized artificial mouth of HATS, at a level of -4.7dBPa at MRP. Two SNRs were used, 3 and 18dB, with the speech level measured according to P.56, and with A-weighting for the noise level. For each handset, the SNR values were set by adjusting the noise level at the primary microphone of the device. Narrowband calls were simulated using a Rohde & Schwarz CMU-200, with speech service provided by AMR-NB codec at 12.2kbps mode rate.

The required signals were captured acoustically at the primary microphone of each device under test, and electrically from the output of the CMU-200. For each device, the output to clean speech was used to estimate the sending frequency characteristic, which then was used to filter the noisy mix. For each device, the time delay of the output signal was estimated using cross-correlation with the input signal, and used to time-align the output signal with respect to the input, prior to processing by the model.

The results for the validation set are shown as scatter plots, in Figures 4a (S-MOS), 4b (N-MOS), and 4c (G-MOS).

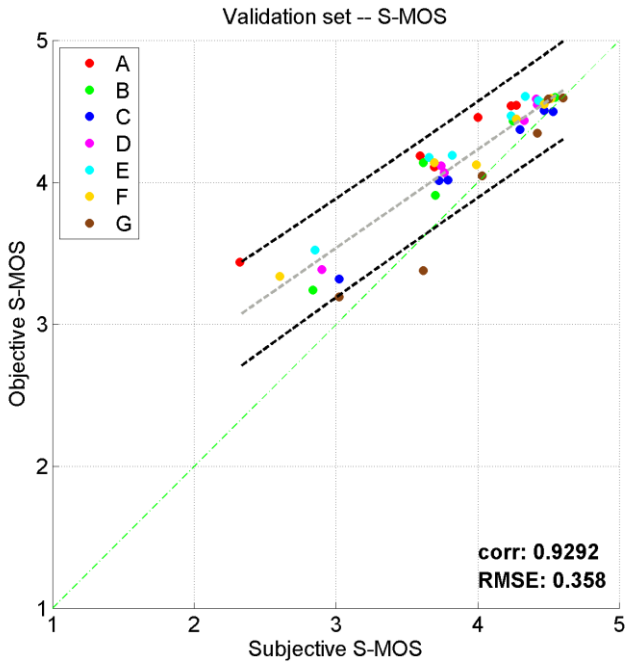

Figure 4a: Validation results, S-MOS, for seven phones under conditions in Table 2. Grey dashed line is best linear fit.

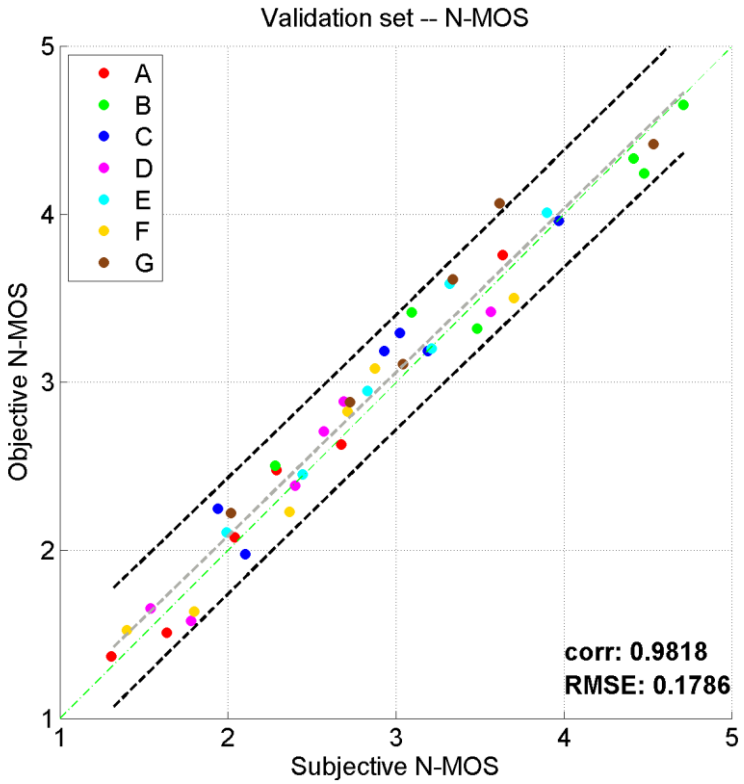

Figure 4b: Validation results, N-MOS, for seven phones under conditions in Table 2. Grey dashed line is best linear fit.

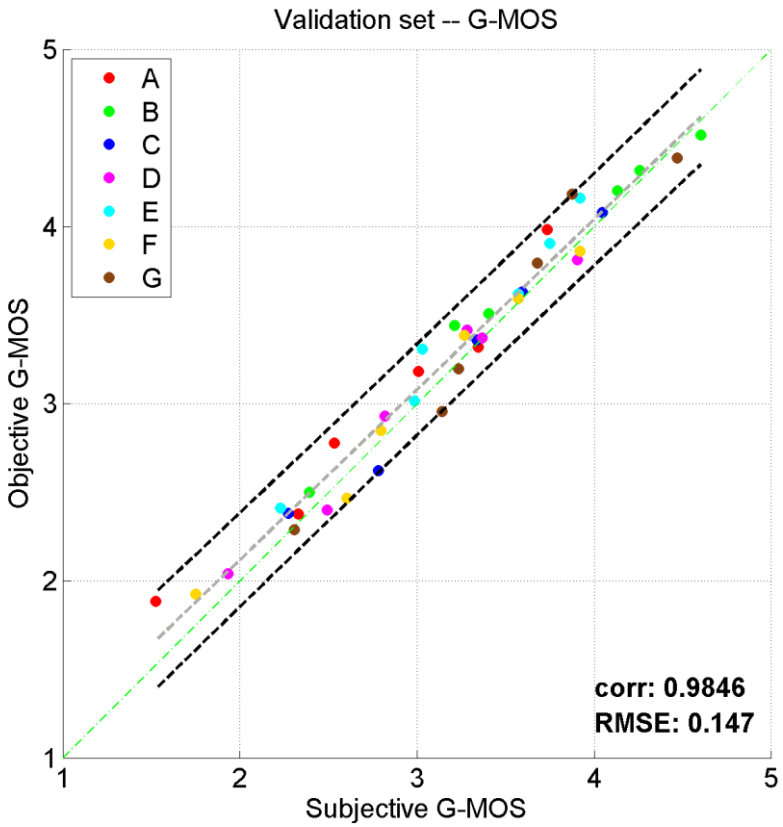

Figure 4c: Validation results, G-MOS, for seven phones under conditions in Table 2. Grey dashed line is best linear fit.

6. Further Work Required

The model appears to have sufficiently good accuracy on the training set, but it is clearly not yet completely adequate on real devices, particularly for S-MOS. For some real devices, the results show an offset which has not yet been accounted for. The immediate future work is to determine the source of the offset and build that into the model. To do that, more controlled validation data will be needed with a larger variety of devices.

As noted earlier, this version of the algorithm does not yet explicitly include features intended to account for aspects of voice processing apart from noise suppression. Such processing would include speech codecs and time-varying gain such as multi-band dynamic range compression. The preliminary validation shows fairly good performance in the presence of one speech codec, AMR-NB 12.2kbps, and for real devices that likely incorporate processing in addition to noise suppression.

Finally, the dataset and algorithm reported here and earlier are narrowband. Extension to wideband is clearly necessary to support deployed wideband telephony systems.

7. Summary

There is a need in the industry for an accurate objective predictor of the performance of high-performance noise suppressors, standardized at ITU-T. This contribution demonstrates feasibility of an approach that can predict SIG, BAK, and OVRL scores (SMOS_LQO, NMOS_LQO, and GMOS_LQO) obtained using the P.835 Amendment 1 Appendix III methodology, with both quasi-stationary and non-stationary distracters at SNRs of 0, 6, 12, and 24dB, with an accuracy of +/- 0.2 MOS on the training set (408 points with a hybrid canceller followed by a constant spectral

subtraction-type suppressor). Reduced absolute accuracy but good monotonicity properties are demonstrated on the preliminary validation set (42 points with a variety of non-constant suppressor strategies implemented in commercially available devices). Further work is needed to collect larger validation data sets and extend to wideband.

References

[1] COM 12 – C 184, P.ONRA contribution – preliminary results from a candidate algorithm. Geneva, January 2011, Geneva, Switzerland.

[2] P.835 Amendment 1, Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. Amendment 1: New Appendix III – Additional provisions for nonstationary noise suppressors (10/2007).

[3] ETSI EG 202 396-1 V1.2.4, Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database. (11/2010).

[4] AH-11-029, Better Reference System for the P.835 SIG Rating Scale, 20-21 June 2011, Geneva, Switzerland.

Appendix: Remapping based on reference conditions

The eight data training data sets were each obtained with different listening panels. Some differences in response patterns can be identified by examining scores for the reference conditions. Figure A1 shows the scores across all eight panels for the reference conditions where only the Noise Suppressor reference is varying, and background noise is not added. Note that the BAK scores tend to be quite high, even at the most distorted NS Levels. This is in contrast to behavior observed for MNRU references, and is the motivation for the proposed NS reference system described in [4].

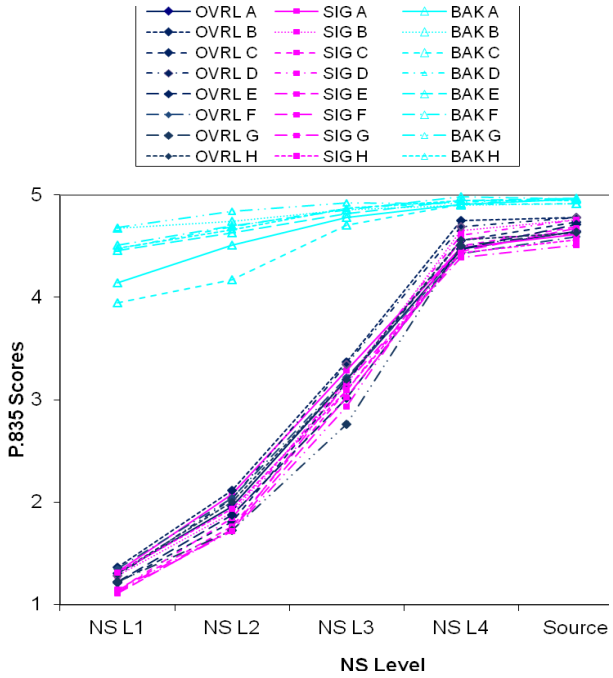

Figure A1. Scores across eight panels, NS Level varies, no additive noise.

Similarly, Figure A2 shows the scores across all eight panels for the reference conditions where pink noise is added. The reduction in SIG at low levels of noise reflects the noise masking noted in Figure 2 above.

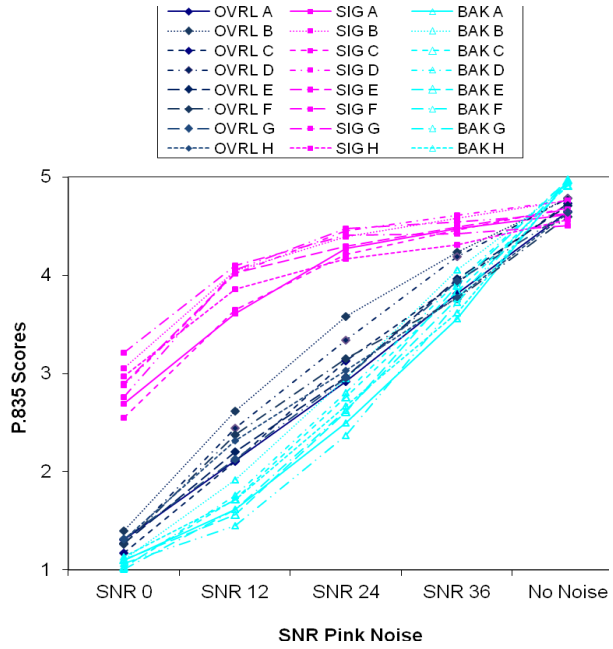

Figure A2. Scores across eight panels, additive noise varies, no NS degradation.

Finally, Figure A3 shows the scores across all eight panels for the reference conditions where NS level and noise co-vary.

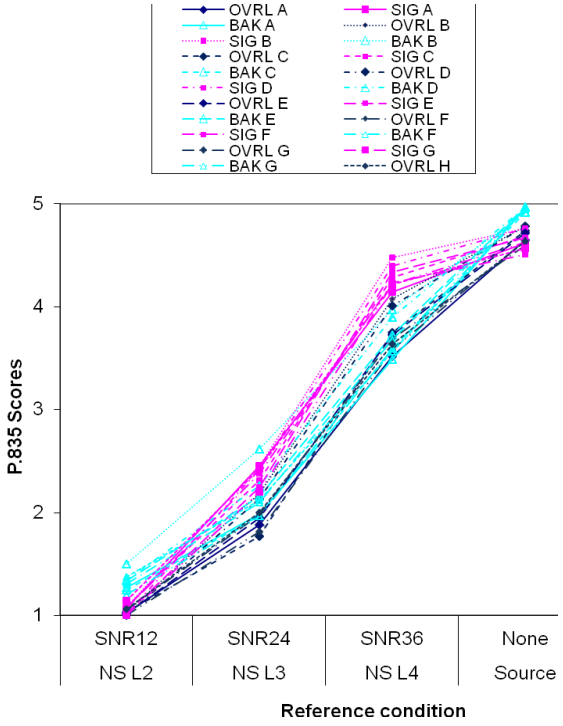

Figure A3. Scores across eight panels, additive noise and NS level co-vary.

While the trends across all panels are consistent, there are some variations between panels. Standard practice in such cases is to treat the variation as random. The simplest remapping is to compute the mean scores for reference conditions across all panels, and then find a linear remapping based on the differences between each panel's responses to reference conditions and the mean response to reference conditions across panels. A remapping is computed separately for SIG, BAK, and OVRL. The same remapping is then applied to responses to test conditions for each panel. No other remappings are used.

This approach is commonly used by Global Analysis Labs charged with combining and analyzing results from multiple Test Labs. While it does require that the reference conditions be common to all tests, it has the advantage of being well-defined and based on observations, rather than approaches that are purely ad hoc or based on hypothesized constructs.

An example of the effect of the remapping is shown in Figure A4, as a scatter plot for G-MOS (OVRL) with mapped scores plotted against raw scores.

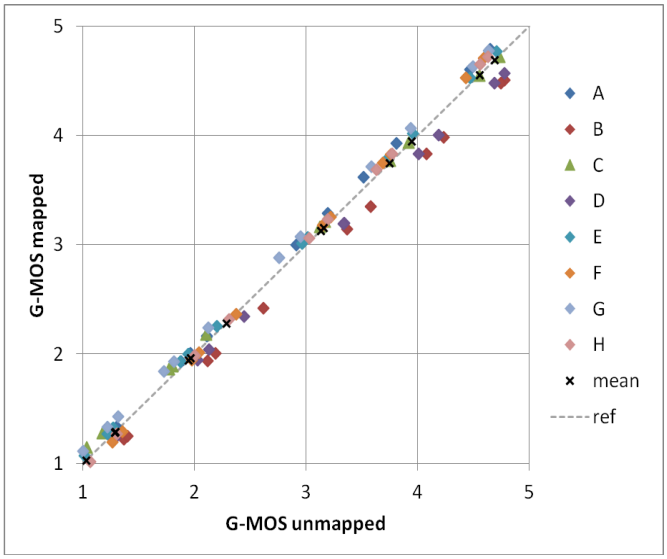

Figure A4. Scatter plot of mapped versus unmapped scores for reference conditions.

As can be seen in Figure A4, the remapping does not affect the mean across-panel ratings. The linear mapping can be seen to generally reduce the overall variation across panels. For scores near limits (1 or 5), the remapping can, in some cases, produce results that would exceed limits, but in these cases the bounding value is used.

_____