INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2009-2012

COM 12 – C 184 – E

January 2011

English only

Original: English

Question(s): 9/12

STUDY GROUP 12 – CONTRIBUTION 184

Source: Audience

Title: P.ONRA contribution – preliminary results from 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 demonstrates early feasibility of an approach that can predict SIG, BAK, and OVRL scores (SMOS_LQO, NMOS_LQO, and GMOS_LQO, respectively) obtained using the P.835 Amendment 1 Appendix III methodology, at SNRs of 0, 6, and 12dB, with an accuracy of +/- 0.2 MOS on the training set (72 points with a constant spectral subtraction-type suppressor using the babble distractor), with reduced absolute accuracy but with good monotonicity properties demonstrated on the preliminary validation set (36 points with a variety of non-constant suppressor strategies with babble, street, and pink noise distractors). Further work is needed to collect larger training and validation data sets, and to extend the algorithm to explicitly handle non-stationary distractors and different noise suppressor strategies such as cancellers and hybrid cancellers/suppressors with time-varying suppression.

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 a candidate approach that 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, with quasi-stationary distractors at SNRs of 0, 6, and 12dB.

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: Lloyd Watts, Ph.D.
Audience Inc.
USA

Tel: +1 650.224.4419
Fax: +1 650.254.1440
Email: lwatts@audience.com

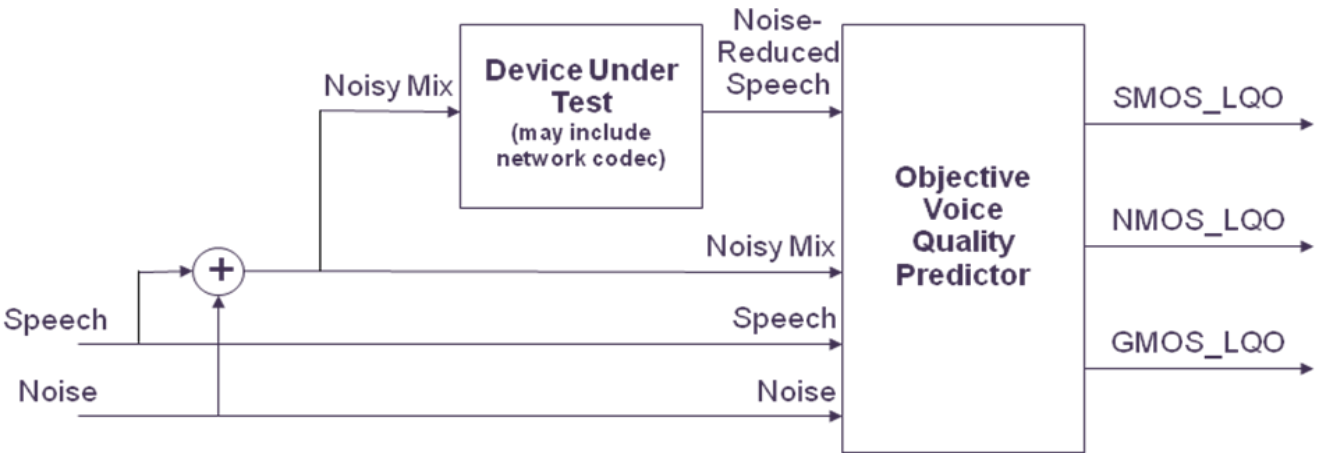


Figure 1: System Diagram

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

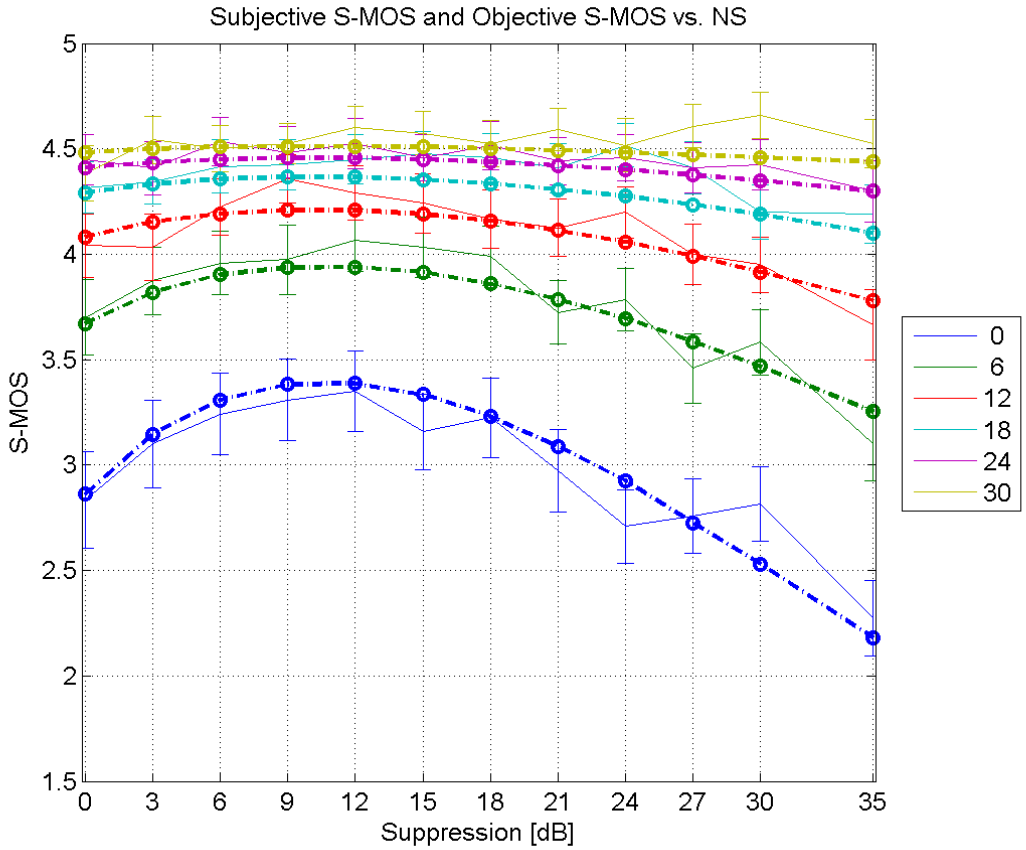
Audio data and Subjective Listening Scores were collected for a broad-ranging sweep of input SNRs and Noise Suppression strength, with SNRs ranging from 0 to 30 dB in 6 dB steps (6 conditions), and NS values ranging from 0 to 30 dB in 3 dB steps, and 35 dB (12 conditions), for a total of 6 x 12 = 72 conditions. Noise Suppression was held constant throughout a given test, and no post-processors (AGC, Post Equalizers) were active. The Noise Suppressor under test was a two-microphone noise suppressor using a multiplicative spectral energy suppression strategy (i.e. pure suppression, no cancellation). A P.835 Amendment I Appendix III methodology was followed, with a listener panel of 32 naïve listeners. The noise source for the training sweep was babble only.

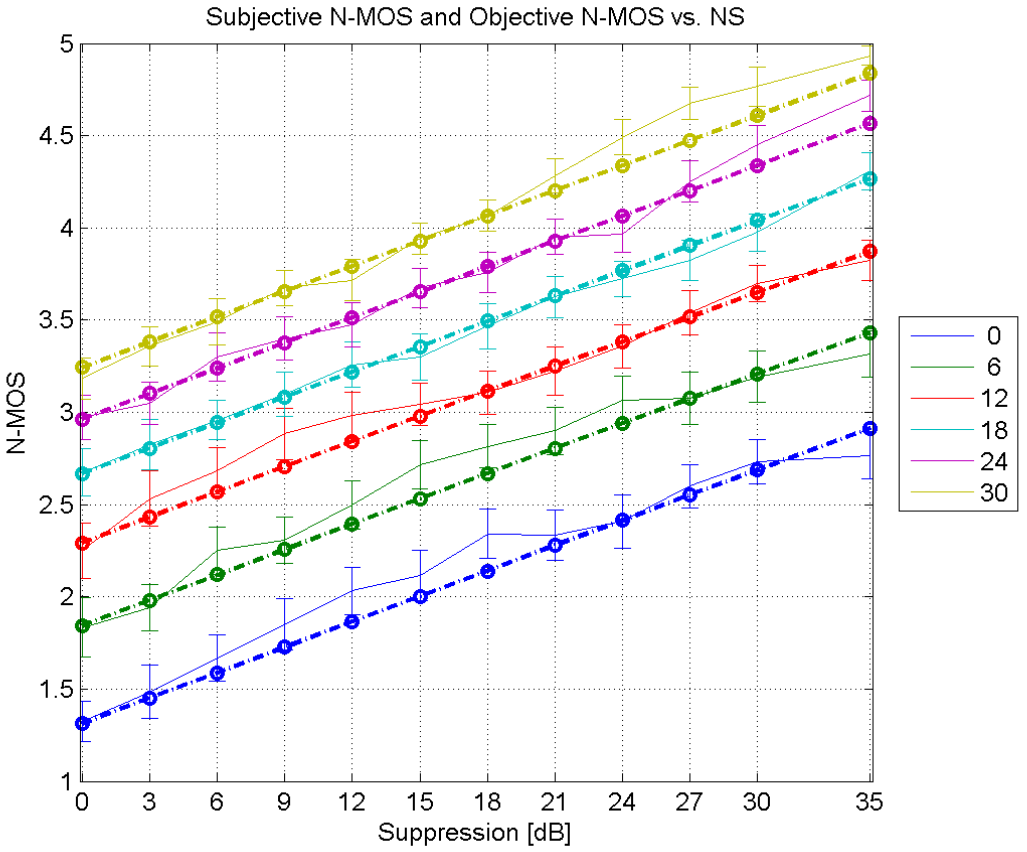
For preliminary validation, other quasi-stationary sources were also considered (street, pink noise), using limited available data collected for other purposes with non-constant suppression and other suppression strategies, including hybrid canceller/suppressors. For this reason, the existing algorithm developed on the training set is not expected to give high absolute accuracy on the validation set – what we are looking for at this early development stage is monotonic performance, as described in the section below.

4. Results – Training Set

The operations described above were performed on the four audio signals for each of the 72 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. Below are the results of the model fit:




Subjective N-MOS and Objective N-MOS vs. NS
N-MOS
Suppression [dB]
0
6
12
18
24
30

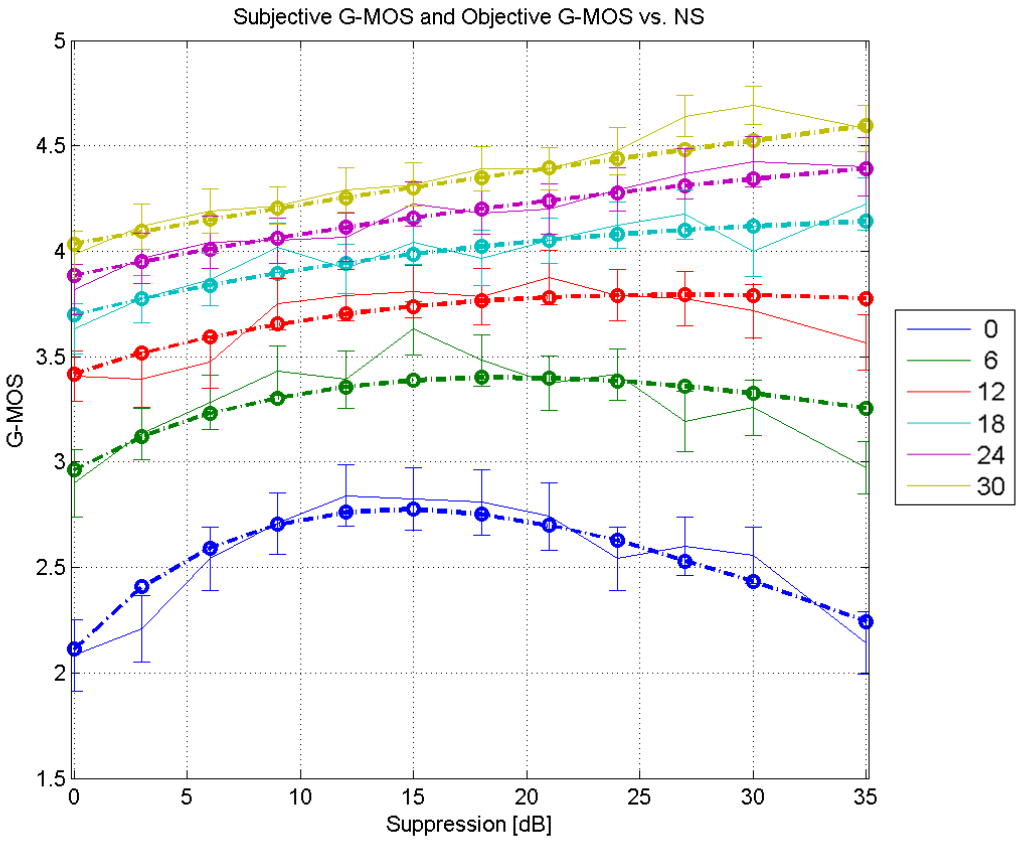
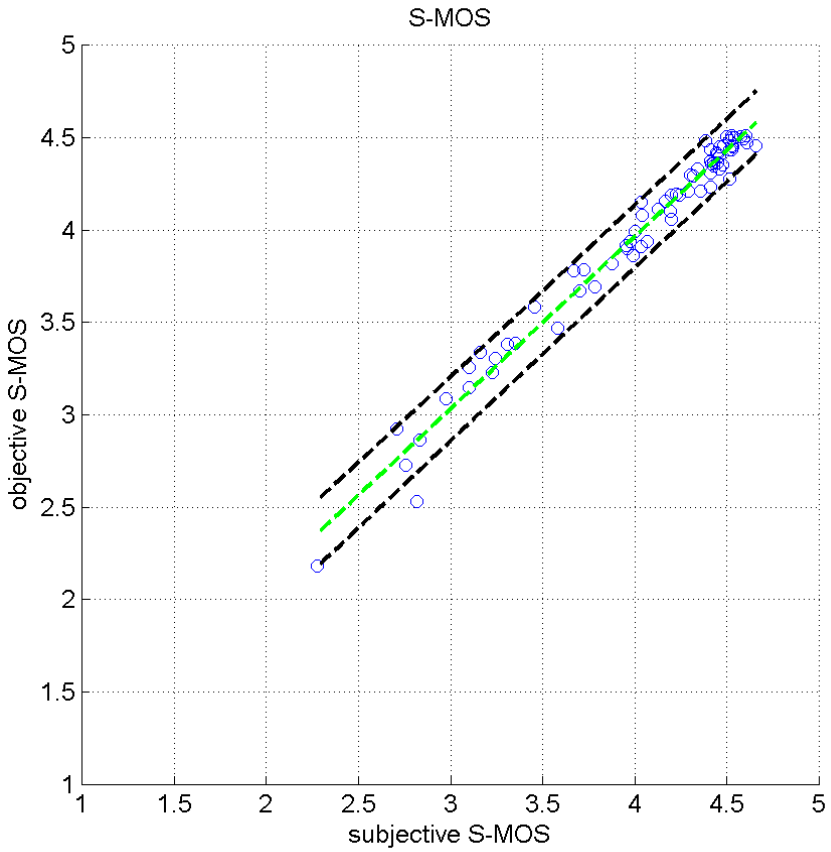
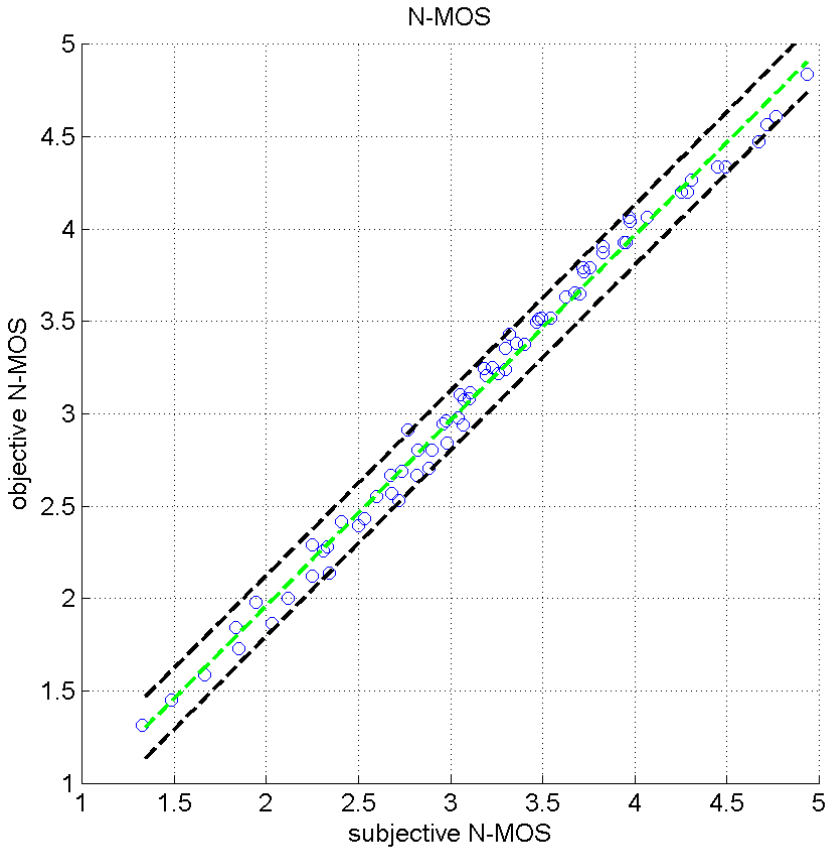

Subjective G-MOS and Objective G-MOS vs. NS
G-MOS
Suppression [dB]
0
6
12
18
24
30

Figure 2: Model Fit on Training Data. Thin lines with error bars are the subjective scores; Bold lines with circle points are the predictions.

The predictions above show that the extracted signals can be used to accurately predict the subjective responses to the audio samples, within approximately +/- 0.2 MOS absolute accuracy.

The same data can be re-plotted in the familiar scatter-plot format, as shown below in Figure 3:


S-MOS
objective S-MOS
subjective S-MOS
1
1.5
2
2.5
3
3.5
4
4.5
5


N-MOS
objective N-MOS
subjective N-MOS
1
1.5
2
2.5
3
3.5
4
4.5
5

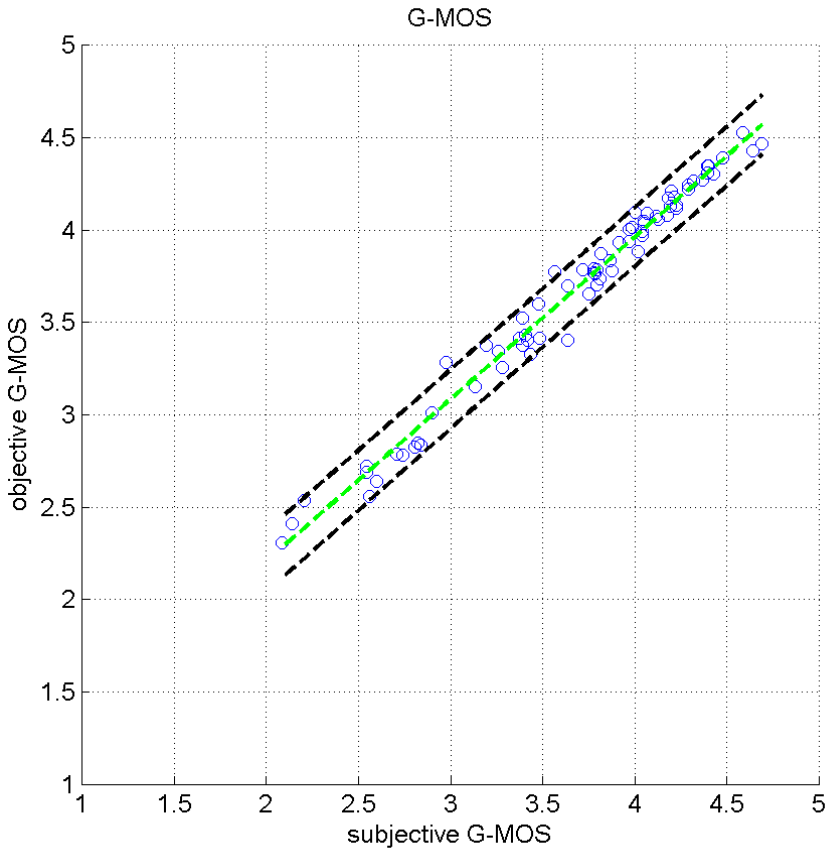
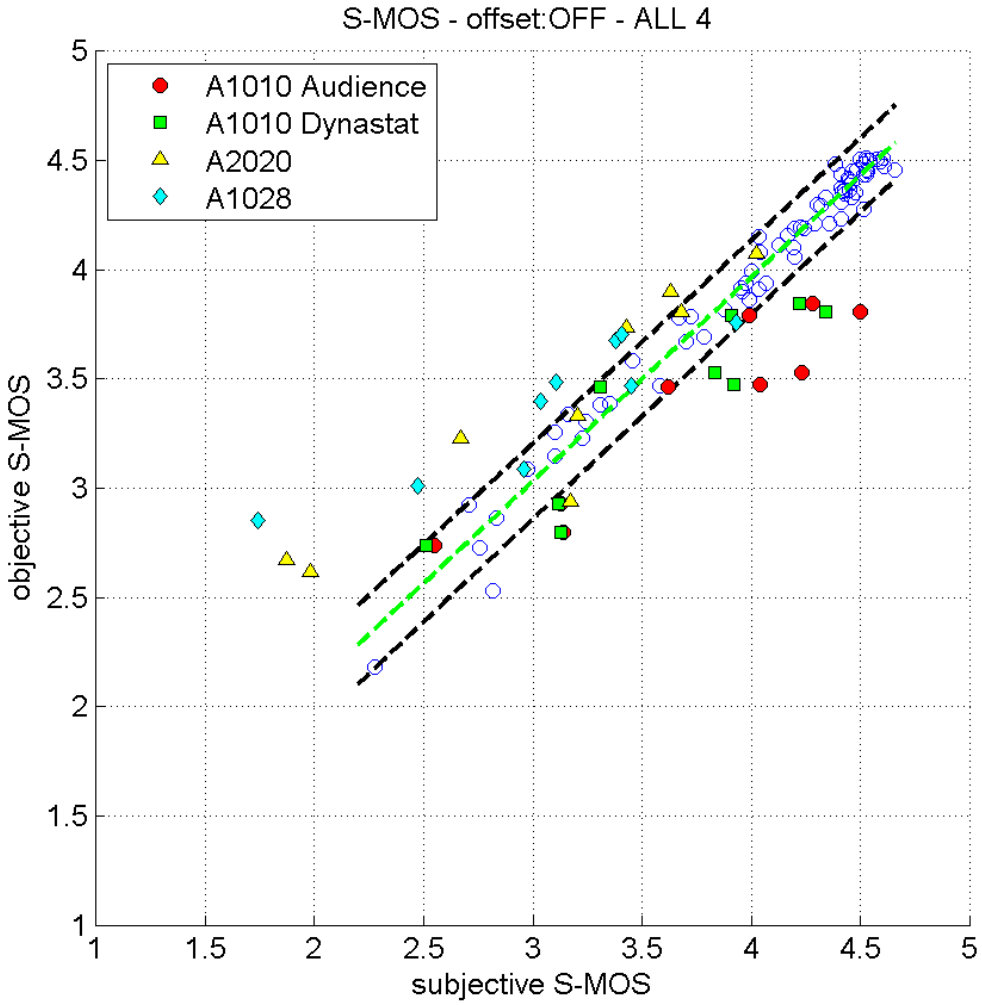
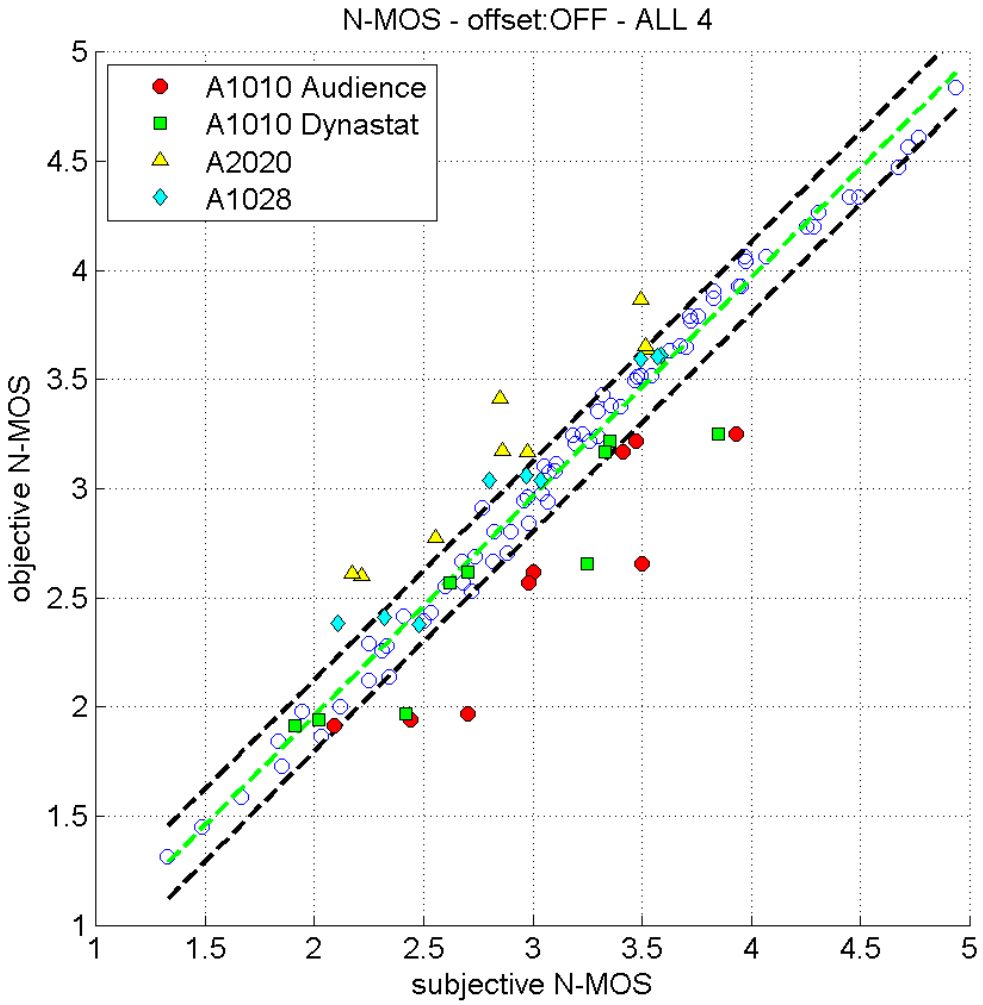


Figure 3: Model Fit on Training Data – Scatter-plot format.

5. Results – Preliminary Validation Set

In Figure 4 below, we show the preliminary validation results on all the available data for three quasi-stationary sources (babble, street, and pink noise), for three non-constant suppressors evaluated in four tests (A1010 with Audience listener panel, A1010 with Dynastat listener panel, A1028, and A2020), and three SNRs (0, 6, 12dB), for a total of 3 x 4 x 3 = 36 test conditions.




N-MOS - offset:OFF - ALL 4
A1010 Audience
A1010 Dynastat
A2020
A1028
objective N-MOS
subjective N-MOS
1
1.5
2
2.5
3
3.5
4
4.5
5

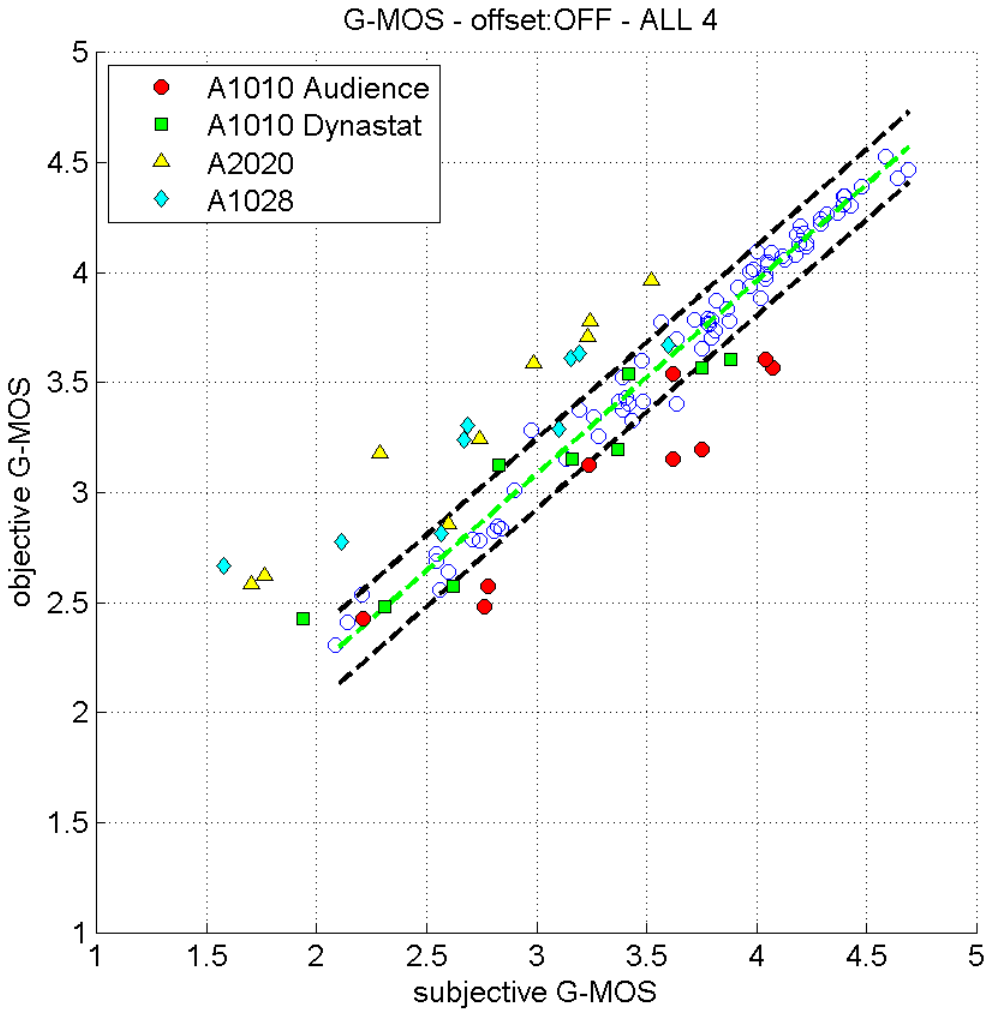


Figure 4: Preliminary Validation Results with multiple suppressor types and quasi-stationary distractors (babble, street, and pink noise).

The validation data clearly shows a larger error (poorer absolute accuracy) than the training data. However, there is an import trend in the pattern of errors that suggests that, for a given suppressor and distractor type, the prediction is monotonically related to the subjective score, i.e., in a given triplet of points (SNR=0, 6, 12dB), when the subjective scores increase, the prediction also increases, and by approximately the same amount. The absolute error comes from a consistent offset for the given triplet of points for suppressor and distractor type – not surprising at this early stage of development, since the model has been developed only for a pure suppressor (no canceller) and babble noise only.

6. Further Work Required

The model appears to have sufficiently good accuracy on the training set, but it is clearly not yet adequate on suppressor types and distractor types for which it has not been trained. These variables appear to introduce 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 data will be needed with a variety of suppressor types and distractor types.

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 early 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 quasi-stationary distractors at SNRs of 0, 6, and 12dB, with an accuracy of +/- 0.2 MOS on the training set (72 points with a constant spectral subtraction-type suppressor), with reduced absolute accuracy but good monotonicity properties demonstrated on the preliminary validation set (36 points with a variety of non-constant suppressor strategies). Further work is needed to collect larger training and validation data sets, and to extend the algorithm to explicitly handle non-stationary distractors and different noise suppressor strategies such as cancellers and hybrid cancellers/suppressors.