2015-01-2335
Published 06/15/2015

doi:10.4271/2015-01-2335
saepcmech.saejournals.org

Quantifying Hands-Free Call Quality in an Automobile

Scott Amman, Francois Charette, Paul Nicastri, John Huber, Brigitte Richardson, Gint Puskorius, Yuksel Gur, and Anthony Cooprider
Ford Motor Co.

ABSTRACT

Hands-free phone use is the most utilized use case for vehicles equipped with infotainment systems with external microphones that support connection to phones and implement speech recognition. Critically then, achieving hands-free phone call quality in a vehicle is problematic due to the extremely noisy nature of the vehicle environment. Noise generated by wind, mechanical and structural, tire to road, passengers, engine/exhaust, HVAC air pressure and flow are all significant contributors and sources of noise. Other factors influencing the quality of the phone call include microphone placement, cabin acoustics, seat position of the talker, noise reduction of the hands-free system, etc. This paper describes the work done to develop procedures and metrics to quantify the effects that influence the hands-free phone call quality. It will be shown that a listening study of using 49 evaluators, indicated that the ETSI EG 202 396-3EG (VoIP Standard) for SMOS (Speech Mean Opinion Score) and NMOS (Noise Mean Opinion Score) correlates better than the ETSI TS 103 106 (Mobile Standard) for speech and noise ratings when quantifying the quality of a hands-free phone call. However, ETSI TS 103 106 was found to correlate better for GMOS (Global Mean Opinion Score). Using these results, MOS scores can be calculated to investigate the influences of factors that can influence the quality of a hands-free call. Two examples are given in this paper. The first example investigates the effect of background noise level in the vehicle cabin on hands-free phone call quality for two suppliers of such systems. The second example explores the effect of noise suppression level, as set by the hands-free system, for two suppliers. The relationships of these effects to SMOS and NMOS are developed and discussed.

CITATION: Amman, S., Charette, F., Nicastri, P., Huber, J. et al., "Quantifying Hands-Free Call Quality in an Automobile," *SAE Int. J. Passeng. Cars - Mech. Syst.* 8(3):2015, doi:10.4271/2015-01-2335.

1. INTRODUCTION

A listening study was conducted to determine the best method for calculating Mean Opinion Scores (MOS) for the quantification of the mobile phone speech quality. The methods investigated produce three MOS values. SMOS (Speech MOS) is intended to describe the quality of the speech produced. NMOS (Noise MOS) is used to give an indication of the noise influence. GMOS (Global MOS) is an indicator of overall quality of the received phone call. All MOS calculations are on a 1 to 5 scale with higher values indicating better performance.

The listening study was conducted in accordance with the ITU-T P.835 Recommendation [1] which produces subjective ratings for speech, noise and overall impressions of the speech in noise. The purpose of this study was to see how these ratings correlate to two different ITU (International Telecommunications Union) standards for objective MOS calculation. The ITU calculations of ETSI EG 202 396-3 [2] and ETSI TS 103 106 [3] were both evaluated and correlated to the subjective ratings of 49 jurors. ETSI EG 202 396-3 was developed using Voice over Internet Protocol (VoIP) COder/DECoders (CODECs) and is the earlier of the two standards. ETSI TS 103 106 utilized mobile phone CODECs for its development and is a more recent standard.

In addition to evaluating the two standards, the P.835 listening study results were used to investigate factors that may influence a customer's perception of the phone call quality. Factors such as: evaluator gender, native language, hearing loss, talker gender and whether or not a subject was involved with Ford's hands-free system (SYNC), were all assessed for significance in affecting the subjective responses of the subjects.

Following this introduction, Section 2 of this paper will discuss vehicle and sound sample selection. Section 3 will address the methodologies used for the P.835 evaluation and how the results correlate to the two different standards for MOS calculation. Factors affecting the subjective scoring of the participants will also be discussed in this section. Section 4 will give two examples of how MOS values were used to explore effects that influence the call quality of hands-free systems. The first example investigates the effect of background noise level in the vehicle cabin on hands-free phone call quality for two suppliers of such systems. The second example explores the effect of noise suppression level, as set by the hands-free system, for two suppliers. The relationships of these effects to SMOS and NMOS are developed and discussed in this section. Finally, Section 5 will complete this paper with some conclusions.

2. VEHICLE AND SOUND SELECTION

All speech recordings were made utilizing a commercially available test system that implements the ITU P1100 recommendation [4] for the assessment of motor vehicle hand-free systems. A component of this system implements the ETSI MOS calculations [2][3].

The typical set-up is shown in Figure 1. The right side of the diagram shows the noise simulation. Recordings of various background noise conditions were made at the hands-free microphone location. The spectral content of these background noises was reconstructed using the speaker and equalizers shown. Further details of the noise simulation can be found in [5]. The MOS algorithms require three signals: 1. a clean signal that is played to the Head And Torso Simulator (HATS), 2. an unprocessed signal that is recorded at the hands-free microphone location, and 3. a processed signal that is collected at the Bluetooth interface using the test system's Bluetooth reference which is paired with the hands-free device. Figure 1 shows the points where these signals are extracted. The entire process is explained in [6]. The recordings that the subjects will listen to are those emitted by the in-vehicle hands-free Bluetooth device in the car. This "processed" signal will include the effects of the noise reduction processing that is performed by the hands-free device. The noise reduction will not only reduce the noise level, but will generally do it at the expense of speech quality. Some systems will perform this trade-off better than others; however, all systems will eventually degrade the speech if the noise reduction is too aggressive.

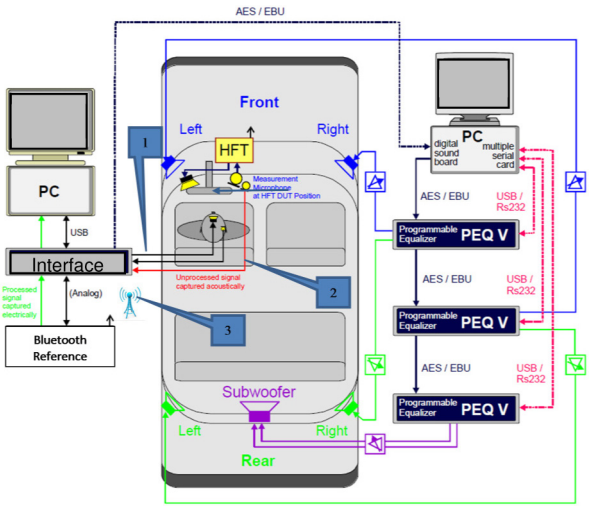

Figure 1. Test set-up for data acquisition for MOS calculation.

For this study, the test condition in which the vehicles were traveling at 100kph on Ford's Dearborn Development Center (DDC) Interior Quietness road surface was chosen. This is a brushed concrete road surface. The HVAC was in the defrost mode in the lowest blower setting. The seat was in the 95th percentile male position (almost all the way rearward). An additional test condition for one of the vehicles was also included and will be explained shortly.

This study used a single female utterance ("You must go and do it at once.") and one male utterance ("He could not remember his name."). These utterances come from the so-called "Harvard Sentences" which are a collection of sample phrases that are used for standardized testing of Voice over IP, cellular, and other telephone systems [7]. They are phonetically-balanced sentences that use specific phonemes at the same frequency as they appear in spoken English.

In order to select competitive vehicles with a range of SMOS and NMOS values, the ETSI EG 202 396-3 standard was calculated on the entire vehicle set for the test condition in which the vehicle is driving at 100kph over a brushed concrete road surface. Those results are plotted in Figure 2. The vehicles chosen for the study are those with the large dots in the plot. SMOS values ranged from 1.2 to 3.1, while NMOS values ranged from 1.3 to 3.3. Clearly, the lower and mid ranges are well represented. In order to include at least one sample with excellent MOS scores, vehicle C with the HVAC in the low position and no road noise was included. This sample produced an SMOS of 4.7 and an NMOS of 4.9. This sample will also give subjects a sample with nearly no noise and very limited speech distortion.

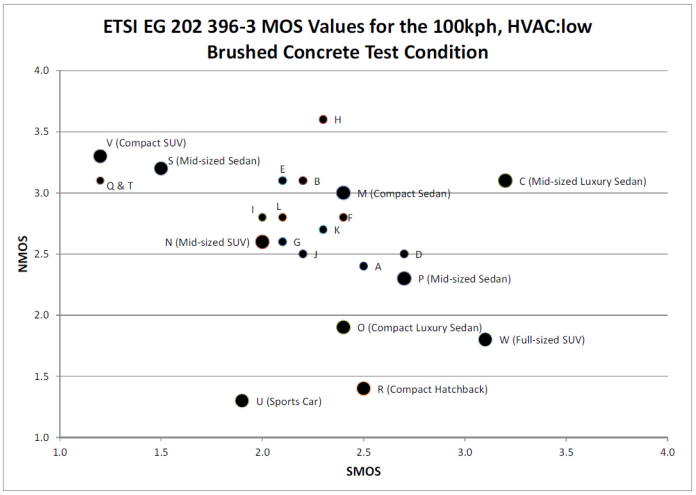

Figure 2. Sound recordings selected (large dots) based on spread in SMOS and NMOS as calculated by ESTI EG 202 396-3.

3. LISTENING EVALUATION

P.835 Evaluation Methodology

The subjective evaluation was conducted in accordance to ITU-T Recommendation P.835 [1]. The rating scales for the background noise, speech distortion and overall speech quality are shown in Figure 3. There were six sessions of participants. The order of sound presentation was randomized by session. As specified in P.835, in half of the sessions the evaluators rated the noise first, while the other half rated the speech first so that order effects would cancel.

Each of the sound samples is the result of noise reduction processing unique to each manufacturer of the in-vehicle hands-free system. As a result, absolute levels of the sound samples can vary. In order to minimize the effect of sound level, the sound samples were all normalized such that the peak dBA level using a 0.125 sec time constant was within 3dB for all the samples. As an example, Figure 4 shows dBA as a function of time for all 11 female-talker sound samples. Please note, C (no road noise) is vehicle C with the low

noise level (HVAC low, no road noise). The sounds were also subjectively assessed to make sure that the voice levels from sound to sound were approximately the same.

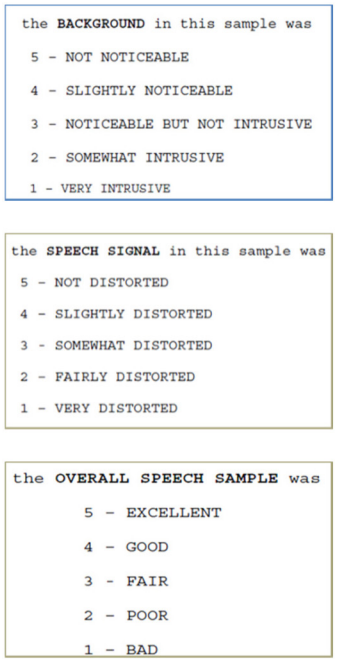

Figure 3. P.835 rating scales for background noise, speech distortion and overall speech quality.

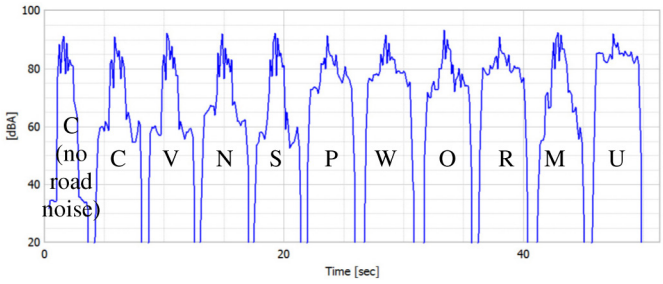

Figure 4. dBA levels for the female talker. dBA time constant was set to 0.125 seconds.

When rating the noise, speech distortion and overall quality of the sound samples, the subjects were played the sound three times for each rating task. A practice block of 11 sounds was first rated in order for the listeners to become accustomed to the sound samples and the process.

3. P.835 LISTENING EVALUATION RESULTS

P.835 Evaluator Ratings

Shown in Figure 5 are the noise, speech and overall rating means and 95% confidence intervals. The female talker, male talker and combined results are shown. Sound samples were also grouped according to significant differences of the combined data. Tukey's HSD post hoc test for significant differences [8] was used to place the samples into significant groups. For all three rating scales, the responses break out into at least five distinct groups. Since it's clear that differences could indeed be detected by the evaluators, the data set lends itself to a further investigation to determine the best standard for calculating NMOS, SMOS and GMOS.

Figure 6 shows the noise, speech and overall fitted means of the ratings for each of the subjects. From these plots it is clear that there is less agreement among subjects with respect to the speech quality than for the noise. This may be an indicator that the subjects were using different subjective criteria or that they were simply using the subjective scale differently. The variation for the overall rating is somewhat in between the speech and noise variation, which one would expect since the overall impression will be influenced by both the noise reduction and speech quality.

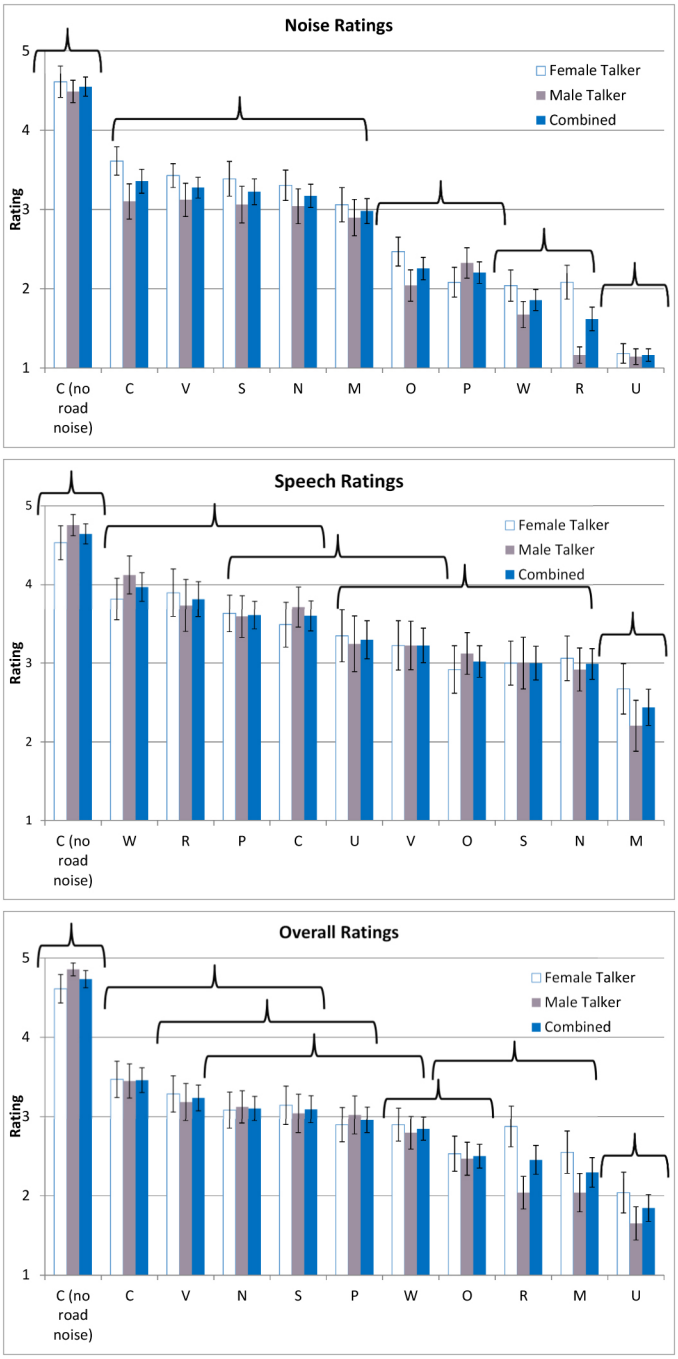

Figure 5. Noise, Speech and Overall rating means and 95% confidence intervals for female talker, male talker and combined male/female results. Significance groups (α=0.05) are indicated by brackets.

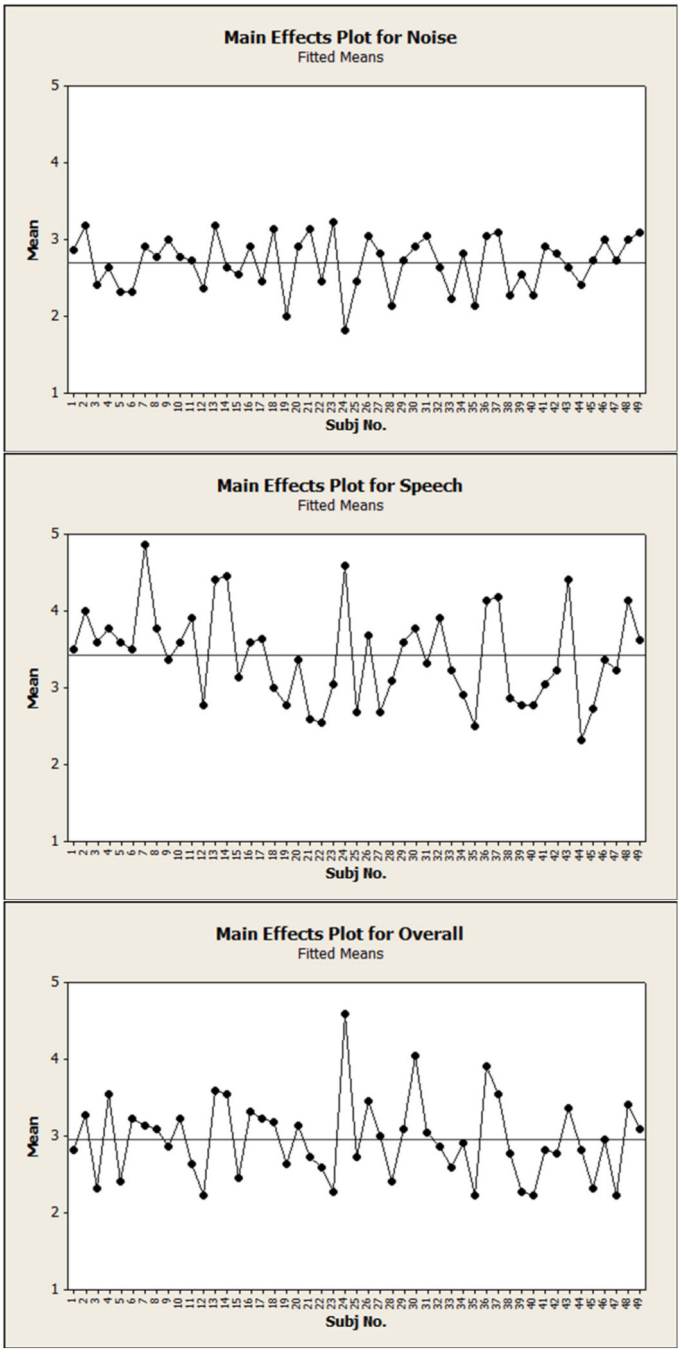

Figure 6. Noise, Speech and Overall ratings for each of the 49 subjects.

P.835 Factor Effects

A number of factors were investigated for their potential impact on the ratings. The factors investigated are summarized below.

Male/Female Talker: In the study, subjects listen to both male and female talkers.

Native English Speaker? : Subjects were asked what was their native language. If the response was something other than U.S. English, they were considered a non-native speaker.

Subject Gender: Subjects indicated their gender.

Age: Age of the participants in 10-year categories.

Subject Involvement with SYNC: Subjects indicated (Y/N) if they were directly involved with Ford's hands-free system (SYNC).

Hearing Loss: Subjects indicated if they had any known hearing loss.

A factor summary is shown in Table 1. In general, even if the factors were statistically significant ($\alpha = 0.05$) they were not large in magnitude. One exception to this was the effect of Age on the ratings. This factor was significant for all three rating scales and had almost a full rating impact on the overall evaluation. Further investigation showed that the ratings generally increased with age, especially for the 60-69 and 70-79 age categories. This may simply be due to age-related hearing loss which may make it more difficult for the older subjects to detect speech distortion.

Table 1. Factor Effects Summary

Factor	Noise	Speech	Overall	Comments
M/F Talker	0.30	none	0.16	Female talker rated higher
Native US English Speaker?	none	0.20	none	Native US English subjects with higher ratings
Subject Gender	none	none	0.25	Female ratings were higher
Age	0.50	0.75	1.00	Ratings increase with age
Subject Involvement with SYNC	0.25	none	0.25	Ratings higher for noise, lower for overall if subject was involved with SYNC program
Hearing Loss	none	0.25	none	Subjects with hearing loss produced higher ratings

P.835 Ratings Correlated to MOS Calculations

The main point of this investigation was to identify the best standard for MOS calculations based on the subjective responses of the 49 subjects. Ratings for noise, speech and overall impressions of the various sound samples showed statistically significant differences justifying a meaningful correlation to the standard MOS calculations. These are shown in Figure 7 along with the corresponding R^2 values.

In general the ETSI 202 396-3 recommendation shows better correlation to the speech and noise ratings; although, neither recommendation shows outstanding speech correlation. Both recommendations correlate quite well to the noise ratings. Objectively quantifying the speech appears to be a more difficult task that quantifying the noise. The ETSI 202 396-3 calculation does show a better relationship to the speech ratings if vehicle M is removed from the analysis (R^2=0.77). Upon listening to vehicle M, artifacts of the noise reduction can be heard in non-speech segments that may have caused evaluators to rate the speech lower than what would be calculated. Regardless, it's clear that the SMOS calculations cannot fully describe the subjective responses. Lack of good correlation may also be due to the fact that the evaluator subjective responses have much more variability as pointed out earlier.

Finally, the GMOS values are plotted against the subjective ratings in the last column of the figure. Both correlations suffer from a heavy influence of a single data point (C with no noise). Even so, the TS 103 106 recommendation provides a better prediction. It should also

be mentioned that this calculation differs from EG 202 396-3 in that it uses a neural network to predict GMOS values from S and NMOS values instead of a quadratic regression approach.

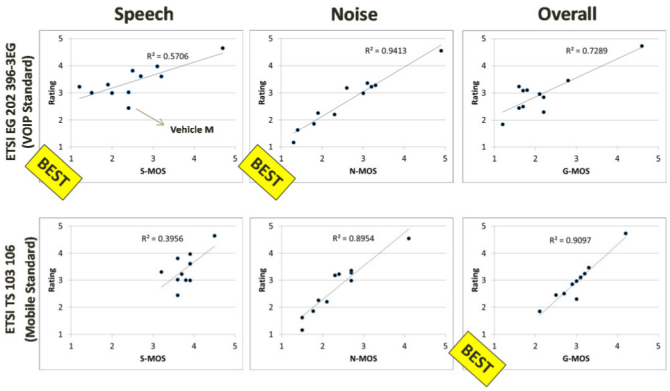

Figure 7. Correlation of MOS scores to the average subjective ratings of the 49 subjects.

4. THE USE OF MOS SCORING FOR HANDS-FREE PHONE QUALITY EVALUATION - EXAMPLES

From the previous subjective study and correlation, it was shown that the ETSI 202 396-3 recommendation shows better correlation to the speech and noise ratings. Having gained confidence that this method of subjective prediction can be used to evaluate the quality of a hands-free calling system, a couple of examples will now be given showing the usefulness of such measures.

Example 1. MOS Values as a Function of Background Noise Level

A multitude of noise reduction strategies exist from various suppliers of hands-free phone systems. How each of those systems perform when exposed to varying degrees noise level, can be a qualifying factor when selecting a system.

Figure 8 shows the results of an experiment in which the recorded background noise of a vehicle was reproduced inside the vehicle as described in [5]. The noise was recorded at the hands-free microphone location while the vehicle was driving over a brushed concrete road surface at 100 kph. 0 dB on the graph corresponds to the nominal noise level in the vehicle. The background noise level was subsequently attenuated and amplified to map out the performance of two hands-free phone systems. All dB levels were calculated over the narrowband speech range of 300-3400Hz commonly used in today's hands-free phone systems.

Some observations can be made with respect to the NMOS values for both systems. Both supplier A and supplier B are equally affected by increases in noise level. Both show about a 0.10 NMOS reduction for every dB increase in noise level. However, supplier B's system shows about a 0.5-0.6 NMOS advantage across all noise levels.

When considering the preservation of the speech quality, the differences between the two systems is small for reduced noise levels (3 dB below nominal). Above those noise levels, supplier A exhibits a 0.3-0.6 SMOS advantage over supplier B. In this region supplier A has clearly chosen to preserve the speech signal at the expense of the noise reduction performance.

Overall conclusions would be the following:

1. In the region in which the noise level is more than 3dB below nominal, supplier B has the advantage due to better noise reduction performance and virtually equivalent speech preservation.
2. In the region where noise levels are greater than 3 dB below nominal, supplier A has superior speech preservation but only at the expense of poorer noise reduction performance. It would be up to the automaker to set the appropriate targets to balance SMOS and NMOS values to achieve proper customer satisfaction and thus determine which was the preferable system.

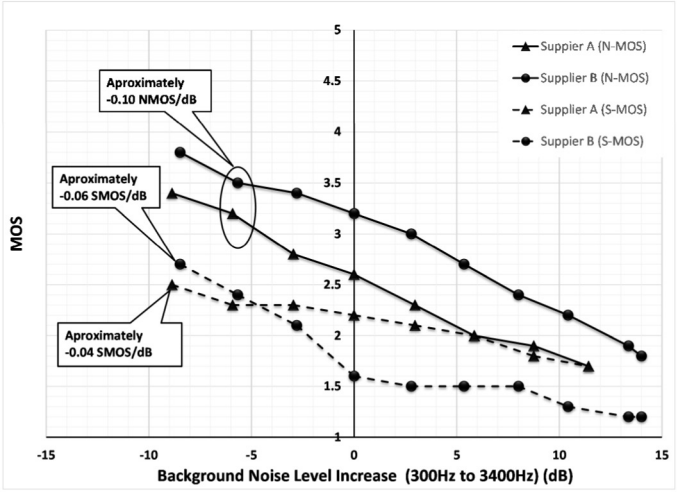

Figure 8. NMOS and SMOS values as a function noise level. 0 dB corresponds to the nominal noise level recorded in the vehicle.

Example 2. MOS Values as a Function of the Hands-Free Noise Suppression Level

This example demonstrates how changing the noise suppression on the hands-free phone system affects the noise reduction (NMOS) and speech preservation (SMOS). Two supplier systems were evaluated. The background noise condition was that of a vehicle driving on an asphalt road at 120 kph. Both systems allowed for adjustment of the noise suppression. The results are shown in Figure 9.

For both systems, the NMOS improvement with noise suppression level is approximately 0.10 NMOS for every db increase in noise suppression. Supplier A does show an advantage at higher suppression levels which can be as much 0.30 NMOS. However, it can be seen that the superior NMOS performance comes at a price with respect to speech preservation as the SMOS values decline more rapidly for supplier A than for supplier C. At lower suppression levels the SMOS values are almost identical for both systems and virtually unaffected by suppression level.

Overall conclusions would be the following:

1. Below 10 dB of noise reduction, supplier A and supplier C performance is nearly identical for both NMOS and SMOS.
2. For settings of 0-10 dB of noise suppression, both supplier systems show improving NMOS with suppression setting with little impact on SMOS.
3. Above 10 dB of noise suppression, supplier A shows superior NMOS values but these are in contrast to poorer SMOS values, once again demonstrating that there is a trade-off between good noise suppression and speech preservation.

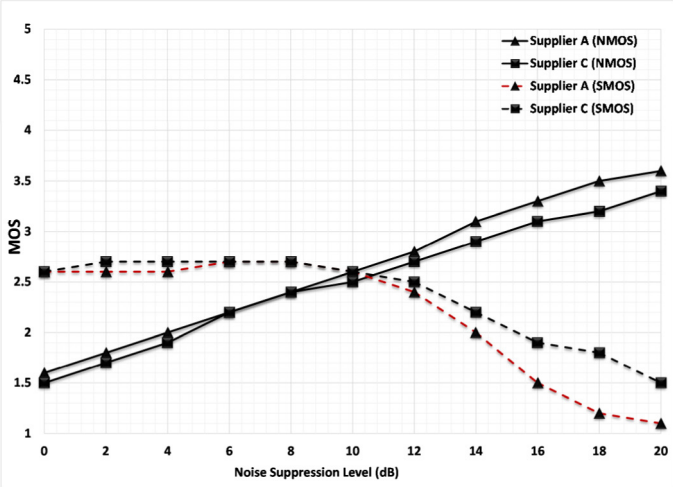

Figure 9. NMOS and SMOS values as a function hands-free noise suppression level.

SUMMARY/CONCLUSIONS

The following are some conclusions that can be drawn from this research:

- The ETSI EG 202 396-3 (VOIP Standard) for S and NMOS correlates best to speech and noise ratings derived from the P.835 evaluation, while the ETSI TS 103 106 (Mobile Standard) for GMOS correlates best to the overall impressions.
- Except for Age, most factor effects were small (i.e., < 0.5). Older subjects tended to produce higher ratings.
- Subject-to-subject variance in ratings was higher for the speech and overall rating tasks than for the noise rating. This would appear to indicate that rating the noise was an easier task than rating the speech or that subjects were in less agreement for the speech rating task.
- In general, MOS values can serve as a valuable tool to identify the quality or specifying a vehicle hands-free phone system. The two examples in this paper demonstrate how these measures can be used. Furthermore, other factors such as seat position, microphone placement, cabin acoustics, etc. could also be investigated in order to quantify their influence on the call quality.

REFERENCES

1. ITU-T Recommendation P.835, Series P: Telephone transmission quality, telephone installations, local line networks, Methods for objective and subjective assessment of quality, Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, 2003.
2. ETSI EG 202 396-3, Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise Part 3: Background noise transmission - Objective test methods, 2008.
3. ETSI TS 103 106, Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods, 2012.
4. ITU-T Recommendation P.1100, Series P: Terminals and subjective and objective assessment methods, Communications involving vehicles, Narrow-band hands-free communication in motor vehicles, 2011.
5. ETSI EG 202 396-1, Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise Part 1: Background noise simulation technique and background noise database, 2008.
6. Head Acoustics Application Note, 3QUEST: 3-fold Quality Evaluation of Speech in Telecommunications Systems, 2008, http://www.headacoustics.de/downloads/eng/application_notes/telecom/Appl_note_3QUEST_e0.pdf
7. Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., Weinstock, M., "IEEE Recommended Practice for Speech Quality Measurements," IEEE Transactions on Audio and Electroacoustics (Volume:17, Issue: 3), 1969.
8. Tukey J., "The Problem of Multiple Comparisons," Unpublished notes, Princeton University, 1953.

CONTACT INFORMATION

Scott Amman
samman@ford.com