# Background Noise Effects on Automotive Hands-free Phone Call Quality and Automatic Speech Recognition

**Francois Charette, John Huber, Scott Amman and Gint Puskorius**

Ford Motor Company, USA
E-mail: fcharett@ford.com

**Abstract:** Hands-free phone calling is the most frequently utilized feature for vehicles equipped with infotainment systems and external microphones which support connection to phones and implement automatic speech recognition (ASR). However, achieving high in-vehicle hands-free phone call quality is challenging due to the extremely noisy nature of a vehicle's interior environment. Noise generated by wind, mechanical and structural sources, tire to road contact, passenger cross-talk, engine/exhaust systems, and HVAC air pressure and flow are all significant contributors to deterioration of hands-free phone call quality. In addition, other factors influencing the in-vehicle phone call quality include microphone placement, cabin acoustics, seat position of the talker, and the noise reduction algorithm of the hands-free system. In this paper, we present results of a study involving over 200 vehicles from a broad range of OEMs and vehicle segments which seeks to quantify the impact of different noise sources (particularly road and HVAC noise) on the quality of hands-free phone calling. We demonstrate that, even though the vehicles studied utilized different noise reduction / signal processing methods and demonstrated unique noise spectral characteristics for the conditions considered, some general relationships between noise level and call quality can nonetheless be observed and quantified with respect to the aggregate performance of the systems. Subsequently, we chose a single vehicle to evaluate the performance of two different suppliers' noise reduction processing capabilities. In this case, the background noise was systematically varied in level to understand the relationship between background noise level and phone call quality when the noise reduction processing and noise spectral content are controlled. The learnings generated from these studies provide a basis by which engineers can establish in-cabin vehicle noise level targets which would result in acceptable hands-free phone call quality. In addition, these processes can be used to quantify the effectiveness of a hands-free system's noise reduction processing under various noise conditions. Finally, we demonstrate that these methods can also be extended to quantifying the performance of in-vehicle automatic speech recognition. In this case, we demonstrate that word error rate (WER) is also influenced in a systematic fashion by background noise level.

**Keywords** Hands-free Call, Automatic Speech Recognition, NMOS, SMOS, Word Error Rate

## 1 Introduction

Interactions between drivers and vehicle infotainment systems are continuously increasing and have become one of the most important sources of customer dissatisfaction when the infotainment system does not operate as expected. J.D. Power Initial Quality Surveys [1] indicate that the in-vehicle infotainment system, which includes audio, communication, entertainment, and navigation sub-systems, is the most frequently cited source of things-gone-wrong for new cars, trucks, SUVs, and vans at 3 months-in-service.

The material presented in this paper is focused on the communication performance of vehicle infotainment systems. Specifically, we consider performance of the hands-free phone call and embedded automatic speech recognition systems.

### 1.1 Hands-free Phone Call

An infotainment system's hands-free phone system is one of the most frequently utilized features of the vehicle infotainment system. The ITU-T P1100 [2] international standard describes a battery of tests and measurements which are used to objectively quantify hands-free phone call quality. This standard covers end-to-end phone call quality, including the in-vehicle infotainment hands-free system, the driver's cellular handset and network carrier, as well as the receiver's cellular handset and network carrier. The end-to-end phone call quality is dependent on a wide range of companies that provide hardware and software services (i.e., cellular handsets and network services). It also includes all potential problems that can arise between the interfaces of hardware and services. Therefore, to keep the testing and measurements manageable and within control of what a vehicle OEM can affect, the scope of hands-free call

quality results presented here are limited to the in-vehicle infotainment hands-free system. Measurements of in-vehicle infotainment hands-free system performance are obtained with a subset of the ITU-T P.1100, which is referred to as 3QUEST [3], [4] throughout this paper. The 3QUEST testing setup and metrics are summarized in section 2 of this paper.

Note that, while a vehicle's hands-free phone call system is a subset of the overall end-to-end phone call system, it is nonetheless one of the most challenging aspects to implement correctly. The performance of the hands-free phone call system is dependent on a wide range of driving conditions, resulting in a very broad range of noisy conditions which can be present in a vehicle cabin. Overall performance is also dependent on microphone design and location, the hands-free system's signal processing algorithms (filtering, noise suppression, etc.), seating position, cabin acoustic properties, and the vehicle's noise-vibration-harshness characteristics (e.g., due to wind, road, and powertrain).

### 1.2 Embedded Automatic Speech Recognition

Automatic speech recognition is another popular feature of a vehicle's infotainment system, which can be used to initiate phone calls (e.g., "dial #"), control music (e.g., "play artist xxx") and obtain navigation directions. In addition, ASR can be used to control certain vehicle features (e.g., "set temperature to 72 degrees"). While modern cloud-based ASR systems (e.g., Siri, Google Now) demonstrate remarkable performance, this cloud-based capability is primarily experienced in relatively quiet environments (e.g., office, home). However, this level of performance is not generally experienced in an automotive embedded ASR system, operating under a broad range of acoustic disturbances.

No established standards exist for evaluating performance of automotive embedded ASR systems. Thus, we adopt the process of recording and playing back various vehicle background noises, combined with pre-recorded close talk utterances convolved with a previously measured vehicle interior impulse response (which is unique for every interior cabin). Using this approach, varying levels of noisy utterances can be synthesized and used to test an embedded ASR system. Word error rate is then used to compare ASR performance at different background noise levels.

## 2 Measurement Setup

The typical laboratory measurement setup is shown in Figure 1. The right side of the diagram shows the noise simulation system used for in-laboratory reproduction of vehicle cabin noise recordings obtained on the road for the given vehicle under test. Recordings of various cabin background noise conditions were made at the hands-free microphone location. The spectral content of these background noises was reconstructed using the speaker system and equalizers as shown in Figure 1. Further details of the noise simulation can be found in [3].

### 2.1 3QUEST Metrics

The 3QUEST metrics from the ETSI standard [4] are presented in units of Mean Opinion Score (MOS). The 3QUEST algorithms produce three MOS metrics: (1) the Speech Mean Opinion Score (S-MOS); (2) the Noise Mean Opinion Score (N-MOS); and (3) the Global Mean Opinion Score (G-MOS). All MOS metrics are reported on a scale from 1 to 5, from lowest to highest quality, respectively. S-MOS quantifies the speech quality; it is low if the measured speech is distorted and is high if the speech sounds natural. N-MOS quantifies the noise portion of the signal; it is low if the noise is annoying and high if the noise is not perceptible. Finally, G-MOS is a metric which combines S-MOS and N-MOS into a single value. Note that these metrics are all based on extensive subjective studies, and provide a model-based prediction of a subjective evaluation. For the remainder of this paper, we will only consider the S-MOS and N-MOS metrics. We have found that G-MOS does not generally reveal any meaningful insights for automotive applications, as the signal processing algorithms are designed to trade off speech versus noise quality.

The 3QUEST algorithms require three input signals: (1) a clean speech signal that is played through to a head and torso simulator; (2) an unprocessed signal which is recorded at the hands-free microphone location; and (3) a processed signal which is collected at the Bluetooth interface between the infotainment system and the paired hands-free device (see [4] for details). The small blue boxes in Figure 1 show the points where these three signals are extracted.

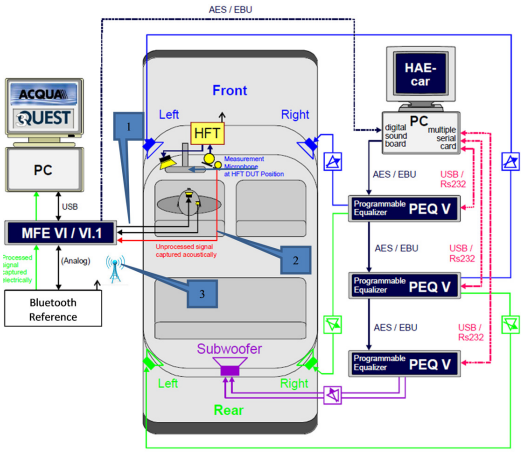



**Figure 1**: Laboratory setup.

# 3 Hands-Free Phone Call Quality Results

Performance for approximately 200 hundred vehicles, from all major OEMs and vehicle segments, was measured using the method described in section 2. Measurements were obtained for multiple background noise conditions, yielding a fairly large amount of data. We analyze all of this data to extract some general relationships of hands-free phone call quality with respect to background noise level.

As indicated above, G-MOS was not considered in this work as we have found that it is more beneficial to set targets based on the individual metrics of S-MOS and N-MOS to ensure that both the speech quality and noise reduction are adequately addressed. It has also been observed that S-MOS is less reliable as a predictor of customer response than N-MOS. Past research in our lab has shown that subjective evaluations of speech quality contain more inherent variability than those for the background noise [6]. Therefore, it would be understandable that an objective metric derived from these results would be less precise. However, S-MOS is a good first order good indicator which allows an engineer to flag potential problems, particularly if a subjective assessment is also employed.

## 3.1 Results Overview

Figure 2 and 3 show scatter plots of S-MOS and N-MOS results as a function of background noise level for road and HVAC fan noise, respectively. All background noise levels presented are reflective of the vehicle cabin background noise as measured at the hands-free system microphone location using an instrument grade microphone of the 3-Quest measurement system (i.e., not the microphone of the hands-free system). In addition, background noise levels were measured in what is defined in the telecommunication industry as the narrow band frequency range of 300-3400 Hz.

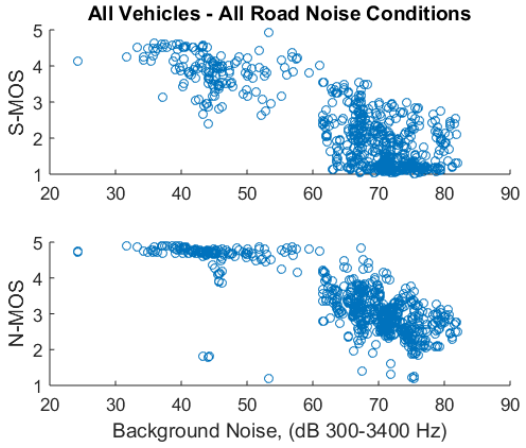



**Figure 2**: All road noise conditions MOS results versus narrow band background noise level.

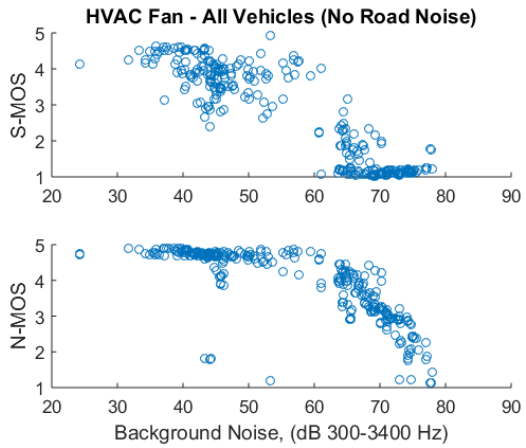



**Figure 3**: HVAC fan noise, all conditions, MOS results versus narrow band noise level.

## 3.2 Results for Road Noise

In this section, the results presented in Figure 2 are separated into the 5 road noise conditions used during testing: (1) 0 KPH (idle or no road noise); (2) 60 KPH on brushed concrete; (3) 80 KPH on brushed concrete; (4) 100 KPH on brushed concrete; and (5) 120 KPH on asphalt; these results are shown in Figures 4 through 8, respectively. In each of these scatter plots, we use a blue circle to denote individual vehicle measurements and a red dot to indicate measurement means. Histograms of MOS and background noise levels are also provided in these figures. Note that there are clear outliers in some of the scatter plots, which may possibly be due to measurements errors. These

outliers were not removed simply because there are few of them and their impact on the overall results or observations presented here are not significant.

It is important to note at this point that each of the MOS score distributions (both S-MOS and N-MOS) represent what the authors call constrained populations (i.e., constrained from 1 to 5). Therefore, as the mean of a given population shifts towards a given end value, the tail of the distribution will tend to collect into that limiting value. This has the effect of producing a bi-modal looking distribution which, given enough shifting of the mean, will become a one sided uni-modal distribution.

In all of these road noise cases, where the HVAC noise was set to low and the vehicle is in motion, the scatter of N-MOS values tends to be centrally located within the range. This therefore results in a distribution of values for N-MOS that follow a uni-modal distribution. One could conclude then, that the vehicle manufacturers were successful at tuning their hands free phone call systems by attempting to remove a reasonable amount of road noise (i.e., by setting an N-MOS target at or near the central value of 3), but stopped at the point where speech would become distorted.

In the case of S-MOS, it was noted previously that the values tend to show more scatter than N-MOS. This could be partly due to the fact that S-MOS is more of a qualitative assessment metric (i.e., speech distortion), and is therefore subject to a range of opinions. Unlike N-MOS, which is derived directly from the level of background noise, S-MOS is not directly related to background noise, but is instead a measure of how much original speech was allowed to pass through the noise suppression algorithm. An additional complication is likely to occur when the background noise is sufficiently loud; here, the subjective rating may be skewed towards a lower value, even if the speech signal is mathematically undistorted from the original clean speech.

In light of these considerations, let's consider the S-MOS distributions in our study group. The scatter of S-MOS values for when the vehicle was in motion appears sufficiently large as to become bi-modal at the constraining boundary of S-MOS equal to 1. In fact, at the highest vehicle speed of 120 kph, the distribution has shifted sufficiently as to result in a one-sided uni-modal distribution at the limit of S-MOS equal to 1. This is clearly a situation where the digital signal processing had difficulty achieving a balance between reasonable noise suppression and speech distortion.

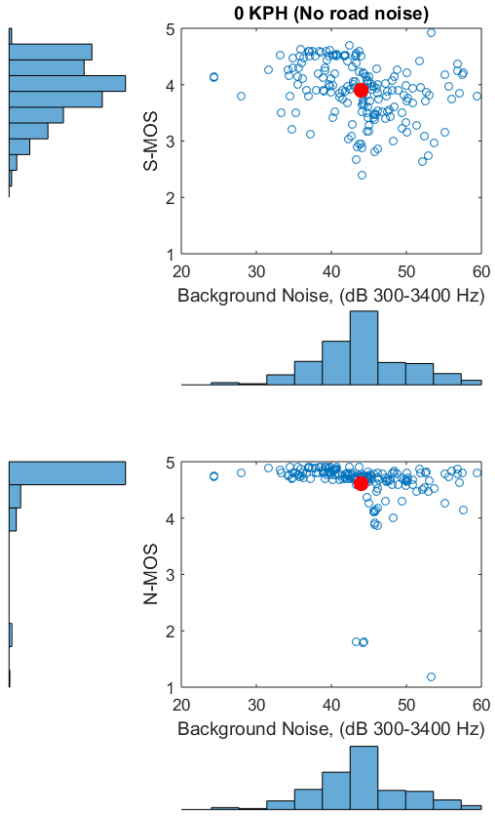



**Figure 4**: 0 KPH (No road noise).

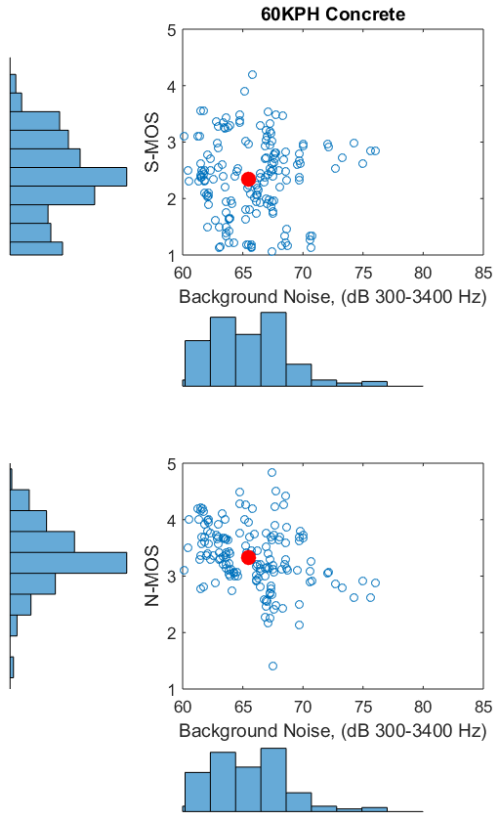



**Figure 5**: MOS results for 60 KPH (concrete).

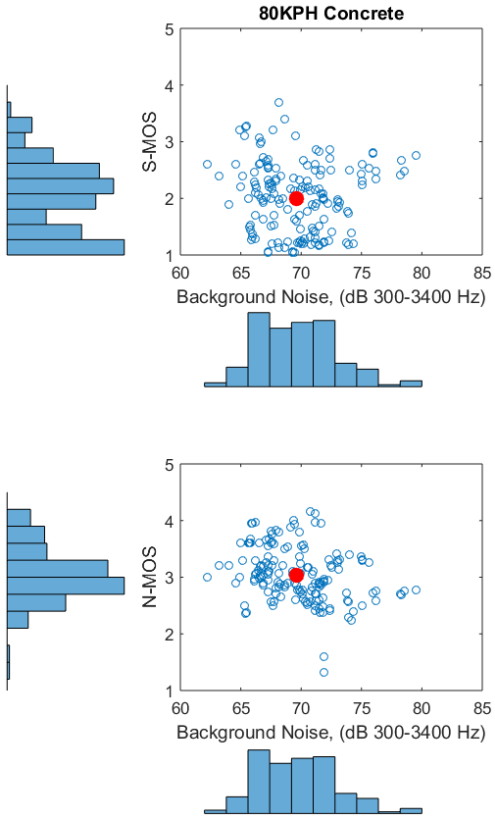



**Figure 6**: MOS results for 80 KPH (concrete).

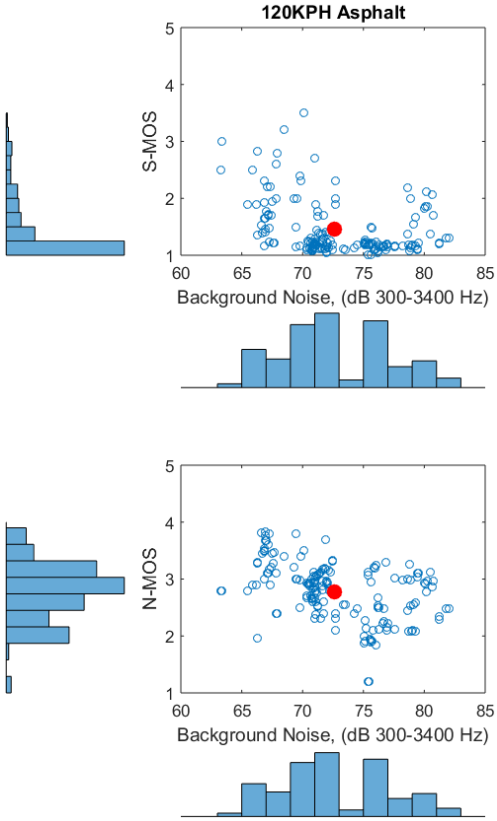



**Figure 8**: MOS results for 120 KPH (asphalt).

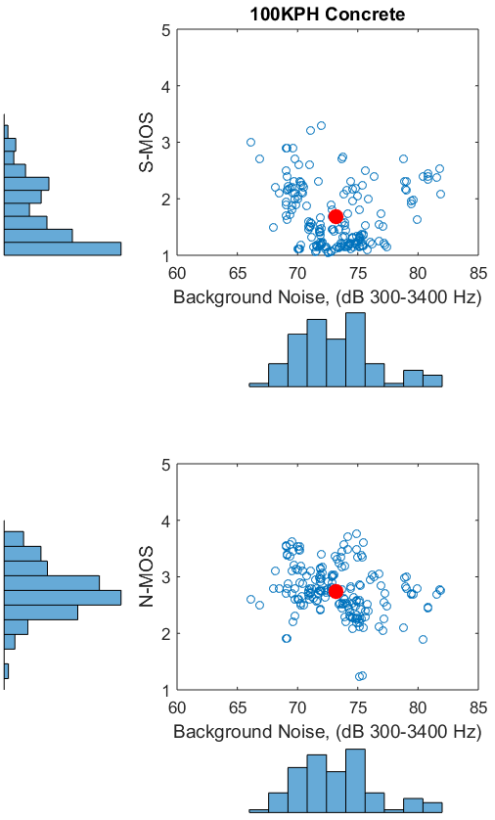
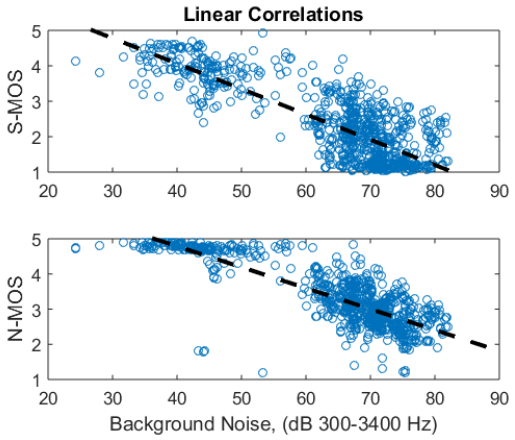



**Figure 7**: MOS results for 100 KPH (concrete).

Figure 9 presents linear regression results for S-MOS and N-MOS as a function of background noise level. While the relationships between MOS scores and background noise level are expected to exhibit some nonlinearity, we use a linear correlation to extract rules of thumb to describe approximate relations between vehicle background noise level and the hands-free phone call system performance. The linear correlations are shown as black dashed lines. The S-MOS linear correlation shows a sensitivity of approximately 0.07 units of MOS lost per 1 dB of increase in background noise. Similarly, the N-MOS analysis shows a sensitivity of approximately 0.06 units of MOS lost per 1 dB of increase in background noise.




| | |
|---|---|
| S-MOS | Linear model Poly1:<br>f(x) = p1*x + p2<br>Coefficients (with 95% confidence bounds):<br>p1 = -0.07153 (-0.07496, -0.06811)<br>p2 = 6.925 (6.699, 7.15)<br><br>Goodness of fit:<br>SSE: 355.9<br>R-square: 0.6475<br>Adjusted R-square: 0.6471<br>RMSE: 0.6233 |
| N-MOS | Linear model Poly1:<br>f(x) = p1*x + p2<br>Coefficients (with 95% confidence bounds):<br>p1 = -0.05957 (-0.0623, -0.05685)<br>p2 = 7.171 (6.992, 7.351)<br><br>Goodness of fit:<br>SSE: 225.8<br>R-square: 0.6675<br>Adjusted R-square: 0.6672<br>RMSE: 0.4965 |

**Figure 9**: Correlation between MOS Results and background noise level of the grouped means.

Figure 10 shows a scatter plot of S-MOS versus N-MOS for all measurements. If N-MOS and S-MOS were thought of as being equally important and balanced, it would seem reasonable to expect that system calibrations would yield a scatter plot result with most points lying along the 45° diagonal. Given that the vast majority of the points lie to the right of the 45° diagonal, we interpret that, in general, the hands-free systems' signal processing algorithms tend to aggressively remove noise at the expense of distorting speech. This could merely be a tuning issue, or possibly a limitation of the signal processing algorithms whereby noise cannot be removed without distorting the speech above a certain noise level threshold. Since the data presented here covers a wide variety of systems, vehicles and algorithms with different tuning, we expect the latter to be the most probable explanation.

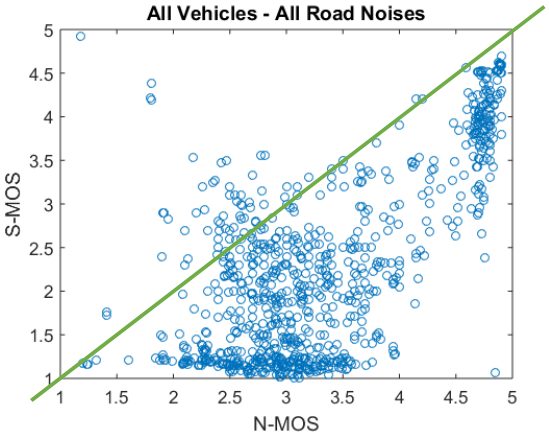



**Figure 10**: N-MOS versus S-MOS.

## 3.3 Results for HVAC

Figure 11 shows measurements for low and high HVAC fan settings in defrost mode, where green and red points denote fan low and high settings, respectively. The overall distribution of N-MOS values is somewhat similar to what was observed for the road noise results. On the other hand, the distribution of S-MOS values indicates two fan-dependent states: (1) when the HVAC fan is low, speech quality is high and essentially unaffected by the fan noise or air flow; and (2) when the HVAC fan is high, the speech quality is degraded due to fan noise and/or air flow.

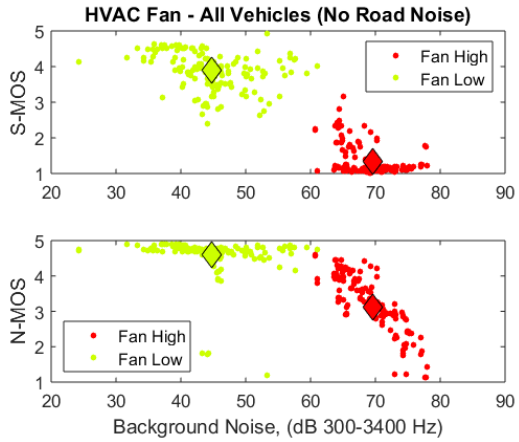



**Figure 11**: S-MOS and N-MOS – Grouped per HVAC fan speed.

## 3.4 MOS Value Variations

A multitude of noise reduction strategies exist from various suppliers of hands-free phone systems. Selecting a specific system for production should certainly be a function of how the systems perform when exposed to varying degrees of noise level.

Figure 12 shows results of an experiment in which the recorded background noise of a vehicle was reproduced inside the vehicle as described in [3]. The noise was recorded at the hands-free microphone location while driving the vehicle over a brushed concrete road surface at 100 KPH. The 0 dB setting on the graph corresponds to the actual road noise level recorded in the vehicle. The background noise level was subsequently attenuated and amplified to map out the performance of two hands-free phone systems. Again, all dB levels were calculated over the narrowband speech range of 300-3400Hz commonly used in today's hands-free phone systems.

We measured the MOS performance of two different suppliers' hands-free phone call systems at different background noise levels (attenuated and amplified). We readily observe some salient characteristics for N-MOS values in these results. Both suppliers' systems are equally affected by increases in noise level, showing an approximately 0.10 decrease in N-MOS for a 1dB increase in noise level. However, supplier B's system has an absolute advantage of 0.5 to 0.6 increase in N-MOS across all noise levels.

When considering the preservation of the speech quality, the difference between the two systems is small for low noise levels (3 dB or more below nominal). Above those noise levels, supplier A's solution exhibits a 0.3 to 0.6 S-MOS absolute advantage over that of supplier B. In this region, the system for supplier A has clearly been designed to preserve speech signal at the expense of noise reduction performance. It would be up to the automotive OEM to set the appropriate targets to balance SMOS and NMOS values to achieve proper customer satisfaction and thus determine which system was preferable.

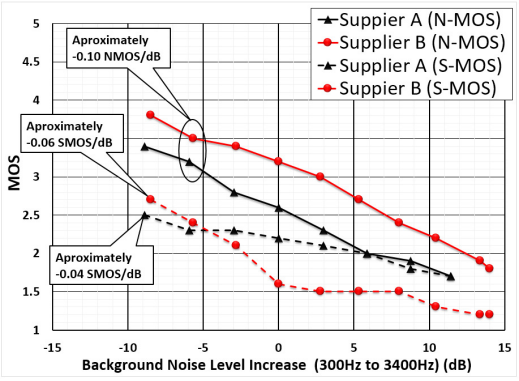



**Figure 12**: MOS Values versus background noise level for a given background noise.

The next example demonstrates how changing the noise suppression setting for the the hands-free phone system affects overall noise reduction (N-MOS) and speech preservation (S-MOS). Again, we evaluated two supplier systems for these purposes. We chose a background noise condition for that of a vehicle driving on an asphalt road at 120 KPH. Both suppliers' systems allowed for adjustment of noise suppression. The results are shown in Figure 13.

For both systems, N-MOS improvement with noise suppression level is approximately 0.10 N-MOS for every dB increase in noise suppression. The system of supplier A does show an advantage of at most 0.30 N-MOS at higher suppression levels. However, it can be seen that superior N-MOS performance is achieved at the expense of speech preservation, as S-MOS values decline more rapidly for the system of supplier A relative to that of supplier C. At lower noise suppression levels (0-10dB), the S-MOS values are nearly identical for both systems and are virtually unaffected by suppression level.

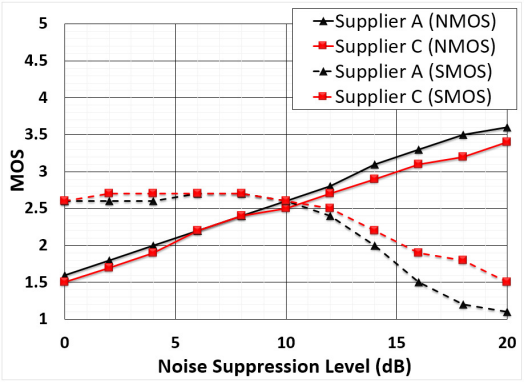



**Figure 13**: MOS Values versus processing noise suppression level setting.

# 4 Embedded Automatic Speech Recognition Results

ASR is another feature of the infotainment system which is regularly used to command a variety of functions. The ASR system uses the same vehicle acoustic path as the hands-free phone system, except that the signal is processed by the infotainment system instead of being transferred to the paired phone device.

It is possible to synthetically create a large quantity of noisy utterances and process them through an embedded ASR engine. Figure 14 shows word error rate (WER) results when the same set of utterances is mixed with five different levels of scaled vehicle background noise. Similar to the previous section, 0 dB is the nominal road noise level in the vehicle at 100 KPH on brushed concrete.

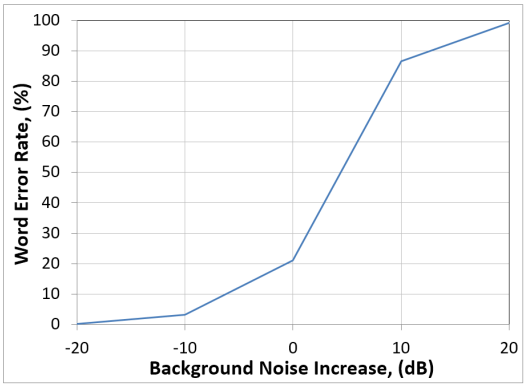



**Figure 14**: WER versus background noise levels.

Figure 14 shows that above the nominal (0 dB), noise level, which is highway speed, the system's performance starts to degrade dramatically

# 5 Conclusion

This paper has shown that existing metrics and methodologies can be applied for quantifying the quality of a hands-free phone call in an automotive environment. The examples provided in this paper show how a large population of vehicle systems behaves under various noise conditions. Two examples were also given on how MOS values can be used to both characterize noise level effects on a given system and how noise suppression level can influence phone call quality. Some of the principles developed for the quantification of hands-free phone call quality can be used as well for ASR evaluation. The simple ASR example studied demonstrates how background noise level can influence speech recognition performance. While this was only a single experiment for one particular road noise conditions, this could easily be extended to a broad variety of noise conditions. Performance could also be measured for other factors such as microphone location, seat position of the talker, cabin volume, etc.

# References

**[1]** Youngs, J., "Modern Infotainment Systems is Top Reported Problem with New Vehicles," J.D. Power 2014 Multimedia Quality and Satisfaction Study (http://www.jdpower.com/cars/articles/car-news/modern-infotainment-systems-top-reported-problem-new-vehicles), August 28, 2014**.**

[2] ITU-T Recommendation P.1100, Series P: Terminals and subjective and objective assessment methods, Communications involving vehicles, Narrow-band hands-free communication.

[3] 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.

[4] 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.

[5] Head Acoustics Application Note, 3QUEST: 3-fold Quality Evaluation of Speech in Telecommunications Systems, 2008, http://www.head-acoustics.de/downloads/eng/application_notes/telecom/Appl_note_3QUEST_e0.pdf

[6] S. Amman, F. Charette, P. Nicastri, J. Huber, B. Richardson, G. Puskorius, Y. Gur, A. Cooprider, "Quantifying Hands-free Call Quality in an Automobile," Proceedings of the SAE Sound and Vibration Conference, Grand Rapids,MI,2015.