Audio Engineering Society

Convention Express Paper 357

Presented at the AES 159th Convention 2025
October 23–25, Long Beach, CA, USA

*This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.*

# Towards Robust Speech Quality Evaluation in Challenging Acoustic Conditions

Gabriele Ravizza[1]

[1] *SenseLab, FORCE Technology, Hørsholm, Denmark*

Correspondence should be addressed to Gabriele Ravizza (gara@forcetechnology.com)

**ABSTRACT**

Assessing speech quality in noisy environments remains a significant challenge for research and development. This study evaluates the robustness of four objective models—3QUEST, VISQOL, NISQA, and DNSMOS P.835—by comparing their predictions with controlled listening test data from 32 participants. Three perceptual dimensions were analyzed: overall quality (G-MOS), speech quality (S-MOS), and noise intrusiveness (N-MOS). Results show that objective metrics reliably capture overall quality trends across diverse background noise scenarios, with correlations often exceeding 0.9. However, ranking system performance proved less consistent, particularly in complex conditions. Intrusive metrics demonstrated more substantial alignment with speech quality judgments. Higher-order ambisonics noise scenes did not weaken metric reliability. These findings highlight the usefulness of objective metrics for broad evaluation while emphasizing the need for hybrid approaches that incorporate targeted perceptual validation.

## 1 Introduction

The dualism between perceptual listening tests and objective audio metrics for speech quality evaluation is a well-known topic in the audio industry. Over the years, numerous metrics have been introduced to meet the need for rapid and repeatable validation of R&D efforts and product performance characterization. In a world where users expect their devices to excel in increasingly challenging conditions, it is valuable to investigate how well objective metrics correlate with listening test data. Multiple factors can contribute to a potential discrepancy between objective and subjective ratings, including training model datasets, shifts in listeners' audio quality expectations, and variability in product performance across the market. This paper examines the correlation between widely adopted objective metrics and real listening data. While similar studies have been conducted previously [1, 2], the novelty of this work lies in its focus on complex background noise conditions and the robust collection of listening data, avoiding reliance on crowdsourced data or aggregated

| | Level L (dBA) | Level R (dBA) |
|---|---|---|
| Call Center | 56.6 | 57.8 |
| Traffic Road | 74.9 | 73.9 |

**Table 1:** Background noise scenes and their respective dBA levels. Data is taken from [4].

results from diverse and potentially inconsistent testing environments.

## 2 Methods

Six in-ear headphones were selected for this test. The devices are anonymized. The microphone path of each device was recorded in silence and in the presence of background noise.. All recordings were normalized to -26 dB FS.

A fully compliant EBU 3276 and ITU-R BS.1116 room is equipped with four 8050A and three 8320A Genelec loudspeakers. Each loudspeaker is calibrated to achieve a flat response at the array's centre. The array is driven by an RME Fireface UC sound card connected to a Windows laptop. All measurements were taken in controlled conditions using Listen Inc.'s SoundCheck 21.

Speech material was selected from the list of phonetically balanced sentences [3]. Four sentences, two male and two female talkers, were selected and played through the mouth simulator of a B&K HATS Type 5128. The mouth simulator was calibrated to achieve a flat response between 100 Hz and 10 kHz. The playback level of the speech material was calibrated to 89 dB SPL at MRP.

### 2.1 ETSI Background Noise Scenarios

Two background noise scenarios were selected from the ETSI ES 202 396-1 database [4]. These binaural audio signals were mapped to the four Genelec 8050A speakers, with each channel duplicated on either side. Speakers were located at 2 m from the array's centre and at $\pm$ 45° and $\pm$ 135°, where 0° corresponds to the HATS' front. Details regarding the calibrated level of the selected noise scenarios are available in Table 1. Real listeners' test data gathered under these background noise conditions are hereinafter referred to as "ETSI listening test data".

### 2.2 Higher-Order-Ambisonics Background Noise Scenarios

Two background noise scenarios were selected from the Higher-Order-Ambisonics sound scene repository (HOA-SSR). The audiovisual HOA-SSR dataset was created in a joint project between FORCE Technology, Bang & Olufsen, Demant, GN Store Nord, Sonova International, WSAudiology, XRHub, and other industrial partners. The scenarios were chosen to be similar, in properties and original recording level, to the ETSI background noise scenarios described in subsection 2.1. The audio signals were captured using the em32 Eigenmike, a spherical microphone array consisting of 32 individual microphones. The recordings were initially produced in a raw 32-channel ambisonic A-format and subsequently processed into a horizontal plane 7-channel 3rd-order ambisonic B-format (AmbiX) using SN3D normalization. All audio files were encoded in PCM format at 1152 kbps per channel, with 24-bit resolution and a 48 kHz sample rate [5]. The scenarios were mapped on the seven Genelec loudspeakers located at 0°, $\pm$45°, $\pm$90°, $\pm$135°. The higher-order ambisonics scenarios were included in the study to investigate whether the audio metrics would remain robust when used to evaluate noise suppression capabilities in more challenging conditions. The same background noise conditions (Call Center and Road) as those selected from the ETSI database were chosen. They were calibrated to achieve the same level. Real listeners' test data gathered under these background noise conditions are hereinafter referred to as "HOA listening test data".

### 2.3 Setup of the Listening Test

An ITU-T P.835 listening test was performed in sound-insulated booths with background noise attenuation conforming to NR10 [6]. The listening test was administered through the SenseLabOnline testing software. The playback headphones were the Sennheiser HD650. Thirty-two naïve listeners performed the test. They all self-reported having normal hearing. All listeners performed two 1-hour listening tests, in which they were asked to rate the speech quality, background noise intrusiveness, and overall audio quality of the same audio samples used to benchmark the objective audio metrics. Thus, each listener provided three distinct ratings on a discrete scale, ranging from 1 to 5, for each sample.

### 2.4 Objective Speech Quality Metrics

Objective audio quality metrics provide quantitative ways to assess perceived sound quality, often based on listener data. Intrusive metrics require both the original and the processed audio, comparing them to measure signal degradation. This makes them more accurate but less practical when a reference is unavailable. Non-intrusive metrics, by contrast, evaluate only the distorted signal, typically using perceptual or machine learning models. While they are more flexible and suitable for real-time or field applications, they are generally less precise because they rely on estimation rather than direct comparison.

### 2.5 Intrusive Metrics

#### 2.5.1 3QUEST

3QUEST is a tool for objective audio evaluation developed by HEAD Acoustics, specifically designed to assess speech quality in various acoustic environments. It is a so-called intrusive metric, as it requires the noise-suppressed signal and two additional signals: a clean reference signal and a noisy reference signal. Using psychoacoustic models, 3QUEST predicts the impact of noise, reverberation, and other distortions on speech intelligibility and quality. These characteristics have made it particularly valuable for developers and researchers in the telecommunications, automotive, and consumer electronics industries. 3QUEST scores have been widely used across the audio industry to provide a repeatable assessment of the audio quality over an audio communication channel. 3QUEST comprises three score sub-categories: G-MOS (Global Mean Opinion Score), assessing overall audio quality; N-MOS (Noise Mean Opinion Score), assessing noise intrusiveness; and S-MOS (Speech Mean Opinion Score), assessing speech distortion. 3QUEST in the ETSI TS 103 106 wideband mode version was used in this study. [7]

#### 2.5.2 VISQOL

VISQOL (Virtual Speech Quality Objective Listener) is a machine learning-based model that evaluates audio quality by mimicking human perception, specifically for speech and audio signals. Initially developed for telecommunication, VISQOL compares a reference signal with a degraded signal to measure perceived quality using a similarity index. It uses a psychometric model to calculate objective quality scores that correlate well with subjective human assessments. VISQOL can assess voice quality in VoIP, audio streaming, and video calls. It can be used for both narrowband (telephone-quality) and wideband (high-fidelity) audio. VISQOL first extracts features from the reference and degraded signals, then compares them through a perceptual model and produces a score that reflects the quality loss. It has been integrated into toolsets for improving speech processing systems and is often used in network quality assurance. The VISQOL version used in the current study was as implemented in Mathworks' MATLAB version 2024a. [8–11]

### 2.6 Non-Intrusive Metrics

#### 2.6.1 NISQA

NISQA (Non-Intrusive Speech Quality Assessment) is a deep learning framework to predict speech quality. The model weights enable the assessment of speech quality transmitted through communication systems, such as telephone or video calls. In addition to predicting overall speech quality, NISQA provides detailed evaluations in other dimensions, including noise, colouration, discontinuity, and loudness, providing insight into the specific factors contributing to quality degradation. The framework was developed on a dataset comprising over 14,000 speech samples, each annotated with subjective quality ratings and labels for each quality dimension. For the current study, only the overall speech quality and the noisiness attributes were used. NISQA v2.0 was used in this study. [12]

#### 2.6.2 DNSMOS P.835

DNSMOS P.835 is a non-intrusive deep learning model developed to assess the perceptual quality of speech that has undergone noise suppression. It predicts three critical scores for audio: speech quality (SIG), background noise quality (BAK), and overall quality (OVRL). DNSMOS P.835 is trained using subjective evaluations based on ITU-T Recommendation P.835, which assesses speech and background noise separately, as well as the overall audio quality rating. This model has been shown to have a high correlation with human ratings in the test set, achieving Pearson's Correlation Coefficient (PCC) values of 0.94 for SIG and 0.98 for both BAK and OVRL. The model uses log power spectrogram features to evaluate audio clips and has been made publicly available as part of Microsoft's Azure

services. DNSMOS P.835 has proved helpful in ranking noise suppressors and evaluating audio processed by different noise reduction models [13]. The DNSMOS P.835 version used in this study was retrieved via the open-universe repository [14].

### 2.7 Analysis

Variable relationships were assessed using Pearson and Spearman correlations, while Root Mean Square Error (RMSE) was applied to quantify absolute prediction accuracy. Together, these measures provide insight into both the strength of the association and the magnitude of the error.

#### 2.7.1 Pearson Correlation Coefficient

The Pearson correlation coefficient measures the linear relationship between two variables X and Y. It ranges from -1 (perfect negative linear correlation) to +1 (perfect positive linear correlation), with 0 indicating no linear correlation. Its formula is based on covariance normalized by the product of standard deviations:

$$r = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

#### 2.7.2 Spearman Correlation Coefficient

The Spearman rank correlation coefficient is a non-parametric measure of monotonic association between two variables, based on their ranks. When there are no ties, it is given by:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where $d_i$ is the difference between the ranks of $x_i$ and $y_i$.

#### 2.7.3 Root Mean Squared Error

The Root Mean Squared Error (RMSE) is a metric to measure the difference between two sets of values, typically representing observed and predicted data. It quantifies the average magnitude of the errors while giving higher weight to larger deviations due to the squaring operation. Mathematically, for two variables $x_i$ and $y_i$ with $n$ paired observations, RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-y_i)^2}.$$

It is worth noting that a RMSE $\geq 1$ moves a rating into a different category on the scale.

## 3 Results

The analysis focused on three key metrics: G-MOS (overall audio quality), S-MOS (speech quality), and N-MOS (noise intrusiveness).

The overall audio quality, as evaluated by real listeners, was compared to 3QUEST's G-MOS, VISQOL's score (since this model provides only one score), NISQA's speech quality measure (to investigate whether it would correlate more with G-MOS or S-MOS listening data results), and DNSMOS P.835's OVRL.

The speech quality, as evaluated by real listeners, was compared to 3QUEST's S-MOS, VISQOL's score (the same one as above), NISQA's speech quality measure (the same one as above), and DNSMOS P.835's SIG. The speech-quality ground-truth listeners' data was analyzed against VISQOL and NISQA scores to investigate whether ground-truth speech scores correlated with these metrics to a higher degree than the overall audio quality real-listeners data.

The noise intrusiveness, as evaluated by real listeners, was compared to 3QUEST's N-MOS, NISQA's noisiness measure, and DNSMOS P.835's BAK. VISQOL wasn't part of this section of the analysis since it does not offer measures of noisiness.

The results of ETSI and HOA background noise scenarios were analyzed and compared.

### 3.1 G-MOS

The overall audio quality scores (G-MOS) showed strong correlation with perceptual listening test data across all metrics. Across ETSI and HOA background noise scenarios, all metrics showed a Pearson correlation coefficient higher than 0.90 (except for 3QUEST with ETSI background, which correlated at 0.89), indicating that these metrics effectively captured the perceived overall quality of the audio across the evaluated scenarios (silence, call center, and road noise). The consistency across background noise conditions suggests

that these objective metrics can reliably predict trends within overall audio quality in real-world listening environments when evaluating a pool of noise suppressors. Slightly different results are obtained when examining the ranking. The Spearman correlation coefficient shows how the metrics rank different systems (devices under test) with higher variability. Overall, the results indicate that the higher-order ambisonics background noise scenarios didn't harm the reliability of the metrics, but rather resulted in an increase in their accuracy. Results are displayed in Figure 1 and Figure 4.

| **ETSI G-MOS** | Pearson | Spearman | RMSE |
|---|---|---|---|
| 3QUEST | 0.89 | 0.76 | 0.70 |
| VISQOL | 0.93 | 0.86 | 0.43 |
| NISQA | 0.94 | 0.90 | 0.39 |
| DNSMOS P.835 | 0.94 | 0.92 | 0.50 |

**Table 2:** G-MOS (Overall Audio Quality) Pearson and Spearman correlation coefficients, and RMSE between each audio metric score and ETSI listening data.

| **HOA G-MOS** | Pearson | Spearman | RMSE |
|---|---|---|---|
| 3QUEST | 0.94 | 0.97 | 0.62 |
| VISQOL | 0.95 | 0.97 | 0.29 |
| NISQA | 0.92 | 0.91 | 0.50 |
| DNSMOS P.835 | 0.90 | 0.93 | 0.42 |

**Table 3:** G-MOS (Overall Audio Quality) Pearson and Spearman correlation coefficients, and RMSE between each audio metric score and HOA listening data.

### 3.2 S-MOS

For speech quality (S-MOS), the correlation with objective metrics was somewhat lower but still significant. Within the ETSI background noise scenarios, all metrics exhibited correlation coefficients higher than 0.8, with VISQOL correlating very well with real listeners' speech quality ratings (r = 0.93). It is an interesting but not wholly unexpected result, as while it's true that VISQOL outputs only one score, the overall audio quality ratings by real listeners correlate with this metric to an even higher degree than the speech quality ratings. Within the HOA background noise scenarios, 3QUEST and VISQOL exhibited Pearson coefficients of 0.88 and 0.87, respectively. NISQA showed a lower correlation of 0.78. It is of interest to note that NISQA's score is a measure of speech quality; thus, a better correlation with the S-MOS than the G-MOS should be expected. Results show an inverse trend, with NISQA speech quality scores correlating significantly better with the overall audio quality ratings than the speech quality ones. This finding is particularly relevant to researchers and development engineers who work with NISQA. DNSMOS P.835 had the weakest correlation (r = 0.69). These results suggest that intrusive metrics, such as 3QUEST and VISQOL, better predict speech quality, while non-intrusive metrics, like DNSMOS P.835, may be less accurate, especially in more complex background noise conditions. Spearman correlation, especially for ETSI background noise conditions, yielded lower results, ranging from 0.63 for 3QUEST to 0.87 for VISQOL.

| **ETSI S-MOS** | Pearson | Spearman | RMSE |
|---|---|---|---|
| 3QUEST | 0.83 | 0.63 | 0.64 |
| VISQOL | 0.93 | 0.87 | 0.64 |
| NISQA | 0.87 | 0.80 | 0.62 |
| DNSMOS P.835 | 0.83 | 0.82 | 0.69 |

**Table 4:** S-MOS (Speech Quality) Pearson and Spearman correlation coefficients, and RMSE between each audio metric score and ETSI listening data.

| **HOA S-MOS** | Pearson | Spearman | RMSE |
|---|---|---|---|
| 3QUEST | 0.88 | 0.91 | 0.46 |
| VISQOL | 0.87 | 0.84 | 0.62 |
| NISQA | 0.78 | 0.75 | 0.76 |
| DNSMOS P.835 | 0.69 | 0.82 | 0.67 |

**Table 5:** S-MOS (Speech Quality) Pearson and Spearman correlation coefficients, and RMSE between each audio metric score and HOA listening data.

### 3.3 N-MOS

Regarding noise intrusiveness (N-MOS) under ETSI background noise conditions, all metrics exhibit a Pearson correlation coefficient higher than 0.8. Spearman correlation results are outstanding for NISQA and DNSMOS P.835, but slightly worse for 3QUEST. Regarding results with HOA background noise scenes, 3QUEST and NISQA performed similarly, with Pearson correlation values of 0.86 and 0.89, respectively. This suggests

that both metrics are well-suited for evaluating the impact of background noise on perceived audio quality. DNSMOS P.835, provided a lower correlation (0.77), indicating that it may be less effective in measuring intrusive background noise. Looking at Figure 3 and Figure 6, it is interesting to qualitatively analyze how, even though metrics and listening data are correlating quite well, the absolute difference between metrics and real listeners' scores is often considerable. The extreme case is represented by noise intrusiveness ratings within the "Road" background noise scenario, which was the more challenging of the two. In this case, both 3QUEST and DNSMOS P.835 overestimated the noise suppression capabilities of the systems under test, resulting in N-MOS predictions that differed significantly from the listeners' data. The NISQA noisiness metric proved to be more accurate, both in terms of correlation and RMSE.

As mentioned earlier, VISQOL wasn't part of the noise intrusiveness results analysis, as it provides only one score and no noisiness ratings.

| **ETSI N-MOS** | Pearson | Spearman | RMSE |
|---|---|---|---|
| 3QUEST | 0.82 | 0.77 | 1.04 |
| NISQA | 0.91 | 0.93 | 0.62 |
| DNSMOS P.835 | 0.81 | 0.95 | 1.01 |

**Table 6:** N-MOS (Noise Intrusiveness) Pearson and Spearman correlation coefficients, and RMSE between each audio metric score and ETSI listening data.

| **HOA N-MOS** | Pearson | Spearman | RMSE |
|---|---|---|---|
| 3QUEST | 0.86 | 0.87 | 1.21 |
| NISQA | 0.89 | 0.87 | 0.64 |
| DNSMOS P.835 | 0.77 | 0.96 | 1.08 |

**Table 7:** N-MOS (Noise Intrusiveness) Pearson and Spearman correlation coefficients, and RMSE between each audio metric score and HOA listening data.

## 4 Discussion and Conclusions

The results reveal insights into the performance of objective audio metrics in complex listening conditions. The high correlation observed between G-MOS scores and objective metrics confirms the efficacy of these metrics in predicting overall audio quality across different background noise scenarios. These findings are consistent with previous research, which suggests that these metrics are robust for general audio quality assessments in controlled and semi-controlled environments.

At the same time, ranking of systems under test proves a more challenging task for audio metrics. As these tools are often used for Research and Development purposes, it is not uncommon for system design decisions to be made based on such data. It will be logical to the reader why this can be problematic, as the results clearly show how one system can outperform another depending on the audio metrics used to evaluate its performance. This highlights certain limitations of objective metrics in general when compared to actual listening data. Metrics will always be as good as, or no better than, the data they were trained on, posing a challenge to ensure that reliable data can be gathered under realistic and controlled conditions.

The role of background noise scenario complexity also emerged as a key factor. Interestingly, the Higher-Order Ambisonics (HOA) conditions did not degrade the reliability of the metrics. More realistic and complex acoustic scenes do not significantly challenge the robustness of objective metrics. Still, this outcome does not minimize the importance of designing evaluation frameworks that closely replicate real-world listening environments, rather than relying solely on simplified test conditions.

From a development perspective, the findings emphasize the dependency of objective metrics on the diversity and quality of their training data. As new codecs [15], algorithms, and machine-learning-based noise suppression models [16, 17] continue to evolve, along with user expectations, retraining and updating these metrics with modern and representative datasets will be essential to ensure their continued validity.

Finally, this study motivates exploration of hybrid evaluation frameworks that combine the efficiency of objective metrics with strategically targeted perceptual evaluation. Such approaches may provide the necessary balance between reproducibility, speed, and perceptual accuracy, ultimately making evaluation pipelines more reliable for iterative R&D.

## References


[1] Torcoli, M., Kastner, T., and Herre, J., "Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, pp. 1530–1541, 2021, doi:10.1109/TASLP.2021.3069302.

[2] Isabelle, S., Francois, H., and Oh, E., "Objective Measures of Voice Quality for Mobile Handsets," in *140th Audio Engineering Society Convention*, 2016.

[3] "IEEE Recommended Practice for Speech Quality Measurements," *IEEE No 297-1969*, pp. 1–24, 1969, doi:10.1109/IEEESTD.1969.7405210.

[4] ETSI, "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," in *ES 202 396-1 V1.8.1*, 2022.

[5] Fela, R. F., Pastor, A., Le Callet, P., Zacharov, N., Vigier, T., and Forchhammer, S., "Perceptual Evaluation on Audio-Visual Dataset of 360 Content," in *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, 2022, doi:10.1109/ICMEW56448.2022.9859426.

[6] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," 2003.

[7] ETSI, "TS 103 106 - V1.1.1 - 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.

[8] Hines, A., Skoglund, J., Kokaram, A., and Harte, N., "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, 2015, 2015, doi:10.1186/s13636-015-0054-9.

[9] Sloan, C., Harte, N., Kelly, D., Kokaram, A., and Hines, A., "Objective Assessment of Perceptual Audio Quality Using ViSQOLAudio," *IEEE Transactions on Broadcasting*, PP, pp. 1–13, 2017, doi:10.1109/TBC.2017.2704421.

[10] Chinen, M., Lim, F. S. C., Skoglund, J., Gureev, N., O'Gorman, F., and Hines, A., "ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric," 2020.

[11] MathWorks, "VISQOL - Objective metric for perceived audio quality," MATLAB 2024a.

[12] Mittag, G., Naderi, B., Chehadi, A., and Möller, S., "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Interspeech 2021*, ISCA, 2021, doi:10.21437/interspeech.2021-299.

[13] Reddy, C. K. A., Gopal, V., and Cutler, R., "DNSMOS P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," 2022.

[14] "Universal Score-based Speech Enhancement with High Content Preservation," in *Proc. Interspeech 2024*, 2024.

[15] Coldenhoff, J., Granqvist, N., and Cerňak, M., "OpenACE: An Open Benchmark for Evaluating Audio Coding Performance," 2024, doi:10.48550/arXiv.2409.08374.

[16] Nogales, A., Caracuel-Cayuela, J., and García-Tejedor, A. J., "Analyzing the Influence of Diverse Background Noises on Voice Transmission: A Deep Learning Approach to Noise Suppression," *Applied Sciences*, 14, p. 740, 2024, doi:10.3390/app14020740.

[17] Healy, E., Johnson, E., Pandey, A., and Wang, D., "Progress made in the efficacy and viability of deep-learning-based noise reduction," *The Journal of the Acoustical Society of America*, 153, p. 2751, 2023, doi:10.1121/10.0019341.

# 5 Appendix 1 - Plots

## 5.1 ETSI Background Noise Scenarios

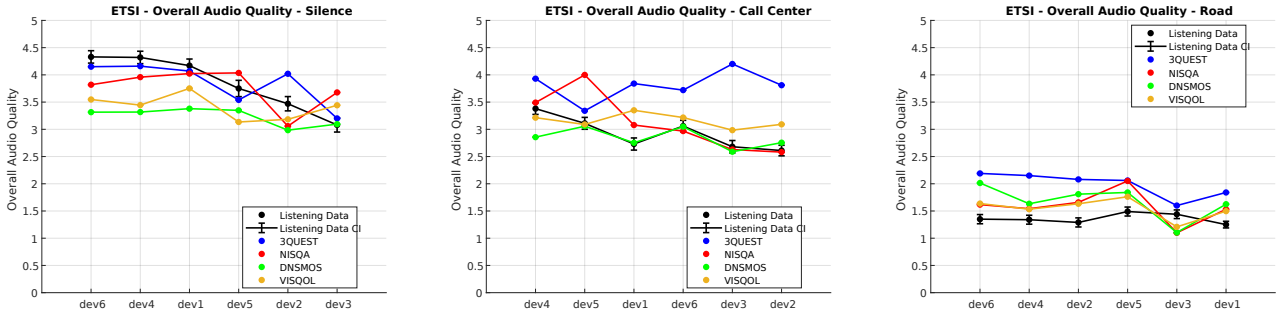



**Fig. 1:** ETSI, Overall Audio Quality

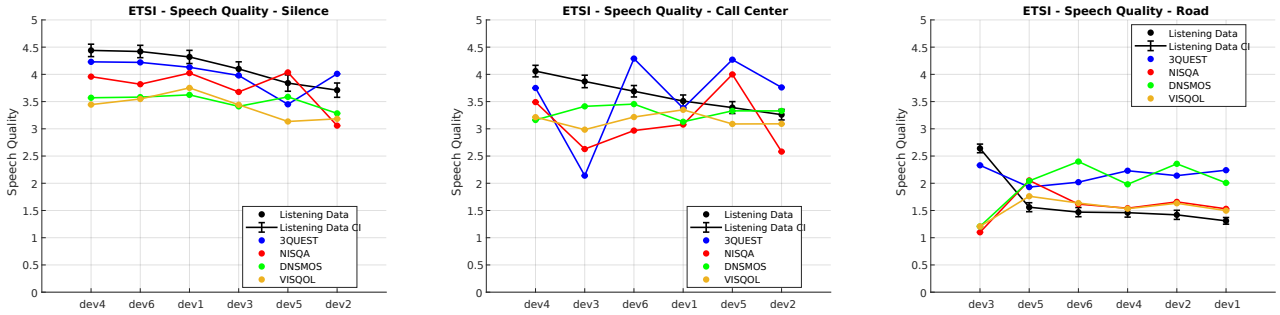



**Fig. 2:** ETSI, Speech Quality

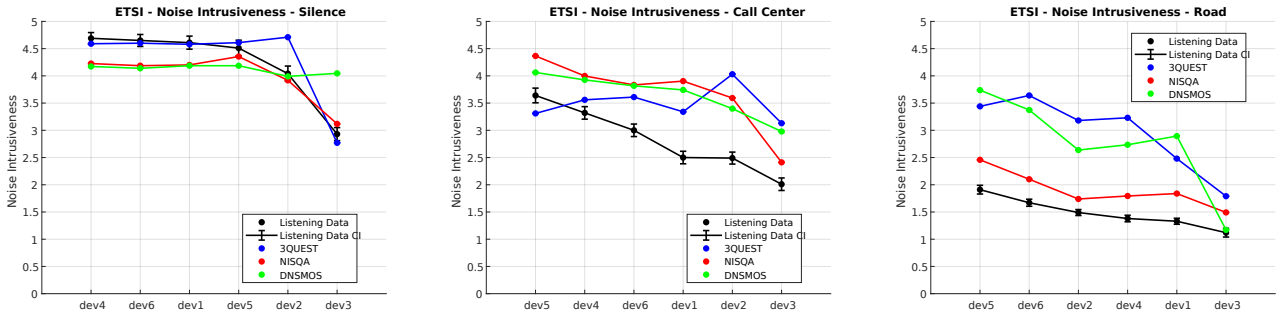



**Fig. 3:** ETSI, Background Noise Intrusiveness

### 5.2 HOA Background Noise Scenarios

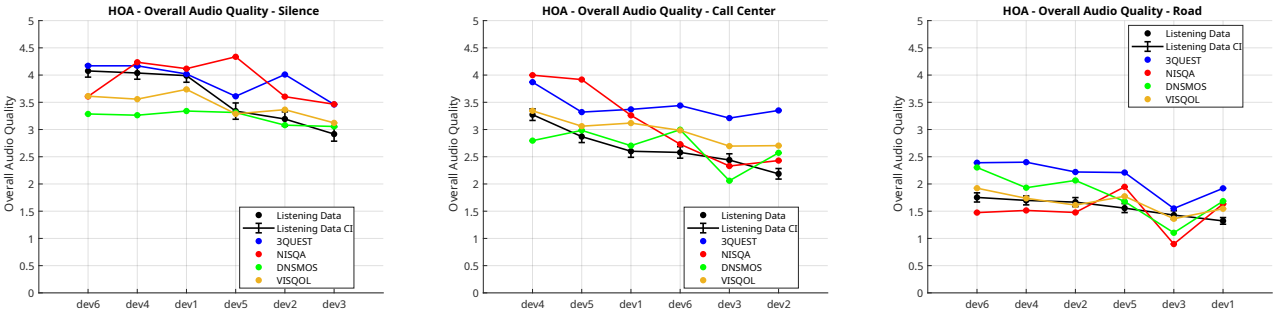



**Fig. 4:** HOA, Overall Audio Quality

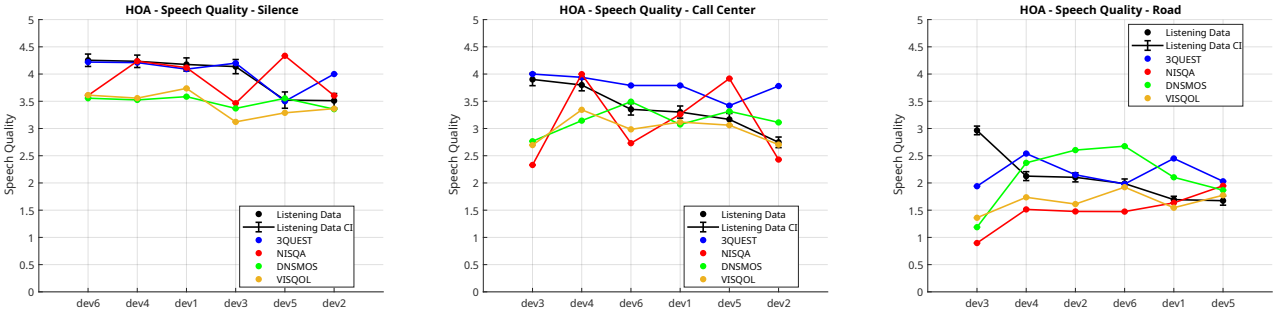



**Fig. 5:** HOA, Speech Quality

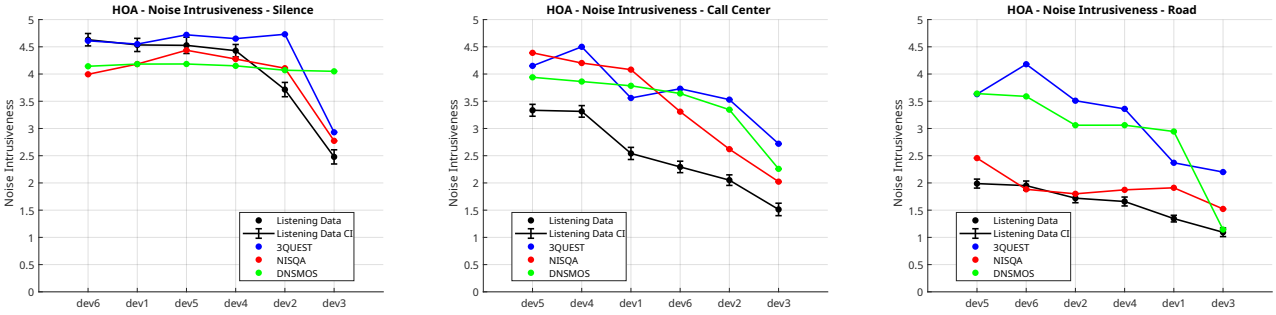



**Fig. 6:** HOA, Background Noise Intrusiveness

# 6 Appendix 2 - Cross-metric Correlation

## 6.1 ETSI Background Noise Conditions

| Metric 1 | Metric 2 | Pearson | Spearman | Kendall |
|---|---|---|---|---|
| NISQA G-MOS | DNSMOS G-MOS | 0.96 | 0.95 | 0.83 |
| NISQA G-MOS | 3QUEST G-MOS | 0.83 | 0.64 | 0.48 |
| NISQA G-MOS | VISQOL G-MOS | 0.93 | 0.82 | 0.71 |
| NISQA N-MOS | DNSMOS N-MOS | 0.84 | 0.93 | 0.80 |
| NISQA N-MOS | 3QUEST N-MOS | 0.74 | 0.69 | 0.50 |
| DNSMOS G-MOS | 3QUEST G-MOS | 0.90 | 0.67 | 0.49 |
| DNSMOS G-MOS | VISQOL G-MOS | 0.96 | 0.87 | 0.75 |
| DNSMOS S-MOS | 3QUEST S-MOS | 0.79 | 0.66 | 0.46 |
| DNSMOS S-MOS | VISQOL G-MOS | 0.95 | 0.81 | 0.63 |
| DNSMOS N-MOS | 3QUEST N-MOS | 0.76 | 0.69 | 0.49 |
| 3QUEST G-MOS | VISQOL G-MOS | 0.95 | 0.75 | 0.58 |
| 3QUEST S-MOS | VISQOL G-MOS | 0.87 | 0.70 | 0.43 |

**Table 8:** Pearson, Spearman, and Kendall correlation coefficients between each audio metric's ratings in ETSI background noise conditions. Only relevant comparisons are included.

## 6.2 HOA Background Noise Conditions

| Metric 1 | Metric 2 | Pearson | Spearman | Kendall |
|---|---|---|---|---|
| NISQA G-MOS | DNSMOS G-MOS | 0.91 | 0.87 | 0.67 |
| NISQA G-MOS | 3QUEST G-MOS | 0.91 | 0.87 | 0.71 |
| NISQA G-MOS | VISQOL G-MOS | 0.95 | 0.90 | 0.75 |
| NISQA N-MOS | DNSMOS N-MOS | 0.77 | 0.87 | 0.70 |
| NISQA N-MOS | 3QUEST N-MOS | 0.79 | 0.76 | 0.57 |
| DNSMOS G-MOS | 3QUEST G-MOS | 0.93 | 0.92 | 0.79 |
| DNSMOS G-MOS | VISQOL G-MOS | 0.93 | 0.92 | 0.79 |
| DNSMOS S-MOS | 3QUEST S-MOS | 0.84 | 0.80 | 0.61 |
| DNSMOS S-MOS | VISQOL G-MOS | 0.91 | 0.91 | 0.78 |
| DNSMOS N-MOS | 3QUEST N-MOS | 0.81 | 0.85 | 0.69 |
| 3QUEST G-MOS | VISQOL G-MOS | 0.98 | 0.97 | 0.89 |
| 3QUEST S-MOS | VISQOL G-MOS | 0.93 | 0.85 | 0.66 |

**Table 9:** Pearson, Spearman, and Kendall correlation coefficients between each audio metric's ratings in HOA background noise conditions. Only relevant comparisons are included.