Audio Engineering Society

Convention Express Paper 430

Presented at the AES 160th Convention
2026 May 28–30, Copenhagen, Denmark

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 The complete manuscript was not peer reviewed.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.

Objective Quality Models for Decision-Making in Speech Coding

Clémence Lamballe, Philippe Gournay, and Roch Lefebvre

Speech and Audio Research Group (GRPA), Department of Electrical and Computer Engineering, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada

Correspondence should be addressed to Clémence Lamballe (Clemence.Lamballe@USherbrooke.ca)

ABSTRACT

Objective quality evaluation is widely used in speech coding, yet objective estimates often exhibit limited agreement with subjective listening-test results. Rather than focusing on absolute score accuracy, this paper evaluates objective speech quality models from a decision-making perspective, defined as their ability to support comparative judgments between speech codecs or codec configurations. A formal ITU-R P.800 Absolute Category Rating (ACR) listening test was conducted with 30 listeners across 24 conditions, covering conventional and neural monophonic speech codecs operating under clear-channel conditions at sampling frequencies from 16 to 48 kHz and bitrates ranging from below 1 kbps to above 16 kbps. The speech material consisted of internally recorded, clean French-language speech that was not used in the development or training of any of the evaluated codecs or objective quality models. Seven objective quality models, namely PESQ, VISQOL Speech, VISQOL Audio, WARP-Q, NISQA, UTMOS, and DistillMOS, were evaluated on the same material. Decision-making performance was assessed by comparing subjective and objective rankings using Kendall's rank correlation coefficient and by assessing the consistency of statistical discrimination across pairwise codec comparisons using two-sided t-tests at a 95% confidence level. The results indicate that some objective quality models are effective for comparing variations within a given speech coding technology under consistent operating conditions, such as fixed sampling frequency. However, all models exhibit limitations, including systematic over- or underestimation for certain coding technologies and reduced reliability when applied across different sampling frequencies. Despite its conventional origins, PESQ remains capable of supporting decision-making when applied to neural speech codecs.

1 Introduction

Quality evaluation plays a central role in speech coding, whether for guiding technology development, assessing performance under realistic operating conditions, or verifying compliance with specifications and standards. Such evaluations can be performed through listening

tests, which provide subjective measures based on human judgment, or through perceptual-model–based software tools, which provide objective estimates of perceived speech quality. In subjective listening tests, such as those defined in ITU-R P.800 using the Absolute Category Rating (ACR) methodology [1], listeners rate speech signals on a discrete scale, and the resulting Mean Opinion Score (MOS) serves as a reference measure of perceived quality.

While subjective evaluations provide the most reliable assessment of perceived quality, they are time-consuming and resource-intensive. As a result, objective quality models have been widely adopted as practical alternatives. These models can be intrusive, requiring a reference signal, or non-intrusive, operating only on degraded signals, and may rely on traditional signal processing or neural network–based approaches. However, objective estimates often exhibit limited agreement with subjective listening-test results, with correlation known to vary across conditions [2, 3]. Several approaches, including regression-based mappings from objective scores to subjective ratings, have been proposed to improve this relationship [4, 2, 5], but their effectiveness remains limited and condition-dependent.

In this study, we adopt a different perspective: rather than evaluating objective estimates solely by their correlation with subjective measures, we assess their practical utility in supporting decisions, such as selecting the most suitable speech-coding technology or configuration for a given application. This shifts the focus from absolute perceptual quality prediction toward decision support.

In this context, decision-making in speech coding is defined as the ability to support comparative judgments between codecs or codec configurations, as performed in subjective listening tests. An objective quality model is considered suitable for decision-making if it yields rankings consistent with subjective evaluations and identifies statistically significant quality differences without producing decisions that contradict subjective evaluations. To assess these properties, ranking agreement is quantified using Kendall's rank correlation coefficient [6], and the consistency of statistical discrimination across pairwise codec comparisons is analyzed using two-sided t-tests at a 95% confidence level.

To investigate these aspects, we conducted a formal listening test on a set of monophonic speech-coding technologies comprising conventional and neural approaches, operating at sampling frequencies from 16 kHz to 48 kHz and bitrates from below 1 kbps to above 16 kbps. The evaluation was performed under controlled conditions using clean speech and a clear transmission channel, enabling a focused analysis of codec behavior and objective-model performance.

The remainder of this paper is organized as follows. Section 2 describes the subjective listening test methodology and the evaluated speech codecs. Section 3 introduces the objective quality models considered in this study. Section 4 presents the evaluation framework used to assess decision-making performance. Section 5 reports the results of the comparative analysis, and Section 6 investigates decision-making behavior under bitrate variations of the same coding technology. Finally, Section 7 summarizes the main findings and practical considerations, and Section 8 concludes the paper.

2 Subjective listening test

This section describes the subjective listening test conducted to provide a reference for evaluating objective speech quality models from a decision-making perspective. The test was designed in accordance with ITU-R P.800 recommendations and focused on clean speech signals evaluated under clear-channel conditions. The following subsections detail the speech material, evaluated speech codecs, test conditions, and subjective evaluation protocol.

2.1 Speech material

A custom speech dataset was created to ensure that none of the speech files used in this study had been used in the development or training of the evaluated speech codecs or objective speech quality models. The speech material was generated from six pairs of sentences spoken by four female and four male speakers. The source speech dataset thus consists of 48 internally recorded, clean monophonic French speech signals sampled at 48 kHz, each with a duration of 8 seconds.

2.2 Evaluated speech codecs

The study includes a set of speech codecs representative of both conventional and neural coding approaches. The 3GPP Enhanced Voice Services (EVS) codec was

selected as a conventional reference, providing a baseline for comparison with neural speech codecs. The neural codecs were chosen based on the availability of pretrained implementations and the perceptual quality of their coded output. Well-established codecs from the literature, such as DAC, EnCodec, and Lyra v2, were included, along with more recent approaches including Mimi, FunCodec, and SNAC. The Mimi codec, which converts audio into discrete tokens, is presented in the Moshi model paper [7], a speech-text foundation model for real-time dialogue. The versions used correspond to the most recent publicly available releases at the time of the experiment. Table 1 summarizes the evaluated codecs and their respective versions.

Table 1: Overview of the evaluated audio codecs

#	Codec	Version
1	EVS [8]	v16.3.0
2	Mimi [7]	Commit 97c4bb2
3	Lyra v2 [9]	v1.3.2
4	EnCodec [10]	0.1.2a3
5	DAC [11]	1.0.0
6	GNC	–
7	FunCodec [12]	Commit 3435d87
8	SNAC [13]	1.2.1

In addition, a simple in-house neural speech codec (GNC) was included. This codec is not intended to represent a competitive coding approach, but rather to examine how objective quality models behave on a codec that was not used in their development or training. Because it is not publicly released, it provides an independent test case with respect to the data used to develop the evaluated models.

2.3 Test conditions

The test conditions, listed in Table 2, are defined by the processing chains illustrated in Figure 1 and 2 and applied to the source speech material to generate the processed speech files. As a first step, the source speech signals were high-pass filtered at 50 Hz to remove low-frequency components that are often poorly handled by speech coding algorithms. The signals were then resampled to the required sampling frequencies (ranging from 16 kHz to 48 kHz) and level-adjusted to approximately -26 dBov at each sampling frequency.

The processed signals were subsequently degraded either by the evaluated speech codecs operating at various bitrates or by Modulated Noise Reference Units (MNRUs). The MNRU conditions were included to provide quality anchors for the listeners.

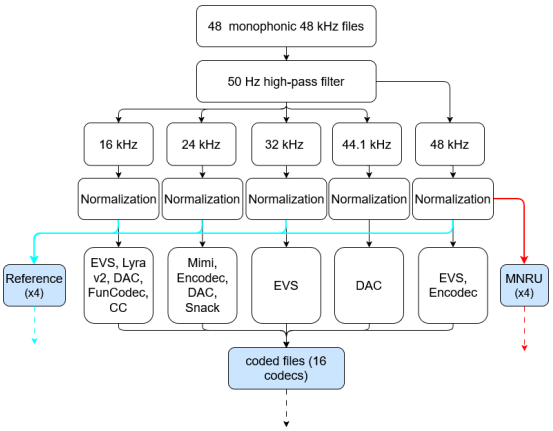


Fig. 1: Pre-processing and coding defining the test conditions

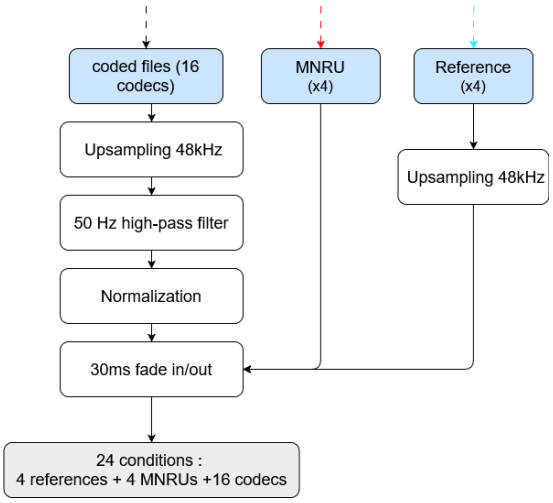


Fig. 2: Post-processing and signal conditioning

Different sampling frequencies were used to reflect the diversity of the evaluated neural speech codecs, in particular those designed to operate at 16 kHz and 24 kHz. This setup allows the evaluation of objective quality models under multiple sampling-frequency conditions. For each sampling frequency, the corresponding undegraded signals were included as reference conditions to assess coding performance.

Table 2: List of reference and coded conditions

#	Description
1	Reference 48 kHz
2	Reference 32 kHz
3	Reference 24 kHz
4	Reference 16 kHz
5	EVS: 16 kHz & 7.2 kbps
6	EVS: 32 kHz & 9.6 kbps
7	EVS: 48 kHz & 16.4 kbps
8	Mimi: 24 kHz & 4.40 kbps
9	Lyra v2: 16 kHz & 3.2 kbps
10	Lyra v2: 16 kHz & 9.2 kbps
11	EnCodec: 24 kHz & 6 kbps
12	EnCodec: 24 kHz & 24 kbps
13	EnCodec: 48 kHz & 12 kbps
14	DAC: 16 kHz & 8 kbps
15	DAC: 24 kHz & 8 kbps
16	DAC: 44.1 kHz & 8 kbps
17	GNC: 16 kHz & 3.6 kbps
18	GNC: 16 kHz & 7.2 kbps
19	SNAC: 24 kHz & 0.98 kbps
20	FunCodec: 16 kHz & 8 kbps
21	MNRU 15 dB, 48 kHz
22	MNRU 25 dB, 48 kHz
23	MNRU 35 dB, 48 kHz
24	MNRU 45 dB, 48 kHz

2.4 Subjective evaluation protocol

Subjective quality scores were collected from 30 listeners following the ITU-R P.800 Absolute Category Rating (ACR) methodology. Listeners were organized into six groups of five, with each group evaluating a distinct subset of speech samples covering all test conditions. For each processed speech file, the subjective score corresponds to the average of the five individual ratings within the group. Therefore, for each test condition, the subjective score corresponds to the average of 240 individual ratings. Subjective test results are reported as MOS LQS (Mean Opinion Score for Listening Quality – Subjective), hereafter referred to as subjective scores.

A short training phase of approximately 3 minutes was conducted prior to the test to familiarize listeners with the evaluation interface and the range of perceived quality. Each group of listeners was presented with a distinct training sequence comprising 12 speech samples covering all speakers. The processing conditions for these samples were manually selected to be representative of the full set of test conditions, spanning the expected quality range, including low- and high-SNR MNRU conditions as well as codecs employing different techniques and bitrates.

A post-screening procedure was applied to verify the consistency of listener ratings. Specifically, for each listener, the subjective score assigned to the 15 dB MNRU condition was required to be lower than that assigned to the 35 dB MNRU condition, and the subjective score for the 25 dB MNRU condition lower than that for the 45 dB MNRU condition. After completion of the listening test and application of this procedure, no listeners were rejected.

3 Objective quality models

Objective quality models aim to estimate perceived speech quality as measured by subjective listening tests. All objective models evaluated in this study are listed in Table 3. These models can be classified as intrusive, requiring an undegraded reference signal, or non-intrusive, operating without a reference. In the present study, NISQA, UTMOS, and DistillMOS belong to the non-intrusive category, the other models being intrusive. The evaluated models further differ in their underlying approach, relying either on traditional signal-processing techniques, as in PESQ, or on neural network–based methods.

Objective quality models were selected based on their open-source availability, their ability to output Mean Opinion Score (MOS) values, and their applicability to speech quality evaluation. Only models originally developed to evaluate the perceptual quality of compressed or transmission-degraded speech (PESQ, NISQA, VISQOL Speech) or the overall perceptual quality of processed speech signals (WARP-Q, UTMOS, DistillMOS) were retained. VISQOL Audio was additionally included, despite not being specifically designed for speech, due to its close relationship to VISQOL Speech, allowing assessment of how a more general audio-oriented model performs on speech signals. PESQ was included as a widely used reference model, despite the withdrawal of ITU-T Recommendation P.862.

Only two models, NISQA and VISQOL Audio, natively support operation at a sampling frequency of 48 kHz. For objective quality models operating at 16 kHz,

Models	fs (kHz)	Content type	Non Intrusive	Version
PESQ [14]	16	Speech	No	pesqc2 v0.0.4
NISQA [15]	48	Speech	Yes	v2.0
VISQOL Audio [16]	48	General audio	No	Commit 38d0b01
VISQOL Speech [16]	16	Speech	No	Commit 38d0b01
WARP-Q [17]	16	Speech	No	v1.5.2
UTMOS [18]	16	Speech	Yes	Utmosv2 v1.2.0
DistillMOS [19]	16	Speech	Yes	0.9.1

Table 3: Details of objective quality models

higher-sampling-rate processed files were downsampled accordingly. For intrusive 16 kHz-based models, the corresponding 48 kHz source files were downsampled to 16 kHz and used as reference inputs.

For WARP-Q, a linear mapping was applied to convert its native 0–1 output range to the 1–5 MOS scale for visualization purposes; this transformation preserves relative ranking relationships and does not affect comparative analyses. For UTMOS, five inference repetitions were performed for each audio file. In this configuration, the model generates predictions using randomly selected frames from the input waveform, and the final score corresponds to the average of the five outputs. Although this procedure is computationally expensive, it reduces prediction variance and improves the stability of codec discrimination. All models were evaluated using the most recent publicly available versions at the time of the experiment.

Throughout this paper, the outputs of objective quality models are reported as MOS LQO (Mean Opinion Score for Listening Quality – Objective) and are hereafter referred to as objective scores.

4 Decision-making performance metrics

This section describes the performance metrics and statistical analyses used to assess the suitability of objective quality models for decision-making in speech coding. Decision-making performance is evaluated by comparing objective and subjective results in terms of both ranking consistency and statistical discrimination between conditions. To this end, Kendall's rank correlation coefficient (τ) is used to assess global agreement between subjective and objective rankings across all conditions, whereas pairwise statistical tests provide a complementary analysis by examining consistency in statistically significant quality differences for individual condition pairs.

4.1 Ranking agreement using Kendall's τ

To assess the overall decision-making performance of objective quality models, we first compare rankings derived from objective scores with those derived from subjective scores using Kendall's τ. Kendall's τ ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates no association, and -1 indicates a complete inversion of rankings. It is mathematically defined as:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{1}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, and n is the number of conditions being ranked. For identical rankings, $n_d = 0$ and $\tau = 1$. While Kendall's τ provides a global measure of ranking agreement, it is computed from mean score differences only and does not account for the statistical significance associated with score variability.

4.2 Statistical discrimination analysis using pairwise t-tests

To complement the ranking analysis, pairwise comparisons are performed using t-tests applied to the subjective and objective score distributions at a 95% confidence level. This makes it possible to determine whether the observed differences between two conditions are statistically significant for both subjective and objective scores. For each pair of conditions, the outcomes of the subjective and objective t-tests are then compared to assess their consistency in terms of statistical discrimination.

The results of the pairwise analyses are categorized as follows:

- **Similar ranking:** Both subjective and objective t-tests indicate a statistically significant difference between the two conditions, and the corresponding mean scores rank the two conditions in the same order.

- **Reversed ranking:** Both subjective and objective t-tests indicate a statistically significant difference between the two conditions, but the corresponding mean scores rank the two conditions in opposite orders.
- **Only the objective t-test is inconclusive:** The subjective t-test indicates a statistically significant difference, whereas the objective t-test does not.
- **Only the subjective t-test is inconclusive:** The objective t-test indicates a statistically significant difference, whereas the subjective t-test does not.
- **Both t-tests are inconclusive:** Neither the subjective nor the objective t-test indicates a statistically significant difference.

This categorization enables a more detailed analysis of the ability of objective quality models to reproduce subjective discrimination decisions, beyond simple ranking agreement.

4.3 Interpretation of decision-making performance metrics

Kendall's τ provides insight into the overall consistency between subjective and objective rankings, while the pairwise t-test analysis reveals how reliably objective models reproduce subjective discrimination decisions. Good decision-making performance is characterized by a high proportion of similar rankings and a low occurrence of reversed rankings.

When only the subjective t-test is inconclusive, the interpretation of the pairwise analysis is ambiguous. This outcome may indicate a misclassification by the objective model, but it may also reflect insufficient statistical power in the subjective evaluation, where a larger number of listeners could potentially reveal a perceptible quality difference.

Therefore, the performance analysis focuses on the most relevant scenarios: similar rankings, reversed rankings, and cases where only the objective scores are inconclusive. The others remain displayed in the graphs for transparency, providing a view of the distribution across all five scenarios.

5 Decision-making performance across coding configurations

This section evaluates the decision-making performance of objective quality models using the metrics introduced in Section 4.

The analysis is conducted separately for conditions at 16 kHz and 24 kHz, which represent the most common sampling frequencies among the evaluated neural codecs, and for a more complete set of conditions spanning multiple sampling frequencies. This separation allows decision-making performance to be examined under matched sampling frequencies as well as in cross-frequency scenarios, where differences in sampling frequency may affect objective model behavior. Since the number of test conditions differs between frequency groups, performance is assessed independently within each category rather than directly across them.

For completeness, the full set of results, including mean scores and confidence intervals for subjective and objective evaluations, is reported in Table 7.

5.1 Decision-making performance for 16 kHz conditions

This subsection examines the decision-making performance of objective quality models for test conditions at a 16 kHz sampling frequency. Among all test conditions, eight correspond to 16 kHz operation, including seven coded conditions and the 16 kHz reference condition. For this subset, all possible pairwise comparisons were analyzed, resulting in 28 condition pairs. Pairwise t-test results for the 16 kHz test conditions are summarized in Figures 3 and 4.

In the following, particular attention is given to two outcome categories that are most informative for decision-making: similar rankings, where subjective and objective evaluations both detect a significant difference and rank conditions in the same order, and reversed rankings, where both detect a significant difference but rank conditions in opposite orders, representing the most detrimental behavior for decision support. These outcome categories are illustrated in Figure 3.

The remaining outcome categories are also shown in Figure 4. Table 4 reports Kendall's τ for each objective quality model for the 16 kHz test conditions. Although Kendall's τ is bounded between –1 and 1, there are no

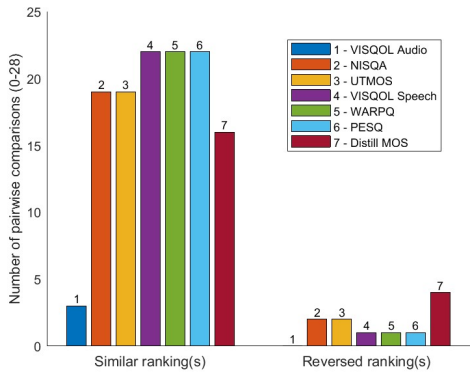


Fig. 3: Number of pairwise comparisons at 16 kHz yielding similar (left) and reversed (right) rankings between objective and subjective evaluations based on t-tests (maximum of 28 comparisons for the eight conditions)

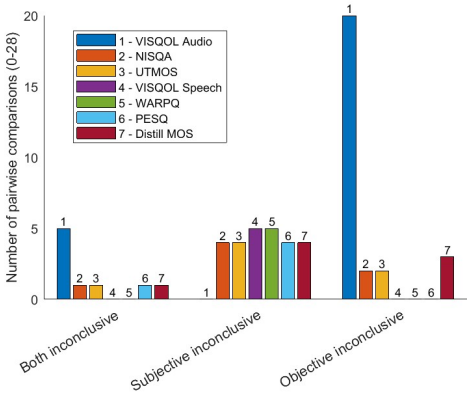


Fig. 4: Number of pairwise comparisons at 16 kHz yielding inconclusive outcomes between objective and subjective evaluations based on t-tests (maximum of 28 comparisons for the eight conditions)

universally accepted thresholds for interpreting its magnitude. For this reason, Table 4 also reports the number of discordant pairs to facilitate model comparison.

VISQOL Speech, WARP-Q, and PESQ

Kendall's τ analysis indicates that VISQOL Speech, WARP-Q, and PESQ achieve the strongest ranking consistency at 16 kHz, a result that is corroborated by the pairwise t-test analysis. These three models yield the same number of similar rankings for condition pairs where statistically significant differences are identified by the t-tests. They also share a statistically significant reversed ranking, between the GNC codec (16 kHz, 7.2 kbps) and Lyra v2 (16 kHz, 3.2 kbps). This suggests an undervaluation of the GNC codec relative to Lyra v2 by these three objective models. While Table 4 reports a total of two discordant pairs for these models, the second, between DAC (8 kbps) and FunCodec (8 kbps), is not statistically significant according to the objective t-test. These findings indicate that the models' ranking errors are surprisingly consistent with one another.

The only notable difference among these three models is observed for the comparison between EVS (16 kHz, 7.2 kbps) and Lyra v2 (16 kHz, 9.2 kbps), as observed in the pairwise comparison differences in Figure 4. In this case, subjective scores and PESQ do not indicate a statistically significant difference, whereas VISQOL Speech and WARP-Q do. It remains unclear which outcome is correct, as a true quality difference may exist but not reach statistical significance in the subjective test considering the available number of listeners.

VISQOL Audio

VISQOL Audio achieves a relatively high Kendall's τ value, suggesting strong ranking consistency when considering ranking consistency alone. However, the pairwise t-test analysis reveals a more nuanced behavior, with only a limited number of condition pairs being statistically discriminated, as illustrated in Figure 4. Examination of the subjective scores in Figure 5 indicates that VISQOL Audio scores are strongly influenced by signal bandwidth, with signals sharing similar effective bandwidths tending to receive similar scores. As a result, at a given sampling frequency—particularly at 16 kHz, where many conditions share similar effective bandwidths—the scores cluster within a narrow range. This reduced separation between conditions limits the model's ability to statistically discriminate between codec configurations, despite the apparent consistency of the rankings based on mean scores.

NISQA, UTMOS and DistillMOS

NISQA, UTMOS, and DistillMOS exhibit lower Kendall's τ values, with DistillMOS performing worst. This ranking behavior is consistent with the pairwise

Model	Kendall's τ	Number of discordant pairs (0-28)
VISQOL Audio	0.714	$n_d = 4$
NISQA	0.571	$n_d = 6$
UTMOS	0.571	$n_d = 6$
VISQOL Speech	0.857	$n_d = 2$
WARPQ	0.857	$n_d = 2$
PESQ	0.857	$n_d = 2$
DistillMOS	0.429	$n_d = 8$

Table 4: Kendall's τ values between subjective and objective rankings for the eight 16 kHz test conditions, together with the corresponding number of discordant pairs (out of 28)

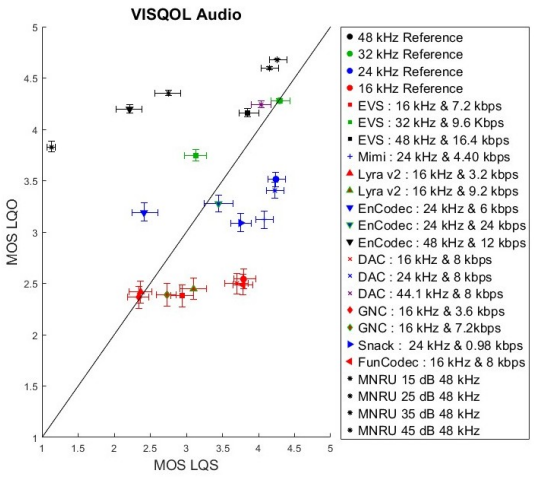


Fig. 5: Scatter plot of VISQOL Audio objective scores versus subjective scores for the 24 test conditions, with 95% confidence intervals shown on both axes

t-test analysis, which reveals fewer similar rankings and a higher incidence of reversed rankings compared to VISQOL Speech, WARP-Q, and PESQ. UTMOS and NISQA exhibit nearly identical reversed rankings, involving EVS at 7.2 kbps and GNC at 3.6 and 7.2 kbps. It remains unclear whether this behavior reflects an overestimation of GNC or an underestimation of EVS. DistillMOS shares this reversed ranking and additionally exhibits two further reversed rankings, both associated with an overestimation of GNC at 3.6 kbps.

Summary of results at 16 kHz

At 16 kHz, VISQOL Speech, WARP-Q, and PESQ exhibit the strongest decision-making performance, achieving the highest ranking consistency with subjective evaluations and the largest number of similar rankings in pairwise comparisons. These models nonetheless share a small number of reversed rankings, indicating occasional failures in correctly ordering codec quality. VISQOL Audio achieves a high Kendall's τ value but demonstrates limited statistical discrimination due to score clustering at fixed sampling frequency. NISQA, UTMOS, and DistillMOS perform less reliably at 16 kHz, with lower ranking agreement and a higher incidence of reversed rankings, limiting their effectiveness for decision-making in this configuration.

5.2 Decision-making performance for 24 kHz conditions

This subsection evaluates the decision-making performance of objective quality models for test conditions at a 24 kHz sampling frequency. Six test conditions correspond to 24 kHz operation, resulting in 15 possible pairwise comparisons. As in Section 5.1, performance is assessed through ranking agreement using Kendall's τ and through pairwise t-test analysis of statistically significant quality differences.

Pairwise t-test results are illustrated in Figures 6 and 7, while ranking agreement based on Kendall's τ is summarized in Table 5.

It should be noted that several objective quality models evaluated in this study—namely VISQOL Speech, WARP-Q, PESQ, and UTMOS—operate at a sampling frequency of 16 kHz. Consequently, when applied to 24 kHz signals, these models do not capture the full signal bandwidth available to subjective listeners or to NISQA, which operates at 48 kHz. The analysis therefore examines whether such 16 kHz-based models can still provide meaningful quality estimates and support correct ranking decisions when 24 kHz signals are downsampled to 16 kHz.

An alternative approach was also investigated. Instead of downsampling the audio to 16 kHz, the file header metadata was modified so that 24 kHz signals were interpreted by the objective quality models as 16 kHz

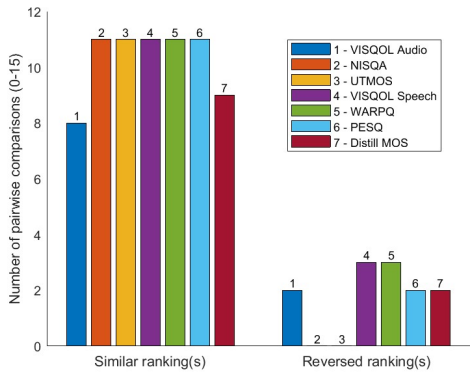


Fig. 6: Number of pairwise comparisons at 24 kHz yielding similar (left) and reversed (right) rankings between objective and subjective evaluations, based on pairwise t-tests (maximum of 15 comparisons for the six conditions)

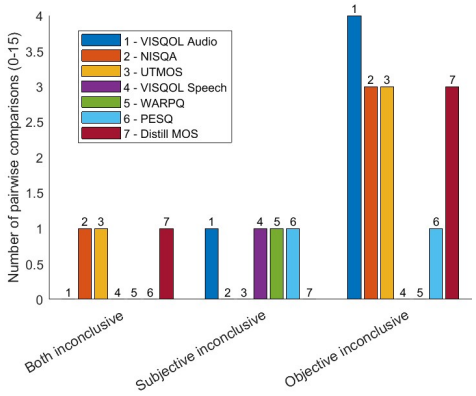


Fig. 7: Number of pairwise comparisons at 24 kHz yielding inconclusive outcomes between objective and subjective evaluations based on t-tests (maximum of 15 comparisons for the six conditions)

signals. This approach effectively exposes the full signal bandwidth to the models, but it also alters the perceived pitch and speaking rate of the signals. Under this configuration, the resulting decision-making behavior remained broadly similar to that obtained with standard downsampling, although the objective scores were systematically shifted toward lower values. Because these shifts were not uniform across codecs, particularly for PESQ and UTMOS, additional ranking errors were observed that were not present under the downsampling condition. Furthermore, this method proved unreliable when the original sampling rate differed substantially from 16 kHz. For example, the 48 kHz conditions produced tightly clustered quality scores. For these reasons, the analysis presented in this paper relies on results obtained using standard downsampling to 16 kHz.

NISQA and UTMOS

NISQA achieves one of the highest Kendall's τ values at 24 kHz, as reported in Table 5. Despite operating at different sampling frequencies, 48 kHz for NISQA and 16 kHz for UTMOS, these two models exhibit remarkably similar decision-making behavior. Their pairwise t-test outcomes are identical, as illustrated in Figures 6 and 7. A closer examination reveals a single difference between the two models for the comparison between the 24 kHz reference condition and DAC at 24 kHz and 8 kbps. In this case, subjective scores and NISQA rank the reference condition higher when comparing mean scores, whereas UTMOS does not. This discrepancy affects Kendall's τ, which depends on ranking order, but not the t-test results, since neither model detects a statistically significant difference between these two conditions.

VISQOL Speech, WARP-Q, and PESQ

VISQOL Speech, WARP-Q, and PESQ exhibit broadly similar ranking behavior for the 24 kHz conditions, as previously observed for the 16 kHz analysis. VISQOL Speech and WARP-Q produce identical reversed rankings for three pairs of conditions: EnCodec (6 kbps) versus SNAC (0.98 kbps), EnCodec (24 kbps) versus SNAC (0.98 kbps), and Mimi (4.40 kbps) versus EnCodec (24 kbps).

PESQ produces the same reversed rankings for the first two pairs of conditions. However, for the comparison between Mimi (4.40 kbps) and EnCodec (24 kbps), PESQ ranks the systems in the same order as the subjective results. As a consequence, PESQ yields a slightly higher Kendall's τ value than VISQOL Speech and WARP-Q, and yields the same Kendall's τ value as NISQA.

It should be noted, however, that PESQ does not detect a statistically significant difference between Mimi (4.40 kbps) and EnCodec (24 kbps). This explains the

Model	Kendall's τ	Number of discordant pairs (0-15)
VISQOL Audio	0.467	$n_d = 4$
NISQA	0.733	$n_d = 2$
UTMOS	0.600	$n_d = 3$
VISQOL Speech	0.600	$n_d = 3$
WARPQ	0.600	$n_d = 3$
PESQ	0.733	$n_d = 2$
DistillMOS	0.467	$n_d = 4$

Table 5: Kendall's τ values between subjective and objective rankings for the six 24 kHz test conditions, together with the corresponding number of discordant pairs (out of 15)

discrepancy observed in the pairwise t-test analysis, where VISQOL Speech and WARP-Q indicate a statistically significant difference for this pair while PESQ does not.

VISQOL Audio and DistillMOS

VISQOL Audio and DistillMOS exhibit the lowest Kendall's τ values among the evaluated models for the 24 kHz conditions. For DistillMOS, the two reversed rankings shown in Figure 6 result from an underestimation of DAC at 8 kbps. However, both models also fail to discriminate significantly between codecs, resulting in a lower number of similar rankings compared to other models. In the case of VISQOL Audio, this behavior is due to its sampling rate bias, as previously explained. In fact, the scores are tightly clustered within a narrow range, which prevents VISQOL Audio from significantly distinguishing quality differences between codecs.

Summary of results at 24 kHz

At 24 kHz, NISQA exhibits the strongest decision-making performance, achieving the highest ranking consistency with subjective evaluations and avoiding reversed rankings in pairwise comparisons. PESQ and UTMOS also show competitive performance, but with different strengths: PESQ achieves higher Kendall's τ values, while UTMOS avoids certain reversed rankings observed with PESQ. VISQOL Speech and WARP-Q display broadly similar behavior but are affected by multiple reversed rankings, which reduces their reliability for decision-making at this sampling frequency. VISQOL Audio and DistillMOS perform least effectively at 24 kHz, exhibiting low ranking consistency and limited statistical discrimination between codec conditions, which limits their usefulness for decision-making in this configuration.

5.3 Decision-making performance across sampling frequencies

In this section, we evaluate the decision-making performance of objective quality models when comparing codecs operating at different sampling frequencies. Unlike the analyses presented in Sections 5.1 and 5.2, which considered fixed sampling rates, this experiment combines conditions across 16 kHz, 24 kHz, and 48 kHz. The objective is to assess whether these models can support consistent comparative judgments in a heterogeneous setting.

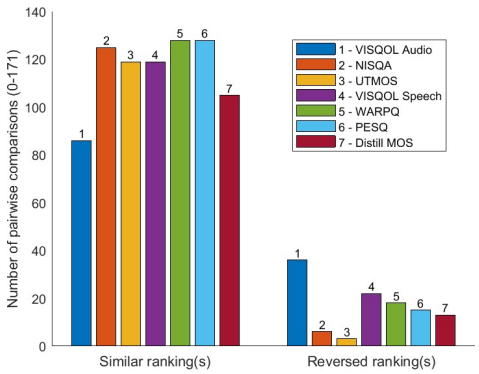


Fig. 8: Number of pairwise comparisons across all frequencies yielding similar (left) and reversed (right) rankings between objective and subjective evaluations based on t-tests (maximum of 171 comparisons for the 19 conditions)

To enable comparison across models operating at different input rates, signals were downsampled to 16 kHz when required. The 48 kHz reference condition was excluded, as intrusive models do not provide scores for it. MNRU conditions were also removed in order to focus on comparisons between codecs. This yields a total of 171 pairwise comparisons.

Across all models, decision-making performance in the cross-frequency analysis is broadly comparable to that

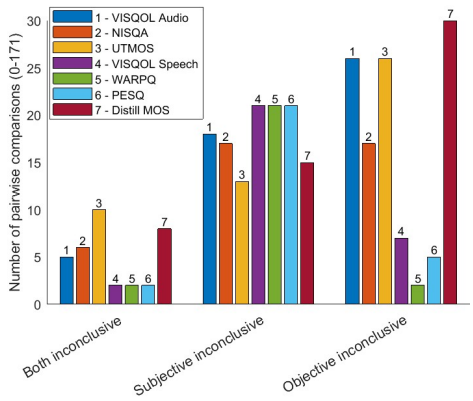


Fig. 9: Number of pairwise comparisons across all frequencies yielding inconclusive outcomes between objective and subjective evaluations based on t-tests (maximum of 171 comparisons for the 19 conditions)

observed in the fixed-frequency analyses. However, the interpretation of the results requires particular care, as differences in performance reflect a balance between ranking consistency and statistical discrimination in pairwise comparisons, as discussed below.

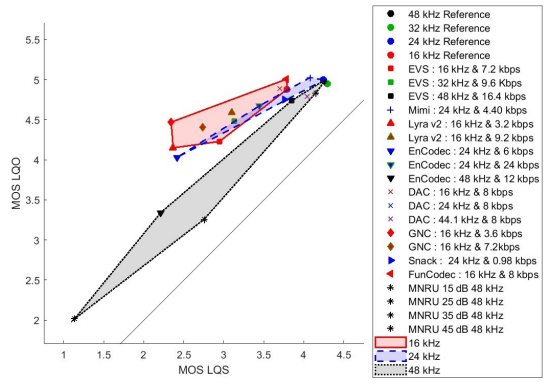


Fig. 10: NISQA subjective scores with convex hull regions illustrating clusters grouped by sampling frequency (16 kHz, 24 kHz, and 48 kHz)

As can be seen from Table 6, UTMOS achieves the highest Kendall's τ (0.719), although the difference relative to NISQA, PESQ, and WARP-Q (all at 0.708) remains small. In contrast, models such as VISQOL Speech (0.591) and VISQOL Audio (0.427) show lower

Model	**Kendall's τ**	**Number of discordant pairs (0-171)**
VISQOL Audio	0.427	$n_d = 49$
NISQA	0.708	$n_d = 25$
UTMOS	0.719	$n_d = 24$
VISQOL Speech	0.591	$n_d = 35$
WARPQ	0.708	$n_d = 25$
PESQ	0.708	$n_d = 25$
DistillMOS	0.614	$n_d = 33$

Table 6: Kendall's τ values between subjective and objective rankings for the 19 test conditions at different frequencies, with the corresponding number of discordant pairs (out of 171)

ranking agreement in this heterogeneous scenario.

A trade-off is observed between ranking consistency, as measured by Kendall's τ, and statistical discrimination, as assessed by pairwise t-tests (Figures 8 and 9). UTMOS, which achieves the highest Kendall's τ, also produces a relatively large number of inconclusive comparisons, indicating limited sensitivity in detecting statistically significant differences between conditions. Conversely, models such as PESQ, WARP-Q, and VISQOL Speech are generally more discriminative, yielding fewer inconclusive results. This comes at the cost of a higher number of reversed rankings. NISQA exhibits intermediate behavior, combining relatively few reversed rankings with a moderate level of statistical discrimination.

VISQOL Audio and DistillMOS exhibit limited discriminative capability, frequently failing to detect statistically significant differences between codecs. This results in a high number of inconclusive pairwise comparisons, as shown in Figure 9.

Interestingly, PESQ, WARP-Q, and UTMOS, which were designed for 16 kHz signals, maintain competitive performance in this cross-frequency analysis despite not having access to higher-frequency content. This suggests that, in this dataset, the absence of information above 16 kHz does not prevent these models from capturing the dominant factors influencing perceived quality. However, their inability to account for high-frequency artifacts should be considered when interpreting the results.

Conditions	Subjective	VISQOL Audio	NISQA	UTMOS	VISQOL Speech	WARPQ	PESQ	Distill MOS
1	4.25 [4.21 – 4.47]	NaN [–]	4.99 [4.91 – 5.04]	3.31 [3.22 – 3.40]	NaN [–]	NaN [–]	NaN [–]	4.35 [4.24 – 4.44]
2	4.29 [4.22 – 4.49]	4.28 [4.26 – 4.30]	4.95 [4.87 – 5.01]	3.35 [3.26 – 3.45]	4.53 [4.48 – 4.57]	4.35 [4.33 – 4.36]	4.63 [4.63 – 4.63]	4.38 [4.27 – 4.46]
3	4.24 [4.22 – 4.50]	3.51 [3.44 – 3.59]	5.00 [4.93 – 5.04]	3.35 [3.25 – 3.45]	4.53 [4.48 – 4.57]	4.34 [4.33 – 4.35]	4.63 [4.63 – 4.63]	4.38 [4.28 – 4.46]
4	3.79 [3.77 – 4.16]	2.54 [2.44 – 2.65]	4.88 [4.81 – 4.93]	3.30 [3.22 – 3.38]	4.53 [4.48 – 4.57]	4.34 [4.32 – 4.35]	4.63 [4.62 – 4.63]	4.32 [4.22 – 4.40]
5	2.95 [3.02 – 3.39]	2.38 [2.28 – 2.48]	4.23 [4.11 – 4.31]	2.63 [2.54 – 2.72]	3.67 [3.60 – 3.74]	3.00 [2.98 – 3.03]	3.91 [3.87 – 3.94]	3.85 [3.74 – 3.96]
6	3.13 [3.29 – 3.62]	3.74 [3.68 – 3.80]	4.49 [4.38 – 4.56]	3.09 [3.01 – 3.16]	3.53 [3.44 – 3.60]	3.12 [3.10 – 3.14]	4.08 [4.05 – 4.11]	4.14 [4.05 – 4.23]
7	3.85 [3.74 – 4.09]	4.16 [4.12 – 4.20]	4.74 [4.66 – 4.79]	3.13 [3.05 – 3.20]	4.12 [4.07 – 4.16]	3.36 [3.34 – 3.38]	4.32 [4.31 – 4.34]	4.15 [4.06 – 4.24]
8	4.08 [4.08 – 4.38]	3.12 [3.03 – 3.20]	5.03 [4.99 – 5.06]	3.39 [3.29 – 3.47]	4.01 [3.94 – 4.07]	3.18 [3.16 – 3.21]	4.12 [4.09 – 4.15]	4.46 [4.36 – 4.52]
9	2.37 [2.51 – 2.84]	2.42 [2.31 – 2.53]	4.15 [4.04 – 4.25]	2.63 [2.54 – 2.73]	3.56 [3.46 – 3.63]	2.75 [2.73 – 2.77]	3.38 [3.34 – 3.42]	3.18 [3.05 – 3.30]
10	3.10 [3.29 – 3.72]	2.45 [2.35 – 2.56]	4.59 [4.53 – 4.65]	3.03 [2.96 – 3.11]	3.91 [3.85 – 3.97]	3.09 [3.07 – 3.11]	3.91 [3.85 – 3.95]	4.06 [3.95 – 4.16]
11	2.41 [2.71 – 3.06]	3.19 [3.10 – 3.27]	4.03 [3.87 – 4.18]	2.53 [2.43 – 2.64]	3.91 [3.82 – 3.99]	2.88 [2.86 – 2.89]	3.57 [3.50 – 3.64]	3.38 [3.22 – 3.50]
12	3.45 [3.72 – 4.13]	3.28 [3.19 – 3.36]	4.67 [4.55 – 4.75]	3.11 [3.01 – 3.20]	4.14 [4.06 – 4.20]	3.25 [3.23 – 3.27]	4.11 [4.06 – 4.16]	4.12 [4.01 – 4.22]
13	2.21 [2.42 – 2.84]	4.20 [4.15 – 4.23]	3.34 [3.17 – 3.49]	2.37 [2.26 – 2.48]	3.83 [3.74 – 3.90]	2.80 [2.76 – 2.83]	3.58 [3.50 – 3.66]	3.05 [2.91 – 3.19]
14	3.70 [3.72 – 4.09]	2.50 [2.39 – 2.60]	4.89 [4.84 – 4.94]	3.22 [3.13 – 3.31]	4.31 [4.27 – 4.35]	3.54 [3.52 – 3.56]	4.39 [4.35 – 4.40]	4.33 [4.24 – 4.40]
15	4.23 [4.19 – 4.42]	3.41 [3.33 – 3.48]	4.98 [4.94 – 5.02]	3.37 [3.28 – 3.45]	4.45 [4.40 – 4.49]	3.99 [3.98 – 4.01]	4.55 [4.54 – 4.56]	4.34 [4.24 – 4.41]
16	4.04 [4.07 – 4.33]	4.24 [4.21 – 4.27]	4.79 [4.71 – 4.86]	3.29 [3.22 – 3.36]	4.24 [4.19 – 4.28]	3.45 [3.43 – 3.47]	4.36 [4.33 – 4.37]	4.40 [4.34 – 4.46]
17	2.34 [2.34 – 2.73]	2.37 [2.26 – 2.48]	4.47 [4.39 – 4.54]	2.91 [2.84 – 2.98]	2.27 [2.19 – 2.34]	2.24 [2.20 – 2.26]	2.44 [2.39 – 2.48]	4.25 [4.18 – 4.31]
18	2.73 [2.85 – 3.18]	2.39 [2.28 – 2.50]	4.41 [4.33 – 4.47]	2.91 [2.84 – 2.97]	2.65 [2.58 – 2.73]	2.49 [2.47 – 2.52]	2.78 [2.71 – 2.87]	4.08 [4.00 – 4.15]
19	3.75 [3.84 – 4.16]	3.09 [3.00 – 3.17]	4.75 [4.65 – 4.83]	3.25 [3.16 – 3.33]	3.18 [3.10 – 3.24]	2.67 [2.65 – 2.69]	3.37 [3.33 – 3.41]	4.44 [4.37 – 4.49]
20	3.78 [3.88 – 4.18]	2.49 [2.38 – 2.59]	5.01 [4.97 – 5.04]	3.42 [3.34 – 3.50]	4.19 [4.14 – 4.24]	3.34 [3.32 – 3.36]	4.15 [4.12 – 4.18]	4.51 [4.44 – 4.56]
21	1.13 [1.09 – 1.26]	3.83 [3.77 – 3.87]	2.01 [1.95 – 2.08]	1.92 [1.83 – 2.02]	3.44 [3.35 – 3.52]	2.76 [2.73 – 2.79]	2.82 [2.74 – 2.91]	1.82 [1.72 – 1.92]
22	2.75 [2.98 – 3.40]	4.35 [4.33 – 4.38]	3.26 [3.06 – 3.44]	2.84 [2.75 – 2.93]	4.29 [4.24 – 4.35]	3.56 [3.53 – 3.59]	3.95 [3.89 – 4.00]	3.71 [3.54 – 3.84]
23	4.15 [4.16 – 4.42]	4.60 [4.58 – 4.61]	4.84 [4.76 – 4.90]	3.20 [3.13 – 3.28]	4.49 [4.44 – 4.54]	4.03 [4.01 – 4.05]	4.44 [4.41 – 4.46]	4.28 [4.16 – 4.37]
24	4.26 [4.22 – 4.48]	4.68 [4.67 – 4.69]	4.98 [4.90 – 5.02]	3.34 [3.24 – 3.43]	4.51 [4.45 – 4.55]	4.22 [4.21 – 4.23]	4.51 [4.47 – 4.54]	4.35 [4.23 – 4.43]

Table 7: Mean scores and corresponding 95% confidence intervals across all 24 conditions for both subjective and objective results

NISQA tends to assign higher MOS-LQO scores to lower-bandwidth signals, indicating a frequency-dependent bias. As shown in Figure 10, this bias appears approximately linear and may affect the reliability of cross-frequency comparisons. Correcting this bias could improve the model's ability to compare codecs operating at different bandwidths.

Overall, none of the evaluated models provides perfectly reliable decisions across sampling frequencies. Even the best-performing model achieves a Kendall's τ of 0.719, corresponding to approximately 14% of codec pairs being ranked in the opposite order compared to subjective evaluations. These limitations are also influenced by model-specific behaviors, such as frequency-dependent biases and restricted bandwidth coverage, which introduce additional uncertainty when comparing codecs across sampling frequencies.

6 Decision-making across bitrates for a single codec

This section evaluates whether objective quality models can correctly discriminate and rank bitrate variations of the same codec, under fixed sampling frequency conditions, in a way that supports subjective decision-making. In this setting, higher bitrates are generally expected to correspond to higher subjective speech quality, making bitrate variation a well-defined decision-making scenario. The analysis considers both codecs evaluated in the listening test and the AMR-WB codec [20] as a reference case.

6.1 Bitrate-based decision-making for evaluated codecs

The analysis focuses on three bitrate comparisons for the same codec: Lyra v2 (9.2 kbps vs. 3.2 kbps), EnCodec (24 kbps vs. 6 kbps), and GNC (7.2 kbps vs. 3.6 kbps). In all three cases, subjective evaluations consistently indicate higher perceived quality at the higher bitrate, establishing a clear decision-making reference. To assess whether objective quality models support this decision-making scenario, pairwise comparisons were performed using two-sided t-tests at a 95% confidence level, combined with an analysis of the resulting ranking directions. VISQOL Speech, WARP-Q, and PESQ successfully discriminated all three bitrate pairs and ranked them consistently with subjective scores. In contrast, NISQA and UTMOS did not detect a statistically significant difference between the two GNC operating points and therefore failed to support bitrate-based decision-making for this codec. VISQOL Audio was unable to detect statistically significant differences across all three codecs; however, its mean scores nonetheless followed the expected bitrate ordering in each case. DistillMOS exhibited the least reliable behavior in this analysis. Although it detected statistically significant differences for all three bitrate pairs, it failed to correctly rank the two GNC bitrates,

illustrating a case where discrimination is achieved but the resulting decision is incorrect.

6.2 Bitrate-based decision-making for AMR-WB

To further examine the behavior of objective quality models across bitrate variations, AMR-WB was selected as a reference codec whose performance over a wide bitrate range is well documented in the literature. As AMR-WB was not included in the subjective listening test, this analysis relies exclusively on objective scores. According to published results, AMR-WB exhibits a progressive increase in perceived quality from 6.60 kbps to 23.05 kbps, followed by a slight degradation at 23.85 kbps [21]. The objective quality models are therefore evaluated on their ability to reproduce this expected behavior. An initial analysis based on mean objective scores shows that NISQA, WARP-Q, and PESQ most closely follow the expected quality trend across bitrates. In contrast, VISQOL Speech, UTMOS, and DistillMOS fail to predict the quality degradation observed at 23.85 kbps relative to 23.05 kbps, although they correctly predict the expected trend at lower bitrates. As shown in Section 5, VISQOL Audio is strongly influenced by sampling frequency. As a consequence, when applied at a fixed sampling frequency, its mean scores tend to cluster across bitrate conditions, limiting its discriminative power in this analysis. To further assess discriminative capability, pairwise two-sided t-tests at a 95% confidence level were conducted across eight adjacent bitrate pairs, from 6.60–8.85 kbps up to 23.05–23.85 kbps. Only WARP-Q and PESQ detected statistically significant quality improvements across all bitrate increases from 6.60 kbps to 23.85 kbps, making them the most discriminative models in this scenario. However, no objective model detected a statistically significant quality degradation between 23.05 kbps and 23.85 kbps. VISQOL Audio performed the worst, yielding inconclusive results for all pairwise comparisons.

6.3 Discussion of bitrate-based decision-making

Overall, these results show that bitrate-based decision-making within a single codec and at a fixed sampling frequency represents a favorable use case for objective quality models. Several models are able to correctly discriminate and rank codec operating points when subjective quality differences follow a clear and monotonic trend with bitrate. However, the analyses also highlight important limitations, including failures to detect small quality differences or to correctly rank operating points, even in this controlled scenario. These findings confirm that while bitrate variation within a codec can be effectively handled by some objective models, reliable decision-making cannot be assumed a priori and remains strongly model-dependent.

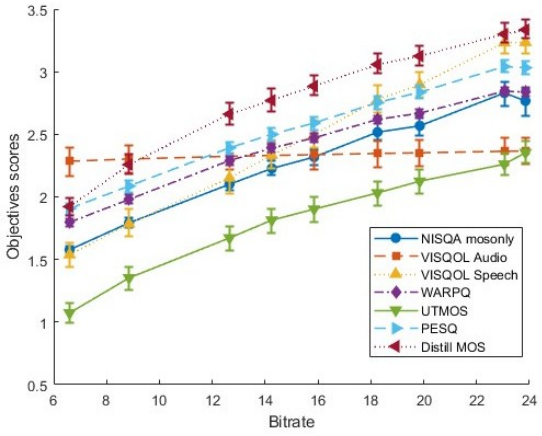


Fig. 11: Model scores at various bitrates for the AMR-WB codec

7 Summary of results and practical considerations

This section summarizes the main findings regarding the decision-making performance of objective quality models across the evaluation scenarios considered in this study.

At 16 kHz, VISQOL Speech, WARP-Q, and PESQ exhibit the strongest decision-making performance, combining high ranking consistency with a large number of similar rankings in pairwise comparisons. In contrast, NISQA, UTMOS, and DistillMOS show lower agreement with subjective evaluations, while VISQOL Audio demonstrates limited statistical discrimination despite a relatively high Kendall's τ value. At 24 kHz, NISQA achieves the highest overall consistency with subjective evaluations and avoids reversed rankings. PESQ and UTMOS also provide competitive performance, with UTMOS not producing reversed rankings and PESQ achieving slightly higher Kendall's τ values.

When comparing codecs operating at different sampling frequencies, objective quality models exhibit limitations in decision-making performance. This is reflected in variations in ranking agreement and in the occurrence of reversed and inconclusive pairwise comparisons, indicating that their reliability may be reduced when applied to heterogeneous codec configurations that differ in sampling frequency and bandwidth.

The analysis of bitrate variations within a single codec and at a fixed sampling frequency shows that several objective models, including VISQOL Speech, WARP-Q, and PESQ, are able to correctly discriminate and rank codec operating points when subjective quality follows a clear trend. However, model behavior may depend on the codec under evaluation. This is illustrated by NISQA and UTMOS, which fail to correctly rank the two operating points of the GNC codec, even though these conditions, whose bitrates differ by a factor of two, exhibit statistically significant differences in subjective quality.

Several model-specific behaviors also emerge from the analysis. UTMOS exhibits high prediction variance when used without repeated inference, which complicates statistical discrimination between conditions. Averaging multiple predictions improves stability but significantly increases computational cost. In addition, UTMOS shows a tendency toward score compression at high quality levels, reducing its discriminative capability in this range.

VISQOL Audio, on the other hand, is strongly influenced by signal bandwidth. This leads to score clustering within a given sampling frequency and limits its ability to discriminate between codec conditions. DistillMOS also exhibits reduced reliability, particularly at higher sampling frequencies, where increased variability further limits statistical discrimination.

From a methodological perspective, the evaluation of 24 kHz signals using models designed for 16 kHz is best performed through downsampling rather than by modifying file metadata. While the latter approach exposes the full bandwidth to the models, it introduces distortions and inconsistent score shifts that degrade ranking reliability.

8 Conclusion

This paper investigated the use of objective speech quality models as tools for supporting decision-making, defined as the ability to produce rankings and statistical discriminations consistent with subjective listening tests. Rather than focusing on the prediction of absolute quality scores, the analysis emphasized the practical use of objective models for comparing and selecting speech coding technologies and configurations.

The results show that some objective quality models are effective for comparing variations of a given speech coding technology when those variations correspond to similar operating conditions, providing meaningful guidance for comparing or selecting codec operating points.

This behavior is consistently observed across both the 16 kHz and 24 kHz analyses. In particular, VISQOL Speech, WARP-Q, and PESQ perform best at 16 kHz, while NISQA achieves the highest ranking consistency at 24 kHz, although it does not significantly outperform some models operating at 16 kHz.

When comparing codecs operating at different sampling frequencies, and more generally across different coding technologies, all evaluated models exhibit noticeable ranking inconsistencies. These inconsistencies limit their reliability for decision-making, particularly in heterogeneous comparison scenarios.

These results highlight the importance of applying objective quality models under the conditions for which they were originally developed, including sampling frequency, bandwidth, and signal characteristics. Deviations from these conditions introduce inconsistencies that affect both ranking reliability and statistical discrimination.

Among the evaluated models, PESQ, despite its conventional design and the withdrawal of its original standard, remains capable of supporting decision-making when applied to neural speech codecs. This result highlights the continued relevance of well-established objective models when used under appropriate operating conditions.

Overall, these findings indicate that objective quality models can serve as useful decision-support tools in speech coding, provided they are applied under consistent and appropriate conditions. Their use for comparing heterogeneous codec configurations, however, should be approached with caution.

Future work should further investigate the decision-making framework introduced in this paper by extending it to broader conditions, such as a wider range of

codecs, different speech languages, and the presence of acoustic background noise or channel impairments, as well as to larger datasets, in order to better characterize the limits and applicability of objective quality models in speech coding. These analyses could in turn support the development of objective quality models that are more robust to variations in sampling frequency and bandwidth, as well as evaluation approaches that better account for heterogeneous codec configurations.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Alliance–Mitacs program, in partnership with VoiceAge Corp.

References

[1] International Telecommunication Union (ITU), "Methods for Subjective Determination of Transmission Quality," Recommendation itu-t p.800, International Telecommunication Union, 2008.

[2] Hu, Y. and Loizou, P. C., "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1), pp. 229–238, 2008, doi: 10.1109/TASL.2007.911054.

[3] Chiang, T.-H. et al., "On the Limitations of Objective Speech Quality Metrics for Neural Speech Enhancement," *arXiv preprint arXiv:2307.04517*, 2023.

[4] ITU-T, "Recommendation P.862.1: Mapping Function for Transforming P.862 Raw Results Scores to MOS-LQO," International Telecommunication Union, 2003.

[5] Falk, T. H. and Chan, W.-Y., "Single-Ended Speech Quality Measurement Using Machine Learning Methods," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pp. I–573–I–576, 2004, doi:10.1109/ICASSP.2004.1325957.

[6] Kendall, M. G., "A New Measure of Rank Correlation," *Biometrika*, 30(1/2), pp. 81–93, 1938, doi:10.1093/biomet/30.1-2.81.

[7] Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., and Zeghidour, N., "Moshi: A Speech-Text Foundation Model for Real-Time Dialogue," *arXiv preprint arXiv:2410.00037*, 2024, submitted on 2024-10-01 (v2).

[8] Heuberger, A., Grill, B., et al., "Enhanced Voice Services (EVS) Codec," Technical paper, Fraunhofer IIS, 2017.

[9] Google, "Lyra V2 – A Better, Faster, and More Versatile Speech Codec," Google Open Source Blog, 2022.

[10] Défossez, A., Copet, J., Synnaeve, G., and Adi, Y., "High Fidelity Neural Audio Compression," *arXiv preprint arXiv:2210.13438*, 2022, submitted on 2022-10-24.

[11] Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., and Kilgour, K., "High-Fidelity Audio Compression with Improved RVQGAN," *arXiv preprint arXiv:2306.06546*, 2023, submitted on 2023-06-11 (v2).

[12] Du, Z., Zhang, S., Hu, K., Liu, S., and Zhang, S., "FunCodec: A Fundamental, Reproducible and Integrable Open-Source Toolkit for Neural Speech Codec," *arXiv preprint arXiv:2309.07405*, 2023, submitted on 2023-09-13 (v2).

[13] Siuzdak, H., Grötschla, F., and Lanzendörfer, L. A., "SNAC: Multi-Scale Neural Audio Codec," *arXiv preprint arXiv:2410.14411*, 2024, submitted on 2024-10-21 (v1).

[14] Torcoli, M., Halimeh, M. M., and Habets, E. A. P., "Navigating PESQ: Up-to-Date Versions and Open Implementations," *arXiv preprint arXiv:2505.19760*, 2025, submitted on 26 May 2025 (v1), last revised 14 Aug 2025 (v2). Accepted for presentation at ITG Conference on Speech Communication 2025, available: https://github.com/audiolabs/pesq.

[15] Mittag, G., Naderi, B., Chehadi, A., and Möller, S., "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," *arXiv preprint arXiv:2104.09494*, 2021, submitted on 19 Apr 2021. Submitted to Interspeech 2021, available: https://github.com/gabrielmittag/NISQA.

[16] 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," *arXiv preprint arXiv:2004.09584*, 2020, submitted on 20 Apr 2020. 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), available: https://github.com/google/visqol.

[17] Jassim, W. A., Skoglund, J., Chinen, M., and Hines, A., "Speech quality assessment with WARP-Q: From similarity to subsequence dynamic time warp cost," *IET Signal Processing*, n/a(n/a), 2022, doi:10.1049/sil2.12151, available: https://github.com/wjassim/WARP-Q.

[18] Baba, K., Nakata, W., Saito, Y., and Saruwatari, H., "The T05 System for The VoiceMOS Challenge 2024: Transfer Learning from Deep Image Classifier to Naturalness MOS Prediction of High-Quality Synthetic Speech," in *IEEE Spoken Language Technology Workshop (SLT)*, 2024, available : https://github.com/sarulab-speech/UTMOSv2.

[19] Stahl, B. and Gamper, H., "Distillation and Pruning for Scalable Self-Supervised Representation-Based Speech Quality Assessment," *arXiv preprint arXiv:2502.05356*, 2025.

[20] 3rd Generation Partnership Project (3GPP), "Technical Specification Group Services and System Aspects; Speech Codec speech processing functions; AMR Wideband Speech Codec; Frame Structure (Release 4)," Technical Specification TS 26.201 V2.0.0, 3GPP, 2001.

[21] European Telecommunications Standards Institute (ETSI), "Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; Performance characterization of the Adaptive Multi-Rate Wideband (AMR-WB) speech codec," Technical Report 3GPP TR 26.976, ETSI, 2020.