

Evaluation of Objective Quality Metrics for Perceptual Coding-Mode Selection in Switched Audio Coding

Fadi Gattoussi, Sanya Tayal, Anika Treffeihn, Goran Marković

Fraunhofer Institut für Integrierte Schaltungen (IIS)

Erlangen, Germany

{fadi.gattoussi, sanya.tayal, anika.treffeihn, goran.markovic}@iis.fraunhofer.de

Abstract—Modern audio codecs often employ switched audio coding, alternating between time-domain coding for speech and frequency-domain coding for music and general audio. Speech/music classification typically guides the mode selection; however, the content type does not always determine which mode listeners perceive as higher quality. As objective audio quality metrics can directly quantify perceptual differences between coded versions, this paper systematically evaluates their reliability for predicting listener preference between coding modes. Using the Enhanced Voice Services (EVS) codec, a dataset of speech and music items was constructed, with each item encoded under both modes and preference labels established through listening tests. Various existing metrics, including classic signal-based measures, standardized perceptual models, and deep-learning-based approaches, are evaluated against these labels across multiple segment durations. Results show that a subset of metrics reliably predicts the listener-preferred mode, and that metrics whose scores exhibit systematic biases achieve high agreement after threshold correction, supporting their use for perceptually guided mode selection.

Index Terms—Objective audio quality metrics, Speech coding, Audio signal classification, Switched audio coding, Transform coding, Enhanced Voice Services (EVS)

I. INTRODUCTION

Switched audio coding is a technique widely adopted in modern audio codecs, such as AMR-WB+ [1], USAC [2], and EVS [3]. This approach dynamically alternates between distinct coding paradigms to optimize for different audio content types. The switching typically occurs between a time-domain coder, such as Algebraic Code-Excited Linear Prediction (ACELP) for speech, and a frequency-domain coder for music and general audio. The EVS codec, for instance, employs an ACELP coder alongside a Transform Coded Excitation (TCX) scheme based on the Modified Discrete Cosine Transform (MDCT) [3]. This hybrid architecture enables EVS to achieve high coding efficiency and perceptual quality across a diverse range of audio signals and bandwidths, particularly at low bitrates, as demonstrated by subjective listening tests [4]. Switched audio coding is also relevant to the Internet of Sounds (IoS), where IoS devices exchange musical and nonmusical audio over bandwidth-limited networks [5] and can benefit from a single codec that handles both content types.

The perceptual performance of a switched audio coding system is highly sensitive to its mode selection mechanism

[6]. State-of-the-art implementations predominantly rely on speech/music classification algorithms to guide this selection. However, a key challenge is that even when a classifier correctly identifies an audio segment’s type (e.g., as “speech”), this does not always correspond to the coding mode that yields the higher perceptual quality [7].

This misalignment, which we confirm with the subjective evaluation reported in Section V, arises from two factors. First, the optimal coding mode depends on the acoustic characteristics of a segment, not on its categorical label. For instance, noisy or reverberant speech may yield higher quality under an MDCT-based coder despite being correctly classified as “speech”, and some monophonic melodies can be better represented by an ACELP-based coder. Second, the classification task itself is limiting: binary “speech” and “music” labels fail to capture the complexity of real-world audio, particularly mixed content such as speech over music, where neither label adequately describes the signal.

Existing approaches typically address these issues by augmenting the primary classifier with signal-based heuristics and decision-smoothing logic [8] rather than relying on direct perceptual criteria. While formal subjective listening tests provide the most reliable measure of coding preference, they are impractical for evaluating large volumes of audio, which motivates the use of objective quality metrics as an alternative. This paper therefore investigates whether objective audio quality metrics can reliably predict which coding mode listeners perceive as higher quality.

Using the EVS codec at 13.2 kbps, each audio segment in a curated dataset was encoded twice: once by forcing all frames into ACELP mode (forced-ACELP), and once into TCX mode (forced-TCX). Subjective listening tests establish the preferred mode for each segment, providing “ground-truth” preference labels. Eighteen objective metrics are then evaluated against these labels across multiple segment durations, examining both the discriminative capacity and the effect of decision threshold calibration on metrics with systematic biases.

The paper is organized as follows: Section II introduces the objective quality metrics under study, Section III describes the dataset construction and subjective listening tests, and Section IV defines the evaluation procedure. Section V presents and discusses the results and their limitations, followed by

conclusions in Section VI.

II. OBJECTIVE AUDIO QUALITY METRICS UNDER EVALUATION

Eighteen intrusive full-reference metrics, covering several distinct categories, were selected for evaluation.

a) SNR-based measures: This group of metrics is derived from the Signal-to-Noise Ratio (SNR). SNR is calculated as the power ratio between a reference signal and the error signal. The Segmental SNR (SegSNR) computes the SNR over short time frames and averages the result to account for temporal variation. The frequency-weighted Segmental SNR (fwSegSNR) extends this by computing a weighted average across mel-frequency bands, with weights derived from the reference signal's energy [9]. We used the implementation described in [10]. Weighted Segmental SNR (WSNR) calculates the SegSNR in the perceptual domain using a weighting filter derived from Linear Predictive Coding (LPC) coefficients, which emulates the perceptual weighting used in the analysis-by-synthesis optimization loop of ACELP coders, as specified for the AMR-WB standard [11].

b) Industry-standard perceptual models: Several standardized metrics for telecommunication quality were evaluated, as they represent the industry benchmarks for assessing speech codec performance. Perceptual Evaluation of Speech Quality (PESQ) models auditory perception to estimate speech quality. ITU-T P.862 specifies the narrowband algorithm (NB_PESQ, 300–3400 Hz) and its raw score [12]; P.862.1 defines the mapping to Mean Opinion Score-Listening Quality Objective (MOS-LQO) [13]. The wideband (WB) extension (PESQ, 50–7000 Hz) is specified separately in P.862.2 [14].

Perceptual Objective Listening Quality Assessment (POLQA), defined in ITU-T P.863 [15], extends speech quality prediction to a wider range of bandwidths and degradations, including time-warping. We used POLQA V3 in the implementation's super-wideband (SWB) setting. Perceptual Evaluation of Audio Quality (PEAQ), defined in ITU-R BS.1387 [16], outputs an Objective Difference Grade (ODG). Its Basic version uses 11 model output variables (MOVs) from an FFT-based ear model. Its Advanced version uses 5 MOVs from FFT- and filter-bank-based ear models. We evaluated three configurations: PQevalAudio's Basic implementation (PEAQ_MCGILL) [17] and GstPEAQ's Basic and Advanced implementations. Input signals for all three implementations were resampled from 16 kHz to 48 kHz. These open-source implementations can deviate from BS.1387 conformance tolerances but provide useful ODG estimates for comparative analysis [17], [18]. For all PEAQ variants, ODG values closer to zero indicate higher predicted quality, whereas more negative ODG values indicate increasing impairment.

Virtual Speech Quality Objective Listener (ViSQOL) v3 is an open-source tool that measures spectro-temporal similarity between a reference and a test signal to produce a MOS-LQO score [19]. In this study, ViSQOL in speech mode (16 kHz) and ViSQOL_AUDIO, which is adapted for general

audio signals (48 kHz), were employed; input signals for ViSQOL_AUDIO were resampled to 48 kHz [20].

c) Intelligibility-focused metrics: Since intelligibility is a core requirement of speech coding, two intelligibility metrics were included. The Short-Time Objective Intelligibility (STOI) metric predicts intelligibility by measuring the correlation between the temporal envelopes of clean and degraded speech spectra across short-time segments [21]. The Extended STOI (ESTOI) improves upon STOI for speech degraded by time-frequency weighted noise by incorporating principles of spectral correlation to overcome some of STOI's simplifying assumptions [22].

d) Modern deep learning-based approaches: We evaluate two metrics based on deep neural networks to determine whether data-driven approaches could capture perceptual nuances missed by traditional models. The Non-Matching Audio Distance (NOMAD) is a perceptual similarity metric that uses deep feature embeddings and can compare a degraded signal against a non-matching clean reference of the same class, guided by the Neurogram Similarity Index Measure [23]. The Speech Contrastive Regression for Quality Assessment (SCOREQ) builds on the same principles but is trained directly on MOS labels rather than a fixed objective measure, aiming to improve generalization for speech quality prediction [24]. Both metrics build their embeddings on a self-supervised wav2vec 2.0 encoder [25], and in our evaluation, both metrics use their deep feature embeddings in a full-reference manner rather than their typical non-matching reference mode. We therefore compute the distance between the embeddings of the coded segment x_{test} and the original segment x_{ref} as $d_{\text{ref,test}} = \|f(g(x_{\text{test}})) - f(g(x_{\text{ref}}))\|_2$, where $g(\cdot)$ yields a sequence of frame-level features at a 20 ms stride, and $f(\cdot)$ averages them over time into a fixed-dimensional embedding, so that different segment durations can be compared. For SCOREQ, we used the natural speech model configuration. For both metrics, quality scores are computed by normalizing embedding distances relative to the maximum observed distance:

$$\text{Score} = \max_{\text{segments}} (d_{\text{ref,test}}) - d_{\text{ref,test}}, \quad (1)$$

where the maximum is taken over all evaluated segments. This formulation ensures that smaller embedding distances (indicating higher perceptual similarity) correspond to higher quality scores.

e) Spectral distortion measures: A group of metrics quantifying raw spectral distortion was included to assess the performance of lower-level signal-based measures. The Log-Likelihood Ratio (LLR) quantifies the prediction error of the processed signal's LPC model on the reference signal, relative to the reference's own LPC model [9]. The Log-Spectral Distance (LSD) computes the root mean square error between the log-magnitude spectra of the reference and test signals [26]. Since LSD is a distance measure where lower values indicate higher quality, its scores were inverted to align with the other metrics. The conversion is analogous to the treatment of the NOMAD and SCOREQ distance measures.

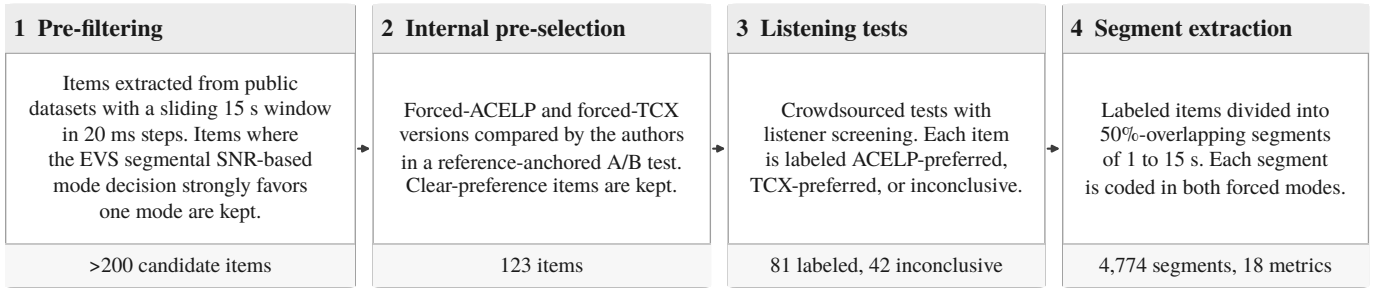


Fig. 1. Extraction of the segments used for the metric evaluation, starting from the public datasets. The footers give the number of items or segments after each stage.

III. DATASET CONSTRUCTION AND SUBJECTIVE REFERENCE ESTABLISHMENT

To create a reliable benchmark for evaluating objective metrics, a new dataset was constructed. For this dataset, the preferred coding mode (ACELP or TCX) for each audio item was established as a subjective reference through listening tests. The pipeline, summarized in Fig. 1, comprises four stages: pre-filtering of candidate items from large corpora by the codec’s segmental SNR-based mode decision over 15-second windows (Section III-B); an internal A/B pre-screen that yields the 123 listening-test items (Section III-B3); crowd-sourced listening tests with listener screening that label each item as ACELP-preferred, TCX-preferred, or inconclusive (Section III-C); and, for the metric evaluation, extraction of 50%-overlapping segments of 1 to 15 seconds from the labeled items (Section IV).

A. EVS Codec Configuration

This study utilizes the 3GPP EVS codec (version 16.3.0) [27] operating at a constant bitrate of 13.2 kbps in WB mode with a 20 ms frame size. This bitrate was selected as a commonly used operating point for EVS-based mobile services [28].

The EVS standard provides two frame-by-frame mode selection mechanisms: At 13.2 kbps, the codec primarily uses a multi-stage speech/music classifier. At 9.6, 16.4, and 24.4 kbps, however, it uses a low-complexity coding mode decision based on estimated segmental SNR [6], switching exclusively between ACELP and TCX modes. This mechanism estimates the coding quality each mode would achieve on a given frame, selects the mode yielding the higher estimated quality, and applies hysteresis to prevent excessive switching. We configured the encoder to restrict the candidate modes to ACELP and TCX and to use this segmental SNR-based decision at 13.2 kbps, thus isolating the binary choice under study.

The reference implementation was further modified to include a forced-mode option that overrides the outcome of the mode decision, including hysteresis, so that every frame of an item is coded in a single mode (ACELP or TCX). This was necessary to create the paired stimuli used in the subjective listening tests and the objective metric evaluation. The two

TABLE I
STUDY CATEGORIES, SOURCE MATERIAL, AND EXPECTED MODE PREFERENCE

Category	Source material	Expected preference
Clean Speech	EARS [29]	ACELP
Instrumental Music	MUSDB18-HQ [30]	TCX
Music Loops	FSLD [31]	ACELP
Noisy Speech	EARS + WHAM! [32]	TCX
Reverberant Speech	EARS + Impulse responses	TCX

coded versions of an item are produced from the same input at the same bitrate and differ only in the coding mode. In addition, the encoder was modified to log the frame-by-frame decisions of the segmental SNR-based mechanism, from which the “ACELP rate” was computed. This measure served as the basis for filtering candidate audio segments from large corpora, as detailed in the next section.

B. Dataset Construction

The five content categories defined for this study are summarized in Table I. Clean Speech and Instrumental Music represent conventional speech-ACELP and music-TCX associations. Music Loops, Noisy Speech, and Reverberant Speech represent inverted associations, i.e. music-ACELP and speech-TCX. The expected modes were used only to guide item selection. The final preference labels were determined from the subjective listening tests.

To identify candidate audio files, we used a filtering strategy based on the segmental SNR-based coding mode decision of the modified EVS encoder. For each segment, we calculated an “ACELP rate”, which represents the proportion of frames for which the SegSNR-based mechanism selected the ACELP mode. By applying high or low thresholds to this rate, we could efficiently isolate segments with strong tendencies toward a specific coding mode. The specific filtering criteria and source materials for each corpus are detailed in the following subsections, while the post-processing steps applied to all candidates are described in Section III-B3.

1) Conventional Preference Scenarios:

a) *Clean Speech*: Source material was drawn from the EARS dataset [29], a collection of monophonic clean speech recordings from 107 speakers. Items at least 15 seconds long were selected and resampled from 48 kHz to 16 kHz. The

ACELP rate was computed over 15-second sliding windows with a one-frame (20 ms) step size. For each speaker, the segment with the highest ACELP rate was selected to maintain speaker diversity, and a threshold of 99% was applied to yield 60 candidates.

b) Instrumental Music: This corpus was sourced from the MUSDB18-HQ dataset [30], using only the 48 tracks with clear Creative Commons licenses. To create instrumental-only content, the “drums,” “bass,” and “other” stems were mixed, then downmixed to mono and resampled to 16 kHz. The ACELP rate was calculated over 15-second sliding windows, and only segments with an ACELP rate below 5% were retained. One segment was randomly selected per song, yielding 48 candidates.

2) Inverted Preference Scenarios:

a) Music Loops: Musical material with a single dominant pitch, and material dominated by isolated sharp onsets, was expected to favor ACELP, since CELP-based coding is known to suit transient-like as well as speech-like signals [6]. This expectation rests on two properties of ACELP. Its long-term predictor converts single-pitch periodicity into prediction gain, and its per-subframe excitation gains keep the coding noise aligned with the temporal envelope. TCX instead quantizes each 20 ms block with a single global gain, so the coding noise spreads over the whole block and can precede an onset as pre-echo [33]. Polyphonic mixtures such as the Instrumental Music items, whose concurrent pitches a single-lag predictor cannot capture, were by contrast expected to favor TCX. To identify such items systematically, we drew source material from the Freesound Loop Dataset (FSLD), which contains approximately 3,000 annotated music loops [31]. All items with a Creative Commons license were selected and standardized by converting to mono, resampling to 16 kHz, and trimming to a duration that is a multiple of 20 ms. Loops were repeated as necessary and then trimmed to meet the required 15-second duration. Only items for which the segmental SNR-based mechanism selected ACELP on every frame were retained as candidates.

b) Noisy and Reverberant Speech: TCX-preferred speech was obtained by degrading clean speech items from the EARS dataset using additive noise and reverberation. For noisy speech, clean speech files longer than 15 seconds were each mixed with a randomly selected real-world noise segment from the WHAM! dataset [32] at an SNR uniformly sampled between 5 and 10 dB (computed using Loudness, K-weighted, relative to Full Scale per ITU-R BS.1770 [34]), and the resulting mixtures were resampled to 16 kHz. Reverberant speech was generated by convolving clean speech items with real full-band room impulse responses (RIRs) sourced from public datasets [35]–[41]. For each speech item, a random RIR with a $T_{60} < 4$ s was applied, and the result was resampled to 16 kHz. For both conditions, the ACELP rate was calculated over 15-second sliding windows, and only segments with an ACELP rate $< 4\%$ were retained. For each unique speaker, the segment with the lowest ACELP rate was selected as a candidate.

3) Common Post-Processing and A/B Pre-Screening: Following the corpus-specific pre-filtering described above, all candidate items underwent two preparatory stages before formal evaluation. First, each item was loudness-normalized to a target of -23 Loudness Units relative to Full Scale (LUFS) and peak-limited to -1 decibels True Peak (dBTP), in accordance with the EBU R 128 [42] and ITU-R BS.1770.

We then performed a pre-selection step: each candidate’s forced-ACELP and forced-TCX encoded versions were compared in an internal reference-anchored A/B test to screen for items that demonstrated a clear and consistent preference for one coding mode. In total, over 200 candidates were reviewed across the four corpora, from which 123 items were selected for formal subjective evaluation: 31 items each from the Clean Speech, Instrumental Music, and Music Loops corpora, and 30 items comprising 14 Noisy Speech and 16 Reverberant Speech items.

C. Subjective Listening Tests

1) Methodology and Interface: Crowdsourced subjective evaluations were conducted on the Amazon Mechanical Turk platform. The workflow, which comprised a qualification pretest, training phase, and main rating session with hidden reliability checks, followed the design principles of ITU-T Recommendation P.808 [43].

Participants were presented with the original reference alongside the two coded versions (forced-ACELP and forced-TCX), labeled randomly as A and B. Participants were instructed to listen to the reference first and could then replay all three stimuli in any order; the platform required each stimulus to be played to the end at least once before a response could be submitted. Listeners rated the quality difference on the seven-point comparison scale defined in ITU-T Recommendation P.800, Annex E (Comparison Category Rating, CCR) [44], ranging from -3 (A much better than B) to $+3$ (B much better than A), with 0 indicating no perceptible difference. Unlike conventional CCR, which compares an unprocessed reference with a single processed sample, each trial here compares two coded conditions against a reference anchor.

2) Participant Screening and Filtering: Prospective participants first underwent a qualification test to verify playback fidelity and hearing acuity. Out of 297 listeners who initiated the study, 204 (68.7%) successfully passed the qualification pretest and proceeded to the training phase and main evaluation. The 123 items were partitioned into eight disjoint listening tests of 15–16 items each, and each test was rated by an independent panel of participants assigned to that test alone. To monitor attention, each test included 3 hidden Gold Standard checks. These reused the same reference/A/B interface as the main task but paired the original reference against a heavily degraded MNRU anchor, making the higher-quality stimulus unambiguous. Participants who failed any Gold Standard check in their assigned test were excluded. Of the 204 listeners who completed the main session, 155 (76%) passed all Gold Standards and contributed to the analysis, yielding an average

of 21 ratings per audio file. The implications of using a screened crowdsourced panel are discussed in Section V-F.

3) *Preference Label Assignment*: Preference labels were derived from the distribution of ratings provided by the remaining valid participants. Raw scores were sign-aligned such that positive values indicate a preference for TCX and negative values a preference for ACELP. For each item, the sign-aligned ratings were averaged across listeners to yield its Comparison Mean Opinion Score (CMOS), and a one-sample t-test determined whether this CMOS differed significantly from zero ($p < 0.05$). Items with $p < 0.05$ were labeled ACELP-preferred or TCX-preferred according to the sign of CMOS. All other items were labeled *inconclusive* and excluded from metric evaluation. Inconclusive denotes insufficient evidence of a preference for either mode. It can reflect ratings near zero, disagreement among listeners, or a weak preference that does not reach statistical significance.

IV. EXPERIMENTAL PROCEDURE FOR METRIC EVALUATION

A. Segment Generation and Objective Metric Computation

Because a codec selects its mode frame-by-frame, a metric guiding this decision must remain reliable on short inputs. We therefore evaluate the metrics on segments of varying length, characterizing how reliability depends on temporal context. Segments were extracted from each audio file with a 50% overlap at eight durations (1, 2, 4, 6, 8, 10, 12, and 15 s), and those with low energy (average power below -50 dBFS for over half the duration) were discarded to focus on perceptually salient content. Durations below 1 s, which would approach the codec’s 20 ms decision granularity, were not evaluated for the reasons given in Section IV-B.

Each retained segment was then encoded at 13.2 kbps using both forced modes. Using the original segment as a reference, an objective quality score was computed for both coded versions. This process was repeated for every metric listed in Section II. Where a metric returned no score for a segment, that segment was excluded from the statistics for that metric and retained for all others. This affects PESQ and NB-PESQ on part of the music material.

B. Preference Label Refinement for Short Segments

Each audio segment initially inherited its preference label (ACELP-preferred or TCX-preferred) from the 15-second parent file it was extracted from. However, perceptual preference can shift at shorter durations: a 15-second clean speech utterance may favor ACELP overall, yet a 1-second excerpt containing only a final sibilant, a breath, and an initial sibilant may favor TCX. To detect such cases, we selected the three metrics with the highest MCC (Section IV-C) on the 15-second segments in an initial evaluation against the inherited labels, namely SCOREQ, NOMAD, and PESQ, and flagged every short segment on which at least two of them disagreed with the inherited label. Two of the authors then independently re-labeled these flagged segments in a blind listening test. A segment was labeled with the mode both raters agreed on.

Segments on which they disagreed were excluded from the evaluation. Although this informal verification is intended to improve label quality, some experimenter bias may persist; formal validation remains for future work. Segments shorter than 1 s were excluded from this study because many of the evaluated metrics do not support sub-second input lengths, and perceptual preference cannot be reliably assessed by listeners at such durations. The final results presented in Section V are calculated against this refined set of preference labels.

For all subsequent analyses, we define the score difference as $\Delta S = S_{\text{TCX}} - S_{\text{ACELP}}$, where S_{TCX} and S_{ACELP} denote the metric scores of the forced-TCX and forced-ACELP versions, respectively.

C. Binary Classification Analysis

For each segment and each metric, a binary prediction was derived by thresholding ΔS at zero. If $\Delta S < 0$, the prediction was “ACELP”, otherwise the prediction was “TCX” (ties at $\Delta S = 0$ were resolved in favor of TCX).

Performance was quantified using the Matthews Correlation Coefficient (MCC), computed from True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). In this study, the “positive” class denotes TCX-preferred segments, and the “negative” class denotes ACELP-preferred segments:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \quad (2)$$

MCC ranges from -1 to $+1$, where $+1$ indicates perfect prediction, 0 indicates performance no better than random chance, and -1 indicates systematic disagreement.

D. Rank-Biserial Correlation Analysis

To complement the binary classification analysis using MCC, we employed the Rank-Biserial Correlation (r_{rb}) [45] to assess the monotonic relationship between the listener preference labels and the objective metric score differences (ΔS). While MCC evaluates a fixed decision rule based on the sign of the score difference (i.e. $\Delta S < 0$ or $\Delta S > 0$), r_{rb} quantifies how consistently TCX-preferred segments yield larger ΔS than ACELP-preferred segments, independent of the absolute calibration of the decision threshold.

The correlation was computed between the binary listener preference variable Y (where $Y = 1$ denotes a preference for TCX and $Y = 0$ for ACELP) and the ranked values of the score differences ΔS , with tied values resolved by mean ranks. For n total segments, let $\bar{R}_1$ and $\bar{R}_0$ denote the mean ranks of the ΔS values for the TCX-preferred ($Y = 1$) and ACELP-preferred ($Y = 0$) groups, respectively. The coefficient is then given by the mean-rank formula of Glass [46]:

$$r_{rb} = \frac{2(\bar{R}_1 - \bar{R}_0)}{n}. \quad (3)$$

This statistic is a linear function of the Mann–Whitney U [45]. Equivalently, it rescales the area under the receiver operating characteristic (ROC) curve (AUC) obtained by thresholding ΔS : $r_{rb} = 2 \text{AUC} - 1$, where AUC is the probability that

TABLE II
SUBJECTIVE PREFERENCE RESULTS: DESCRIPTIVE STATISTICS AND
PREFERENCE COUNTS

Category	N	μ	Med.	σ	ACELP- preferred	TCX- preferred	Inconclusive
Clean Speech	31	-1.34	-2.00	1.73	24	0	7
Music Loops	31	-1.15	-2.00	2.00	17	0	14
Instr. Music	31	+1.36	+2.00	1.82	0	23	8
Noisy Speech	14	+1.00	+2.00	1.75	0	10	4
Rev. Speech	16	+0.58	+1.00	1.84	0	7	9

μ , Med., σ : mean, median, and standard deviation of the sign-aligned listener ratings pooled per category.

a randomly chosen TCX-preferred segment is ranked above a randomly chosen ACELP-preferred segment, with ties counted as one-half [47]. The resulting coefficient ranges from -1 to $+1$. A value of $+1$ indicates perfect separation, where every TCX-preferred segment yields a higher ΔS than any ACELP-preferred segment, while a value near 0 implies no association between the score difference and listener preference. Cross-class ties in ΔS attenuate $|r_{rb}|$ but are infrequent in our data (Section V-C). This analysis separates metrics that fail due to a fundamental lack of signal (low r_{rb}) versus those that fail due to poor calibration or thresholding (high r_{rb} , low MCC).

V. RESULTS

Section V is organized in two parts. We first summarize the listening test outcomes (Section V-A) and the effect of the short-segment label refinement (Section V-B). We then evaluate the objective metrics in three stages: classification performance across segment durations (Section V-C), the relationship between discriminative capacity and decision boundary calibration (Section V-D), and content-type robustness on 15-second segments (Section V-E).

A. Listening Test Results

Of the 123 items, 41 (33.3%) were ACELP-preferred, 40 (32.5%) TCX-preferred, and 42 (34.1%) inconclusive. The detailed breakdown is summarized in Table II.

The results for conventional scenarios confirmed the traditional association of speech with time-domain coding and music with frequency-domain coding. Listeners preferred the ACELP-coded version of Clean Speech in 24 of 31 items and the TCX-coded version of Instrumental Music in 23 of 31 items. Both categories produced sharply concentrated ratings with strong median preferences and few inconclusive items.

The inverted scenarios exhibited higher variance in listener ratings. The Music Loops category generally favored ACELP, with 17 of 31 items showing a significant preference. However, the distribution of ratings was broader compared to Clean Speech, resulting in 14 inconclusive items. Reverberant Speech proved the most ambiguous category: its ratings were widely spread with no clear majority. While the mean score was positive ($\mu = +0.58$), the median was only $+1.00$, and the majority of items (9 of 16) were inconclusive. This high variance indicates that for reverberant signals at 13.2 kbps, listeners do not have consistent preference between the two

modes. Conversely, Noisy Speech showed a stronger alignment with the expected TCX preference, with 10 of 14 items significantly favoring the frequency-domain coder.

Across all categories, no item exhibited a statistically significant preference opposite to the prior expectation assumed during dataset construction.

B. Effect of Short-Segment Label Refinement

Exhaustively relabeling 4,774 extracted segments is infeasible. The metric disagreement criterion of Section IV-B therefore restricts manual relabeling to the segments most likely to have shifted from their inherited 15-second label, flagging 98 segments, 73 of them at 1 s. The inherited label was confirmed in 79 cases, and was changed to TCX-preferred for two 1-second Clean Speech segments. Seventeen segments were discarded.

C. Classification Performance Across Segment Durations

Fig. 2(a) presents MCC values for all evaluated metrics across segment durations from 1 to 15 s after label refinement. At the full 15-second duration, the metrics separate into three broad performance tiers: a top tier led by SCOREQ with perfect classification, followed by NOMAD, PESQ, NB_PESQ, LSD, and POLQA; a middle tier of moderate-to-high MCC including PEAQ_GST_BASIC, WSNR, ESTOI, PEAQ_GST_ADV and SegSNR; and a lower tier with weaker agreement. As the analysis in the next section reveals, however, some metrics in the middle and lower tiers possess strong discriminative ability that is obscured by decision boundary miscalibration rather than an actual inability to separate the two preference classes. Ties at $\Delta S = 0$ accounted for $\leq 0.2\%$ of segments for most metrics but reached $\sim 5\%$ for POLQA, all at its MOS ceiling of 4.75; resolving them in favor of TCX (Section IV-C) did not materially change POLQA's MCC at 15 s.

Three temporal profiles are evident from the duration analysis. Neural metrics and standardized perceptual models (SCOREQ, NOMAD, PESQ, POLQA, and LSD) show a pronounced positive relationship between duration and reliability, indicating a benefit from additional temporal context. Signal-level metrics such as WSNR, SegSNR, and intelligibility metrics (STOI and ESTOI) remain relatively stable across duration, as their discriminative signal is captured by local spectral or temporal envelope properties. The remaining metrics exhibit persistently low MCC regardless of duration; VISQOL and fwSegSNR are notable examples, though as shown in the next section, their underlying causes differ fundamentally.

D. Discrimination Capacity and Decision Boundary Calibration

The rank-biserial correlation (r_{rb}), presented in Fig. 2(b), quantifies how consistently each metric's score differences ΔS separate the two preference classes in rank order, independently of the fixed zero-threshold decision boundary used for MCC computation. Comparing r_{rb} with MCC reveals three different metric categories:

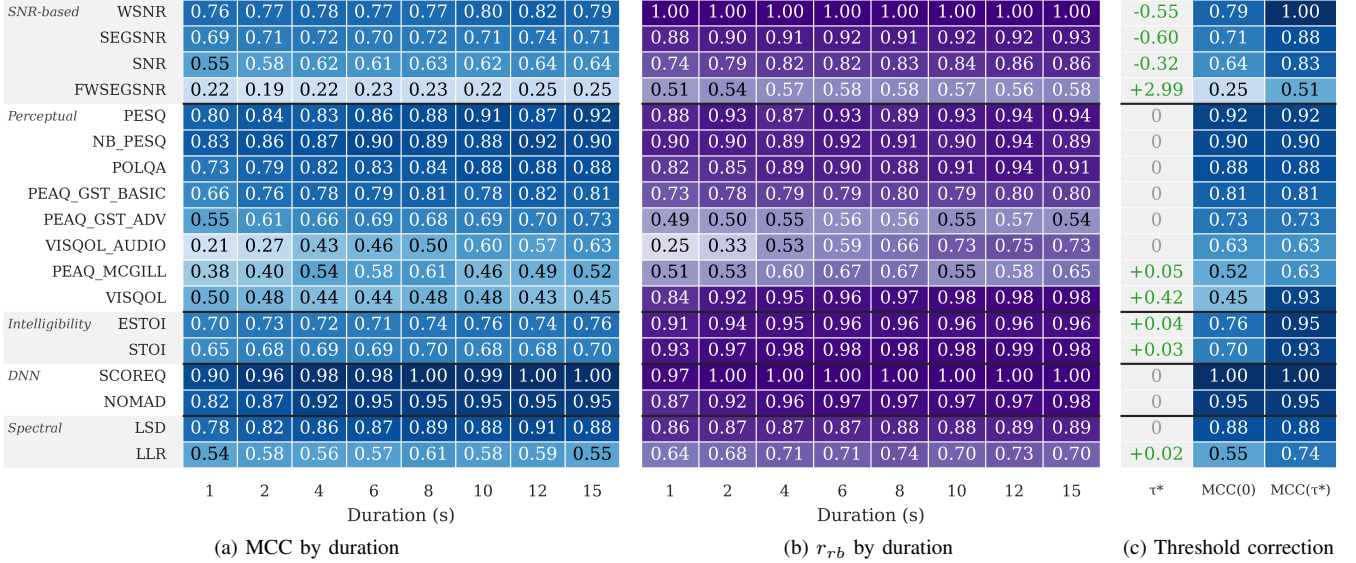


Fig. 2. (a) Binary classification performance and (b) ordinal discrimination capacity across segment durations from 1 to 15 seconds, measured using the rank-biserial correlation r_{rb} defined in (3). (c) MCC-optimal threshold correction for 15-second segments: τ^* is the optimized decision threshold, on the same per-metric scale as the score differences. MCC(0) and MCC(τ^*) are the MCC before and after correction. Zero indicates no correction was applied (i.e., MCC improvement ≤ 0.05). Row labels in (a) apply to all panels; rows are grouped by the metric categories of Section II (SNR-based, perceptual, intelligibility, DNN, spectral) and ordered by decreasing MCC at 15 s within each group.

1) *Jointly High Discrimination and Calibration*: SCOREQ achieves both perfect rank separation and a well-calibrated zero decision boundary from 2 s onward. NOMAD similarly reaches high r_{rb} and MCC at durations of 4 s and above. Among non-neural metrics, PESQ and LSD both exhibit high discrimination and calibration at 15 s. For these four metrics, both the ordinal ranking of segments and the zero decision boundary are well-aligned with listener preference on this dataset.

2) *Miscalibrated Decision Boundary*: Several metrics exhibit high r_{rb} paired with substantially lower MCC, indicating strong discriminative capacity undermined by a misaligned zero threshold. Since this ordinal separation implies that an appropriate threshold can distinguish the two preference classes, we locate that threshold by selecting the value τ^* that maximizes MCC over the 15-second segments:

$$\tau^* = \arg \max_{\tau} \text{MCC}(\tau), \quad (4)$$

where $\text{MCC}(\tau)$ is computed by classifying segments as TCX-preferred when $\Delta S > \tau$ and ACELP-preferred otherwise. A correction is reported (a nonzero τ^*) only when it improves MCC by more than 0.05.

Fig. 2(c) presents the correction results for all metrics for 15-second segment duration. The largest gains are observed for VISQOL, whose ΔS distribution is systematically shifted above zero for both preference classes, and for STOI and ESTOI, which exhibit a similar positive offset. WSNR achieves perfect r_{rb} at every evaluated duration, meaning every TCX-preferred segment receives a higher ΔS than every ACELP-preferred segment; a negative threshold correction recovers

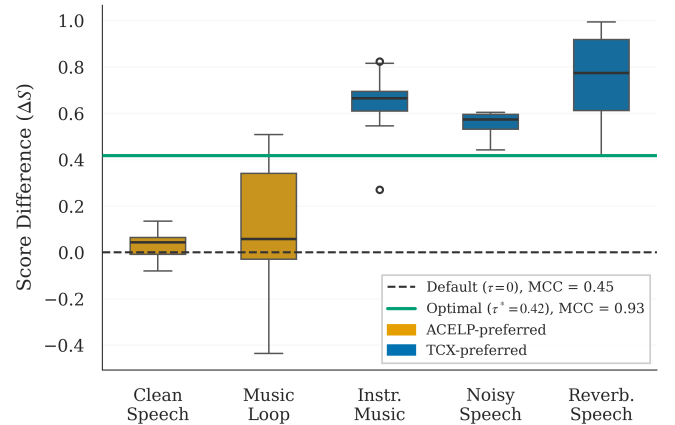


Fig. 3. Score difference (ΔS) distribution for ViSQOL by audio content type at 15 s. The dashed line shows the default zero threshold; the solid line shows the MCC-optimal threshold τ^* . The corrected threshold recovers high MCC by accounting for the systematic positive offset in ΔS .

perfect in-sample MCC. SegSNR and SNR display analogous ACELP-favoring offsets corrected by negative thresholds.

Fig. 3 illustrates the correction for VISQOL, where all five content categories are shifted above zero, causing the default threshold to misclassify the majority of ACELP-preferred items.

The strong rank separation of WSNR and SegSNR is consistent with the dataset construction methodology. The coding mode decision used to compute the “ACELP rate” selects modes by comparing estimated segmental SNRs in the perceptually weighted LP domain [6], which is the same domain

in which WSNR operates. The dataset therefore predominantly contains items for which this class of metrics already exhibits large inter-mode differences, and their generalization to independently selected material warrants further investigation.

The threshold τ^* is selected on the same set on which it is evaluated, so the corrected MCC is an in-sample optimum rather than an estimate of achievable performance. We apply it not to quantify deployment accuracy but to remove each metric’s global calibration offset prior to the per-content-type analysis in Section V-E. Whether a specific τ^* transfers to independently selected material remains to be verified.

3) *Fundamentally Poor Discrimination*: fwSegSNR exhibits low r_{rb} across all durations, and even at its MCC-optimal threshold its MCC reaches only 0.51. Unlike the miscalibrated metrics, whose MCC recovers to high values once their offset is removed, fwSegSNR remains a weak predictor after correction; this confirms that its performance is constrained by limited discriminative capacity rather than by threshold miscalibration. VISQOL_AUDIO and PEAQ_MCGILL likewise exhibit limited discriminative capacity regardless of threshold selection.

E. Content-Type Robustness

Having established which metrics are miscalibrated, this section examines per-category accuracy using corrected thresholds where applicable. Fig. 4 presents prediction accuracy by audio content type for the 15-second segments. Metrics identified as miscalibrated in the previous section are evaluated at their MCC-optimal threshold τ^* . The five categories comprise two ACELP-preferred categories: Clean Speech ($n = 24$) and Music Loops ($n = 17$), and three TCX-preferred categories: Instrumental Music ($n = 23$), Noisy Speech ($n = 10$), and Reverberant Speech ($n = 7$). The relatively small sample sizes, particularly for Noisy Speech and Reverberant Speech, should be considered when interpreting these results. PESQ and NB_PESQ are evaluated on their scored segments only. Both time-align the reference and coded signals utterance by utterance, so a segment without a detected utterance receives no score. This affects six music segments for PESQ and four for NB_PESQ, and no speech segment.

SCOREQ achieves perfect accuracy across all five categories. NOMAD, PESQ, and LSD reach between 80% and 88% on Music Loops and near-perfect accuracy on the other four categories. Although PESQ, NB_PESQ, POLQA, STOI, ESTOI, and SCOREQ are intended for evaluating only speech, each reaches at least 80% accuracy on both music categories. WSNR also reaches perfect accuracy across all five categories after threshold correction, but Section V-D relates its strong rank separation to the dataset construction.

Because the correction applies a single global threshold per metric, identical across all content categories, it cannot induce differences between categories; the remaining per-category failures therefore reflect genuine content-type limitations rather than a global calibration offset. Music Loops, the inverted-preference category where musical content is perceptually superior when encoded with ACELP, remain the

SNR-based	WSNR	1.00	1.00	1.00	1.00	1.00
	SEGSNR	1.00	0.82	0.95	1.00	0.86
	SNR	1.00	0.88	0.82	1.00	0.86
	FWSEGSNR	0.96	0.76	0.55	1.00	0.29
Perceptual	PESQ	1.00	0.80	1.00	1.00	1.00
	NB_PESQ	1.00	0.81	0.95	1.00	1.00
	POLQA	1.00	0.82	0.91	1.00	1.00
	PEAQ_GST_BASIC	1.00	0.59	0.95	1.00	1.00
	PEAQ_GST_ADV	0.96	0.35	1.00	1.00	1.00
	VISQOL_AUDIO	0.88	0.82	0.77	0.80	0.71
	PEAQ_MCGILL	1.00	0.41	0.77	1.00	1.00
	VISQOL	1.00	0.88	0.95	1.00	1.00
Intelligibility	ESTOI	1.00	0.88	1.00	1.00	1.00
	STOI	1.00	0.82	1.00	1.00	1.00
DNN	SCOREQ	1.00	1.00	1.00	1.00	1.00
	NOMAD	1.00	0.88	1.00	1.00	1.00
Spectral	LSD	0.96	0.82	0.95	1.00	1.00
	LLR	0.96	0.53	0.91	1.00	1.00
		CS	ML	IM	NS	RS

Fig. 4. Prediction accuracy by audio content type for 15-second segments. CS = Clean Speech, ML = Music Loops, IM = Instrumental Music, NS = Noisy Speech, RS = Reverberant Speech. Row grouping and order as in Fig. 2; orange and blue column labels mark ACELP- and TCX-preferred categories, respectively.

primary source of errors. Even with corrected thresholds, PEAQ_GST_ADV and PEAQ_MCGILL drop below 50% accuracy on this category. For PEAQ_GST_ADV, the 11 misclassified Music Loop segments have a mean ΔS of 1.75 ODG, more than four times the mean ΔS of the TCX-preferred segments (0.41 ODG), and seven of them exceed the largest TCX-preferred ΔS (0.67 ODG), so no global threshold can separate them. This explains why PEAQ_GST_ADV pairs a moderate MCC (0.73) with a low r_{rb} (0.54) at 15 s in Fig. 2(a) and (b), and why threshold correction cannot remove these errors. PEAQ_GST_ADV thus systematically favors frequency-domain coding on this musical material, contrary to listener preference. PEAQ_MCGILL, by contrast, errs in both directions on music: the ΔS values of its 10 misclassified loops (mean 0.47 ODG) are interspersed with those of the TCX-preferred segments, and it also misclassifies 23% of the Instrumental Music segments, consistent with the limited discrimination noted in Section V-D.

F. Scope and Limitations

The reported findings are bounded by the construction of the evaluation set. Candidate items were retained only at ACELP rates close to 0% or 100% under the codec’s segmental SNR-based decision (Section III-B) and then pre-screened for a clear A/B preference. Consequently, every conclusive item carries the preference label expected for its category at construction time (Table II). The reported accuracies are therefore likely to overestimate performance on unfiltered material, as are the threshold-corrected scores, whose threshold is calibrated on

the evaluation set itself (Section V-D). By the same selection, the segmental SNR-based decision that pre-filtered the items agrees with the subjective label of every conclusive item, so it cannot serve as a baseline that a metric could demonstrably exceed. The results establish which metrics recover known listener preferences, not whether any metric outperforms that decision.

Screening improves the reliability of crowdsourced ratings, but does not establish agreement with an expert panel under controlled conditions. A previous study using multi-stimulus quality ratings found differences in rating level and variability between experienced and inexperienced listeners, while the relative ranking of conditions was largely preserved [48]. Whether these findings extend to the reference-anchored pairwise comparisons used here remains to be verified.

The observed rankings characterize performance on WB material. Upsampling the inputs for VISQOL_AUDIO and PEAQ to 48 kHz does not restore content above 8 kHz, so their results cannot be assumed to generalize to full-band audio. POLQA nevertheless achieves strong discrimination with WB input in its SWB setting. For metrics developed only for speech, the music-category results demonstrate useful transfer beyond their original domain. SCOREQ and NOMAD identify the preferred mode reliably on the selected music items, whereas the unscored music segments limit the use of PESQ and NB_PESQ for this task.

VI. CONCLUSION

This paper investigated whether intrusive objective audio quality metrics can reliably predict the listener-preferred coding mode using the EVS codec at 13.2 kbps. We evaluated eighteen metrics against subjective preference labels across five content categories, including both conventional and inverted preference scenarios. A subset of metrics, led by the deep-learning metrics SCOREQ and NOMAD, reliably identifies the preferred mode for segment durations as short as one second.

The analysis further establishes that discriminative capacity and decision boundary calibration are distinct properties of a metric: ViSQOL, STOI, and other metrics that fail under the default zero threshold nonetheless carry strong discriminative signal, as shown by their high rank-biserial correlation. Correcting a global threshold offset confirms that this latent signal can be exploited. This broadens the set of metrics that show potential in reliably predicting or labeling the listener-preferred mode at scale.

Four directions for future work emerge. First, the findings should be validated on independently selected material, in particular items with intermediate ACELP rates, where the segmental SNR-based decision is split between the modes and a comparison against it becomes meaningful, and extended to the inconclusive items once reliable labels can be established for them, to confirm generalizability beyond the current evaluation set. Second, the evaluation should be extended to sub-second segment durations approaching the

frame-level granularity of the codec. This requires adapting metrics for shorter inputs and investigating alternative subjective preference labeling strategies; the informal short-segment label assignment used in this study should be formally verified through dedicated listening tests. Third, categories such as ACELP-preferred music, where most metrics still struggle, warrant targeted investigation to strengthen content-type robustness, including exploration of composite objective functions that combine multiple metrics. Fourth, extending the evaluation to other switched codecs, bitrates, and mode pairs would establish how broadly these findings apply. Together, these steps aim to advance mode preference prediction toward practical deployment in adaptive codec mode selection.

ACKNOWLEDGMENT

The authors would like to thank Dr. Christian Uhle and Markus Schnell for their insightful suggestions and valuable comments, which helped improve the paper.

REFERENCES

- [1] J. Mäkinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, “AMR-WB+: A new audio coding standard for 3rd-generation mobile audio services,” in *Proc. ICASSP*, vol. 2, 2005, pp. II/1109–II/1112.
- [2] M. Neuendorf *et al.*, “Unified speech and audio coding scheme for high quality at low bitrates,” in *Proc. ICASSP*, 2009, pp. 1–4.
- [3] M. Dietz *et al.*, “Overview of the EVS codec architecture,” in *Proc. ICASSP*, 2015, pp. 5698–5702.
- [4] A. Rämö and H. Toukoma, “Subjective quality evaluation of the 3GPP EVS codec,” in *Proc. ICASSP*, 2015, pp. 5157–5161.
- [5] L. Turchet *et al.*, “The Internet of Sounds: Convergent trends, insights, and future directions,” *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11 264–11 292, Jul. 2023.
- [6] E. Ravelli, C. R. Helmrich, G. Fuchs, and M. Multus, “Low-complexity and robust coding mode decision in the EVS coder,” in *Proc. ICASSP*, 2015, pp. 5888–5892.
- [7] J. Mäkinen, A. Lakaniemi, and P. Ojala, “Low-complex audio encoding for mobile multimedia,” in *Proc. IEEE Veh. Technol. Conf. (VTC)*, 2006, pp. 461–465.
- [8] V. Malenovsky, T. Vaillancourt, Z. Wang, K. Choo, and V. Atti, “Two-stage speech/music classifier with decision smoothing and sharpening in the EVS codec,” in *Proc. ICASSP*, 2015, pp. 5718–5722.
- [9] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
- [10] ModelScope Team, “ClearerVoice-Studio: An AI-powered speech-processing toolkit,” <https://github.com/modelscope/ClearerVoice-Studio>, 2024, accessed: 2025-09-11.
- [11] B. Bessette *et al.*, “The adaptive multirate wideband speech codec (AMR-WB),” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 8, pp. 620–636, Nov. 2002.
- [12] ITU-T, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” Int. Telecommun. Union, Rec. P.862, Feb. 2001.
- [13] —, “Mapping function for transforming P.862 raw result scores to MOS-LQO,” Int. Telecommun. Union, Rec. P.862.1, Nov. 2003.
- [14] —, “Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs,” Int. Telecommun. Union, Rec. P.862.2, Nov. 2007.
- [15] —, “Perceptual objective listening quality prediction,” Int. Telecommun. Union, Rec. P.863, Mar. 2018.
- [16] ITU-R, “Method for objective measurements of perceived audio quality,” Int. Telecommun. Union, Rec. BS.1387-2, May 2023.
- [17] P. Kabal, “An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality,” Dept. Elect. Comput. Eng., McGill Univ., TSP Lab Tech. Rep. v2, Dec. 2003.

- [18] M. Holters and U. Zölzer, “GstPEAQ – an open source implementation of the PEAQ algorithm,” in *Proc. Int. Conf. Digit. Audio Effects (DAFx)*, Trondheim, Norway, Nov. 2015.
- [19] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O’Gorman, and A. Hines, “ViSQOL v3: An open-source, production-ready objective speech and audio metric,” in *Proc. QoMEX*, 2020.
- [20] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, “ViSQOLAudio: An objective audio-quality metric for low-bitrate codecs,” *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. EL449–EL455, Jun. 2015.
- [21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
- [22] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
- [23] A. Ragano, J. Skoglund, and A. Hines, “NOMAD: Unsupervised learning of perceptual embeddings for speech enhancement and non-matching-reference audio-quality assessment,” in *Proc. ICASSP*, 2024, pp. 1011–1015.
- [24] —, “SCOREQ: Speech quality assessment with contrastive regression,” in *Proc. NeurIPS*, vol. 37, 2024.
- [25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Proc. NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.
- [26] A. Erell and M. Weintraub, “Estimation using log-spectral-distance criterion for noise-robust speech recognition,” in *Proc. ICASSP*, Albuquerque, NM, USA, 1990, pp. 853–856.
- [27] 3GPP, “Codec for Enhanced Voice Services (EVS); ANSI C code (floating-point),” 3GPP, Tech. Spec. 26.443, V16.3.0, Jan. 2022, release 16.
- [28] Nokia, “3GPP Enhanced Voice Services (EVS) codec,” Nokia White Paper SR1702006833EN, 2017.
- [29] J. Richter *et al.*, “EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation,” in *Proc. Interspeech*, 2024.
- [30] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bitner, “MUSDB18-HQ: An uncompressed version of MUSDB18,” <https://doi.org/10.5281/zenodo.3338373>, Dec. 2019.
- [31] A. Ramires *et al.*, “The Freesound loop dataset and annotation tool,” in *Proc. ISMIR*, 2020.
- [32] G. Wichern *et al.*, “WHAM!: Extending speech separation to noisy environments,” in *Proc. Interspeech*, Sep. 2019.
- [33] G. Fuchs, C. R. Helmrich, G. Marković, M. Neusinger, E. Ravelli, and T. Moriya, “Low delay LPC and MDCT-based audio coding in the EVS codec,” in *Proc. ICASSP*, 2015, pp. 5723–5727.
- [34] ITU-R, “Algorithms to measure audio programme loudness and true-peak audio level,” Int. Telecommun. Union, Rec. BS.1770-5, Nov. 2023.
- [35] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, “Estimation of room acoustic parameters: The ACE challenge,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 10, pp. 1681–1693, 2016.
- [36] M. Jeub, M. Schäfer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2009.
- [37] K. Prawda, S. J. Schlecht, and V. Välimäki, “Robust selection of clean swept-sine measurements in non-stationary noise,” *J. Acoust. Soc. Am.*, vol. 151, no. 3, pp. 2117–2126, 2022.
- [38] D. Fejgin, W. Middelberg, and S. Doclo, “BRUDEX database: Binaural room impulse responses with uniformly distributed external microphones,” in *Proc. ITG Conf. Speech Commun.*, 2023, pp. 126–130.
- [39] D. Di Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, “dEchorate: A calibrated room impulse response dataset for echo-aware signal processing,” *EURASIP J. Audio Speech Music Process.*, 2021.
- [40] S. V. Amengual Garí, B. Sahin, D. Eddy, and M. Kob, “Open database of spatial room impulse responses at Detmold University of Music,” in *Proc. Audio Eng. Soc. Conv. 149*, 2020.
- [41] University of Kent, “A Sonic Palimpsest: Revisiting Chatham historic dockyards – impulse responses,” <https://research.kent.ac.uk/sonic-palimpsest/impulse-responses/>, 2022, accessed: 2026-05-29.
- [42] EBU, “Loudness normalisation and permitted maximum level of audio signals,” Eur. Broadcasting Union, Rec. R 128, Nov. 2023.
- [43] ITU-T, “Subjective evaluation of speech quality with a crowdsourcing approach,” Int. Telecommun. Union, Rec. P.808, 2018.
- [44] —, “Methods for subjective determination of transmission quality,” Int. Telecommun. Union, Rec. P.800, 1996.
- [45] E. E. Cureton, “Rank-biserial correlation,” *Psychometrika*, vol. 21, no. 3, pp. 287–290, 1956.
- [46] G. V. Glass, “Note on rank biserial correlation,” *Educ. Psychol. Meas.*, vol. 26, no. 3, pp. 623–631, 1966.
- [47] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
- [48] N. Schinkel-Bielefeld, N. Lotze, and F. Nagel, “Audio quality evaluation by experienced and inexperienced listeners,” in *Proc. Meetings Acoust. (ICA)*, vol. 19, 2013, p. 060016.