Perceptual MDCT Audio Coding with Budget-Consistent Psychoacoustic Bit Allocation

Li-Hong Guo
Data Science Degree Program,
National Taiwan University and Academia Sinica,
Taipei, 106, Taiwan
r13946015@ntu.edu.tw

Abstract

Bit-allocation gains depend on comparison design and evaluation target. We examine these dependencies in a modified discrete cosine transform (MDCT) prototype with complete serialized-byte accounting. Experiments on 500 clips from the Environmental Sound Classification dataset with 50 classes (ESC-50) compare four allocation objectives at four rates and a constant-resolution baseline; a central ablation extends to all 2000 clips. At 191.73 kilobits/s, masking-guided allocation improves segmental signal-to-noise ratio (segSNR) over a width-weighted comparator, yet removing masking-related weights raises segSNR further. Waveform and predicted audio-quality metrics also rank the comparators differently. A feature-regularized search, tuned on 50 clips and tested on 100 clips from a separate fold, increases log-mel cosine similarity by 0.0137 over a zero-penalty control using the same search budget. This gain comes at a 0.91-decibel segSNR cost, with no demonstrated classification gain. These findings motivate three evaluation practices: match transmitted bytes, isolate allocation factors, and verify feature improvements against the intended application endpoint.

Keywords: perceptual audio coding, bit allocation, rate accounting, environmental sound, acoustic features

1 Introduction

Perceptual audio coding assigns quantization precision according to auditory relevance. Established systems combine transform coding, psychoacoustic modeling, entropy coding, and detailed bitstream syntax (Brandenburg and Stoll, 1994; Bosi and Goldberg, 2003; Painter and Spanias, 2000). Research prototypes make allocation choices inspectable, but interpreting an observed gain requires both a complete rate constraint and a comparison that isolates the proposed mechanism. Signed coefficient symbols, quantizer scales, activity flags, and alignment all contribute to the transmitted rate.

We study two design questions under a fixed serialized-byte budget. Does an apparent reconstruction gain come from masking-related weights or from other differences in the comparator? Can an explicit acoustic-feature penalty improve feature retention beyond the effect of additional allocation search? We separate masking margin, tonality, and coefficient-count weighting, and compare feature refinement with an identical-search zero-penalty control. Environmental sounds provide diverse events for examining waveform, spectral, feature, predicted quality, and classification outcomes.

We provide a packet cost model and four-rate comparisons, with a full-corpus check of the central ablation. Three findings guide allocation design. First, the segmental signal-to-noise ratio (segSNR) gain over a width-weighted comparator does not establish a masking benefit: controls removing masking margin, or margin and tonality, attain higher mean segSNR. Second, segSNR and predicted quality rank the comparators differently. Third, feature-regularized search improves acoustic-feature agreement at a segSNR cost under equal search and byte budgets. All codec scores use decoded serialized streams. Together, these findings motivate factor-isolating comparisons and application-specific evaluation.

2 Related Work and Scope

The modified discrete cosine transform (MDCT) and overlap-add synthesis use time-domain aliasing cancellation (Princen and Bradley, 1986). Psychoacoustic transform coding estimates band en-

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 59–66
November 6-7, 2026 ©2026 Association for Computational Linguistics

ergies, masking thresholds, and noise tolerance (Painter and Spanias, 2000). We use simplified, inspectable versions of these established components. Advanced Audio Coding (AAC) and Opus (Valin et al., 2012) provide the broader codec context; our experiments hold a simplified raw-packet syntax fixed to study allocation choices.

Task-directed compression predates this study. Digalakis et al. (1998) allocate bits among quantized cepstral subvectors using word error rate; Kuznetsova et al. (2025) train residual vector quantizers for intermediate neural features with task-specific losses. Both transmit features for machine inference. Our refinement retains waveform decoding and asks whether a handcrafted log-mel penalty changes allocation under an exact serialized-byte constraint. The refinement selects offline from finite packet proposals, using the same proposals and search budget for a zero-penalty control. This design isolates the penalty's contribution while retaining decoded audio for separate reconstruction, feature, and classification measurements.

Evaluation must match the domain. The Environmental Sound Classification dataset with 50 classes (ESC-50) contains diverse non-speech events (Piczak, 2015). Short-time objective intelligibility (STOI) predicts speech intelligibility (Taal et al., 2011); here it is only an exploratory diagnostic. Handcrafted log-mel statistics measure acoustic-feature agreement, not learned semantic similarity. We additionally use Virtual Speech Quality Objective Listener (ViSQOL) in general-audio mode (Chinen et al., 2020). Its model-predicted quality score complements, but does not replace, a listening experiment such as Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) (International Telecommunication Union, 2015).

3 Codec and Complete Rate Accounting

Figure 1 separates allocation from representation and evaluation. Every comparator shares the transform, quantizer, input preprocessing, and packet syntax.

3.1 Transform and Quantization

Mono pulse-code modulation (PCM) audio is resampled to $F_s = 48{,}000$ Hz. Each long window contains $2N = 1024$ samples and advances by $H = N = 512$ samples. With

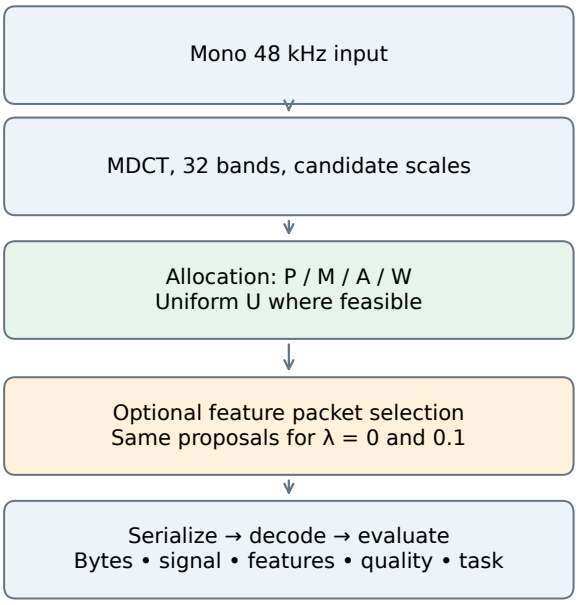

Figure 1: Controlled signal path. P, M, A, and W change allocation weights; U uses a common resolution. Optional feature refinement selects among complete frame packets. Evaluation always follows serialization and decoding.

$w[n] = \sin[\pi(n + 1/2)/(2N)]$, the transform is

$$X_i[k] = \sum_{n=0}^{2N-1} x[iH + n]w[n] \times \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2} + \frac{N}{2}\right)\left(k + \frac{1}{2}\right)\right], \quad (1)$$

for $0 \le k < N$. The inverse transform and overlap-add use the same sine window, whose interior perfect-reconstruction condition is $w^2[n] + w^2[n + N] = 1$. The implementation uses a fast Fourier transform (FFT) backend and has a matrix reference for numerical checks. Boundary handling is shared: only trailing fill is added, with no leading half-window padding. Consequently, the first and final unpaired half-windows have boundary reconstruction error even without quantization, and that error remains in the reported scores. No claim of short-window transient control is made.

The N coefficients form $B = 32$ contiguous Bark-like bands. Band b contains w_b coefficients. A normalized coefficient $u = X/s_b$ uses resolution parameter r_b and a symmetric dead-zone quantizer. For $r_b \ge 1$, let $Q = 2^{r_b - 1}$, $\delta = 1.5$, $\Delta = (Q + \delta/2)^{-1}$, and $z_0 = \delta\Delta/2$. Then

$$q(u) = \begin{cases} 0, & |u| \le z_0, \\ \operatorname{sgn}(u)\min(Q, j(u)), & \text{otherwise}, \end{cases} \quad (2)$$

where $j(u) = 1 + \lfloor (|u| - z_0)/\Delta \rfloor$. Nonzero indices reconstruct as $s_b \operatorname{sgn}(q)[z_0 + (|q| - 1/2)\Delta]$; zero reconstructs to zero. An inactive band ($r_b = 0$) is zero. For $r_b < 4$, scale search tests 0.5, 0.75, and 1.0 times the band's peak magnitude and minimizes mean squared error (MSE); at higher resolutions only the peak scale is used. The maximum resolution is 16.

3.2 Controlled Allocation Objectives

Let $e_b(r)$ denote mean squared coefficient error after candidate quantization. All adaptive conditions minimize a surrogate of the form

$$D_b(r) = \alpha_b w_b e_b(r). \tag{3}$$

The masking-guided condition P uses

$$\alpha_b^{\mathrm{P}} = (0.25 + m_b)(1 + 0.75 t_b)\beta_b. \tag{4}$$

Here $t_b \in [0, 1]$ derives from spectral flatness; m_b is the positive band-energy margin in decibels (dB) over spread masking power plus an approximate absolute threshold of hearing (ATH); and $\beta_b \geq 1$ is a transient-dependent high-frequency boost. Masking spreads with 12 dB attenuation per Bark distance. Without calibrated listening sound-pressure levels, these are allocation heuristics rather than a validated audibility model.

Three controls separate the weighting choices:

$$\begin{aligned} \alpha_b^{\mathrm{M}} &= 0.25(1 + 0.75 t_b)\beta_b, \\ \alpha_b^{\mathrm{A}} &= \beta_b, \\ \alpha_b^{\mathrm{W}} &= w_b \beta_b. \end{aligned} \tag{5}$$

M removes the masking margin while retaining tonality. A removes both margin and tonality weights. All retain the same effective transient factor. The implementation applies the boost twice to scheduled non-long frames even though short transforms are disabled; β_b includes this shared factor. These controls therefore do not remove every auditory-inspired mechanism.

W is the historical code option named `uniform_allocation`, but its objective is $\beta_b w_b^2 e_b(r)$ and its resolutions are adaptive. We call it *width-weighted allocation*. P–W changes both masking-related and width weighting, so it cannot isolate a masking benefit. A separate true uniform condition U uses the largest common positive resolution that fits all quantizable bands within each frame. U never drops a valid non-silent band to fit the packet.

3.3 Serialized Byte Constraint

The original nominal constraint $\sum_b w_b r_b \leq R_{\mathrm{nom}}$ omits representation costs. Since the quantizer has $2^{r_b} + 1$ indices, raw signed representation needs $r_b + 1$ bits per active coefficient. Consequently, $R_{\mathrm{nom}} = 1024$ corresponds to 96 kb/s of nominal coefficient cost at 93.75 frames/s, not a complete 96 kb/s stream. The original P/W estimates were 201.529/196.610 kb/s, a 2.50% difference, and omitted parts of the container and alignment. Those historical results are retained in the companion audit rather than used for matched-rate conclusions.

Our MCR1 syntax has a 26-byte file header and fixed-size frame packets of P_f bytes. A frame begins with a four-byte activity mask; each active band stores a one-byte resolution, four-byte floating-point scale, and byte-aligned raw signed indices. Before final padding, the cost is

$$\begin{aligned} C_i^{\mathrm{byte}} &= 4 + \sum_{b \in \mathcal{A}_i} c_b(r_b) \leq P_f, \\ c_b(r) &= 5 + \left\lceil \frac{w_b(r+1)}{8} \right\rceil, \end{aligned} \tag{6}$$

where $\mathcal{A}_i$ is the active-band set. Numerically silent bands without a valid scale under the shared 10^{-12} peak cutoff remain inactive. The greedy allocator selects positive distortion reduction per marginal byte, including activation cost, then pads the unused capacity. Candidate scale tables use double precision; the chosen scale is rounded to its transmitted 32-bit value before actual quantization. Thus, allocation uses a surrogate distortion table while rate and evaluation use the exact serialized representation. The search is not globally optimal.

Each five-second clip produces 468 packets. Packet sizes 128, 192, 256, and 370 bytes give complete measured rates of 95.8880, 143.8112, 191.7344, and 277.0976 kb/s. The P/M/A/W matrix uses all four sizes. With every band active, U needs 304 bytes for $r = 1$, 370 for $r = 2$, and 425 for $r = 3$. We therefore evaluate U at 370 bytes; a true all-active uniform comparator is infeasible at 256. Shared near-silent inactive bands can allow a larger common resolution. Equality of total bytes includes padding and need not mean equal useful payload.

4 Feature-Regularized Packet Search

Weighted coefficient MSE does not directly preserve log-mel statistics. We implement an of-

fline search that adds a full-clip acoustic-feature penalty. For each frame i, the finite proposal set comprises the complete packets produced by P, M, A, and W at 256 bytes. Starting from all-P packets, the search performs one chronological coordinate sweep, evaluating the other three proposals at each frame and accepting only a decrease in

$$J_\lambda(\mathbf{z}) = \frac{D(\mathbf{z})}{\max(D_{\mathrm{P}}, \epsilon)} + \lambda \frac{F(\mathbf{z})}{\max(F_{\mathrm{P}}, \epsilon)}, \quad (7)$$

where $\epsilon = 10^{-12}$, $\mathbf{z}$ selects one proposal per frame, and

$$\begin{aligned} D(\mathbf{z}) &= \sum_{i,b} \alpha_{ib}^{\mathrm{P}} \sum_{k \in b} (X_i[k] - \widehat{X}_{i,\mathbf{z}}[k])^2, \\ F(\mathbf{z}) &= \|\phi(x) - \phi(\widehat{x}_{\mathbf{z}})\|_2^2. \end{aligned} \quad (8)$$

D_{P} and F_{P} are fixed per-clip baseline values. Distortion is computed from parsed transmitted scales and indices, not the candidate-table approximation. The 80-dimensional feature vector ϕ concatenates temporal means and standard deviations of 40 log-mel power bands. Its analysis uses 2048-sample Hann windows, hop 512, and $\log_{10}(E + 10^{-10})$ for mel-band power E.

Replacing one synthesis frame affects overlapping feature windows. We update all affected windows and the full-clip sufficient statistics; independent full decoding and feature extraction verify the final result. Packet concatenation preserves the exact file length. The $\lambda = 0$ control receives the identical proposals and one-sweep budget, separating the feature penalty from extra search. This is a restricted full-reference, offline allocation search, with no claim of a causal encoder or global optimum.

5 Experimental Protocol

5.1 Data and Parameter Selection

The controlled matrix uses the original deterministic seed-42 selection of two clips per class from each official ESC-50 fold: 500 five-second clips and 481 source recordings. The codec has no learned parameters; five-fold coverage is not five codec-training runs. We also evaluate P versus A at 256 bytes on all 2000 clips, representing 1524 source recordings. This full-corpus comparison includes the 500-clip subset and is a coverage expansion, not an independent replication sample.

Feature refinement uses 50 development clips, one per class from the selected fold-1 clips, taking the first filename in lexicographic order. All 100 selected fold-5 clips are used for testing. Development and test source recordings do not overlap. Following a two-clip development pipeline check, we fixed the grid to $\lambda \in \{0, 0.001, 0.01, 0.1, 1, 10\}$. The two small positive values were added after this development-only check, before formal selection or testing. We select the candidate with the lowest mean development feature error among those whose mean segSNR drops by at most 1 dB relative to P. This selects $\lambda = 0.1$; the development drop is 0.852 dB. The constraint is a selection criterion, not a per-clip or test-set guarantee. Earlier aggregate results motivated the extension, so fold 5 is held out from this parameter selection, not a previously unseen dataset.

5.2 Measurements and Source Separation

Signal-to-noise ratio (SNR) is $10 \log_{10}[(\sum x^2 + 10^{-20})/(\sum (x - \widehat{x})^2 + 10^{-20})]$. SegSNR averages per-segment SNR values in dB over non-overlapping 2048-sample segments, including the final partial segment. We neither clip scores nor remove silence. Spectral error is the nonnegative pooled root-mean-square (RMS) difference of dB magnitudes:

$$E_{\mathrm{RMS}} = 20 \sqrt{\frac{1}{TK} \sum_{t,k} \left[\log_{10} \frac{|S_{tk}| + \eta}{|\widehat{S}_{tk}| + \eta} \right]^2}, \quad (9)$$

with Hann window 2048, hop 512, $\eta = 10^{-12}$, T frames and K frequency bins. Lower is better. This differs from the historical log-spectral MSE field, whose dB transform can be negative. The log-mel cosine similarity ρ_{mel} is the cosine similarity between $\phi(x)$ and $\phi(\widehat{x})$. STOI is reported separately as a speech-derived diagnostic.

For general-audio quality prediction we use the official ViSQOL v3.1.0 audio model, with its support-vector regression mapping to a mean opinion score–listening quality objective (MOS-LQO).[1] The implementation passes three official audio conformance fixtures. We evaluate P/W at 256 bytes on the same 500 clips. The command-line reader requires 16-bit PCM; each reference/P/W triplet receives one common gain to avoid clipping before conversion at 48 kHz. Gain, quantization error, and file hashes are recorded. This preprocessing is shared within each paired

[1] Official ViSQOL v3.1.0 source and audio model.

comparison. Five-second environmental clips can contain silence or atypical events, limiting interpretation of the predicted score; it is not a measured listener opinion.

The downstream probe uses the same 80-dimensional features and a 200-tree random forest with seed 42, trained only on clean audio. Each of the five official folds is tested using a model trained on the other selected folds. Two source recordings cross official fold boundaries in this subset. We purge training clips sharing any test source: training sizes are 400 for fold 1 and 399 for folds 2–5. The same fitted model evaluates every condition within a fold. Feature-refinement classification reuses the source-purged fold-5 model and never supplies class labels or classifier predictions to the allocator.

5.3 Inference and Verification

All contrasts pair the same clips. We report means, mean differences, and 95% percentile bootstrap confidence intervals (CIs), with 10,000 draws. Codec, feature, and ViSQOL analyses use seed 20261002; classifier intervals use seed 42. The source-cluster bootstrap resamples complete recording-source groups. Each replicate mean is the sum of sampled clip-level differences divided by the sampled clip count. These pointwise CIs cover conditional sampling uncertainty, not subset selection or model/tuning uncertainty.

Supplementary tables also report clip/class bootstrap checks and paired standardized effects for the codec matrix. Clip-level Wilcoxon signed-rank tests assess paired distributions, not specifically means, and are not source-adjusted tests.

The codec matrix has a fixed Holm family of 65 tests: P versus M/A/W at four rates and P versus U at 370 bytes, each over five metrics. Full-corpus P–A uses a separate five-test family. For P/W at 256 bytes, the ViSQOL analysis treats its quality endpoint and the five reconstruction/feature endpoints as one six-test family. Feature refinement has 12 tests: six endpoints, including directly optimized feature error, against P and $\lambda = 0$.

Classification uses clip-level exact McNemar tests, without adjustment for source dependence: three P–M/A/W contrasts on 500 clips, Holm-adjusted together, and one selected-versus-zero contrast on 100 clips. These are distinct, declared families rather than one experiment-wide error guarantee.

All encoded signals are decoded from complete byte streams. Checks cover signed extrema, resolutions 1–16, quantizer thresholds, silence, costs, and equal stream lengths. Optimized packing/parsing matches the original encoder bytes and decoder samples. Previously validated fixed-packet P/W outputs are reproduced exactly at 256 bytes. Hashes, input manifests, execution-time source snapshots, validation records, and numerical tables are retained; no failed sample is replaced by an estimated score.

Supplementary material. The accompanying archive `ROCLING2026_paper20_revision_bundle.zip` contains numerical tables, protocols, code, and validation records. Its root file `SUPPLEMENTARY_README.md` maps each experiment to its results and statistical comparisons, and links to reproduction commands. The separate LaTeX source archive contains typesetting files only; neither archive includes audio data.

6 Results

6.1 Controlled Weights and Rate Dependence

Table 1 shows that the comparator determines the apparent benefit. At 256 bytes/frame, P exceeds W in segSNR but has higher spectral error and lower log-mel cosine similarity. Removing the masking margin (M), or margin and tonality (A), improves mean scores on all five metrics relative to P.

The P–W difference in segSNR is $+3.583$ dB (source-cluster 95% CI $[3.46, 3.71]$); the difference in log-mel cosine similarity is -0.0432 ($[-0.0491, -0.0372]$). The P–M difference in segSNR is -0.419 dB (source-cluster 95% CI $[-0.47, -0.37]$); the difference in log-mel cosine similarity is -0.0092 ($[-0.0113, -0.0072]$). The P–A difference in segSNR is -0.437 dB (source-cluster 95% CI $[-0.49, -0.39]$); the difference in log-mel cosine similarity is -0.0070 ($[-0.0091, -0.0051]$).

For P, mean metadata, raw payload, and padding are 55874.2, 53130.2, and 10829.5 bytes per clip. The metadata term includes the file header. Equal file length thus leaves different allocations of useful payload and side information.

Figure 2 reports all predeclared rate points.

Across the four rates, P–M segSNR differences range from -0.42 to -0.39 dB; log-mel cosine similarity differences range from -0.0131 to -0.0035. For P–A, the corresponding segSNR dif-

Method	segSNR ↑	SNR ↑	E_{RMS} ↓	ρ_{mel} ↑	STOI (exploratory)
P	25.487	24.936	26.938	0.9062	0.9422
M	25.906	25.634	26.163	0.9154	0.9442
A	25.924	25.651	26.454	0.9132	0.9461
W	21.905	22.341	20.198	0.9494	0.8732

Table 1: Mean scores on 500 clips at exactly 191.7344 kb/s (119834 bytes per clip). SNR and spectral quantities are in dB. M removes masking margin; A removes margin and tonality; W changes width weighting. STOI is a speech-derived diagnostic, not environmental-audio quality.

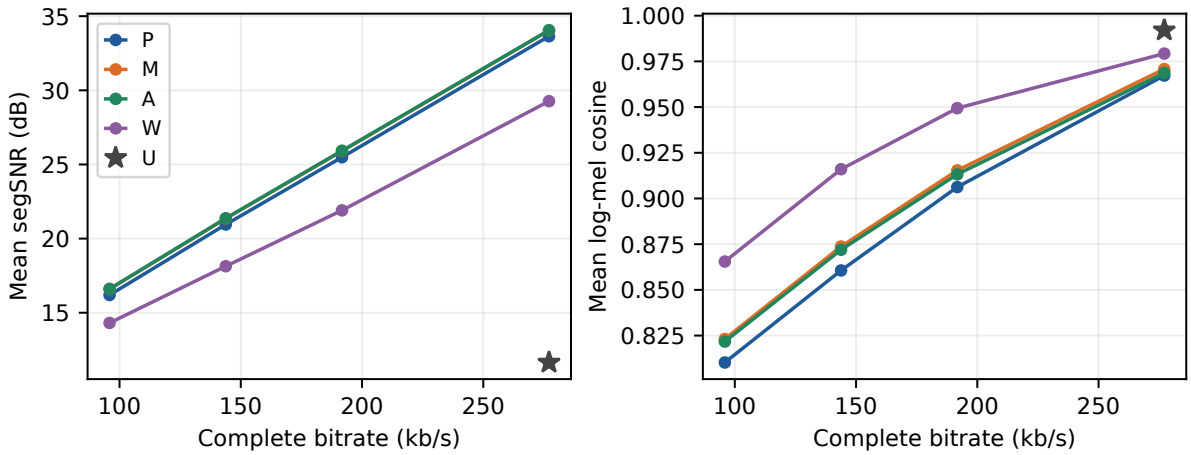

Figure 2: Four-rate means for the same 500 clips. Lines connect measured points; U is evaluated only at 277.10 kb/s, where a common positive resolution fits the shared packet syntax. The panels measure different objectives and do not imply a perceptual ranking.

ferences range from -0.44 to -0.40 dB; log-mel cosine similarity differences range from -0.0115 to -0.0014. Against the width-weighted comparator, P–W segSNR differences range from $+1.88$ to $+4.38$ dB; log-mel cosine similarity differences range from -0.0554 to -0.0120.

At 370 bytes, U reaches 11.64 dB segSNR and 0.9920 log-mel cosine similarity. P–U differences are $+22.00$ dB (source CI $[21.04, 22.97]$) and -0.0248 ($[-0.0276, -0.0221]$), respectively. This is a true common-resolution comparison within the same packet format.

Of the 65 declared clip-level tests, 63 have Holm-adjusted $p < 0.05$; the supplementary tables indexed in Section 5.3 provide all endpoint-specific effects and adjusted values. Statistical significance does not establish perceptual relevance.

At 370 bytes, P–A spectral RMS and log-mel cosine similarity contrasts have Holm-adjusted $p = 0.945982$ and 0.942774, respectively, although their source-cluster mean CIs exclude zero. P–U STOI has mean difference -0.00036 (source CI $[-0.006484, 0.005160]$), yet Holm-adjusted signed-rank $p = 4.70 \times 10^{-23}$. Mean CIs and signed-rank tests target different aspects of paired differences; the CIs are pointwise, whereas the tests receive 65-test multiplicity correction.

Full-corpus control. On all 2000 clips, the P–A comparison at 256 bytes covers 1524 source recordings. The mean segSNR difference is -0.44 dB (source CI $[-0.47, -0.41]$). The mean log-mel cosine similarity difference is -0.0071 (source CI $[-0.0082, -0.0060]$). The mean spectral RMS difference is $+0.47$ dB (source CI $[0.39, 0.55]$). This comparison expands corpus coverage for the central ablation while retaining the same encoder and syntax. The supplementary tables in Section 5.3 include all five endpoints and their separate five-test Holm correction.

6.2 Predicted Audio Quality and Classification

ViSQOL audio-mode MOS-LQO averages 3.977 for P and 4.449 for W. The P–W difference is -0.471 (source-cluster 95% CI $[-0.504, -0.438]$; six-endpoint Holm-adjusted $p = 4.67 \times 10^{-81}$).

This predicted general-audio quality score favors W, despite P's higher segSNR. The P/W ranking therefore depends on the evaluation endpoint even at an identical transmitted rate. The classifier trained on clean audio achieves accuracies of 30.2% on clean audio, 24.0% on P, 24.6% on W, and 25.0% on each of M and A. The P–W difference is -0.6 percentage points (source-cluster 95% CI $[-3.60, 2.40]$; Holm-adjusted exact McNemar $p = 0.788$). P–M and P–A each differ by -1.0 percentage points, with source intervals spanning zero and Holm-adjusted $p = 0.539$. None of the three contrasts establishes a classification advantage. These intervals condition on the fixed fitted models; absence of significance is not equivalence.

6.3 Held-Out Feature Refinement

Table 2 shows that the selected penalty improves log-mel cosine similarity at the cost of segSNR on the fold held out from parameter selection.

Relative to $\lambda = 0$, log-mel cosine similarity changes by $+0.0137$ (source-cluster 95% CI $[0.0111, 0.0165]$) and squared feature error by -22.47 (CI $[-27.50, -17.87]$). The associated reconstruction changes are -0.91 dB in segSNR (CI $[-1.12, -0.71]$) and -2.46 dB in spectral RMS error (CI $[-3.01, -1.96]$).

The 12 feature-study tests have a maximum Holm-adjusted $p = 8.90 \times 10^{-11}$, including 2 directly optimized feature-error comparisons. The selected search changes a mean of 120.15 of 468 frame packets; its mean segSNR drop from P is 0.907 dB. STOI also decreases, from 0.9421 for the zero-penalty control to 0.9364 for the selected penalty.

On the same 100 clips, classification accuracy is 30.0% for the selected penalty and 32.0% for the zero-penalty control. The difference is -2.0 percentage points (source CI $[-5.05, 0.00]$; exact McNemar $p = 0.50$). Only 2 clips differ in classification correctness between the conditions; neither a classification benefit nor equivalence is established.

7 Discussion

7.1 Implications for Allocation Design

Separate rate control from effect attribution. Complete packet accounting controls transmitted cost; the M and A ablations separate weighting choices. P improves segSNR over W, yet removing masking margin and tonality yields still higher mean segSNR. The P–A direction persists across four rates and the full-corpus comparison at one rate. Evaluating a proposed weight therefore requires keeping coefficient-count weighting and packet syntax fixed while removing that weight. Byte matching and factor separation answer complementary questions.

Choose endpoints for the intended use. At 191.73 kb/s, segSNR favors P over W, whereas ViSQOL, spectral error, and log-mel cosine similarity favor W. Thus, rate control alone does not determine which allocation to choose. Report reconstruction, quality prediction, or task performance according to the intended use, alongside the common rate measure.

Control the search when adding a feature loss. Sharing proposals and sweep budgets separates the feature penalty's effect from additional search. Relative to the zero-penalty control, the selected penalty improves log-mel cosine similarity and spectral RMS error while lowering segSNR. The classification probe establishes no gain. This supports selecting the penalty with a development-set reconstruction tolerance, then testing task performance separately from feature agreement.

7.2 Limitations and Future Work

The findings concern this prototype's weights, greedy allocator, and packet syntax. Uncalibrated hearing thresholds, shared transient boosts, boundary error, and absent mixed-window control limit generalization. Substantial metadata and padding mean that equal file lengths do not imply equal useful payload. Historical AAC/Opus runs used target rates, so comparative codec efficiency remains untested. The ablations identify an attribution problem but do not explain its bandwise mechanism.

Full-corpus coverage is limited to P–A at one rate; the broader matrix uses 500 clips, and feature refinement uses one test fold, finite proposals, and one sweep. Source clustering and training-source purging address recording reuse; intervals remain conditional on the corpus, models, and tuning protocol. The single-seed classifier's low clean accuracy limits task evidence. Log-mel optimization is not semantic validation. ViSQOL covers P/W only, STOI remains exploratory, and no human listening or speech-recognition experiment was performed.

Condition	segSNR (dB) ↑	SNR (dB) ↑	E_{RMS} (dB) ↓	ρ_{mel} ↑	Feature error ↓
P	25.529	24.429	27.476	0.9075	109.46
Search, $\lambda = 0$	25.537	24.431	27.460	0.9076	109.27
Search, $\lambda = 0.1$	24.622	24.379	25.000	0.9213	86.80

Table 2: Held-out fold-5 feature refinement: 100 clips, 96 source recordings, equal-length 119834-byte streams. Feature error is the directly optimized squared Euclidean loss F, not a perceptual or semantic score. The zero-penalty control uses the same proposal set and sweep budget.

Future work should trace bandwise bytes and errors to explain the weighting effect, assess feature-refined streams with quality prediction and listening tests, and evaluate task transfer with stronger classifiers, multiple seeds, and speech data. These steps connect the observed tradeoffs to mechanisms and applications.

8 Conclusion

Controlled MDCT comparisons show that a segSNR advantage over a width-weighted comparator does not establish a masking benefit, and that allocation rankings depend on the evaluation endpoint. Feature-regularized search increases log-mel cosine similarity at equal rate and search budgets, with a 0.91 dB segSNR cost and no demonstrated classification gain. These results support a concrete evaluation practice: control transmitted bytes, isolate the proposed allocation factor, and assess its tradeoffs against the application's intended endpoint.

References

Marina Bosi and Richard E. Goldberg. 2003. *Introduction to Digital Audio Coding and Standards*. Kluwer Academic Publishers, Boston, Massachusetts.

Karlheinz Brandenburg and Gerhard Stoll. 1994. ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio. *Journal of the Audio Engineering Society*, 42(10):780–792.

Michael Chinen, Felicia S. C. Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. 2020. ViSQOL v3: An open source production ready objective speech and audio metric. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6.

Vassilis Digalakis, Leonardo Neumeyer, and Manolis Perakakis. 1998. Product-code vector quantization of cepstral parameters for speech recognition over the WWW. In *5th International Conference on Spoken Language Processing (ICSLP 1998)*. Paper 0940.

International Telecommunication Union. 2015. Method for the subjective assessment of intermediate quality level of audio systems. Recommendation ITU-R BS.1534-3, International Telecommunication Union.

Anastasia Kuznetsova, Inseon Jang, Wootaek Lim, and Minje Kim. 2025. Task-specific audio coding for machines: Machine-learned latent features are codes for that machine. In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE.

Ted Painter and Andreas Spanias. 2000. Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4):451–513.

Karol J. Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1015–1018, Brisbane, Australia. Association for Computing Machinery.

John P. Princen and Alan B. Bradley. 1986. Analysis/synthesis filter bank design based on time domain aliasing cancellation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(5):1153–1161.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136.

Jean-Marc Valin, Koen Vos, and Timothy B. Terriberry. 2012. Definition of the Opus audio codec. Request for Comments 6716, Internet Engineering Task Force.