# CARE-VOICE: CONSTRAINT-AWARE RELIABLE EVALUATION VIA SELECTIVE DECISIONS FOR VOICE DESIGN

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

An open-ended voice-design brief may specify attributes to change, properties to preserve, permitted exceptions, and conflicting goals, so no single scalar score can carry the evaluation. We introduce CARE-VOICE (Constraint-Aware Reliable Evaluation via Selective Decisions for Voice Design), a constraint-aware framework that compiles a brief into a Voice Constraint Graph (VCG), records typed evidence and provenance, audits tool health, and delegates the final PASS/FAIL/ABSTAIN decision to a frozen selective scorer. The evaluator may acquire evidence adaptively, but it cannot rewrite the decision contract after observing a candidate. The evidence we assemble supports a reliability claim, not a human-level perceptual claim.

CARE-VOICE is evaluated on complementary evidence tiers: 96 controlled counterfactual conditions with exact labels, 40 parallel semantic pairs, 90 single-attribute and 52 multi-constraint outputs from a deployed service, and existing blinded-listening studies; no new human ratings are collected, and the underpowered multi-constraint study serves only as exploratory failure analysis. Three findings summarize the evidence. First, global RMS, a common shorthand for vocal energy, agrees with the human majority on only 25% of semantic pairs, whereas typed evidence with selective refusal reaches 100% selective agreement at 42.5% coverage. Second, against strict human majorities on single-attribute pairs, CARE-VOICE retains comparable agreement to a strong direct judge (84.0% selective at 75.8% coverage versus 81.8% forced), with overlapping uncertainty and therefore no superiority claim. Third, abstentions and position-inconsistency checks flag half of the pairs on which listeners fail to reach a majority, exposing ambiguity that forced labels conceal.

Constraint structure prevents unsupported penalties, evidence authority blocks weak proxies from deciding alone, and selective refusal turns judge instability, identity drift, and tool failures into recorded, auditable events.

## 1 INTRODUCTION

Modern text-to-speech and voice design systems accept instructions such as *younger, calm, trustworthy, slightly breathy, but preserve pronunciation and avoid an advertising tone*. Evaluating such outputs is difficult for three reasons. First, a brief contains multiple kinds of constraints: properties to change, properties to preserve, properties that are unconstrained, and exceptions that should not be penalized. Second, many convenient audio measurements are only proxies for the intended construct—global RMS, for example, measures recording amplitude, not perceived vocal energy. Third, perceptual judges and external tools may be unavailable, inconsistent, or unable to access the audio. A system that always emits a scalar score therefore converts epistemic uncertainty and tool failures into false precision.

Recent work has substantially improved instruction-following TTS evaluation through multi-dimensional benchmarks, model-based judges, and pairwise audio protocols (Huang et al., 2025; Manku et al., 2025; Jeon et al., 2026; Li et al., 2026). In parallel, agent-as-a-judge methods decompose evaluation into planning and tool use (Zhuge et al., 2024). We build on these directions but address a different question: *how should an evaluator act when the available evidence has different validity, cost, and failure modes?*



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We propose CARE-VOICE, a constraint-aware active evaluator for voice design whose agentic component is deliberately narrow: it parses a brief, identifies required evidence, executes tools, records provenance, and escalates uncertain cases, while a deterministic scoring contract owns the final label and cannot be rewritten at run time. The separation preserves auditability while retaining the flexibility of tool-using agents.

Our contributions are:

- We formulate a *constraint–evidence contract*: the Voice Constraint Graph represents target, preservation, unconstrained, conflict, and exemption relations independently of the tools used to inspect audio.
- We define evidence authority, provenance, tool-health auditing, and a frozen selective rule separating technical failure from perceptual uncertainty, with first-class ABSTAIN.
- We construct deterministic counterfactual, fault-injection, and metamorphic tests of invariance, monotonicity, reversibility, path dependence, and technical-gate behavior without additional human listening.
- We reanalyze existing blinded-listening data as a reliability benchmark: human majority labels evaluate accepted decisions, while unresolved votes evaluate whether automated warning signals expose perceptual ambiguity.

## 2 RELATED WORK

**Evaluation of controllable speech generation.** InstructTTSEval evaluates complex natural-language instruction following across heterogeneous speech attributes (Huang et al., 2025). EmergentTTS-Eval focuses on prosody, expressiveness, and linguistic challenges using a model-as-a-judge protocol (Manku et al., 2025). ParaPairAudioBench studies pairwise paralinguistic judgments by large audio-language models (Jeon et al., 2026), while AnyAudio-Judge uses dynamically generated rubrics for general audio instruction following (Li et al., 2026). These works show that fixed waveform metrics are insufficient for open-ended instructions. CARE-VOICE is complementary, focusing on constraint compilation, evidence provenance, tool failure, and selective risk rather than a universal perceptual judge.

**LLM and agent-based judges.** LLM-as-a-judge systems can correlate with human preference but inherit prompt sensitivity, position bias, and access failures (Zheng et al., 2023). Agent-as-a-Judge decomposes evaluation into an agentic workflow that can inspect intermediate artifacts and use tools (Zhuge et al., 2024). Our design adopts decomposition but constrains authority: the agent may select evidence but may not modify the scoring contract or silently replace missing evidence.

**Selective prediction.** Selective classification lets a system reject examples in exchange for lower risk on the accepted subset (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019)—a natural fit for expensive perceptual evaluation, where coverage and selective accuracy are reported together. We apply it to voice design and make the rejection reason auditable.

## 3 PROBLEM FORMULATION

Let a voice brief be $b$, an audio candidate be $x$, and the requested comparison or acceptance decision be $q$. A conventional evaluator directly predicts a score $s = f(b, x, q)$. We instead compile $b$ into a graph $G = (V, E)$ and collect an evidence set

$$\mathcal{E} = \{(z_i, r_i, p_i, h_i)\}_{i=1}^{m}, \tag{1}$$

where $z_i$ is an observation, $r_i$ is its evidence role, $p_i$ is provenance, and $h_i$ is tool-health metadata. The role $r_i$ is one of *primary perceptual evidence*, *direct proxy*, *weak proxy*, or *diagnostic*. A frozen scorer $S(G, \mathcal{E})$ returns one of $\{\text{PASS}, \text{FAIL}, \text{ABSTAIN}\}$.

For a dataset of $n$ labeled comparisons, let $g_i \in \{0, 1\}$ indicate whether the evaluator accepts the case and $c_i \in \{0, 1\}$ indicate correctness. We report coverage and selective risk:

$$\text{Coverage} = \frac{1}{n}\sum_i g_i, \qquad \text{Risk} = 1 - \frac{\sum_i g_i c_i}{\max(1, \sum_i g_i)}. \tag{2}$$

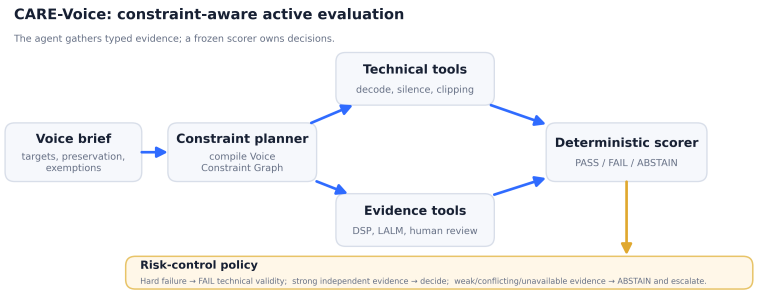



Figure 1: CARE-VOICEseparates agentic evidence acquisition from deterministic scoring. Missing or conflicting semantic evidence triggers abstention and escalation rather than an invented score.

Overall accuracy, reported only as a secondary measure, counts abstentions as incorrect so that low coverage cannot hide behind high accuracy on a small accepted subset. Conversely, overall accuracy is bounded by coverage: the two must be read together, with selective risk on accepted cases as the primary measure.

# 4 METHOD

## 4.1 VOICE CONSTRAINT GRAPH

The VCG represents five relation types. **Targets** specify desired changes such as higher pitch or greater semantic energy. **Preservation constraints** specify attributes such as linguistic content, speaker identity, or accent that should remain stable. **Unconstrained attributes** are explicitly excluded from scoring. **Conflicts** record mutually incompatible requirements. **Exemptions** define conditions under which a rule should not apply. Each edge contains a priority, expected direction, tolerance, and admissible evidence roles. This representation avoids a common evaluation error: penalizing a candidate for changing an attribute that the user never asked to preserve.

As a concrete example, the brief from the introduction—*younger, calm, trustworthy, slightly breathy, but preserve pronunciation and avoid an advertising tone*—compiles to four target edges (perceived age ↓, arousal ↓ for *calm*, breathiness ↑ with a small tolerance capturing *slightly*, advertising tone *absent*), one preservation edge on linguistic content, and one recorded conflict: *younger* and *trustworthy* compete for pitch if both are operationalized through median F0, so the edge priorities are stored and the scorer reports the conflict instead of silently averaging the two goals. Attributes the brief never mentions, such as speaking rate, default to unconstrained and are excluded from scoring. The admissible-evidence field then constrains what may decide each edge: perceived age admits human or qualified LALM comparisons, pitch admits median F0, and no energy-related edge may be decided by RMS alone. The full edge schema, the aggregation rules that map a graph and an evidence set to a decision, and their correspondence to the audit reason codes are specified in Appendix B.

## 4.2 TYPED EVIDENCE AND ACTIVE ACQUISITION

Evidence is grouped by construct validity, not by whether it is numerical. Primary perceptual evidence directly answers the requested construct, for example a blinded human comparison of perceived age. Direct proxies have a stable and narrow relation to the target, such as median F0 for pitch or effective voiced duration for pace under parallel text. Weak proxies correlate with the construct but are insufficient alone; RMS for perceived energy is the motivating example. Diagnostics explain failures but do not enter the score.

The agent executes a bounded policy: (1) run technical gates; (2) collect low-cost direct proxies; (3) test whether independent evidence families agree beyond frozen thresholds; (4) request an audio-capable model or human review only when the expected value of information is positive; and (5) abstain if required evidence is unavailable. The released implementation uses a deterministic policy rather than a learned controller, so every tool call and stopping decision is reproducible.

### 4.3 HARD GATES AND FROZEN SCORING

Technical gates check decodability, duration, clipping, silence, and DC offset before semantic scoring. For controlled pitch and pace comparisons, the default thresholds are 1 semitone and 8% effective-pace change, respectively. In the semantic experiment, automatic acceptance requires both pitch and timing proxies to exceed their thresholds in the same direction. Weak proxies cannot independently produce PASSor FAIL. If direct proxies disagree or remain below threshold, the system returns ABSTAIN. Tool errors (missing audio attachments, rate limits) are stored as evidence-health events, not converted into low candidate scores.

For a weak-proxy conflict in the real-service pilot, primary perceptual evidence may override the proxy only if the judge demonstrates access to both files, returns the same canonical direction after swapping A/B order, reaches a frozen confidence of 0.80, and has no high-severity evidence-audit error. The audit checks explicit duration claims against decoded waveform durations and flags rationale statements that conflict with shared preservation constraints. Findings are graded: fabricated access claims and duration mismatches beyond a frozen tolerance are high-severity and block the override, while preservation-claim conflicts and smaller duration mismatches are medium-severity annotations recorded alongside the decision.

## 5 EXPERIMENTAL DESIGN

We use complementary experiments with different validity profiles. Controlled transformations provide exact labels for evaluator mechanics. Parallel semantic speech provides both corpus-defined relations and existing listener votes for studying construct validity. A deployed VoiceDesign service supplies 90 single-attribute outputs and 52 outputs for multi-constraint briefs. Existing blinded-listening responses provide perceptual majority labels and unresolved cases; no new listening data are collected. Speaker-held-out splits, signal attacks, Praat comparison, repeated LALM calls, and metamorphic relations test generalization and tool reliability. We package these settings as the versioned CARE-Voice Reliability Benchmark and report them separately rather than merging them into a single leaderboard score.

### 5.1 CONTROLLED COUNTERFACTUAL BENCHMARK

We select one utterance from each of eight speakers in the Mini LibriSpeech dev-clean subset (Panayotov et al., 2015) and create 12 conditions per source using FFmpeg—a normalized base, pitch shifts of $\pm 3$ and $\pm 0.3$ semitones, tempo factors of $0.8$ and $1.25$, a fast condition padded with silence, loudness shifts of $\pm 6$ dB, clipping, and near-silence—yielding 96 audio files, 40 pairwise cases, and 24 technical-gate cases in four slices: **strong** cases with effects well beyond the decision threshold, **ambiguous** pitch cases that are expected ties, **silence confounds** that preserve fast articulation while reversing a naive total-duration comparison, and **technical failures**. Because the transformations are known, this experiment evaluates implementation correctness rather than human preference.

### 5.2 PARALLEL SEMANTIC ACTIVATION

We use EmoV-DB, a parallel emotional speech corpus (CC BY-NC 4.0) (Adigwe et al., 2018): 20 utterance identifiers shared by the *sleepy* and *amused* recordings of two speakers, yielding 40 same-speaker, same-content pairs, converted to mono 16-kHz signed PCM without loudness normalization. We treat the corpus relation sleepy $<$ amused as a proxy label for semantic activation, not a human gold rating of a continuous energy scale.

For each pair we measure median F0 and effective voiced duration (direct proxies from independent pitch and timing families), global RMS and spectral centroid (weak proxies), and F0 range and RMS dynamic range (diagnostic). We compare RMS-only, pitch-only, pace-only, forced fusion without abstention, fusion without evidence roles, and CARE-VOICE, with thresholds frozen at 1 semitone and 8% effective-pace change and bootstrap 95% intervals from 5,000 resamples.

### 5.3 Real-Service Generation

We first run a 10-sample integration pilot, then scale to 90 Mandarin samples: three texts, three target attributes, two directions, and five stochastic replicates. This study measures directional prompt control, not reference-conditioned speaker editing, because the service does not expose a reusable voice identity. Two judge models are used: the service's incumbent operational judge, Gemini 3 Flash Preview, and a second judge, Gemini 3.5 Flash, admitted under the frozen gate of §5.4. Unless noted otherwise, single-attribute and human-agreement results use the admitted judge, and multi-constraint operational results use the incumbent judge; each judge receives both candidate URLs in one request. We compare a direct question, a dynamically derived rubric, and active position-balanced judging. We additionally generate baseline and brief-conditioned candidates for 26 non-conflicting multi-constraint briefs (52 samples); the remaining four briefs contain explicit contradictions and are evaluated only in the structural VCG benchmark.

### 5.4 Second-LALM Qualification

We qualify Gemini 3.5 Flash as a *second* audio-language evidence source—second to the incumbent operational judge—using a frozen qualification gate on public counterfactual audio. The model receives both complete candidate files in one request and is evaluated in both presentation orders. The 24 calls comprise 16 strong pitch/pace judgments, four repeated forward judgments, and four invalid-audio diagnostics. Every valid response must demonstrate access by transcribing both candidates and return a strict JSON decision, confidence, evidence, and quality warning. Admission requires at least 95% audio access, strict parsing, and returned-model identity, and at least 75% on strong-direction accuracy, pitch accuracy, pace accuracy, position consistency, and repeat stability; invalid-audio rejection is diagnostic because deterministic DSP owns technical validity. Any failed gate blocks the model from the formal decision chain.

### 5.5 Existing Blinded-Listening Data

We reuse three completed studies under their original quality-control rules. The semantic study contains 40 sleepy/amused pairs; nine of ten listeners pass the attention gate. The single-attribute study contains 45 higher/lower pairs over pitch, pace, and energy; all ten listeners qualify, and a label is resolved only when one response receives a strict majority on the requested dimension. This yields 33 resolved pairs and 12 unresolved pairs. The multi-constraint study contains 26 baseline/designed pairs, but only six of ten listeners satisfy the preregistered repeat-consistency rule (eight required), so it serves only as exploratory failure analysis.

The unit of analysis is an audio pair, not an individual rating. We report Wilson intervals for agreement, listener-vote entropy, consensus-stratified sensitivity analyses, and ambiguity-detection precision/recall, treating an unresolved strict-majority outcome as the target event. Consensus thresholds and ambiguity analyses were specified after inspecting completed aggregate data and are therefore retrospective; they do not replace the primary strict-majority analysis.

## 6 Results

### 6.1 Controlled Transformations

Table 1 shows that the robust deterministic metrics recover all 40 known pair relations; a naive total-duration baseline fails all eight silence-confounded pace comparisons. CARE-VOICEaccepts 27 of 40 cases and is correct on all accepted cases. The 13 abstentions concentrate on the two slices the frozen contract refuses to decide: deliberately ambiguous pitch pairs, whose effects sit inside the tie band, and loudness pairs, whose only automatic evidence (global RMS) is typed as a weak proxy and cannot decide alone. The technical gate correctly classifies all 24 clean, clipped, and silent cases; without it, gate-case accuracy falls to 33.3% (8/24) as clipped and near-silent audio enters semantic scoring.

Table 1: Controlled counterfactual benchmark. Abstentions are typed refusals on evidence-insufficient cases, not silent errors; overall accuracy counts each refusal as an error. The naive baseline is defined only on the eight silence-confounded pace cases.

| Evaluator | Cases | Coverage | Selective acc. | Overall acc. |
|---|---|---|---|---|
| Naive total duration | 8 | 100.0 | 0.0 | 0.0 |
| Robust DSP | 40 | 100.0 | 100.0 | 100.0 |
| CARE-VOICE | 40 | 67.5 | 100.0 | 67.5 |

Table 2: Agreement with the human energy majority on 40 parallel EmoV-DB pairs. Selective accuracy is measured only on non-abstained cases; overall accuracy counts each refusal as an error and is bounded by coverage, so a selective system cannot win the overall column by construction.

| Evaluator | Coverage | Selective acc. | Overall acc. |
|---|---|---|---|
| RMS only | 100.0 | 25.0 | 25.0 |
| Pitch only | 100.0 | 90.0 | 90.0 |
| Pace only | 100.0 | 80.0 | 80.0 |
| No abstention | 100.0 | 95.0 | 95.0 |
| No evidence roles | 100.0 | 85.0 | 85.0 |
| CARE-VOICE | 42.5 | 100.0 | 42.5 |

### 6.2 SEMANTIC ACTIVATION AND EVIDENCE VALIDITY

Table 2 exposes a construct-validity failure. Even though global RMS is often used as shorthand for vocal energy, RMS-only evaluation agrees with the human majority on only 25% of the 40 pairs—and the reason is directional: amused speech has a mean RMS *2.20 dB lower* than sleepy speech (95% bootstrap interval $[-3.28, -1.09]$), while median F0 is 3.16 semitones higher and effective pace 31.31% faster. The human labels are far from unanimous: all 40 pairs have an amused $>$ sleepy majority, but only 255/360 individual votes (70.8%) support it—six listeners almost always select amused, three frequently select sleepy—and Fleiss' $\kappa = -0.103$ despite the uniform pair-level majority. A majority label, a construct definition, and an acoustic proxy are distinct objects.

The no-abstention fusion makes two errors (95% agreement); giving global RMS equal authority adds four more (85%). The frozen CARE-VOICE rule accepts only the 17 pairs where both independent direct proxies exceed threshold in the same direction, reaching 100% selective agreement at 42.5% coverage (Figure 2). This supports evidence typing and selective refusal, not a claim that the rule captures every listener's interpretation of energy.

### 6.3 SCALED REAL-SERVICE EVALUATION

All 90 generated samples pass the technical gates. Pitch and pace are controllable but stochastic: 10/15 pitch and 11/15 pace pairs exceed the frozen DSP threshold in the intended direction. The nominal high-energy prompt produces higher global RMS in only 2/15 pairs and is lower in 8/15, again demonstrating that RMS cannot serve as the sole semantic judge. Full DSP outcomes appear in Appendix E.

The direct and dynamic-rubric Gemini 3.5 Flash judges access all 45 pairs and agree with the prompt-defined direction on 95.6% (43/45) and 93.3% (42/45), respectively. These values diagnose the generation protocol and judge behavior; they are not perceptual accuracy. The active policy routes 26/45 pairs to a position-balanced judge; 19 remain consistent after swapping. Frozen fusion produces 29 PASS, 2 FAIL, and 14 ABSTAIN outcomes (full operational results in Appendix E).

Before interpreting attribute agreement, human ratings answer a question the pipeline cannot: did the service edit a voice, or sample two different voices? Mean speaker similarity between paired candidates is 2.49/5, with 48.4% of ratings at most 2, and under the strict local-edit definition—the requested attribute changes while both non-target attributes remain equivalent—*none of the 45 pairs succeeds*. Attribute agreement below therefore compares two independently generated voices, not identity-preserving editing.

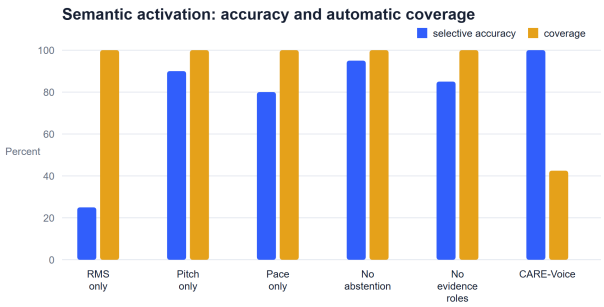



Figure 2: Semantic-activation accuracy versus coverage. Evidence typing denies the misleading RMS proxy equal authority; abstention removes the remaining forced-decision errors.

Table 3: Agreement with the strict human majority on the requested dimension (Gemini entries: the admitted Gemini 3.5 Flash judge). Coverage is relative to 33 resolved pairs; intervals are Wilson 95%; DSP ties are kept as ties. Selective agreement on accepted pairs is the primary metric; overall agreement counts abstentions as incorrect by construction.

| Evaluator | Coverage | Selective agreement | 95% CI | Overall agreement |
|---|---|---|---|---|
| DSP | 100.0% | 63.6% (21/33) | 46.6–77.8 | 63.6% (21/33) |
| Gemini direct | 100.0% | 81.8% (27/33) | 65.6–91.4 | 81.8% (27/33) |
| Gemini dynamic | 100.0% | 78.8% (26/33) | 62.3–89.3 | 78.8% (26/33) |
| CARE-VOICE | 75.8% | 84.0% (21/25) | 65.4–93.6 | 63.6% (21/33) |

Read under this interpretation, Table 3 gives the primary comparison against strict human majorities. Direct Gemini 3.5 Flash reaches 81.8% agreement on all 33 resolved pairs; CARE-VOICE reaches 84.0% on 25 accepted pairs at 75.8% coverage. The Wilson intervals overlap substantially, so the result does not establish superiority over direct judging; the frozen policy retains comparable observed agreement while refusing a subset of cases.

Two qualifications apply. First, the frozen contract yields a single risk–coverage operating point rather than a tunable curve; we therefore also report forced-answer *overall* agreement, under which CARE-VOICE agrees on 21/33 resolved pairs (63.6%), identical to the forced DSP baseline: refusal converts uncertain forced choices into explicit escalations at a measurable coverage cost rather than manufacturing accuracy. Second, eight of the fourteen abstentions fall on resolved pairs whose majority label is in principle decidable, so the frozen rule sacrifices decidable cases alongside genuinely ambiguous ones.

Agreement strengthens with listener consensus in a retrospective sensitivity analysis (Appendix I), and pitch and energy remain the hardest dimensions for listeners: only 8/15 pitch and 11/15 energy pairs reach a strict majority, versus 14/15 pace pairs.

Can automated signals identify the 12 pairs on which listeners fail to form a strict majority? CARE-VOICE abstains on 14/45 pairs, including 6/12 unresolved pairs (42.9% precision, 50.0% recall); a direct-confidence flag below 1.0 captures 9/12 but marks 29 pairs (31.0% precision); position inconsistency marks seven pairs, four unresolved (57.1% precision, 33.3% recall). No current signal solves ambiguity detection, but position balancing and evidence audits expose failure information that a forced label omits. Active acquisition sends 57.8% of pairs (26/45) to the LALM (relative-cost accounting: Appendix E); no policy parameter is tuned on human data.

### 6.4 MULTI-CONSTRAINT VCG EVALUATION

On 30 exact structural cases, VCG obtains 100% constraint recall, zero false penalties on unconstrained or exempted attributes, and 100% conflict detection; a static rubric has full recall but a 66.67% false-penalty rate and 75% conflict detection; an unstructured baseline has 80% recall, the

same false-penalty rate, and no conflict detection. This validates the representation and parser, not perceptual compliance.

All 52 generated samples pass technical gates. The admitted second judge passes every frozen admission gate but detects only 1/4 invalid-audio cases, so deterministic tools retain technical authority (per-task results in Appendix D). For the primary operational judge (the incumbent Gemini 3 Flash Preview), direct judging prefers the brief-conditioned candidate in 46.2% of pairs (12/26) and a dynamic rubric raises this to 73.1% (19/26). Position-balanced VCG judging is order-consistent on only 9/26 pairs. The second judge likewise reaches only 50.0% (13/26) direct and 65.4% (17/26) dynamic prompt agreement, with 50.0% VCG position consistency. The VCG therefore acts as a structural stress test and abstention trigger, not an accuracy-improving prompt recipe.

Benchmark users should measure position consistency per judge and task family *before* trusting multi-constraint judgments, excluding order-unstable decisions by abstention. Here this retains only 9/26 primary-judge decisions, converting unreliable forced choices into explicit escalation—the behavior the frozen contract is designed to produce.

The listening data do not support a formal leaderboard (only six listeners pass the preregistered gate). Exploratory brief-fit majorities favor the designed candidate on seven pairs, the baseline on twelve, and neither on seven; direct and dynamic Gemini each agree with 8/19 resolved majorities, the stable VCG subset with 5/9; mean speaker similarity is 2.38/5. The original preservation item encoded the wrong success direction, so preservation success is removed from all main claims. Full votes and protocol deviations appear in Appendix I.

### 6.5 Generalization and Tool Validation

Speaker-disjoint threshold selection on EmoV-DB yields 62.5% mean test coverage with 100% selective agreement across two folds; the gap to the frozen 42.5% operating point shows thresholds do not transfer uniformly across speakers; we retain frozen thresholds because a fixed contract, not maximal coverage, is under test. Signal attacks, a Praat comparison (0.35-semitone median F0 error; 94.4% direction agreement), and metamorphic tests confirm evaluator mechanics, with recorded exceptions—pitch under one reverb slice, pace under loudness-affecting attacks, and 75% path independence (full results in Appendix G).

## 7 Discussion

**Scope of the agentic component.** The novelty is not an unconstrained model rating audio: the evaluator translates a brief into a testable graph, chooses evidence by validity, audits tool health, and escalates under cost and uncertainty, while a deterministic scorer prevents post-hoc rule rewriting. Delegating scoring to a stronger LALM introduces transport, prompt, version, and position dependencies and cannot replace technical gates: the admitted Gemini 3.5 Flash still reaches only 50% multi-constraint position consistency and 1/4 invalid-audio detection—admission establishes bounded competence, not blanket authority.

**Benchmark implications.** Voice design benchmarks should report exact-counterfactual correctness, technical-gate performance, selective risk–coverage, position consistency, repeated-call variance, and tool success; a single score can reward proxy misuse or conceal judge outages. Twelve of 45 single-attribute pairs lack a strict majority—pitch and energy are less stable than pace—so human disagreement is an evaluation target: benchmarks should measure agreement on resolved labels and detection of unstable cases. The present rule recalls only half of unresolved pairs, but abstention, position checks, and audits expose the shortfall.

**Majority labels are not perceptual ground truth.** The semantic study illustrates the risk: all 40 pairs have an amused majority, yet Fleiss' $\kappa = -0.103$ because three of nine listeners systematically prefer sleepy—the majority aggregates listeners who may not share the construct. There is also partial circularity: the corpus sleepy $<$ amused relation is realized largely through pitch and timing, so a rule accepting exactly the large-effect pairs aligns with the corpus direction partly by construction; the 100% selective agreement is consistency with a pair-level majority, not perception-level correctness.

The single-attribute study partially mitigates both issues—prompt-defined directions, independent listeners, unresolved pairs—so majority labels serve as a reliability reference with measurable instability.

## 8 Limitations and Responsible Use

The semantic study has two English speakers, the single-attribute benchmark 45 pairs and ten non-expert Mandarin listeners, and the multi-constraint study is underpowered; intervals are wide and consensus analyses retrospective. Independent candidate generation and low speaker similarity create identity confounding.

Effective pace is sensitive to gain, normalization, and noise; even direct proxies degrade under uncontrolled conditions. Speaker-held-out threshold selection shifts coverage from 42.5% to 62.5%; we keep frozen thresholds at the cost of that instability. Both admitted judges are Gemini-family deployments whose private versions may change, so shared failure modes remain possible; three further models failed admission (Appendix H), so instability evidence reflects one admitted family plus admission failures, not a broad cross-family comparison. The human comparison shows no superiority over direct judging; under the forced-answer convention CARE-Voice only matches the DSP baseline. The policy is hand-written, costs are relative units (Appendix E), and reference-conditioned identity preservation is untested.

Responsible-use considerations appear in the Ethics Statement.

## 9 Conclusion

We presented CARE-Voice, a constraint-aware architecture for reliable voice-design evaluation. It separates flexible evidence acquisition from frozen selective scoring, verified by exact transformations, fault injection, metamorphic relations, and existing listening data. On resolved pairs it matches a strong direct judge's agreement at lower coverage, while abstentions and position checks expose half of human disagreement. Unsupported decisions, identity confounds, weak-proxy failures, and judge instability become recorded events with reason codes, not hidden error terms.

## Ethics Statement

Voice design evaluation can encode demographic and cultural stereotypes, particularly for perceived age, gender presentation, professionalism, and personality. A VCG should therefore preserve the user's explicit intent while avoiding protected-attribute inference that is unrelated to the task. The present system does not clone or identify real speakers, and we do not claim human-equivalent perceptual judgment. The EmoV-DB subset is used under its CC BY-NC 4.0 license for non-commercial research only. The listening studies were run under their original consent and exclusion protocols, and this paper reuses their completed responses without collecting new ratings. We do not release internal service credentials.

## Reproducibility Statement

The frozen scoring contract, thresholds, and evidence-role assignments appear in Appendix A; the VCG edge schema and aggregation rules in Appendix B; and the qualification, robustness, and human-review protocols in Appendices D–I. We release the versioned CARE-Voice Reliability Benchmark—normalized task records with compiled VCGs, validation schemas, reference submissions, and the unified evaluator—together with the analysis code that regenerates every table and figure, as described in Appendix K.

## AI Use Statement

The Gemini models discussed in this paper are subjects of the evaluation. AI assistants were used for language polishing; the research design, experiments, analyses, and all scientific claims are the authors' own, and the authors take full responsibility for the paper's content.

## REFERENCES

Adaeze Adigwe, Noë Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. In *Proceedings of the 9th International Conference on Speech Prosody*, 2018.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2151–2159, 2019.

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. Instructttseval: Benchmarking complex natural-language instruction following in text-to-speech systems. *arXiv preprint arXiv:2506.16381*, 2025.

Jisu Jeon, Seungyeon Jwa, Joosung Lee, Jinhyeon Kim, Woojin Chung, Hwiyeol Jo, Jeonghoon Kim, Jonghyun Choi, and Soyoon Kim. Parapairaudiobench: Paralinguistic pairwise audio benchmark for lalm-as-a-judge. *arXiv preprint arXiv:2606.24648*, 2026.

Haitao Li, Tian Tan, Yuguang Yang, Shan Yang, and Xie Chen. Anyaudio-judge: A dynamic rubric-based benchmark and evaluator for audio instruction following. *arXiv preprint arXiv:2606.03116*, 2026.

Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. Emergenttts-eval: Evaluating tts models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. *arXiv preprint arXiv:2505.23009*, 2025.

OpenMOSS Team. MOSS-Audio Technical Report. *arXiv preprint arXiv:2606.01802*, 2026.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*, 2024.

## A REPRODUCIBILITY DETAILS

### A.1 FROZEN THRESHOLDS AND EVIDENCE ROLES

Table 4 lists the frozen contract of the released implementation. Thresholds were selected before the EmoV-DB analysis from the original pilot protocol. The semantic experiment does not tune them per speaker or per utterance.

The effective pace effect for parallel text is

$$100\left(\frac{d_{\text{sleepy}}}{d_{\text{amused}}}-1\right), \tag{3}$$

where $d$ is effective voiced duration after excluding frames below $-45$ dBFS. F0 is estimated with a windowed autocorrelation tracker over the range 60–400 Hz. The estimator is intentionally dependency-light and all intermediate metrics are serialized.

Table 4: Frozen evidence contract used by the released implementation.

| Construct | Observation | Threshold | Role |
|---|---|---|---|
| Technical validity | clipping ratio | $\leq 0.001$ | hard gate |
| Technical validity | silence ratio | $\leq 0.85$ | hard gate |
| Pitch direction | median F0 | 1 semitone | direct proxy |
| Pace direction | voiced duration | 8% | direct proxy |
| Semantic energy | global RMS | 1.5 dB | weak proxy |
| Semantic energy | spectral centroid | 8% | weak proxy |
| Semantic energy | F0 range | 1 semitone | diagnostic |

### A.2 Controlled transformation matrix

For each of eight Mini LibriSpeech utterances, the benchmark contains the following generated conditions:

- normalized base audio;
- pitch at $-3$ and $+3$ semitones;
- weak pitch at $-0.3$ and $+0.3$ semitones, labeled as a tie under the 1-semitone threshold;
- tempo at $0.8\times$ and $1.25\times$;
- the $1.25\times$ condition with added trailing silence;
- loudness at $-6$ and $+6$ dB;
- a clipped condition and a near-silent condition.

This produces five pairwise relations per utterance: strong pitch, ambiguous pitch, strong pace, silence-confounded pace, and strong loudness. Three gate cases per utterance evaluate clean, clipped, and silent audio.

### A.3 Speaker-level semantic results

Table 5: Speaker-level accuracy and coverage on the EmoV-DB subset.

| Speaker | Evaluator | Selective acc. | Coverage |
|---|---|---|---|
| Jenie | No evidence roles | 95.0 | 100.0 |
| Jenie | CARE-Voice | 100.0 | 40.0 |
| Josh | No evidence roles | 75.0 | 100.0 |
| Josh | CARE-Voice | 100.0 | 45.0 |

The degradation caused by equal-weight RMS is concentrated in the Josh subset, where RMS agrees with the activation ordering in only two of 20 pairs. The speaker variation reinforces that an evidence role should reflect construct validity, not observed accuracy on a single corpus partition.

## B VCG Specification

This appendix specifies the Voice Constraint Graph as implemented in the released artifact: the edge schema, the aggregation rules that map a graph and an evidence set to a Pass/Fail/Abstain decision, and their correspondence to the audit reason codes. The specification describes only what the released implementation executes; it is not a design sketch of extensions.

### B.1 Edge schema

A VCG is a set of typed edges over the attributes mentioned in a brief. Attributes the brief does not mention are unconstrained by default and carry no edge. Table 6 defines the edge fields.

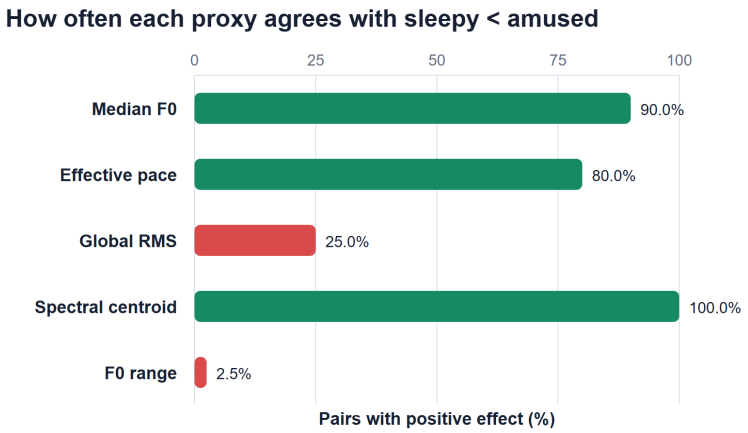



Figure 3: Fraction of semantic pairs with a positive effect under the sleepy $<$ amused relation. The proxy families disagree sharply.

Table 6: VCG edge schema. Runtime edges carry all six fields; admissible roles are drawn from the four-level evidence typing of §4.2. Serialized benchmark records store the relation, attribute, and direction slots, while priorities, tolerances, and admissible roles are resolved from the frozen contract (Table 4); the counterfactual track additionally records the per-item evidence role and minimum effect.

| Field | Domain | Meaning |
|---|---|---|
| Relation | {target, preservation, unconstrained, conflict, exemption} | how the attribute enters scoring |
| Attribute | identifier from the brief's attribute vocabulary | the property under test |
| Priority | ordered integer | resolution order when edges conflict |
| Expected direction | {increase, decrease, preserve, absent} | the requested effect |
| Tolerance | per-construct magnitude with units | below it, a change is a tie |
| Admissible roles | subset of evidence roles | which evidence may decide this edge |

As an example, the brief *younger, calm, trustworthy, slightly breathy, but preserve pronunciation and avoid an advertising tone* compiles to: target edges on perceived age (decrease; primary perceptual evidence only), arousal for *calm* (decrease), breathiness (increase; small tolerance for *slightly*), and advertising tone (absent); a preservation edge on linguistic content (preserve; equivalence within tolerance; ASR-based and human evidence admissible); and a conflict edge recording that perceived age and trustworthiness compete for pitch when both are operationalized through median F0, with the two target priorities stored for reporting. Speaking rate, never mentioned, receives no edge and is excluded from scoring.

### B.2 Aggregation rules

Given a compiled VCG and an evidence set, the frozen scorer applies the following rules in order. Each rule emits the audit reason code in parentheses.

1. **Technical gate.** If any technical gate fails (decodability, duration, clipping, silence, DC offset), the decision is FAIL on technical validity and no semantic edge is scored (`technical_gate_failed`).
2. **Exclusion.** Unconstrained edges and edges covered by an exemption are removed from scoring before any evidence is consulted; penalizing them is itself logged as an evaluation error.

3. **Per-edge decision.** An edge is decided only by evidence carrying an admissible role. Weak proxies never decide an edge alone. For controlled pitch and pace comparisons, acceptance requires both independent direct proxies to exceed their frozen thresholds in the same direction (`independent_direct_proxies_agree`); agreement in the opposite direction yields FAIL (`independent_direct_proxies_reverse`).

4. **Refusal.** If a required edge has no admissible evidence, the admissible evidence is conflicting or below threshold, or the graph contains an active conflict edge, the decision is ABSTAIN (`direct_proxies_weak_or_conflicting`). Conflict edges are reported with their stored priorities and are never resolved by averaging the competing goals.

5. **Aggregation.** The pair-level decision is the conjunction of edge outcomes under a frozen selective rule: all required edges satisfied yields PASS; a decisive violation yields FAIL; otherwise ABSTAIN. The rule is fixed before any candidate is observed and cannot be rewritten at run time.

**Scope.** Free-text parsing is outside the scope of this specification. The released benchmark makes scoring reproducible without a parser: in the structural track, each task record pairs the raw brief text with a compiled constraint graph—a `gold` record containing targets with expected directions, preservation, unconstrained, exemption, and conflict slots—so constraint-level scoring never depends on a parser at evaluation time.

## C REAL-SERVICE PILOT DETAILS

The fixed Mandarin preview text contains neutral declarative speech, an emphasized phrase, and a sentence-final cadence. All conditions share a young-adult female presentation, standard Mandarin, clear articulation, and a professional non-advertising style. Only one target attribute is changed per condition. Table 7 lists the paired outcomes.

Table 7: Directional-control pilot with 10 generated audio samples.

| Axis | Replicates | Hits | Pooled effect | Agent outcome |
|---|---|---|---|---|
| Pitch | 2 | 1 | $+2.24$ semitones | 1 pass, 1 abstain |
| Pace | 2 | 1 | $+7.03\%$ | 1 pass, 1 abstain |
| Energy | 1 | 0 | $-4.00$ dB RMS | 1 pass, 0 abstain |

The first pace replicate is accepted because direct timing evidence and the position-balanced perceptual judgment agree. The second remains below the 8% threshold and its rationale contains a high-severity duration error, so the system abstains. For energy, the position-consistent perceptual judgment overrides contradictory RMS because RMS is explicitly typed as a weak proxy.

Table 8: Position-balanced Gemini judgments and rationale audit. The expected prompt direction is B.

| Pair | Decision | Order consistent | Confidence | Audit finding |
|---|---|---|---|---|
| Pitch 1 | B | yes | 0.95 | preservation claim conflict |
| Pitch 2 | abstain | no | 0.00 | preservation claim conflict |
| Pace 1 | B | yes | 0.80 | duration claim mismatch |
| Pace 2 | B | yes | 0.95 | high-severity duration mismatch |
| Energy 1 | B | yes | 0.80 | none in scored evidence |

All five pairs demonstrate audio access. Four are order-consistent and all four decided labels follow the prompt-defined direction. The audit does not automatically equate a warning with a wrong label: it separately records whether the categorical decision is stable and whether the supplied explanation contains independently checkable errors.

## D SECOND-LALM QUALIFICATION DETAILS

Two judge models are used throughout this paper. Gemini 3 Flash Preview is the incumbent operational judge of the deployed service; Gemini 3.5 Flash is the second judge admitted through the frozen qualification gate. Unless a table states otherwise, the reported single-attribute judgments and the human-agreement comparison use Gemini 3.5 Flash, while the multi-constraint operational results use Gemini 3 Flash Preview. Gemini 3.5 Flash is accessed through an API endpoint whose URL and key the executable artifact reads from environment variables (`GEMINI_RESPONSES_URL`, `GEMINI_RESPONSES_API_KEY`). Remote audio uses `fileData`; local WAV files use Base64 `inlineData`. Raw audit records replace Base64 payloads with length metadata and never store authorization headers. The request name contains spaces (`Gemini 3.5 Flash`); the returned service identifier is `gemini-3.5-flash`.

```
POST <endpoint-url>
Authorization: Bearer <redacted>
{"model":"Gemini 3.5 Flash",
 "contents":{"role":"USER","parts":[
  {"fileData":{"fileUri":"<audio-A>","mimeType":"audio/wav"}},
  {"fileData":{"fileUri":"<audio-B>","mimeType":"audio/wav"}},
  {"text":"<frozen-pairwise-prompt>"}]}}
```

The qualification protocol matches the intended pairwise evidence task. Four speakers contribute one strong pitch and one strong pace pair. Each of the eight pairs is judged in forward and reversed order, producing 16 strong-direction calls and eight position-consistency checks. Four forward calls are repeated once for stochastic stability. Four additional base-versus-invalid pairs test silence and severe clipping as diagnostics. Each successful response must include transcriptions for both candidates as an audio-access check, plus a strict JSON winner, confidence, evidence, and quality warning.

Admission requires at least 95% audio access, strict JSON parsing, and returned-model identity, and at least 75% strong-direction accuracy, pitch accuracy, pace accuracy, position consistency, and repeat stability. Invalid-audio rejection is not an admission gate because deterministic DSP retains technical authority. All 24 calls complete and pass every admission gate. Strong-direction accuracy is 87.5% (14/16): pitch reaches 75% (6/8) and pace reaches 100% (8/8). Position consistency is 75% (6/8), repeat stability is 100% (4/4), and invalid-audio detection is 25% (1/4). The final result is `admitted=true`.

On 45 single-attribute pairs, direct and dynamic-rubric prompt-direction agreement are 95.6% (43/45) and 93.3% (42/45). Active routing selects 26 pairs; 19 survive the forward/reverse consistency check and 16 of those 19 follow the prompt-defined direction. Three-call repeatability is unanimous for 8/9 pairs. On 26 multi-constraint pairs, direct and dynamic-rubric agreement are 50.0% (13/26) and 65.4% (17/26); the VCG protocol retains 13 decisions, with 53.8% (7/13) prompt agreement and 50.0% position consistency. These values qualify the model for use as a second evidence source, but remain operational protocol diagnostics; agreement with blinded listeners is analyzed separately on the completed human studies.

## E SCALED VOICEDESIGN PROTOCOL

The scaled protocol crosses three Mandarin texts, three controlled attributes (pitch, pace, and energy), two directions, and five stochastic replicates, producing 90 samples and 45 pairs. All samples pass the frozen technical gates. Direct and dynamic-rubric judgments are single-order baselines. The active policy selects pairs when evidence is weak, tied, non-primary, or below the confidence threshold, and then evaluates both presentation orders. Consequently, 26 selected pairs require 52 individual judgments. The paper reports both selected-pair fraction and actual judgment count.

The active-cost table uses relative units rather than prices: one DSP evaluation costs 1 unit, one LALM judgment costs 30 units, and one human pair costs 300 units. These assumptions are intentionally exposed in the artifact and can be changed without modifying decisions. The full-stack comparison assumes two LALM orders and human review for every pair; the fixed baseline uses one LALM order and no human review. Table 11 instantiates the accounting on the 45-pair study.

Table 9: DSP outcomes on the 45-pair generated-audio study. Below denotes an effect below the frozen magnitude threshold.

| Axis | Pairs | Direction hit | Reverse | Below / pooled effect |
|---|---|---|---|---|
| Pitch | 15 | 10 | 1 | 4 / $+1.965$ st |
| Pace | 15 | 11 | 0 | 4 / $+16.92\%$ |
| Energy RMS | 15 | 2 | 8 | 5 / $-1.543$ dB |

Table 10: Operational Gemini 3.5 Flash diagnostics on the 45-pair study.

| Judge | Pairs | Access | Prompt agreement |
|---|---|---|---|
| Direct | 45 | 100% | 95.6% (43/45) |
| Dynamic rubric | 45 | 100% | 93.3% (42/45) |
| Active + position swap | 26 | 100% | 84.2% (16/19) |

The table makes the trade-off explicit. Under these unit costs the active policy is *more* expensive than the single-order fixed baseline, because position balancing doubles the cost of every routed pair (52 judgments on 26 pairs versus 45 single-order judgments). The policy therefore buys auditability and refusal, not savings, relative to the cheapest automatic protocol; its cost advantage materializes only against protocols that include human review or universal dual-order judging. The conclusion is sensitive to the LALM:DSP cost ratio: as LALM calls become cheaper, the routing overhead grows relative to simply judging every pair in one order, while a higher human cost strengthens the case for selective escalation.

## F Multi-Constraint Protocol

The structural benchmark contains 30 briefs derived from ten templates and three variants. Each case explicitly lists target attributes, preservation constraints, unconstrained attributes, exemptions, and conflicts. Four cases contain contradictions and are planning-only. The remaining 26 briefs are synthesized with one neutral baseline and one brief-conditioned sample each. Candidate B is therefore the intended generated direction, not perceptual ground truth.

The direct judge answers which sample better fits the brief. The dynamic judge first derives three audible criteria. The VCG judge receives the explicit graph, is instructed not to penalize unconstrained or exempted attributes, and is run in both presentation orders. A canonical decision is retained only if the forward decision equals the reversed-order decision after relabeling. This procedure produces 9 stable decisions from 26 pairs, demonstrating that the abstention mechanism is operational but also that the tested LALM remains position-sensitive.

## G Robustness and Measurement Validation

The attack suite applies $+8$ dB gain, loudness normalization, 48 kbps MP3 compression, 20 dB additive noise, synthetic reverberation, leading silence, trailing silence, and a $+3$-semitone pitch shift to eight source utterances. The F0 comparison uses Parselmouth/Praat on 186 generated and corpus conditions. Metamorphic tests evaluate monotonicity, reversibility, pitch–pace commutativity, and path independence on four sources. These tests are diagnostics for evaluator mechanics; they do not measure naturalness or preference.

**Results.** On 64 signal attacks, pitch decisions remain correct under all attacks except one reverb slice (87.5%). Effective pace is robust to leading/trailing silence, MP3 compression, and pitch shift, but falls to 50% under loudness normalization, 25% under $+8$ dB gain, and 12.5% under 20 dB noise because the fixed-energy VAD changes its speech mask. Against Praat over 186 conditions, the lightweight F0 estimator has 0.35-semitone median absolute error, 1.06-semitone P90 error, and 94.4% direction agreement over 18 pairs (17/18). Metamorphic tests attain 100% pitch and pace monotonicity, reversibility, and pitch–pace commutativity, but only 75% path independence: applying gain before a tempo change alters the voiced mask relative to the reverse order, so loudness-affecting operations do not commute exactly with the pace measurement. This is a property of the measurement

Table 11: Relative-cost accounting on the 45-pair single-attribute study (1 DSP = 1 unit, 1 LALM judgment = 30 units, 1 human pair = 300 units).

| Protocol | DSP calls | LALM calls | Human pairs | Total |
|---|---|---|---|---|
| Fixed baseline (single order) | 45 | 45 | 0 | 1,395 |
| CARE-VOICE active (both orders, 26 pairs) | 45 | 52 | 0 | 1,605 |
| Full-stack (both orders + human) | 45 | 90 | 45 | 16,245 |

Table 12: Primary operational judge on 26 multi-constraint baseline/designed pairs. Candidate B is a prompt direction, not human truth.

| Judge | Access | Prompt agreement | Position consistency |
|---|---|---|---|
| Direct | 100% | 46.2% (12/26) | – |
| Dynamic rubric | 100% | 73.1% (19/26) | – |
| VCG + position swap | 100% | 66.7% (6/9 decided) | 34.6% (9/26) |

pipeline rather than nondeterminism in the frozen scorer, and it is why the evidence record serializes the operation order alongside each observation.

## H UNADMITTED JUDGE ATTEMPTS

Three further models—Gemini-3.1-Flash-Lite, Qwen3-Omni, and MOSS-Audio (OpenMOSS Team, 2026)—were attempted under the same frozen qualification protocol and were not admitted into the decision chain. Their failures occurred before the accuracy gates, at the transport, audio-access-demonstration, or strict-parsing stages: the candidate could not ingest two complete audio files in one request through the endpoint available to us, or did not return the required strict JSON decision together with transcriptions of both candidates. Because admission requires passing the access and parsing gates first, no per-gate accuracy, position-consistency, or stability scores exist for these models, and their raw responses are retained only as historical failure records. Table 13 summarizes the outcome at the level the audit records support.

Table 13: Judge models attempted but not admitted. All three failed before the accuracy gates, at the transport, audio-access-demonstration, or strict-parsing stages; the retained records do not support finer per-model attribution, and no accuracy numbers exist.

| Model | Outcome |
|---|---|
| Gemini-3.1-Flash-Lite | not admitted |
| Qwen3-Omni | not admitted |
| MOSS-Audio | not admitted |

We report these attempts for two reasons. First, they make the scope of our instability evidence explicit: the quantitative position-consistency and repeat-stability results in this paper come from two deployments of one model family, complemented by admission-stage failures of three further models rather than by a broad cross-family comparison. Second, the failures are themselves evidence for the transport, access, and version dependencies discussed in the main text: an evaluator that silently substitutes a different model or endpoint when the intended judge is unavailable would convert exactly these failures into hidden evidence shifts, which is what the evidence-health record is designed to prevent.

## I HUMAN REVIEW PROTOCOL

The repository generates a local blinded pairwise review page for every protocol. A participant identifier deterministically randomizes both item order and A/B order, while copied audio assets use opaque filenames and the condition mapping remains in a coordinator-only file. Reviewers must play both samples before rating energy, pitch, pace, whole-brief fit, target-constraint fit, or

preservation fit as configured. They may choose tie or uncertain, score naturalness for both samples, report confidence and speaker similarity where applicable, and provide short audible evidence. The page stores progress locally and exports JSON without remote transmission. Each study contains two hidden repeat trials and one identical-audio attention trial. Preregistered exclusions cover missing consent, non-headphone playback, potentially interfering hearing difficulty, incomplete ratings, insufficient playback, failed attention checks, and repeat consistency below 50%. The aggregator canonicalizes A/B order, records all exclusions, and computes strict-majority labels, Fleiss' $\kappa$, nominal Krippendorff's $\alpha$, and per-condition naturalness. We target at least eight qualified listeners per question.

The completed collection contains ten submissions for each study. Nine listeners qualify for the semantic study, all ten qualify for the single-attribute study, and six qualify for the multi-constraint study. The latter does not meet the target of eight qualified listeners per question and is therefore exploratory. Participant identifiers and answer-level files are not included in the submission package; the released artifact contains de-identified pair-level votes, quality-control counts, and aggregate reliability statistics.

### I.1 Single-attribute vote distributions

Table 14 summarizes the requested dimension. Pace is the only consistently resolved control axis. Pitch has seven unresolved pairs, and energy has four unresolved and four majority reversals. The strict local-edit criterion additionally requires the two non-target dimensions to remain equivalent; no pair satisfies this criterion.

Table 14: Strict-majority outcomes by requested attribute, out of 15 pairs per attribute. Vote support is the fraction of qualified listener votes on the requested dimension that favor the designed direction.

| Attribute | Designed majority | Reverse majority | Unresolved | Vote support |
|---|---|---|---|---|
| Pitch | 6 | 2 | 7 | 55.3% |
| Pace | 14 | 0 | 1 | 87.3% |
| Energy | 7 | 4 | 4 | 54.0% |

### I.2 Consensus sensitivity and ambiguity flags

The primary analysis uses every strict-majority pair. As a retrospective sensitivity check, we additionally require the majority to contain at least 70% or 80% of listener votes. Results are shown in Table 15. The higher-consensus subsets are easier for both direct and selective judges; they must not be interpreted as newly preregistered test sets.

Table 15: Retrospective sensitivity to listener consensus. Entries are correct/decided.

| Evaluator | All resolved | $\geq 70\%$ | $\geq 80\%$ |
|---|---|---|---|
| Gemini direct | 27/33 | 25/28 | 19/21 |
| Gemini dynamic | 26/33 | 24/28 | 19/21 |
| CARE-Voice | 21/25 | 20/22 | 16/17 |

We also treat the 12 pairs without a strict majority as an ambiguity-detection target. CARE-Voice abstention flags 14 pairs and captures six unresolved cases. Direct confidence below 1.0 flags 29 and captures nine. Position inconsistency flags seven and captures four; an evidence-audit warning flags nine and captures four. These analyses use only existing model outputs and existing votes.

### I.3 Protocol deviations discovered after collection

First, successful preservation usually means that two candidates are perceptually equivalent on the preserved property. The original aggregator instead assigned the designed/B direction as the expected answer for `preservation_fit`. We retain the collected choices but do not use preservation "success" in any main claim. Second, no listener selected `uncertain` in any of the valid dimension

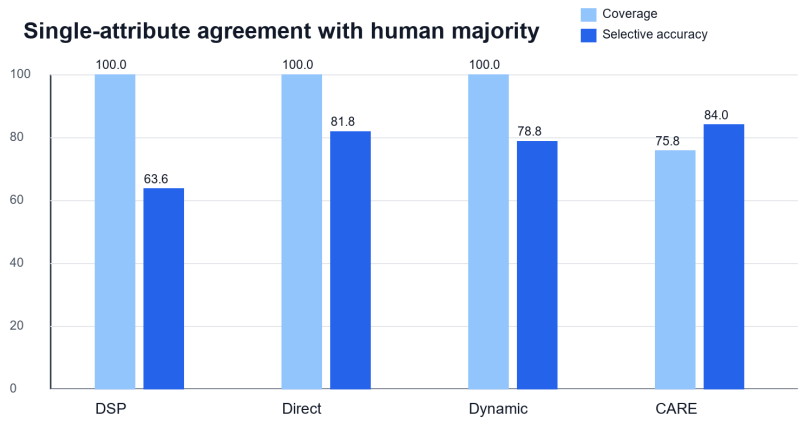



Figure 4: Coverage and selective agreement on the 33 resolved single-attribute pairs.

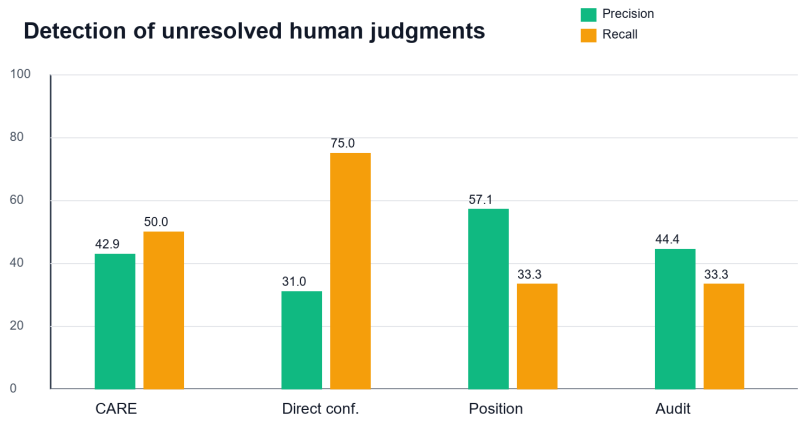



Figure 5: Retrospective detection of unresolved single-attribute judgments.

ratings. Because the interface requested additional free text after selecting uncertain, this option likely underestimates subjective uncertainty; strict-majority resolution and vote entropy are used instead. Third, the low speaker-similarity ratings make identity drift an explicit confound rather than a hidden generation detail.

## J AUDIT SCHEMA

Every condition stores its source or generation request, local audio path, tool version when available, and technical metrics. Every pair stores the expected relation, evidence observations, evidence roles, thresholds, tool-health state, final decision, confidence, and reason code. External model failures preserve the raw response in a separate audit directory. The current reason codes include `technical_gate_failed`, `independent_direct_proxies_agree`, `independent_direct_proxies_reverse`, and `direct_proxies_weak_or_conflicting`.

## K ARTIFACT AVAILABILITY

The standalone benchmark package can be inspected and scored without installing the method implementation:

```
cd benchmark
python3 run_benchmark.py list
python3 run_benchmark.py validate \
  --submission baselines/care_voice_reference.json
python3 run_benchmark.py evaluate \
  --submission baselines/care_voice_reference.json \
  --tracks all --output outputs/care_voice_reference_report.json
```

Every experiment in this paper is driven by a versioned, machine-readable protocol file and regenerates from the repository: counterfactual generation, semantic analysis, pairwise judging and fusion, second-LALM qualification, scaled and multi-constraint benchmarks, human-vote aggregation, figure regeneration (`analysis/build_artifacts.py`), and manuscript compilation. The complete command reference ships with the repository README and the method-code snapshot rather than in this appendix.

The public counterfactual source is Mini LibriSpeech under CC BY 4.0. The semantic source is EmoV-DB under CC BY-NC 4.0. The submission artifact packages the evaluated audio as local WAV files and uses paths relative to the artifact root; no object-storage URL is required. Raw service responses, credentials, signed object-storage URLs, and participant-level listening files remain local. Gemini-3-Flash-Preview, the incumbent operational judge, completed the service's operational protocols including the multi-constraint operational judging reported in the main text, and Gemini 3.5 Flash passed the frozen second-model qualification and completed the formal 45-pair and 26-pair evaluations. Gemini-3.1-Flash-Lite, Qwen3-Omni, and MOSS-Audio attempts remain historical failure records. Reference-conditioned editing remains excluded because no authorized target profile or reusable voice identity was available. The submission bundle contains local audio stimuli, de-identified aggregate human results, compact pair-level predictions, machine-readable protocols and summaries, analysis code, figures, a method-code snapshot, and the exact LaTeX source.