ONE REAL ENDPOINT IS ENOUGH: REAL-SPEECH ANCHORING FOR DATA-EFFICIENT S2S POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

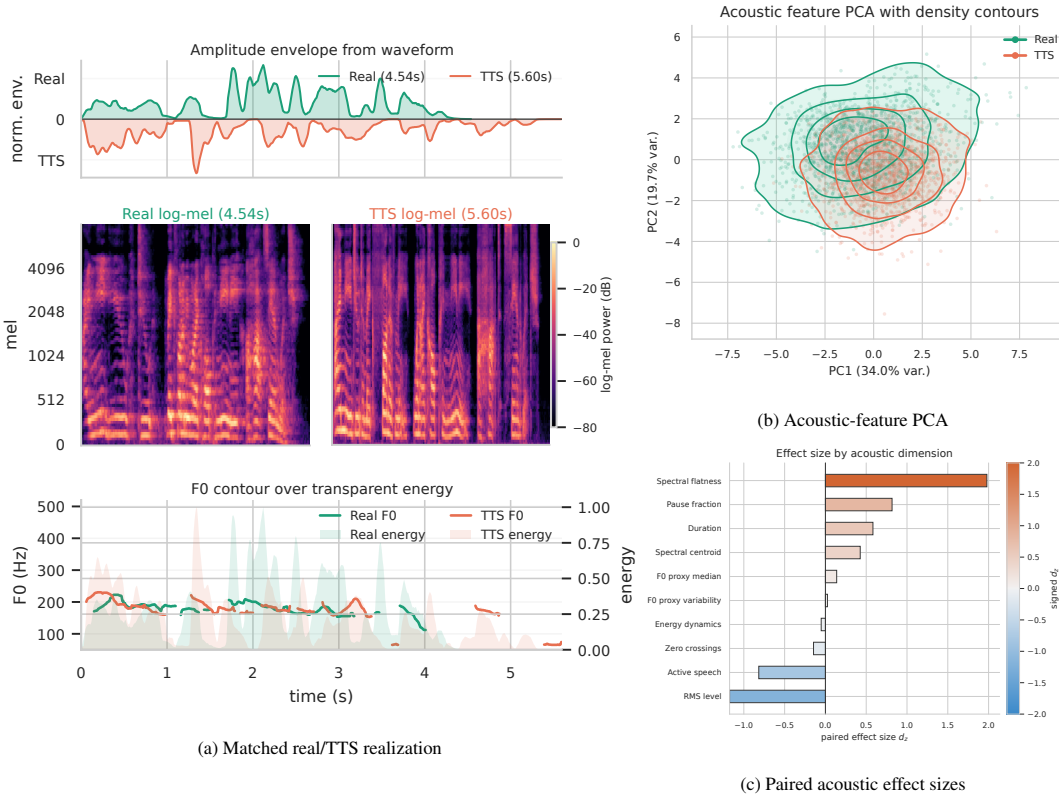
Speech-to-speech (S2S) assistants must adapt to how a request is spoken and reply with appropriate prosody, yet their training supervision rarely carries this information: human-recorded dialogue is scarce and misaligned with assistant behavior, while scalable synthetic dialogue bounds every speech realization by the synthesizer. We call this shortfall the Non-Semantic Information Gap and trace it to the unit both paradigms share, the complete dialogue. At the level of a single exchange, one verified real recording suffices when kept in the role whose realization is the supervision signal, with only its counterpart synthesized. **RAST** (Real-speech Anchored S2S Training) applies this principle in two directions: user-side anchoring (RAST-U) pairs real input cues with condition-appropriate responses, and assistant-side anchoring (RAST-G) places recorded prosody in the generation targets. We instantiate RAST as 284,660 exchanges spanning 2,914 hours, and introduce **RASTBench**, a transcript-controlled benchmark that scores response audio directly and separates completing the request from adapting to the speech condition. Against Pure-TTS controls that differ only in the provenance of the anchored endpoint, retaining real speech improves RASTBench by 4.72, 4.52, and 8.81 points for U, G, and their mixture, all statistically significant, with gains concentrated on condition-adaptation criteria. The two directions yield complementary capability profiles, their mixture achieves the strongest response adaptation, and a quarter of the corpus recovers most of the full-data gains.

1 INTRODUCTION

Speech-to-speech (S2S) interaction requires two capabilities beyond producing a correct textual answer: *speech-conditioned response adaptation*, which adapts the response strategy to non-semantic cues in the user's speech, and *expressive speech generation* (Schuller & Batliner, 2013; Ao et al., 2024; Lin et al., 2025). A user saying "I'm fine." cheerfully may invite a brief acknowledgment, whereas the same words spoken through sobs may warrant a supportive check-in delivered with suitable pacing and intonation.

Both capabilities are acquired from post-training supervision, and neither can be learned from transcripts alone: a model must observe non-semantic variation in its inputs to condition on it, and in its targets to generate it. Existing construction paradigms fall short in complementary ways. Fully human-recorded dialogue preserves speech realization at both endpoints but is scarce, costly, and rarely aligned with the behavior expected of an assistant (Guo et al., 2021; Clark et al., 2019). Fully synthetic dialogue is scalable and controllable (Fang et al., 2025), but its speech realizations are bounded by the synthesis system: covering an emotion or speaking-style category does not establish coverage of its acoustic realizations. Emotional TTS treats cross-category control and within-category prosodic variation as separate modeling objectives (Luo et al., 2021), and emotion recognizers fine-tuned on synthesized speech perform well on synthetic test utterances yet generalize poorly to human recordings, even when the synthetic speech is generated from human transcripts and checked for perceptible emotional expression (Ge et al., 2026). Because scalable S2S post-training has relied predominantly on the synthetic route (Fang et al., 2025; Xie & Wu, 2024), S2S supervision inherits this mismatch. We refer to the resulting shortfall as the **Non-Semantic Information Gap**, characterized in Section 2.

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.



(a) Matched real/TTS realization

(b) Acoustic-feature PCA

(c) Paired acoustic effect sizes

Figure 1: Transcript-matched speech realizations. A paired example, acoustic-feature distributions, and paired effect sizes show variation beyond lexical content. The diagnostic motivates retaining speech information during supervision; paired training comparisons evaluate its downstream value.

Our approach starts from the observation that this dilemma is tied to the paradigms' shared unit of construction: a complete dialogue. If supervision is instead constructed per exchange, one verified real recording suffices. The recording is kept in the role where its speech realization is consumed, as the user input whose cues should shape the response or as the assistant target whose prosody should be reproduced, and only the missing conversational partner is synthesized. This places real non-semantic information exactly where it supervises learning while inheriting the scalability and content control of synthesis for the other side. We instantiate this principle as **RAST** (Real-speech Anchored S2S Training): user-side anchoring (RAST-U) supervises speech-conditioned response adaptation, assistant-side anchoring (RAST-G) supervises expressive speech generation, and their union (RAST-Mix) provides both. Section 2 derives this design from the supervision requirements, and Section 3 realizes it as a corpus of 284,660 exchanges totaling 2,914.0 hours.

Testing whether the real anchor matters requires distinguishing speech-conditioned behavior from simply answering the transcript. We therefore introduce **RASTBench**, a transcript-controlled benchmark that directly evaluates generated response audio and whose rubrics distinguish completing the literal request from meeting condition-specific requirements on response strategy or spoken realization (Section 4), complemented by WavBench for speech understanding and reply relevance and DeEAR for generated-speech expressiveness (Li et al., 2026; Lin et al., 2025). Fine-tuning Kimi-Audio-7B-Instruct, we compare each quarter-scale RAST configuration with a Pure-TTS control that replaces only the real-speech-origin endpoint with cloned audio while holding dialogue texts, synthesized counterparts, and training horizon fixed, isolating the contribution of retained real speech from that of constructed dialogue content (Section 5). RAST-U, RAST-G, and RAST-Mix each improve RASTBench over their controls with statistically significant paired differences, and the gains extend to condition-specific criteria rather than literal task completion alone. The two directions yield distinct capability profiles at every data fraction: RAST-U leads on RASTBench, whereas RAST-G leads on DeEAR expressiveness.

Our contributions: (i) **RAST**, an S2S post-training strategy and corpus that use one verified recording per exchange to provide user-side or assistant-side speech supervision; (ii) **RASTBench**, a transcript-controlled benchmark whose rubrics distinguish task completion from condition-specific adaptation in response strategy or spoken realization; and (iii) paired evidence for the value of retaining real speech, with analyses of how anchoring direction and data scale affect.

2 Data Construction as Supervision Design

Data construction determines which speech information reaches the model and in which role it supervises assistant behavior. Figure 1 makes the gap concrete: real utterances and TTS re-syntheses of the same transcripts differ in timing, pitch, and energy, both in the paired example and across matched utterances in the acoustic-feature distributions and effect sizes, so matching transcripts does not ensure equivalent speech realizations. We define the **Non-Semantic Information Gap (NSIG)** as a shortfall in the diversity and accuracy of task-relevant non-semantic information in S2S training supervision relative to the requirements of real interaction. On the input side, supervision must connect non-semantic cues to condition-appropriate responses; on the output side, it must include speech realizations for expressive speech generation.

Table 1 compares five construction configurations against these two requirements together with corpus expansion and control over assistant behavior. The symmetric configurations trade one against the other: fully recorded dialogue satisfies both realization requirements but not scalability, and fully synthetic dialogue satisfies scalability and control but neither realization requirement. The two asymmetric configurations show that the requirements are separable: a real user input paired with a synthesized response satisfies the input-side requirement, and a synthesized query paired with a real response satisfies the output-side requirement, each at the cost of only one synthesized endpoint.

Table 1: S2S data-construction configurations and their supervision trade-offs. R: real-speech-origin audio (including processed recordings); S: synthesized speech. RAST-Mix pairs (R, S) or (S, R) across different exchanges.

Category	User input	Assistant output	Advantages and limitations
Fully recorded	R	R	Advantage: Preserves human non-semantic information at both endpoints (Guo et al., 2021; Lee et al., 2023). Limitation: Expensive to scale; original responses may not match assistant goals (Clark et al., 2019).
Fully synthetic	S	S	Advantage: Scalable construction with controllable task coverage and response strategies (Fang et al., 2025). Limitation: Expressiveness of both endpoints depends on the synthesis system (Li et al., 2023).
User-side (RAST-U)	R	S	Advantage: Pairs real input cues with constructed responses for speech-conditioned response adaptation (Ao et al., 2024). Limitation: Requires reliable speech-condition annotations; output expressiveness remains TTS-dependent.
Assistant-side (RAST-G)	S	R	Advantage: Supplies real-speech-origin targets for expressive speech generation (Lin et al., 2025; KimiTeam et al., 2025). Limitation: Fixed response content constrains task coverage.
Mix (RAST-Mix)	R / S	S / R	Advantage: Covers both supervision roles across exchanges with one real anchor per exchange, retaining scalability.

The comparison motivates using a standalone recording, rather than a recorded dialogue pair, as the unit of supervision construction: RAST preserves the recording in the role where its speech realization is used and synthesizes a compatible partner. We hypothesize that retaining these realizations provides supervision beyond transcript-matched synthesis, with RAST-U and RAST-G supporting speech-conditioned response adaptation and expressive speech generation respectively, and RAST-Mix covering both roles; Section 5 tests this hypothesis.

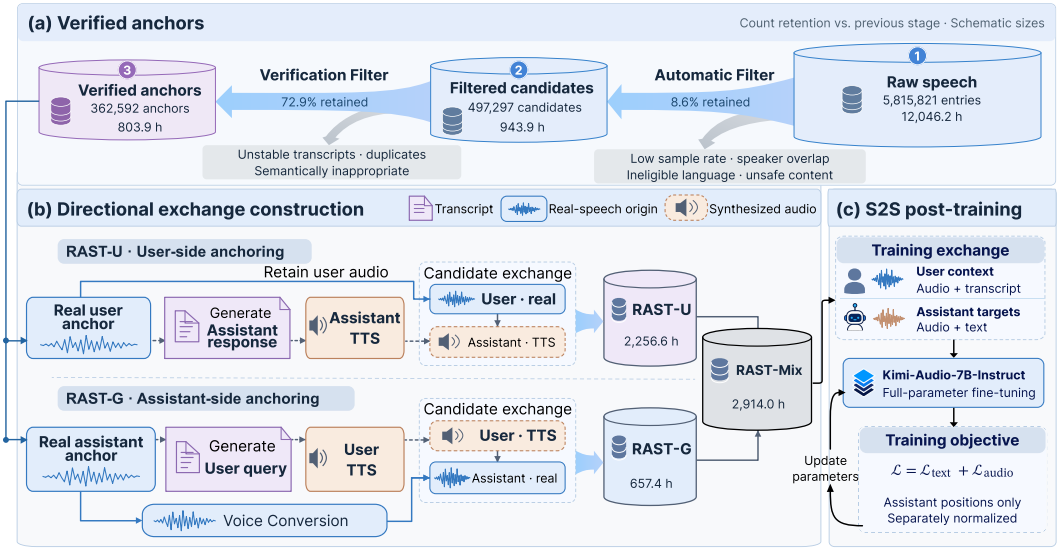


Figure 2: **Overview of RAST. (a)** Raw speech passes automatic and transcript-verification filters to yield verified anchors. **(b)** Each anchor is completed into a two-turn exchange: RAST-U retains the real user turn and synthesizes the assistant response; RAST-G keeps the real assistant turn as target and synthesizes a preceding query. Both pass exchange-level quality control; their union forms RAST-Mix. **(c)** Kimi-Audio-7B-Instruct is fine-tuned with a joint text and audio loss at assistant positions.

3 Real-Speech Anchored S2S Training

RAST instantiates the design of Section 2. Figure 2 gives an overview: raw speech is filtered into verified anchors (a), each anchor is completed into a two-turn exchange in one of two directions (b), and the resulting corpus is used for full-parameter post-training (c).

3.1 User-side and assistant-side anchoring

User-side anchoring. For RAST-U, a verified real utterance is retained as user audio x_u. Its transcript and supported speech-condition annotations guide generation of the assistant transcript t_a, synthesized in a consistent assistant voice to obtain x_a. The response must address the literal request and reflect the speech conditions when relevant, without exposing annotation labels.

Assistant-side anchoring. For RAST-G, a verified real utterance is the source of the assistant response. Its transcript t_a and supported annotations guide generation of a plausible preceding query t_u, which is synthesized to obtain x_u. The query must make the fixed response conversationally appropriate without exposing annotation metadata. The response audio is denoised and voice-converted to the Kimi-Audio assistant timbre, and the resulting x_a serves as the target of the audio loss.

Mixed anchoring. RAST-Mix pools exchanges from both directions without altering them, so it covers both supervision roles while keeping one real-speech-origin endpoint per exchange.

3.2 Verification and scalable instantiation

The pipeline asks a single question of every recording: can this utterance, on its own, anchor one side of a conversation whose other side we then construct? Three progressively stricter stages answer it before any counterpart is generated, and the completed pair is checked afterwards.

From recordings to anchors. Raw speech is collected from YouTube, spanning 12,046.2 hours of in-the-wild recordings. Recordings are standardized, vocals isolated, and speaker diarization with

voice activity detection yields single-speaker segments of 3–30 s, screened for language and audio quality. Transcripts are verified by cross-checking two ASR systems; only verified segments proceed to anchor selection, where a two-stage text filter identifies self-contained questions or answers and rejects utterances that depend on unseen images or missing prior dialogue, are truncated, or are unsafe. For the retained anchors, audio-only and audio-video models annotate affect, speaking style, prosody, voice quality, speaker presentation, and ambient context, with specialized classifiers providing cross-checks. Labels require supporting evidence; unreliable labels are cleared while otherwise usable anchors are retained. Four research-team experts review the annotation stack on expert-labeled examples before it is frozen. The funnel yields 497,297 filtered candidate segments and 362,592 verified anchors (Appendix A.1).

From anchors to exchanges. The missing side is then constructed, conditioned on the anchor's transcript and surviving annotations. Qwen3-235B-A22B-Instruct-2507 (Qwen Team, 2025; Yang et al., 2025) generates the counterpart text with separate delivery instructions, so that what is said and how it is spoken are specified independently. Step-Audio-EditX (Yan et al., 2025) synthesizes the counterpart, using a canonical assistant voice for responses and a balanced speaker pool for user queries; Seed-VC (Liu, 2024) converts assistant-side anchors to the assistant timbre. The completed pair is inspected as a dialogue: the counterpart must answer or license the anchored transcript, avoid factual errors and explicit mention of hidden labels, and pass audio checks for synthesis artifacts and, for converted anchors, preserved intelligibility and expressive structure (Appendix A.2). The pipeline retains 284,660 exchanges from 362,592 verified anchors, a yield of 78.5% (Table 2), totaling 2,914.0 hours: 774.8 real-speech-origin and 2,139.2 synthesized. The two pools are kept at their post-filtering sizes without rebalancing. Raw anchors are not redistributed, and source identifiers allow deletion requests to propagate to all derived assets.

Table 2: Final statistics of RAST training data.

Direction	Exchanges	User (h)	Assistant (h)
User-side anchoring	203,118	373.2	1,883.4
Assistant-side anchoring	81,542	255.8	401.6
Total	**284,660**	**629.0**	**2,285.0**

3.3 Post-training objective

We represent an exchange as $d = (x_u, t_u, x_a, t_a)$ with user and assistant audio x_u, x_a and transcripts t_u, t_a. Given the corpus $\mathcal{D}_{\mathrm{RAST}}$, we fine-tune all parameters θ with the sum of assistant-side text and audio next-token losses,

$$\mathcal{L}(\theta; \mathcal{D}_{\mathrm{RAST}}) = \mathcal{L}_{\mathrm{text}}(\theta; \mathcal{D}_{\mathrm{RAST}}) + \mathcal{L}_{\mathrm{audio}}(\theta; \mathcal{D}_{\mathrm{RAST}}), \tag{1}$$

normalized separately and evaluated only at assistant positions. User tokens provide context without contributing to the loss, so in RAST-U the real anchor acts purely as conditioning, whereas in RAST-G it is the target of $\mathcal{L}_{\mathrm{audio}}$ (Appendix C.1).

4 RASTBench: Transcript-Controlled Evaluation of Non-Semantic Grounding

RASTBench evaluates task completion and condition-appropriate responses under transcript-controlled speech variation, targeting *non-semantic grounding*: using acoustic and paralinguistic evidence beyond the transcript to guide response strategy and spoken realization. Existing spoken-dialogue benchmarks either score response transcripts rather than audio (Ao et al., 2024; Cheng et al., 2025) or evaluate audio without same-transcript contrasts and without separating task completion from adaptation (Li et al., 2026; Jiang et al., 2026; Yang et al., 2026). RASTBench combines controlled synthetic conditions with human speech and directly evaluates the generated audio.

Contrast-group design. The unit of RASTBench is a *contrast group*: several audio realizations of one transcript differing only in a designated speech condition. In the TTS Bench, a *baseline* sample carries the neutral condition and *target* samples the contrasting conditions, with all non-target metadata fixed. In the Human Bench, each group pairs an authentic recording with a voice-cloned

TTS control synthesized from its verified transcript, sharing the condition description, intended response policy, and rubric. Human Bench scores are computed on the authentic side; the cloned side supports a paired real-versus-TTS diagnostic (Appendix D.3).

Benchmark construction and coverage. The TTS Bench covers emotion, tone, age, gender, and ambient sound. Requests are constructed so that the transcript does not reveal the target condition, yet the condition can legitimately change the appropriate response through a task-relevant mechanism; MiMo-V2.5-TTS (Xiaomi MiMo, 2026) renders the same transcript under each condition, and groups are rejected for weak response-policy contrast, lexical leakage of the condition, or changes to non-target metadata. The Human Bench covers emotion and tone with authentic utterances from a source pool disjoint from the RAST training anchors; cloned controls use the same TTS system. Table 3 reports the inventory and Appendix B the construction protocol.

Scoring. Each group carries a rubric with two kinds of binary criteria. *Baseline* criteria state what any response to the transcript must do; *displacement* criteria state how response strategy, framing, or spoken realization should shift under the target condition. Criteria describe behavioral outcomes rather than reference wording, and merely naming the condition earns no displacement credit. The audio-capable judge `gemini-3.5-flash` scores each response from its audio alone, given the sample role, condition description, and applicable criteria; it receives neither the user audio, a reference answer, nor a response transcript. For item i with applicable criteria $\mathcal{C}_i$, the judge returns $r_{i,c} \in \{0, 1\}$, and the benchmark score is

$$S_{\text{RASTBench}} = 100 \sum_{d \in \mathcal{D}} \frac{w_d}{|\mathcal{I}_d|} \sum_{i \in \mathcal{I}_d} \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} r_{i,c}, \tag{2}$$

where $\mathcal{I}_d$ is the item set of dimension d and w_d is $1/10$ for the five TTS dimensions and $1/4$ for the two Human dimensions, so the two sections contribute equally. We also report B-Macro and D-Macro, the baseline and displacement pass rates macro-averaged over the TTS dimensions. On 455 expert-labeled criteria from 131 responses, the judge reaches 88.13% criterion-level agreement and $\rho = 0.761$ with expert majority labels; the benchmark is constructed with a different model family from the judge (Appendix B.3).

Table 3: Statistics of RASTBench dataset.

Section / dim.	Groups	Avg Dur. (s)
TTS Bench	**2,028**	**10.98**
Emotion	512	8.26
Tone	372	12.82
Age	322	16.86
Gender	310	17.17
Ambient	512	6.60
Human Bench	**1,270**	**7.15**
Emotion	917	7.51
Tone	353	6.20
Total	**3,298**	**9.75**

5 EXPERIMENTS

We ask three questions: whether retaining real speech improves post-training beyond the contribution of the constructed dialogue content (Section 5.2), how the two anchoring directions shape the resulting capabilities (Section 5.3), and how the benefits vary with data scale (Section 5.4). Table 4 summarizes absolute scores, and Figure 3 presents the paired TTS-replacement effects.

5.1 EXPERIMENTAL SETUP

We fine-tune Kimi-Audio-7B-Instruct (KimiTeam et al., 2025) with U, G, or Mix supervision at four nested data fractions (25%, 50%, 75%, and 100%) on eight NVIDIA H20 GPUs, using AdamW with learning rate 5×10^{-6} and global batch size 32. Each quarter-scale configuration has a Pure-TTS control that replaces only its real-speech-origin endpoint, keeping the dialogue texts, synthesized counterpart, and training horizon fixed within the pair (Appendix C). Paired comparisons between a RAST configuration and its control are therefore controlled, differing only in the provenance of one endpoint; comparisons across directions or data fractions are descriptive, since U, G, and Mix differ in corpus size, response content, and optimization steps. We write U, G, and Mix for RAST-U, RAST-G, and RAST-Mix, and U-TTS, G-TTS, and Mix-TTS for their controls.

Table 4: **Post-training results.** (A) RASTBench condition scores; Human Bench uses authentic recordings. (B) TTS Bench baseline/displacement macros (B/D), WavBench understanding and reply relevance (Rel.) with their average, and DeEAR Expressive. Quarter-scale and full-data use 25% and 100% of the corpus. Scores are 0–100 (higher is better); bold marks column maxima. Paired differences with 95% CIs are in Figure 3 and Appendix C.3.

A. RASTBench	TTS Bench					Human Bench			
Method	Emo	Tone	Age	Gender	Ambient	Emo	Tone	Avg	Overall
Base	70.48	76.00	71.75	60.45	64.01	27.14	49.56	38.35	53.45
Quarter-scale									
U-TTS	77.98	74.31	75.63	64.02	63.92	71.13	68.47	69.80	70.48
RAST-U	80.71	82.02	80.89	71.70	70.11	73.75	72.87	73.31	75.20
G-TTS	66.93	71.14	71.31	57.41	56.33	59.87	63.19	61.53	63.08
RAST-G	67.43	80.86	77.63	66.83	59.83	63.14	66.22	64.68	67.60
Mix-TTS	74.74	71.55	73.44	59.25	59.27	65.08	66.06	65.57	66.61
RAST-Mix	70.87	85.16	81.76	70.39	67.43	76.59	74.84	75.72	75.42
Full-data									
RAST-U	**81.63**	83.27	81.55	70.60	**71.88**	75.93	75.87	75.90	76.84
RAST-G	69.21	83.17	77.47	66.79	61.24	63.03	66.44	64.73	68.16
RAST-Mix	75.33	**88.16**	**82.88**	**73.42**	69.21	**79.61**	**77.05**	**78.33**	**78.06**

B. Diagnostics	TTS criteria		WavBench					DeEAR
Method	B-Macro	D-Macro	Und. Emo	Und. Gender	Und. Age	Reply Rel.	Avg	Expr.
Base	82.85	7.23	74.95	11.00	42.00	80.26	52.05	17.16
Quarter-scale								
U-TTS	82.99	23.02	82.10	10.00	44.00	89.87	56.49	27.67
RAST-U	89.49	27.89	**87.43**	27.50	60.33	**96.62**	67.97	26.24
G-TTS	75.79	18.32	78.67	55.00	53.33	79.48	66.62	22.17
RAST-G	82.13	23.73	77.14	58.00	57.33	76.36	67.21	29.08
Mix-TTS	78.86	20.34	79.20	25.00	35.33	89.09	57.16	19.19
RAST-Mix	87.13	28.01	85.52	44.50	**62.67**	91.69	**71.09**	26.52
Full-data								
RAST-U	**90.00**	28.55	85.52	20.50	54.33	95.32	63.92	27.19
RAST-G	83.22	25.10	76.00	**61.50**	57.33	78.96	68.45	**29.36**
RAST-Mix	89.62	**31.65**	84.57	33.50	54.67	94.29	66.76	26.94

RASTBench evaluates task completion and condition-appropriate responses, with baseline and displacement macros (B and D) separating the two criterion types. DeEAR Expressive assesses generated-speech expressiveness. On WavBench, we evaluate emotion, gender, and age understanding together with reply relevance, reporting component scores and their composite average. Differences are reported with 95% bootstrap intervals.

5.2 Effects of real-speech anchoring

Real-speech anchors improve condition-appropriate responses. The quarter-scale paired comparisons improve RASTBench by about five points for U and G and nine for Mix, with all three 95% difference intervals excluding zero (Figure 3). Because these comparisons retain the dialogue texts and synthesized counterparts, the gains are attributable to the retained speech realizations rather than additional dialogue content. Improvements extend across both TTS Bench and Human Bench, and D-Macro also improves for all three configurations, so the benefit reaches condition-specific response requirements rather than shared task completion alone.

Input-side anchoring improves recognition and reply relevance. RAST-U exceeds U-TTS on all three WavBench understanding tasks and on reply relevance, with the largest gains in gender and age understanding, so the composite improvement is not driven by a single task. Since U and U-TTS share the same synthesized assistant targets, only the user-side realization of the training pairs

changes; together with the RASTBench improvement, this supports preserving input-side speech information when teaching condition-appropriate responses.

Output-side anchoring supplies effective expressive targets. RAST-G makes the complementary change: user inputs and response transcripts are fixed, while the assistant audio targets retain source-based realizations. DeEAR improves by roughly seven points, with increases in arousal, prosody, and predicted naturalness (Appendix D), so stronger expression does not come at a cost in naturalness. These benefits are obtained with source recordings converted to the assistant timbre (Section 3), supporting their use as generation targets within the target voice interface.

5.3 Capability profiles

Anchoring directions favor different capabilities, and mixing rebalances them. Across all four data fractions, U scores higher on RASTBench, whereas G scores higher on DeEAR, consistent with their supervision roles: U pairs real input cues with condition-appropriate responses, while G places real-speech-origin realizations in generation targets. Mix achieves the highest full-data RASTBench score but not G's expressiveness. The pools differ in speech provenance and response form: U provides longer model-generated replies, whereas G supplies shorter real-speech-origin responses (Section 3.2); mixing thus yields a different capability profile rather than preserving each direction's strongest result, and does not isolate response length from provenance or corpus content. Within the mixed corpus, retaining real-speech-origin endpoints still improves both external benchmarks over Mix-TTS.

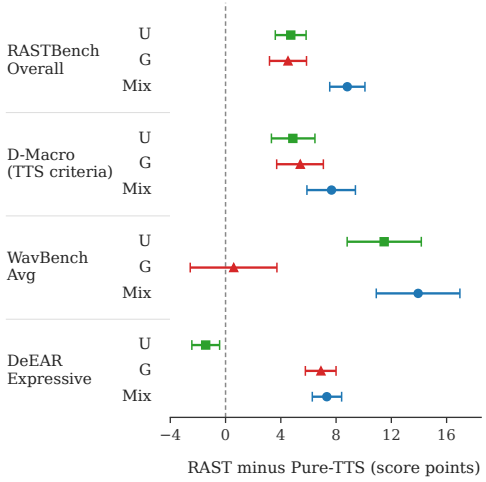


Figure 3: Quarter-scale RAST-minus-Pure-TTS differences on RASTBench, D-Macro, WavBench, and DeEAR; bars show 95% joint-bootstrap CIs.

5.4 Data scaling

Most RASTBench gains emerge by quarter scale. All evaluated RAST configurations improve over Base with positive paired intervals. Quarter-scale runs retain 89.3–96.2% of their full-data RASTBench gains, computed as $(S_{25} - S_{\text{Base}})/(S_{100} - S_{\text{Base}})$, so most of the improvement appears before using the full training pool. Full-data training still benefits U and Mix, whose quarter-scale runs remain below their full-data counterparts with nonzero paired differences, and the directional profiles of Section 5.3 persist at every fraction.

Scaling dialogue supervision yields task-dependent gains. From quarter scale to full data, Mix improves RASTBench, D-Macro, and reply relevance, while its WavBench composite declines mainly through lower gender and age understanding (Figure 4). RAST expands supervision for generating condition-appropriate replies rather than for attribute classification, so its understanding scores measure how dialogue supervision transfers to recognition, and that transfer does not scale monotonically with the targeted behavior. Because the training horizon is fixed in epochs within each family, larger fractions also receive more optimization steps, and the scale comparisons reflect both changes. Complete scores and paired scale comparisons are in Appendix D.2.

6 Related Work

Speech-native dialogue models. AudioPaLM, SpeechGPT, Moshi, and Kimi-Audio connect speech understanding and generation with language modeling through different acoustic representations and training designs (Rubenstein et al., 2023; Zhang et al., 2023; Défossez et al., 2024;

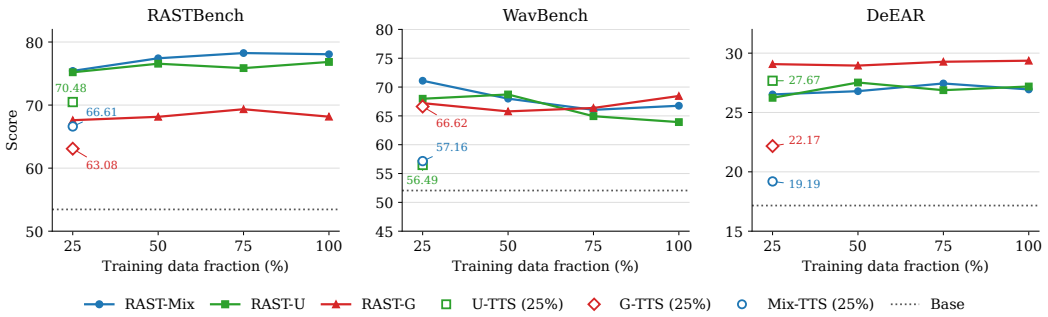


Figure 4: Capability profiles across nested training-data fractions. Lines connect the evaluated RAST configurations; dotted lines show Base. Open markers show the quarter-scale U-TTS, G-TTS, and Mix-TTS controls. WavBench summarizes speech understanding and reply relevance. Paired scale comparisons are reported in Appendix C.3.

KimiTeam et al., 2025). RAST focuses on the supervision provided by post-training exchanges within an existing S2S architecture.

Condition-aware response supervision. The training corpus accompanying SD-Eval pairs speech with response texts generated from transcripts and paralinguistic or environmental attributes (Ao et al., 2024); GOAT-SLM combines attribute-aware instruction tuning, speech–text alignment, and expressive generation training (Chen et al., 2025); and ParaS2SAlign uses supervised warm-up, reward-model distillation, and GRPO for paralinguistic-aware S2S interaction (Yang et al., 2026). These works teach models to respond to speech conditions rather than only recognize them; RAST-U constructs such supervision around verified real user speech.

Evaluation beyond transcripts. SD-Eval and VoxDialogue assess response appropriateness under varied paralinguistic and environmental conditions (Ao et al., 2024; Cheng et al., 2025). WavBench covers explicit acoustic understanding and generation tasks and implicit acoustic interaction (Li et al., 2026); S2S-Arena uses reference-free pairwise comparisons of generated speech, and ParaS2SBench evaluates the compatibility of input and output content and speaking styles (Jiang et al., 2026; Yang et al., 2026). RASTBench instead evaluates response audio within transcript-controlled groups and reports shared-task and condition-specific criterion satisfaction separately.

7 Conclusion

RAST constructs S2S supervision around one verified real-speech-origin endpoint per exchange, and RASTBench evaluates task completion and condition-appropriate responses under transcript-controlled speech variation. Paired TTS replacements show that the retained real speech improves response adaptation across user-side, assistant-side, and mixed training; user-side recipes favor response adaptation and assistant-side recipes favor expressiveness; and quarter-scale runs retain 89.3–96.2% of full-data RASTBench gains.

Limitations

RAST inherits the distributional structure of naturally occurring web speech rather than a predefined label distribution, resulting in uneven coverage of non-semantic conditions and limiting conclusions about rare or weakly represented speech phenomena. In addition, assistant-side anchors tend to be relatively concise compared with the longer, more structured responses expected from general-purpose assistants, so RAST-G primarily provides real-speech-origin supervision for prosody and expressive realization within a shorter-turn regime.

AI USE STATEMENT

Generative AI tools assisted language editing and LaTeX format migration. AI models also form part of the research pipeline: they support speech transcription and annotation, dialogue generation, speech synthesis and voice conversion, data filtering, benchmark and rubric construction, and automated response evaluation. The models, prompts, quality-control procedures, and scope of human review are all described in Paper and appendix.

The authors reviewed the AI-assisted manuscript revisions and remain responsible for the study design, data quality and validation procedures, reported analyses, and conclusions.

REFERENCES

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 56898–56918. Curran Associates, Inc., 2024. doi: 10.52202/079017-1813. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/681fe4ec554beabdc9c84a1780cd5a8a-Abstract-Datasets_and_Benchmarks_Track.html.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520/.

Hongjie Chen, Zehan Li, Yaodong Song, Wenming Deng, Yitong Yao, Yuxin Zhang, Hang Lv, Xuechao Zhu, Jian Kang, Jie Lian, Jie Li, Chao Wang, Shuangyong Song, Yongxiang Li, Zhongjiang He, and Xuelong Li. GOAT-SLM: A spoken language model with paralinguistic and speaker characteristic awareness, 2025. URL https://arxiv.org/abs/2507.18119.

Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, zehan wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, and Zhou Zhao. Voxdialogue: Can spoken dialogue systems understand information beyond words? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 288–303, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/011df158529ddceb5f2f7a65f2732a5a-Paper-Conference.pdf.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300705. URL https://doi.org/10.1145/3290605.3300705.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue, 2024. URL https://arxiv.org/abs/2410.00037.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PYmrUQmMEw.

Yuan Ge, Haishu Zhao, Aokai Hao, Junxiang Zhang, Bei Li, Xiaoqian Liu, Chenglong Wang, Jianjin Wang, Bingsen Zhou, Bingyu Liu, Jingbo Zhu, Zhengtao Yu, and Tong Xiao. On the emotion understanding of synthesized speech, 2026. URL https://arxiv.org/abs/2603.16483.

Haohan Guo, Shaofei Zhang, Frank K. Soong, Lei He, and Lei Xie. Conversational end-to-end tts for voice agents. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 403–409, 2021. doi: 10.1109/SLT48900.2021.9383460.

Feng Jiang, Zhiyu Lin, Yiyang Liu, Liumeng Xue, Fan Bu, Yuhao Du, Xiangying Chen, Benyou Wang, and Haizhou Li. S2S-arena: Evaluating paralinguistic instruction following in speech-to-speech models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34962–34978, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1615. URL https://aclanthology.org/2026.acl-long.1615/.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025. URL https://arxiv.org/abs/2504.18425.

Keon Lee, Kyumin Park, and Daeyoung Kim. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095751.

Yangzhuo Li, Shengpeng Ji, Yifu Chen, Tianle Liang, Haorong Ying, Yule Wang, Junbo Li, Jun Fang, and Zhou Zhao. Wavbench: Benchmarking reasoning, colloquialism, and paralinguistics for end-to-end spoken dialogue models, 2026. URL https://arxiv.org/abs/2602.12135.

Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Zhiyu Lin, Jingwen Yang, Jiale Zhao, Meng Liu, Sunzhu Li, and Benyou Wang. Decoding the ear: A framework for objectifying expressiveness from human preference through efficient alignment, 2025. URL https://arxiv.org/abs/2510.20513.

Songting Liu. Zero-shot voice conversion with diffusion transformers, 2024. URL https://arxiv.org/abs/2411.09943.

Xuan Luo, Shinnosuke Takamichi, Tomoki Koriyama, Yuki Saito, and Hiroshi Saruwatari. Emotion-controllable speech synthesis using emotion soft labels and fine-grained prosody factors. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 794–799, 2021.

Qwen Team. Qwen3-235B-A22B-Instruct-2507. https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507, 2025. Hugging Face model.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. Audiopalm: A large language model that can speak and listen, 2023. URL https://arxiv.org/abs/2306.12925.

Björn Schuller and Anton Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, 2013. doi: 10.1002/9781118706664.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026. doi: 10.48550/arXiv.2601.21337. URL https://arxiv.org/abs/2601.21337.

Xiaomi MiMo. Speech synthesis (MiMo-V2.5-TTS Series). Xiaomi MiMo API documentation, 2026. URL https://mimo.mi.com/docs/en-US/quick-start/usage-guide/audio/speech-synthesis-v2.5. Official API documentation for mimo-v2.5-tts, mimo-v2.5-tts-voicedesign, and mimo-v2.5-tts-voiceclone; accessed 2026-07-14.

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming, 2024. URL https://arxiv.org/abs/2408.16725.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Chao Yan, Boyong Wu, Peng Yang, Pengfei Tan, Guoqiang Hu, Li Xie, Yuxin Zhang, Xiangyu Tony Zhang, Fei Tian, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, Shuchang Zhou, and Gang Yu. Step-audio-editx technical report, 2025. URL https://arxiv.org/abs/2511.03601.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shu-wen Yang, Ming Tu, Ting-Wei Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. ParaS2S: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/734b2e77222680728b9ce78c573eae1e-Abstract-Conference.html.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1055. URL https://aclanthology.org/2023.findings-emnlp.1055/.

A RAST Data Construction

RAST construction applies quality control to the source utterance, its speech-condition annotations, and the completed dialogue. We first select recordings that can serve as standalone conversational endpoints, then reconcile their annotations before generating the missing endpoint. Text and audio checks assess the resulting exchanges before training. The following sections describe this pipeline, its stage-wise retention, and the release and privacy policies.

A.1 Anchor selection and annotation

Audio preparation. Speech segments are extracted from YouTube. Audio preparation standardizes the recordings, isolates vocals, and uses speaker diarization and voice activity detection to obtain

Table 5: Automatic filtering gates for real-speech anchor candidates.

Gate	Check	Parameters
Format standardization	Sampling rate, bit depth, channel count, loudness	$<$ 24 kHz discarded; target 44.1 kHz, 16-bit, mono, -20 dBFS
Vocal isolation	Source separation and vocal-track retention	UVR-MDX-NET-Inst_HQ_3
Speaker diarization	Single-speaker likelihood and overlap removal	Speaker-embedding cosine similarity > 0.7
VAD segmentation	Speech activity, silence splitting, merge and trim	Silero VAD at 16 kHz; duration 3–30 s; merge gap $<$ 2 s
Language filtering	Language ID and confidence	Whisper; {zh, en, fr, ja, ko, de}; confidence > 0.8
Safety filtering	Private, toxic, sexual, violent, or unsafe content	Text-based and audio-based filters

single-speaker segments. Segmentation follows speech and silence boundaries, merging nearby segments from the same speaker within the duration limit. Table 5 summarizes the audio-preparation and screening criteria. Source identifiers and timestamp offsets preserve the link to the original recording for deduplication and provenance tracking.

ASR verification and anchor selection. We transcribe and normalize the segments using Qwen3-ASR (Shi et al., 2026) and Qwen3-Omni (Xu et al., 2025), checking agreement before anchor selection. Segments with unstable transcripts are removed or manually reviewed; only those with verified transcripts proceed to the text-based filter.

Using these verified transcripts, a two-stage text filter checks whether an utterance can function as a conversational endpoint. It first identifies a self-contained question or answer, considering the whole utterance before extracting a continuous passage. Separate question and answer validators check whether a request is understandable and answerable, or whether a response conveys clear, useful information. These checks reject fragments that require an unseen image or missing preceding dialogue, truncated sentences, promotional scripts, and unsafe or inappropriate content. This step connects audio usability to dialogue usability: a clear recording alone is insufficient if its meaning depends on context absent from the training exchange.

Evidence-based annotation. For the selected anchors, the annotation inventory in Table 6 covers speech conditions supported by observable acoustic or visual evidence. Annotation prompts distinguish a salient property from an uncertain or uninformative description: age and gender may be unknown, and tone or ambient descriptions may be absent when the evidence is weak.

Initial annotations are reconciled using dimension-specific checks before dialogue completion. Age and gender labels require agreement between the multimodal annotation and the corresponding specialized classifier; conflicts leave the label unset. Voice-quality labels use the intersection of the audio and audio-video annotations. Emotion, tone, and contextual descriptions use their respective verification rules. Descriptions marked for removal by these rules are cleared. Missing or cleared fields are passed to the dialogue generator as `None`, so an unreliable attribute need not remove an otherwise usable anchor. This separates retention of the real utterance from retention of each conditioning label.

We select annotation components using expert-labeled validation examples, considering label accuracy, abstention, and detectable failure modes. Audio-only models supply affect, style, and voice-quality evidence; audio-video models contribute context when releasable video is available. Specialized classifiers provide consistency checks, and acoustic features support filtering. Four research-team experts review component outputs, thresholds, prompts, and failure cases before the annotation stack is frozen and used across data variants.

Table 6: Non-semantic speech annotation inventory and evidence requirements.

Category	Representative labels	Evidence requirement
Affect	neutral, happy, sad, angry, anxious, frustrated, excited	Prosody, pitch contour, energy, rhythm, facial or contextual evidence when available
Speaking style	urgent, whisper, hesitant, confident, explanatory, conversational	Timing, volume, articulation, pause structure, interactional stance
Prosody and voice quality	fast, slow, high pitch, low pitch, breathy, hoarse, monotone	Acoustic features and audible perceptual evidence
Speaker presentation	child-like, adult, elderly, masculine-presenting, feminine-presenting	Treated as uncertain presentation labels, not identity attributes
Ambient context	traffic, music, public space, office, household, emergency sounds	Background acoustic evidence or video-context evidence
Safety and privacy	private information, unsafe request, sensitive content	Text and audio policy filters with conservative removal

A.2 Directional dialogue completion

Conditioned text completion. Text-side exchange construction uses Qwen3-235B-A22B-Instruct-2507 (Qwen Team, 2025; Yang et al., 2025). Both directions condition on the anchor transcript and retained speech-condition annotations. User-side completion generates an assistant response using age, gender, emotion, tone, voice quality, ambient sounds, speaker description, and audio-source descriptions. These descriptions inform response content and delivery, while the response must address the user's request. Assistant-side completion uses emotion, tone, voice quality, and speaker description to infer a preceding query for the fixed real response. Its prompt considers explicit instructions, situational context, and the user's implied state, rather than merely converting the answer into a question.

Both templates return a clean counterpart transcript together with TTS instructions and emotion/style tags. Keeping the transcript separate from delivery instructions allows the same generated content to determine what is spoken while the instructions specify its spoken realization. The complete templates below show how the retained annotations enter each direction.

Synthesis and voice normalization. Step-Audio-EditX (Yan et al., 2025) generates synthesized counterparts with a canonical assistant timbre for responses and a balanced Common-Voice-derived speaker pool (Ardila et al., 2020) for user queries, subject to license and split constraints. When an assistant-side real-speech-origin anchor requires voice normalization for the target interface, we use Seed-VC (Liu, 2024). The converted recording remains the anchored endpoint; synthesis supplies the opposite endpoint. Audio is standardized in sampling rate, channel count, and loudness and checked for transcript-duration consistency.

Exchange-level quality inspection. The completed pair is checked as a dialogue, in addition to the earlier checks on the anchor alone. Text-level inspection evaluates whether the question is self-contained, whether the question and answer use a consistent language, and whether the answer contains obvious factual errors, hallucinations, or nonsensical content. Directional relevance also matters: a generated response must answer the real query, and a generated query must be compatible with the fixed real response. Table 7 summarizes these checks alongside condition use, safety, and audio-quality criteria.

Audio inspection covers both synthesized counterparts and voice-normalized anchors, including truncation, mispronunciation, severe artifacts, and loss of intelligibility or expressive structure after conversion. Only complete anchor–counterpart pairs enter exchange-level inspection; unmatched anchors are excluded from the training exchanges. The stage-wise counts in Table 8 distinguish completion yield from subsequent quality retention.

Table 7: Exchange-level filtering criteria after synthetic endpoint completion.

Failure type	Rejection criterion	Applied to
Response irrelevance	Question requires missing context, or the counterpart does not answer, elicit, or conversationally license the anchored transcript.	Both directions
Factuality failure	Response contains clearly false, hallucinated, or nonsensical content.	Both directions
Condition overuse	Generated text explicitly mentions hidden labels, annotation fields, or acoustic metadata without conversational justification.	Both directions
Safety failure	Generated or anchored endpoint contains unsafe advice, privacy leakage, or policy-violating content.	Both directions
TTS failure	Mispronunciation, truncation, severe artifacts, wrong language, or non-speech output.	Synthetic endpoint
Voice-normalization failure	Converted assistant-side anchor loses intelligibility, timing, or expressive structure, or violates assistant identity constraints.	Assistant-side anchors
Duration failure	Endpoint duration is too short, too long, or inconsistent with the transcript after synthesis or conversion.	Both directions
Metadata inconsistency	Transcript, language, labels, audio file, source fields, or direction tags conflict.	Both directions

The complete templates are as below. Named placeholders are filled from the anchor text and annotation values, with missing descriptions represented as `None`.

Shared system message. Both directions use the following system message; the corresponding complete template is supplied as the user message.

> You are an advanced Conversational AI assistant.

User-side anchoring (complete prompt).

> You are an advanced **Conversational AI** with high context awareness. Your task is to generate a verbal response that is both intellectually substantive and nonsemantical precise to a user's question by analyzing multi-dimensional descriptive input.
>
> **1. Input Data**
> **User's Question:** "{question_text}"
> **Contextual Descriptions:**
> • **Age & Gender:** {age}, {gender}
> • **Emotion Description:** {emotion_desc}
> • **Tone Description:** {tone_desc}
> • **Voice Quality Description:** {voice_quality_desc}
> • **Environment Sounds Description:** {env_desc}
> • **Combined Speaker Description (User):** {speaker_combined} *(Format: [Role] at [Scenario] doing [Activity])*
> • **Audio Source Description:** {audio_source_desc}
>
> **2. Reasoning Logic (Deep Inference & Synthesis)**
> **Phase 0: Language Consistency (MANDATORY)**
> • The `inferred_response` **MUST** be in the **same language** as the input `User's Question`. For example, if the User's Question is in Chinese, the generated response must also be in Chinese; if in English, the response must be in English. Do NOT translate or switch languages.
> **Phase 1: Demographic & Persona Alignment**
> • **Age Analysis:**

• **Child:** MUST use simple vocabulary, warm, supportive, and educational tone, and trigger **Safety Rules (avoid inappropriate information)** for PII or physical risks.
• **Elderly:** MUST be patient, considerate, and use clear, structured sentences.
• You should determine the limitations of the response process based on users' age.

Phase 2: Emotional & Tone Strategy (Decoupled Logic)

• **The Non-Assumption Principle:** Acknowledge the *state* described in {emotion_desc} and {tone_desc}, but do not invent a cause casually.
• *Good:*(Static Hooks) "I can tell you're feeling a bit frustrated..." OR
• *Good:*(Dynamic Hooks) Start the response by directly addressing the *core* of the user's argument or the *intensity* of their emotion. Use a "Mirroring Statement" that reflects the user's specific vocabulary or energy level without being repetitive.
• **Tone Matching:** - **Negative User Tone (Angry/Tense):** AI stays **Professional, De-escalating, and Calm**. Do NOT mirror negativity.
• **Positive User Tone (Excited/Warm):** AI mirrors with **Enthusiasm**.
• **Vulnerable User Tone (Sad/Worried):** AI acts as a **Steady, Healing presence**.

Phase 3: Environmental & Physical Constraints

• **Acoustic Environment:** Parse {env_desc} for:
• **Urgency:** (e.g., sirens, alarms) Prioritize safety checks and immediate acknowledgment.
• **Physical Constraints:** If descriptions imply "noisy/crowded," keep responses short and punchy. If "quiet/private" (e.g., library), imply a softer, discreet tone.
• **Voice Quality & Source:** If {voice_quality_desc} indicates difficulty speaking (e.g., "breathless" or "sore throat"), minimize the user's need for long follow-up answers.

Phase 4: Content Enrichment (The "Fuller" Response)

• **Semantic Depth:** If the user presents a complex argument, respond with a nuanced perspective that matches their intellectual or emotional level.
• **Non-Semantic Integration:** Use the tone_desc and voice_quality to determine the "pacing" of your thoughts. A "punctuated" or "hard" user tone requires an AI response that is firm, structured, and equally weighty—not soft or dismissive.

3. TTS Instruction Generation

Step 1: AI Stance Mapping

• Do not always "stay calm." Match the **Intellectual Energy** of the user while maintaining a **Professional/Constructive Stance** by using appropriate emotion or speaking style.
• **If User is Accusatory/Passionate:** AI should sound **Grave, Earnest, and Resonant** (Use [serious] for a formal/earnest tone, or [authority] to maintain a commanding, or [empathy] to show understanding, steady resolve that de-escalates the tension.).
• **If User is Vulnerable:** AI should sound **Gentle and Grounded**. Can use [empathy] to show understanding or [comfort] for a reassuring presence.

Step 2: TTS Tag Selection

• **Valid Emotion List:** [happy], [sad], [angry], [surprised], [fear], [disgusted], [empathy], [embarrass], [excited], [depressed], [admiration], [coldness], [humour]
• **Valid Speaking Style List:** [serious], [authority], [whisper], [warm], [comfort], [chat], [older], [child], [generous], [ethereal]
• **Valid Paralinguistic List (inline only):** [Breathing], [Laughter], [Surprise-oh], [Confirmation-en], [Uhm], [Surprise-ah], [Surprise-wa], [Sigh], [Question-ei], [Dissatisfaction-hnn]

Step 3: Tag Usage Rules

• **Emotion & Style:** Select at most ONE emotion tag and ONE style tag. These go into tts_labels_used field.
• **Paralinguistics:** Embed directly into tts_instructions text to create **Rhythm**. Use [Breathing] before a heavy statement; use [Confirmation-en] to acknowledge a point. Do NOT include paralinguistics in tts_labels_used.

4. Output Format

CRITICAL: Output ONLY valid JSON. No markdown code blocks, no extra text.

```
{{
  "inferred_response": "Clean text answer without any tags",
  "tts_instructions": "Text with inline [Paralinguistic] tags Only",
  "captions_used": ["age", "emotion_desc", "tone_desc", "env_desc",
"speaker_combined", "voice_quality_desc"],
  "tts_labels_used": {{"emotion": "selected_style_tag_or_null",
"style": "selected_style_tag_or_null"}},
  "reason": "Brief explanation of the reasoning process, specifically
how the natural language descriptions led to the inferred response."
}}
```

Assistant-side anchoring (complete prompt; Chinese examples translated).

`Task Description`:
You are an AI specializing in conversational analysis and generation. Your task is to reverse-engineer a **User-AI conversation**. You will be given an **AI Model's answer**, which includes its text and several **natural language paralinguistic descriptions**.
Your goal is to generate the **Human User's question** that prompted this response. This generated question must demonstrate deep reasoning based on both the AI's answer text **and** the nuances provided in the descriptive labels (emotion, tone, speaker profile, etc.).
Constraints:
1. The AI is a speech assistant and has **no visual capabilities**. The generated user question must reflect this limitation.
2. Provide Text-to-Speech (TTS) instructions for synthesizing the **User's question**, reflecting the *inferred* emotional and paralinguistic state of the user *when they were asking the question*.
3. **Language Consistency:** The `tts_instructions` **MUST** be in the **same language** as the input `Answer Text`. For example, if the Answer Text is in Chinese, the generated question must also be in Chinese; if in English, the question must be in English. Do NOT translate or switch languages.

Input: (AI Model's Response Data)
- **Answer Text:** "${answer_text}"
- **Labels:**
  - **Emotion Description:** ${emotion_desc}
  - **Voice Quality Description:** ${voice_quality_desc}
  - **Tone Description:** ${tone_desc}
  - **Speaker Comprehensive Description:** ${speaker_combined} *(Note: This describes "Who is doing what in which scenario")*

Generation Process:
1. Data Density Evaluation:
Before reasoning, assess the richness of the provided labels:
- **High Density:** All labels (Role, Tone, Emotion, Voice) are specific and characteristic. → *Should give a "Complex Instruction" (Multi-constraint), not just a "Simple Instruction if appropriate".*
- **Medium Density:** Only some labels are specific (e.g., specific Tone but generic Role). → *Aim for a "Situational/Relational" inference.*
- **Low Density:** Most labels are generic or missing. → *Aim for a "Contextual Natural Inquiry" (Focus on semantic content BUT still add depth and context).*

! CRITICAL - Minimum Complexity Rule (Apply to ALL density levels):
Regardless of data density, the generated question **MUST NOT** be a trivial, shallow, or single-clause question. Specifically:
- **[X] FORBIDDEN:** Simple "what is X?" / "What is X?" one-liners that merely rephrase the answer topic.
- **[X] FORBIDDEN:** Questions that only ask for a single fact already fully stated in the answer (e.g., "What is the main purpose of this intensive training?" or "How does a submarine observe its surroundings while underwater?").
- **[X] FORBIDDEN:** Questions that simply convert the answer into an interrogative sentence without adding any new context.
- **[OK] REQUIRED:** Even for Low Density data, enrich the question by applying at least TWO of the following strategies:
  1. Adding **context or background** (e.g., "I heard that... / I recently saw a news report saying...; could you analyze it in detail...")
  2. Requesting **multiple dimensions** (e.g., "Tell me about X, and also explain why Y" / "I not only want to know X but also want to understand the reasons behind it")
  3. Embedding a **scenario or motivation** (e.g., "I'm writing a report on... / I am preparing material about...; help me organize...")
  4. Asking for **deeper analysis, comparison, or explanation** beyond surface-level facts
  5. Incorporating **follow-up or multi-part sub-questions** to demonstrate genuine curiosity
- **[OK] Target Length:** The question should generally be **at least 2-3 sentences or clauses**, not just a single short sentence.

2. Select the Optimal Reasoning Path:
- **Path A: Explicit Multi-Constraint**
  - Use this if the AI's response is highly stylized and specific.
  - *Logic:* The user explicitly requests the persona and style.
  - *Example:* "Tell me about [Topic] using a [Tone] tone. Please act as a [Role] and make your voice sound [Voice Quality]."

- **Path B: Implicit Contextual Trigger**
  - Use this if the `Speaker Description` provides a clear professional or situational context.
  - *Logic:* The user sets a scene where the AI *naturally* adopts the role without being told to "act".
  - *Example:* "I'm in the middle of a [Scenario], give me an update on [Topic]." (This naturally triggers a Reporter/Professional role).
- **Path C: Relational/Reactive Inference**
  - Use this if the AI's `Emotion` or `Tone` is the primary feature (e.g., very reassuring, apologetic, or excited).
  - *Logic:* The AI's style is a *reaction* to the user's implied state.
  - *Example:* If AI is "extremely apologetic," the user might have expressed "frustration" or "found a mistake."

3. Formulate the User's Question (The Goal):

- **Integration:** The question **MUST** synthesize the `Emotion`, `Speaker Description`, `Tone`, and `Voice Quality`.
- **Example of Complexity (do not copy the text, you should add variations):** "Can you tell me [Content]? Please act like a [Role] in a [Scenario]. I need your tone to be [Tone] and [Emotion], and it would be great if your voice had a [Voice Quality] edge to it. Do not always use the fixed tricks, be creative."
- To ensure variety, choose from these (or similar) patterns:
  - **Direct Command:** "Tell me about...", "Explain the...", "Describe..."
  - **Situational:** "I'm currently [doing X], help me with [Y]..."
  - **Inquiry:** "What's the deal with...", "How does [X] even work?", "Is it true that..."
  - **Emotional:** "I'm so [excited/worried] about...", "This is incredible! Tell me more!"
  - **Role-Play:** "Imagine you're my [Role], how would you handle..."

! Self-Check Before Finalizing: Re-read your generated question and ask:

- "Is this question something a real person would actually ask an AI assistant, or is it just a textbook quiz question?"
- "Does this question have enough context and depth, or could a 5-year-old have asked it?"
- "Am I just flipping the answer into a question?" $\rightarrow$ If YES, rewrite with added context, motivation, or sub-questions.

4. Generate TTS Instructions (User's State):

- Infer the user's intent. If they are giving complex instructions, they might sound `Serious` or `Focused`.
- **Allowed Tags:**
  - Emotion: [happy], [sad], [angry], [surprised], [fear], [disgusted], [empathy], [embarrass], [excited], [depressed], [admiration], [coldness], [humour]
  - Style: [serious], [authority], [whisper], [warm], [comfort], [chat], [older], [child], [generous], [ethereal]
  - Paralinguistics (inline only): [Breathing], [Laughter], [Surprise-oh], [Confirmation-en], [Uhm], [Surprise-ah], [Surprise-wa], [Sigh], [Question-ei], [Dissatisfaction-hnn]
- **Tag Usage Rules:**
  - **Emotion & Style:** Select at most ONE emotion tag and ONE style tag. These go into `tts_labels_used` field as a structured object.
  - **Paralinguistics:** Embed directly into `tts_instructions` text to create natural speech rhythm. Do NOT include paralinguistics in `tts_labels_used`.

Output Format:

```
{
  "inferred_question":  "The clean text of the user's question without any tags.",
  "tts_instructions":  "Text with inline [Paralinguistic] tags Only",
  "captions_used":  [
    "Specific phrases from Emotion/Voice/Speaker/Tone labels"
  ],
  "tts_labels_used":  {"emotion":  "selected_style_tag_or_null", "style": "selected_style_tag_or_null"},
  "reason":  "Brief explanation of the reasoning process, specifically how the natural language descriptions led to the inferred question."
}
```

Table 8: RAST construction stages and final endpoint composition. Counts refer to source segments before completion, exchanges after completion, and individual audio endpoints in the final two rows. Durations sum the corresponding audio.

Stage / component	Count	Duration (h)	Role in construction
Raw collection pool	5,815,821	12,046.2	Source pool for conservative anchor mining
Filtered candidate segments	497,297	943.9	Single-speaker, language-eligible, safety-screened candidates
Validated authentic anchors	362,592	803.9	Unique transcript-stable, deduplicated, annotation-supported source anchors
Complete candidate S2S exchanges	327,063	–	One real-speech-origin anchor paired with its synthesized counterpart
After exchange-level quality inspection	284,660	2,914.0	Final exchanges passing semantic, factuality, safety, audio-quality, and duplicate checks
real-speech-origin endpoint audio	284,660	774.8	Final real-origin endpoints
synthesized counterpart audio	284,660	2,139.2	Synthesized completion of the missing conversational endpoint

A.3 Corpus statistics

Table 8 summarizes construction stages and final endpoint composition. Validation retains 72.9% of filtered candidates and 85.2% of their duration. Completion yields one synthesized counterpart for 90.2% of validated anchors; the remaining 35,529 unmatched anchors are excluded before exchange-level inspection. Final exchanges retain 87.0% of complete candidates, or 78.5% of validated anchors. Each contains one real-speech-origin anchor and one synthesized counterpart, for 569,320 endpoint-audio instances in total.

The final corpus comprises 203,118 user-side and 81,542 assistant-side exchanges, with durations by conversational role reported in Table 2. RAST-Mix combines these two pools.

A.4 Release and ethics

Table 9 summarizes release eligibility and licensing. RASTBench assets include a benchmark card, metadata schema, splits, scoring scripts, and judge prompts, subject to source restrictions.

Source identifiers enable deletion or opt-out requests to propagate to segments, annotations, and derived assets. Private information, identifying sensitive details, and unsafe content are removed before annotation and exchange construction. Age and gender annotations describe uncertain speaker presentation rather than verified identity.

B RASTBench Construction and Evaluation

RASTBench measures task completion and condition-appropriate responses under transcript-controlled speech variation. This appendix describes its construction, quality control, and audio-native scoring protocol; complete model results are reported in Appendix D.

B.1 Benchmark construction

Construction proceeds from scene cards specifying a condition-neutral request and a speech-dependent mechanism, through contrast blueprints and matched samples, to group-level shared-baseline and displacement rubrics. We use `gpt-5` for text construction and rubric generation, `gemini-3-pro-preview` for age and ambient critics, and `gpt-5` for tone and gender crit-

Table 9: Release status, licensing, and reproducibility assets.

Asset type	Release status	License basis	Notes
Raw real anchors	Not released	Speaker privacy constraints	No raw real-speech audio is redistributed.
Synthetic counterparts	Released when permitted	Project-generated TTS or voice-normalized outputs, subject to model and voice licenses	Includes synthesized user queries and assistant responses used to complete RAST exchanges.
Metadata and labels	Released	Derived annotations	Includes transcripts, normalized transcripts, non-semantic labels, confidence fields, source categories, timestamps, and release eligibility flags.
Construction scripts	Released	Apache-2.0	Includes filtering scripts, prompt templates, synthesis configuration, validation code, and version-pinned dependencies.
RASTBench assets	Partially released	Benchmark-specific license and source restrictions	TTS Bench assets are released when distributable; Human Bench audio is access-controlled, with scripts and judge prompts released.
Model checkpoints	Released for research use	Subject to base-model and fine-tuning licenses	Fine-tuned checkpoints and evaluation recipes are released with model-card documentation.

ics. Age critics operate at blueprint and sample stages; ambient critics operate at the blueprint stage. Emotion additionally uses deterministic balance, leakage, and duplicate-text checks.

The TTS Bench uses MiMo-V2.5-TTS (Xiaomi MiMo, 2026) for emotion, tone, age, gender, and ambient-sound contrasts. Each group shares a generated transcript verbatim and holds non-target condition metadata fixed while varying the designated condition: affect or delivery, perceived voice presentation, or background sound. Ambient groups contain a quiet realization and two mixtures of the same clean carrier.

The Human Bench draws emotion and tone examples from authentic YouTube utterances with source transcripts and speech-attribute annotations. Candidates must have a question label, valid audio, language-appropriate transcript length, and source-corpus WER at most 0.2, followed by rule-based and `gpt-4o` text-suitability filtering. MiMo-V2.5-TTS synthesizes a voice-cloned control (A) from the retained transcript and a reference voice clip, paired with the authentic recording (B). Both sides share the transcript, condition label, intended response policy, and complete rubric.

Table 10 gives the dimension-wise benchmark inventory. Its authentic utterances come from a source pool separate from the training-anchor pool. All realizations of a transcript group remain in the same benchmark split, and raw and normalized transcripts are checked for lexical leakage.

The following translated excerpts illustrate sample construction and final rubric generation. The preceding scene-card and blueprint stages require condition-neutral wording, a task-relevant mechanism, distinct justified response paths, and preservation of child and older-adult autonomy. Complete Chinese prompts accompany the construction code.

Sample and criterion generation (age; excerpt).

> **System:** Construct `sample_A`, `sample_B`, and `sample_C`. A/B/C share the exact input text and all metadata except `speaker_age`. A carries the shared base-

Table 10: Dimension-wise RASTBench inventory. Audios are model inputs; Avg Dur. and Hours denote mean and total input duration, respectively.

Section / Dim.	Groups	Audios	Avg Dur. (s)	Hours	Labels
TTS Bench					
Emotion	512	1,536	8.26	3.52	neutral, happy, excited, anxious, fearful, sad
Tone	372	744	12.82	2.65	urgent, hushed_private
Age	322	966	16.86	4.52	child, adult, elder
Gender	310	620	17.17	2.96	male, female
Ambient	512	1,536	6.60	2.82	quiet, subway, driving, construction, office, emergency, rain, restaurant, children_home, gym
Subtotal	2,028	5,402	10.98	16.47	
Human Bench					
Emotion	917	1,834	7.51	3.83	841 descriptions (e.g., anxious, calm & soothing)
Tone	353	706	6.20	1.22	109 descriptions (e.g., hesitant, direct & assertive)
Subtotal	1,270	2,540	7.15	5.04	
Total	**3,298**	**7,942**	**9.75**	**21.51**	

line check; B and C carry condition-specific displacement checks. Each criterion tests one observable judgment, avoids fixed answers and generic labels.

User: Generate the contrast samples from the supplied blueprint, including its objective mechanism, expected displacement, and three action paths. Verify transcript identity, absence of age-leak words, distinct positive mechanisms for displacement criteria, and preservation of child and older-adult autonomy. Return one JSON object per blueprint.

Group-level rubric generation (all dimensions; excerpt).

Generate one group-level rubric for a transcript-controlled non-semantic grounding benchmark. `shared_baseline` describes the common task; `displacement` describes the target sample's condition-supported adaptation. A baseline sample uses only the shared baseline. A target sample uses the shared baseline plus its displacement. Criteria must describe high-level behavioral outcomes, not keywords, a complete reasoning chain, or a unique answer.

B.2 Quality control

Filtering operates on whole contrast groups. Model critics assess mechanism validity, condition relevance, response-policy separability, rubric quality, and stereotype risk; deterministic validators check transcript identity, schemas, condition taxonomies, metadata, leakage, and audio validity. Table 11 specifies the rejection rules.

Age and gender denote perceived voice presentation rather than verified identity, anatomy, competence, or social role. Adaptation requires a task-relevant mechanism: age may support conditional developmental or accessibility guidance, while a urinary-pain response may discuss sex-linked possibilities without inferring anatomy from voice. Prompts and critics preserve child and older-adult agency and prohibit stereotypes about politeness, emotionality, competence, spending, family roles, or preferred detail.

Researchers define the taxonomies and acceptance rules and review development examples, critic disagreements, and disputed cases. Among three annotators, two independently review each Human candidate for audio quality, criteria quality, and usability. We exclude 24 groups whose cloned control is both more than twice the authentic recording's duration and longer than 30 seconds. TTS items receive no exhaustive human acceptance pass, and there is no separate exhaustive post-hoc fairness audit of the age and gender groups.

Table 11: Operational group-level filtering rules. Deterministic checks enforce identity, schema, taxonomy, leakage, and file validity; model critics assess mechanism and semantic separability.

Failure	Concrete rejection rule	Examples
Weak contrast	Reject a group if its action paths are substantively the same, no objective mechanism connects the condition to the requested task, or the critic rates the response displacement as weak. Emotion additionally requires evidence for both readings and rejects `displacement_separability=weak`; ambient requires distinct B/C conditions and an ambient-sensitivity score of at least 4/5.	Only shortening an answer for a child, mirroring emotion without changing help, or naming background noise without changing the action path.
Semantic leakage	Reject if the shared transcript names or uniquely implies the target condition. Dimension-specific lexicons catch direct emotion words, age self-reports and life-stage terms, gendered self-identification or anatomy, explicit secrecy/urgency instructions for tone, and named locations or sounds for ambient. The normalized transcript string must also be identical across all sample records in a group.	"I am ten years old," "because I am a woman," "please whisper," or "I am driving" in the common transcript.
Confounded contrast	Reject if any non-target metadata field changes within a TTS group, a target voice realization contains an unintended emotion/tone instruction, an ambient pair does not reuse the same clean carrier, or an audio file is missing, unreadable, empty, or has a synthesis/duration anomaly.	Changing speaker and emotion together, using different wording across A/B/C, or mixing B/C from different clean utterances.

For tone, age, and gender, critic rejection excludes a record, while rewrite recommendations are advisory. Failed or unparseable generator outputs produce no accepted item. Invalid age and ambient critic outputs are rejected; tone and gender critic parse failures use a permissive fallback subject to deterministic validation. Records marked `critic_failed`, hard validation failures, and invalid audio are excluded. Emotion retains accepted blueprint criteria if preliminary-criterion generation fails; every evaluated item requires a valid final group-level rubric.

B.3 Scoring and judge validation

We use the audio-capable API model `gemini-3.5-flash` with `temperature=0.0` and `max_tokens=16384`; `top_p`, `top_k`, and API seed are unspecified. Each response is judged independently from its audio, baseline/target role, condition description, and applicable shared-baseline and displacement rubrics. The judge receives neither the original user audio or raw transcript, a reference answer, other group responses, nor an ASR transcript of the response. Separately requested response transcripts serve only diagnostic logging. Each item contributes one successful judge call; failed requests are retried, and missing or unparseable judgments are excluded.

Audio-native judge prompt. The judge receives the response audio with the sample role, condition description, and applicable rubric. The excerpt below retains the criterion-level scoring rules.

Table 12: Audio-native Gemini 3.5 Flash judge calibration against expert majority labels. Agreement is reported as a percentage; intervals are for label correlation.

Judgment	Agreement	Correlation	95% CI of correlation
Semantic task completion	86.25	0.724	[0.636, 0.800]
Speech-conditioned response adaptation	94.34	0.756	[0.548, 0.916]
All criterion labels	88.13	0.761	[0.692, 0.825]
Sample-level pass	95.28	0.841	[0.703, 0.948]

Evaluate the attached model-response audio using the supplied sample role, speech condition, and criteria. Score each applicable criterion independently as 1 (satisfied) or 0 (not satisfied), with a reason of 10–30 characters. Copy each criterion in full.
For baseline samples, evaluate only the shared task baseline; condition adaptation is not required, and `displacement_criteria_scores` must be empty. For target samples, evaluate both baseline and displacement criteria; an ordinary generic response is insufficient.
Do not require fixed keywords or assign 0 for minor wording defects. If a response merely names the condition without an actual change in strategy, assign 0 to the corresponding displacement criterion. If it is clearly unsafe, stereotyped, or severely over-adapted, assign 0 to the affected criteria.
Score the audio directly; transcription is separate. Return valid JSON only, with `baseline_criteria_scores` and `displacement_criteria_scores` containing `criterion`, `score`, and `reason` fields. Do not calculate aggregate scores.

The original deployed Chinese prompt and scoring implementation accompany the evaluation code. Each model's scores use all its available valid judgments. Item scores average applicable binary criteria, so baseline and displacement contribute according to their criterion counts; items then receive equal weight within each dimension. TTS baseline items carry only baseline criteria, while TTS target items and both Human sides carry the full rubric. Overall gives equal weight to the five-dimension TTS macro-average and the two-dimension authentic-human B-side macro-average, following Section 4. B-Macro and D-Macro separately average each item's corresponding criterion pass rate within each TTS dimension, then equally average the five dimensions; D-Macro includes only items with displacement criteria. All reported scores use a 0–100 scale.

The Real–Control Gap averages authentic-minus-cloned score differences within the two Human dimensions, where $M_{d,s}$ is the mean valid-item score for dimension d and side s:

$$\Delta_{\mathrm{RH}} = \frac{1}{2} \sum_{d \in \{\mathrm{Emo}, \mathrm{Tone}\}} (M_{d,B} - M_{d,A}). \tag{3}$$

Positive values favor authentic speech. The gap compares two acoustic realizations that may differ in several speech properties; it does not isolate an individual cue's causal effect. Per-item rubric scores measure condition-appropriate responses, rather than the magnitude of differences between paired responses. Appendix D.3 reports the A/B results, and Appendix C.3 describes uncertainty estimation.

Judge calibration uses expert majority labels for 455 criteria from 131 responses. Table 12 reports agreement and label correlation; correlation intervals use 10,000 contrast-group bootstrap replicates stratified by the seven dimensions (seed 42), retaining each group's labeled responses and criteria. Sample-level pass requires every applicable criterion to pass and uses the 127 responses with complete annotations.

C Experimental Protocol

The experiments compare anchoring directions, training-data fractions, and paired Pure-TTS controls. We describe their training and decoding settings, define the external evaluation metrics, and specify how scores and uncertainty are estimated. Complete results are collected in Appendix D.

C.1 Training, controls, and decoding

We fine-tune all parameters of Kimi-Audio-7B-Instruct with AdamW in BF16, using a learning rate of 5×10^{-6}, a 2% warm-up, and a maximum sequence length of 3,072 tokens. The loss sums text and audio next-token losses, each normalized over assistant positions; user tokens provide conditioning. The four data fractions are nested subsets of a fixed random ordering. Each training file is shuffled with seed 42 and split into 95% training and 5% validation data.

Pure-TTS controls. Each control uses the corresponding quarter-scale corpus and replaces only the real-speech-origin anchor with TTS generated from its original transcript: user audio in U-TTS, assistant audio in G-TTS, and the corresponding anchor in each construction direction in Mix-TTS. Dialogue texts, the synthesized counterpart, and the training horizon are retained. G-TTS uses Step-Audio-EditX, with the original assistant text and a Kimi-Audio voice reference.

Decoding. Base and the twelve RAST checkpoints use text decoding with temperature 0.0 and top-$k = 5$, and audio decoding with temperature 0.8 and top-$k = 10$ at 24 kHz. Each evaluated item contributes one valid generated response; empty text and missing or unreadable audio are excluded from scoring.

C.2 External evaluation metrics

WavBench measures speech understanding and reply relevance, while DeEAR scores the expressiveness of generated response speech. RASTBench uses the scoring protocol in Appendix B.3.

WavBench. Our composite combines percentage accuracies on emotion, gender, and age understanding with reply relevance. The three understanding tasks contain 525, 200, and 300 items, respectively. Reply relevance is evaluated on 77 implicit examples: 22 age, 24 pitch, 19 speed, and 12 understanding examples. The text judge rates the relevance of the reply to the user text on a 1–5 scale. Writing the mean rating as $\mathrm{ImplicitText}_{77}$, we report $\mathrm{Rel.} = 20\,\mathrm{ImplicitText}_{77}$ and combine the four components with equal weight:

$$S_{\mathrm{WB}} = \frac{\mathrm{UndEmo} + \mathrm{UndGender} + \mathrm{UndAge} + 20\,\mathrm{ImplicitText}_{77}}{4}. \tag{4}$$

DeEAR. We apply the released DeEAR_Base checkpoint and XGBoost fusion model to generated responses to the 1,270 authentic-human inputs in RASTBench: 917 emotion and 353 tone items. We report the three regression outputs, Arousal, Prosody, and Nature, and the fused Expressive score, each averaged over available valid responses and multiplied by 100. Base's score of 17.16 is recomputed from our archived Kimi-Audio-7B-Instruct per-response scores on these inputs.

C.3 Statistical estimation

Scores and differences use each model's available valid judgments. We use 10,000 joint bootstrap replicates with seed 42 and percentile 95% intervals. Shared evaluation units receive the same draws across models; differences are recomputed in every replicate. The intervals describe evaluation-sample uncertainty for fixed model and judge outputs. Differences are computed before final display rounding. RASTBench retains its released two-decimal rounding at the dimension, section, macro, and Overall stages, including dimension-wise Human B-minus-A differences, for both point estimates and bootstrap replicates.

Table 14 lists the resampling units. RASTBench resamples transcript groups within the five TTS and two Human dimensions, retaining valid items from partially observed groups and recomputing scores with the weights defined in Section 4. Group membership is fixed by the benchmark inventory; the A/B diagnostic resamples both realizations together. WavBench resamples items within its three understanding tasks and four implicit subsets, preserving the subsets' item-count weights and reply relevance's one-quarter composite weight. DeEAR resamples source-item IDs.

Table 13: Quarter-scale replacement effects: RAST minus its corresponding Pure-TTS control, in score points, with 95% bootstrap intervals below. Intervals belong to the displayed difference. Values retain the source's precision; differences for external metrics are computed before display rounding.

Direction	RASTBench	Adaptation (D)	WB composite	DeEAR Expr.
U	+4.72 [3.60, 5.83]	+4.87 [3.32, 6.47]	+11.48 [8.80, 14.17]	−1.44 [−2.44, −0.43]
G	+4.52 [3.18, 5.86]	+5.41 [3.70, 7.08]	+0.59 [−2.56, 3.72]	+6.90 [5.78, 8.00]
Mix	+8.81 [7.54, 10.09]	+7.67 [5.89, 9.40]	+13.94 [10.91, 16.96]	+7.33 [6.28, 8.40]

Table 14: Evaluation metrics and bootstrap units.

Benchmark	**Resampling unit**	**Reported metrics**
RASTBench	contrast group	criterion pass rates, dimension scores, Overall
WavBench	item within task/subset	understanding, reply relevance, composite
DeEAR	item	Arousal, Prosody, Nature, Expressive

D Complete Results and Additional Analyses

We report complete benchmark scores across training-data fractions, then give uncertainty estimates for the main comparisons. The final subsection separates training-time Pure-TTS controls from test-time Human A/B diagnostics and the acoustic analysis that motivates real-speech anchoring.

D.1 Complete benchmark results

Tables 15, 16, and 17 report all configurations. RASTBench includes the five TTS dimensions and the authentic-human B side of the two Human dimensions; WavBench reports the four components of Eq. 4; DeEAR reports its regression outputs and fused Expressive score.

D.2 Statistical comparisons

Table 18 gives RASTBench score intervals and differences from Base. Table 19 brings together the Pure-TTS and data-scale comparisons. Table 20 separates the TTS-section, authentic-human, baseline-criterion, and displacement-criterion effects. All estimates follow Appendix C.3.

D.3 Additional diagnostics

Human A/B results. Table 21 reports authentic-human B scores and their voice-cloned A controls across all configurations. The Real–Control Gap follows Eq. 3. Relative to their Pure-TTS controls, the changes in this gap are -0.51 ($[-3.17, +2.12]$) for U, $+0.97$ ($[-2.09, +4.17]$) for G, and -0.84 ($[-3.65, +1.97]$) for Mix. All three intervals include zero: the comparisons do not show a significant change in the authentic-minus-cloned score gap.

Quarter-scale Pure-TTS diagnostics. Table 22 gives the section scores and criterion diagnostics underlying the quarter-scale comparisons. Pure-TTS controls replace the training anchor; Human A/B changes the test-time input realization. Their effects address different comparisons, and Table 20 reports the training-control differences.

Transcript-matched acoustic diagnostic. Figure 1 compares real utterances with TTS resyntheses of the same transcript. After transcript normalization and silence trimming, we extract duration, voiced and pause fractions, F0 median and variability, RMS energy statistics, spectral centroid and flatness, and summary cepstral statistics. Features are z-scored before PCA visualization.

Table 15: RASTBench dimension scores. Items have equal weight within each dimension; RH dimensions use the authentic-human B side.

Config.	Scale	Emo	Tone	Age	Gender	Ambient	RH-Emo	RH-Tone	Overall
Base	–	70.48	76.00	71.75	60.45	64.01	27.14	49.56	53.45
RAST-Mix	100%	75.33	88.16	82.88	73.42	69.21	79.61	77.05	78.06
RAST-Mix	75%	74.03	89.37	82.70	73.31	69.19	79.61	77.93	78.25
RAST-Mix	50%	73.36	87.52	82.74	73.28	68.46	79.38	76.15	77.42
RAST-Mix	25%	70.87	85.16	81.76	70.39	67.43	76.59	74.84	75.42
RAST-U	100%	81.63	83.27	81.55	70.60	71.88	75.93	75.87	76.84
RAST-U	75%	79.73	82.62	81.07	71.44	69.73	76.60	72.98	75.86
RAST-U	50%	80.52	83.61	81.31	72.20	70.32	76.48	74.60	76.56
RAST-U	25%	80.71	82.02	80.89	71.70	70.11	73.75	72.87	75.20
RAST-G	100%	69.21	83.17	77.47	66.79	61.24	63.03	66.44	68.16
RAST-G	75%	68.04	83.07	77.86	66.38	61.07	64.17	70.62	69.34
RAST-G	50%	68.62	82.81	77.66	67.08	60.90	64.10	65.66	68.14
RAST-G	25%	67.43	80.86	77.63	66.83	59.83	63.14	66.22	67.60
U-TTS	25%	77.98	74.31	75.63	64.02	63.92	71.13	68.47	70.48
G-TTS	25%	66.93	71.14	71.31	57.41	56.33	59.87	63.19	63.08
Mix-TTS	25%	74.74	71.55	73.44	59.25	59.27	65.08	66.06	66.61

Table 16: WavBench understanding accuracies, reply relevance (Rel., 1–5 rating $\times 20$), and the composite average from Eq. 4.

Config.	Scale	UndEmo	UndGender	UndAge	Rel.	Avg
Base	–	74.95	11.00	42.00	80.26	52.05
RAST-Mix	100%	84.57	33.50	54.67	94.29	66.76
RAST-Mix	75%	83.62	35.00	54.67	90.91	66.05
RAST-Mix	50%	86.10	38.50	53.33	94.03	67.99
RAST-Mix	25%	85.52	44.50	62.67	91.69	71.09
RAST-U	100%	85.52	20.50	54.33	95.32	63.92
RAST-U	75%	85.71	28.00	51.00	95.06	64.94
RAST-U	50%	87.81	33.50	59.00	94.55	68.71
RAST-U	25%	87.43	27.50	60.33	96.62	67.97
RAST-G	100%	76.00	61.50	57.33	78.96	68.45
RAST-G	75%	75.62	56.50	52.33	81.04	66.37
RAST-G	50%	74.48	57.50	54.00	77.14	65.78
RAST-G	25%	77.14	58.00	57.33	76.36	67.21
U-TTS	25%	82.10	10.00	44.00	89.87	56.49
G-TTS	25%	78.67	55.00	53.33	79.48	66.62
Mix-TTS	25%	79.20	25.00	35.33	89.09	57.16

Paired effect sizes are computed as $d_z = \bar{\Delta}/s_\Delta$ over matched real–TTS differences. This analysis characterizes acoustic variation that remains after transcript matching and serves as a diagnostic alongside the training comparisons.

Table 17: Complete DeEAR results ($\times 100$). Arousal, Prosody, and Nature are the three neural regression heads; Expressive is the XGBoost-fused main indicator scoring the expressiveness of generated response speech.

Config	**Scale**	**Arousal**	**Prosody**	**Nature**	**Expressive**
Base	—	28.50	36.21	55.42	17.16
RAST-Mix	100%	35.96	40.44	65.40	26.94
RAST-Mix	75%	35.18	39.98	65.86	27.45
RAST-Mix	50%	35.44	39.86	65.22	26.80
RAST-Mix	25%	35.59	39.84	65.07	26.52
RAST-U	100%	35.51	40.04	65.57	27.19
RAST-U	75%	35.21	40.07	65.27	26.88
RAST-U	50%	36.09	40.32	65.79	27.53
RAST-U	25%	35.20	39.48	64.71	26.24
RAST-G	100%	37.81	41.88	66.55	29.36
RAST-G	75%	38.50	42.25	66.62	29.27
RAST-G	50%	37.84	41.50	66.42	28.95
RAST-G	25%	37.80	41.81	66.40	29.08
U-TTS	25%	35.77	40.09	65.97	27.67
G-TTS	25%	34.12	40.48	61.23	22.17
Mix-TTS	25%	33.87	38.38	58.46	19.19

Table 18: RASTBench scores and 95% group-bootstrap intervals. Score intervals and intervals for differences from Base are shown separately.

Config.	Scale	Score	95% CI	Δ vs Base	95% CI
Base	–	53.45	[52.26, 54.63]	–	–
RAST-Mix	100%	78.06	[77.10, 79.00]	+24.61	[+23.29, +25.97]
RAST-Mix	75%	78.25	[77.29, 79.17]	+24.80	[+23.47, +26.15]
RAST-Mix	50%	77.42	[76.47, 78.38]	+23.97	[+22.61, +25.33]
RAST-Mix	25%	75.42	[74.42, 76.41]	+21.97	[+20.64, +23.34]
RAST-U	100%	76.84	[75.85, 77.84]	+23.39	[+22.08, +24.73]
RAST-U	75%	75.86	[74.86, 76.82]	+22.41	[+21.08, +23.77]
RAST-U	50%	76.56	[75.56, 77.57]	+23.11	[+21.78, +24.52]
RAST-U	25%	75.20	[74.12, 76.24]	+21.75	[+20.39, +23.12]
RAST-G	100%	68.16	[66.94, 69.34]	+14.71	[+13.19, +16.24]
RAST-G	75%	69.34	[68.19, 70.49]	+15.89	[+14.40, +17.40]
RAST-G	50%	68.14	[66.92, 69.32]	+14.69	[+13.20, +16.23]
RAST-G	25%	67.60	[66.38, 68.79]	+14.15	[+12.67, +15.65]

Table 19: Score differences and 95% bootstrap intervals. Panel A compares quarter-scale RAST with its corresponding Pure-TTS control; Panel B compares quarter-scale and full-data RAST. Differences use $S_1 - S_2$.

Benchmark	Contrast	S_1	S_2	Δ	95% CI
A. Real-speech anchoring versus Pure-TTS					
RASTBench	U25 vs U-TTS	75.20	70.48	+4.72	[+3.60, +5.83]
RASTBench	G25 vs G-TTS	67.60	63.08	+4.52	[+3.18, +5.86]
RASTBench	Mix25 vs Mix-TTS	75.42	66.61	+8.81	[+7.54, +10.09]
WavBench	U25 vs U-TTS	67.97	56.49	+11.48	[+8.80, +14.17]
WavBench	G25 vs G-TTS	67.21	66.62	+0.59	[-2.56, +3.72]
WavBench	Mix25 vs Mix-TTS	71.09	57.16	+13.94	[+10.91, +16.96]
DeEAR	U25 vs U-TTS	26.24	27.67	-1.44	[-2.44, -0.43]
DeEAR	G25 vs G-TTS	29.08	22.17	+6.90	[+5.78, +8.00]
DeEAR	Mix25 vs Mix-TTS	26.52	19.19	+7.33	[+6.28, +8.40]
B. Quarter-scale versus full-data training					
RASTBench	U25 vs U100	75.20	76.84	-1.64	[-2.62, -0.70]
RASTBench	G25 vs G100	67.60	68.16	-0.56	[-1.86, +0.76]
RASTBench	Mix25 vs Mix100	75.42	78.06	-2.64	[-3.67, -1.64]

Table 20: Quarter-scale diagnostic differences (RAST minus its corresponding Pure-TTS control) with 95% group-bootstrap intervals.

Metric	Direction	Δ	95% CI
TTS Bench	U	+5.92	[+5.04, +6.77]
TTS Bench	G	+5.90	[+4.89, +6.88]
TTS Bench	Mix	+7.47	[+6.49, +8.44]
RH-B	U	+3.51	[+1.40, +5.59]
RH-B	G	+3.15	[+0.69, +5.70]
RH-B	Mix	+10.15	[+7.85, +12.52]
B-Macro	U	+6.50	[+5.56, +7.44]
B-Macro	G	+6.34	[+5.25, +7.41]
B-Macro	Mix	+8.27	[+7.20, +9.35]
D-Macro	U	+4.87	[+3.32, +6.47]
D-Macro	G	+5.41	[+3.70, +7.08]
D-Macro	Mix	+7.67	[+5.89, +9.40]

Table 21: Human A/B diagnostics. Gap averages the dimension-wise B-minus-A differences.

Config.	Scale	RH-Emo B	RH-Tone B	RH-A	RH-B	Gap
Base	–	27.14	49.56	47.20	38.35	-8.85
RAST-Mix	100%	79.61	77.05	76.82	78.33	+1.51
RAST-Mix	75%	79.61	77.93	77.97	78.77	+0.80
RAST-Mix	50%	79.38	76.15	78.21	77.77	-0.45
RAST-Mix	25%	76.59	74.84	76.84	75.72	-1.12
RAST-U	100%	75.93	75.87	75.19	75.90	+0.71
RAST-U	75%	76.60	72.98	76.06	74.79	-1.28
RAST-U	50%	76.48	74.60	75.11	75.54	+0.43
RAST-U	25%	73.75	72.87	75.21	73.31	-1.90
RAST-G	100%	63.03	66.44	66.82	64.73	-2.09
RAST-G	75%	64.17	70.62	65.83	67.40	+1.56
RAST-G	50%	64.10	65.66	64.00	64.88	+0.88
RAST-G	25%	63.14	66.22	65.11	64.68	-0.44
U-TTS	25%	71.13	68.47	71.19	69.80	-1.39
G-TTS	25%	59.87	63.19	62.94	61.53	-1.41
Mix-TTS	25%	65.08	66.06	65.84	65.57	-0.28

Table 22: Quarter-scale diagnostics. RH A/B compares test-time audio realizations; Pure-TTS controls replace training audio. Gap averages the dimension-wise B-minus-A differences.

Model	TTS Bench	RH-A	RH-B	Gap	B-Macro	D-Macro
RAST-U	77.09	75.21	73.31	-1.90	89.49	27.89
U-TTS	71.17	71.19	69.80	-1.39	82.99	23.02
RAST-G	70.52	65.11	64.68	-0.44	82.13	23.73
G-TTS	64.62	62.94	61.53	-1.41	75.79	18.32
RAST-Mix	75.12	76.84	75.72	-1.12	87.13	28.01
Mix-TTS	67.65	65.84	65.57	-0.28	78.86	20.34