# VOICEDIFF: EXPLAINING VOICE DIFFERENCES WITH SPEECH LANGUAGE MODELS

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

## ABSTRACT

As real-time spoken dialogue and multi-party assistants expand voice-based interaction, speech language models need to understand information about speakers beyond verbal content and use it in their responses. Because one person can sound different across utterances and different people can share similar characteristics, voice similarity must be distinguished from speaker identity. We introduce Voice Difference Explanation, a task covering speaker profiles such as gender and age, voice characteristics such as pitch and timbral brightness, and recording conditions such as noise and reverberation. It connects individual voice judgments to comparisons of shared and differing attributes, integrating them into a natural-language response alongside a separate identity verdict. Public speech metadata, acoustic measurements, and controlled recording conditions provide training targets and VoiceDiff-Bench. To study useful representations for these judgments, we introduce VoiceDiff, a reference model whose hierarchical speaker tokenizer combines a frozen speaker encoder's pooled speaker embedding with temporally resolved frame-level features. Task-specific supervision improves general Audio-LLMs on multiple voice judgments, while VoiceDiff achieves stronger pitch, timbral-brightness, and identity results than the evaluated task-adapted general models. At a matched audio-token budget, speaker embeddings favor identity judgments, frame-level features favor detailed acoustic judgments, and their combination supports both. Relation-description training integrates comparisons, summaries, and identity into one response with category-summary accuracy generally close to rule-based composition of individual outputs. Overall conclusions largely follow the comparisons stated in the response; error analysis identifies constituent comparison accuracy as a key basis for more accurate relation descriptions. The benefits of task-specific supervision also extend to selected external cases, unseen during task training, in which spoken claims conflict with vocal cues. These findings support task-specific supervision and complementary speech representations for grounding linguistic judgments and explanations in the voice itself.

## 1 INTRODUCTION

Voice-first systems are expanding spoken interaction from real-time dialogue to multi-party assistance and speech-conditioned action. Full-duplex speech models enable real-time spoken exchange (Défossez et al., 2024), while multi-party assistants and speech-conditioned robots distinguish users to inform dialogue and actions (Fu et al., 2025; Zhao et al., 2025). These systems need to understand not only verbal content but also what the voice reveals about the speaker. Without visual cues or separate user information, a voice provides speaker-related cues absent from the words alone. Understanding these cues extends beyond distinguishing speakers to recognizing characteristics such as pitch and timbre and their similarities and differences across utterances. To use this understanding in language, speech language models must select voice information relevant to a question and bring together characteristics and relationships across utterances in their responses.

Recent audio language model research is expanding from describing individual sounds to understanding and explaining relationships between them. ADIFF introduces natural-language explanations of audio differences (Deshmukh et al., 2025), and PolyAudio studies multi-audio comparison and cross-recording speaker linking with brief justifications (Kumar et al., 2026). Voice-related work includes descriptive profiling (Baali et al., 2025), free-form attribute comparisons (Pandey et al.,



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.

2025), and attribute-vector interpretations of similarity (Lee & Lee, 2025), as well as characteristic descriptions and rationales for speech evaluation and identity judgments (Manakul et al., 2026; Thebaud et al., 2026). For human voices, this relational perspective must account for both within-person variation and between-person resemblance. The same speaker can sound different across utterances (Lavan et al., 2019), while different speakers can share characteristics such as pitch or timbre. A voice-relation task should therefore explain shared and differing characteristics while keeping similarity distinct from speaker identity.

Connecting voice understanding to language responses requires studying what to extract from speech and how to train models to use it. Speaker-embedding similarity supports identity comparison (Yakovlev et al., 2024), but a single score does not reveal which characteristics agree or which voice is higher pitched or brighter. These judgments require speaker-discriminative cues alongside detailed acoustic information, while distinguishing recording conditions such as noise and reverberation that can obscure observations. Moreover, accepting audio input does not ensure that responses use vocal cues. VoxParadox (Pang et al., 2026) shows that several Audio-LLMs can follow spoken claims rather than acoustic evidence when they conflict, motivating direct evaluation of whether voice information supports responses. We study three linked questions: **(i) supervision:** which capabilities improve when models are explicitly trained to judge voice characteristics and differences; **(ii) representation:** which speech representations support different voice judgments and identity; and **(iii) integration:** how accurately and consistently learned comparisons and identity judgments can be conveyed in a single description.

We introduce **Voice Difference Explanation** to study these questions. The task covers **speaker profiles**—gender, age, and regional background; **voice characteristics**—pitch and timbral brightness; and **recording conditions**—noise and reverberation. **Voice Difference Understanding** identifies these attributes in individual voices, compares category agreement or the direction and magnitude of numerical differences, and judges speaker identity. **Voice Relation Description** integrates shared and differing characteristics while separately reporting recording conditions and identity. Metadata, acoustic measurements, and controlled recording conditions from public corpora provide training targets and **VoiceDiff-Bench**. Our reference model **VoiceDiff** uses a frozen speaker encoder's **speaker embedding**, pooled over the utterance, and temporally resolved **frame-level features**. A hierarchical speaker tokenizer maps both to language-model input tokens. The results show that supervision on the defined attributes and comparisons improves multiple voice judgments in general Audio-LLMs; VoiceDiff achieves stronger pitch, timbral-brightness, and identity results than the evaluated task-adapted general models. Comparing speaker-tokenizer adapters with the encoder and training protocol fixed reveals complementary strengths: speaker embeddings favor identity, and frame-level features favor detailed acoustic judgments. Their combination supports both at the same audio-token budget. Relation-description training integrates comparisons, summaries, and identity into one response with category-summary accuracy generally close to rule-based composition. Overall conclusions largely follow the stated comparisons; error analysis points to constituent comparison accuracy, rather than applying the summary rule, as the main target for improvement. Task-specific supervision also improves judgments on selected external cases, unseen during task training, where spoken claims conflict with vocal cues.

Our contributions are summarized as follows:

- **Voice Difference Explanation.** Voice Difference Explanation links individual attributes to explanations of shared and differing characteristics while distinguishing similarity, identity, and recording conditions. Explicit target criteria, training data, and VoiceDiff-Bench provide a common basis for learning and evaluating these capabilities.
- **Complementary speaker representations.** Controlled speaker-tokenizer adapter comparisons reveal distinct strengths of speaker embeddings and frame-level features for identity and detailed acoustic judgments. Our reference model VoiceDiff combines them to support both at a matched audio-token budget.
- **Utility and transfer of voice-centered supervision.** Training on the defined attributes and comparisons improves multiple voice judgments in general Audio-LLMs and transfers to selected external cases with conflicting spoken claims. Analysis of integrated descriptions further identifies constituent comparison accuracy as a key foundation for accurate relation explanations.

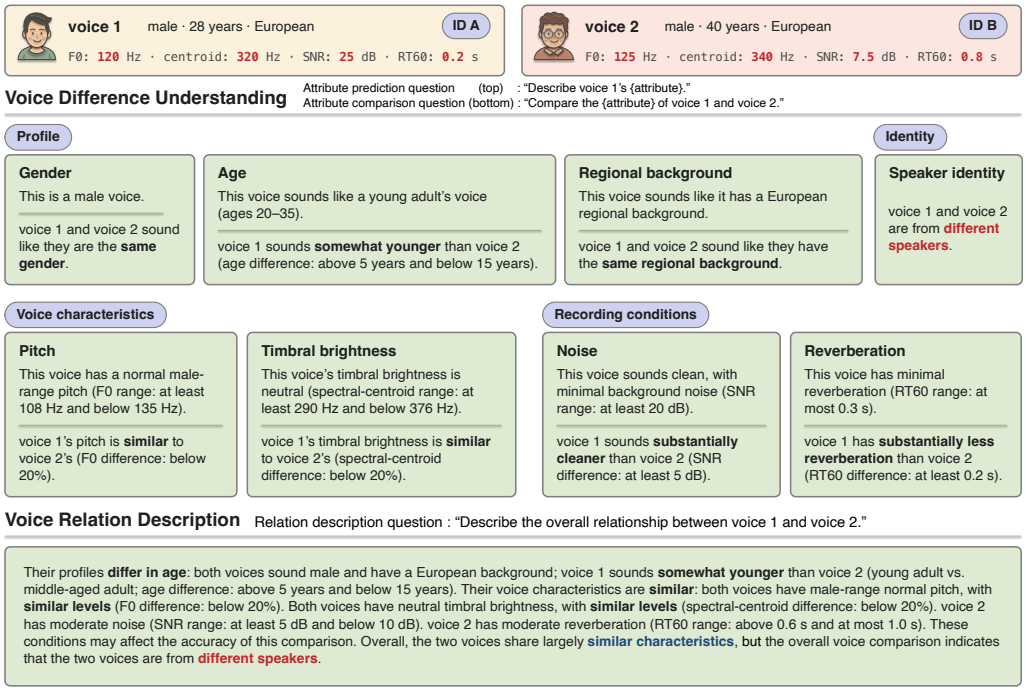



Figure 1: **Voice Difference Explanation, illustrated with target-generator outputs.** Attribute cards show single-voice targets for voice 1 above the divider and pairwise-comparison targets below; identity is pairwise only. The relation-description panel organizes these comparisons into shared and differing characteristics, recording conditions, and an overall relation alongside identity. Responses are unabridged generated targets, not model predictions. Illustrative metadata and measurements are reader references, not model inputs.

## 2 Voice Difference Explanation

**Voice Difference Explanation** asks models to identify a voice's speaker profile, voice characteristics, and recording conditions, compare these attributes between two voices, and integrate the comparisons into a natural-language description alongside a separately assessed speaker-identity verdict.

The attributes fall into three categories: **speaker profile** (gender, age, regional background), **voice characteristics** (pitch, timbral brightness), and **recording conditions** (noise, reverberation). The task distinguishes characteristic similarity from speaker identity: different speakers may share several characteristics, while utterances from the same speaker may differ in voice characteristics and recording conditions.

Figure 1 illustrates two complementary capabilities. **Voice Difference Understanding** comprises *Voice Attribute Prediction* for one voice, *Voice Attribute Comparison* for two voices, and speaker-identity judgment. Understanding questions can request one or several attributes. **Voice Relation Description** organizes these judgments into a requested category-level description or a full account alongside speaker identity. The figure shows target-generator examples constructed from illustrative metadata and measurements; the model receives only audio and a question.

### 2.1 Task Design

Metadata, acoustic measurements, and prespecified comparison and aggregation rules define the target expressions and evaluation criteria.

**Targets for Voice Attribute Prediction** are defined by mapping metadata and acoustic measurements to attribute-specific labels. Profile targets use available gender, age, and country metadata, with

countries grouped into broad regions. These age and regional references are metadata categories, not independent listener judgments of apparent age or accent. Pitch is represented by median F0 over voiced frames, and timbral brightness by a power-weighted spectral centroid, an acoustic proxy related to brightness (Schubert & Wolfe, 2006). Pitch and brightness label boundaries are fixed from training recordings before evaluation. Noise and reverberation are categorized by signal-to-noise ratio (SNR) and reverberation time (RT60), respectively. Sentence-form responses for voice characteristics and recording conditions include the interval and unit associated with each label, making its reference criterion explicit.

**Targets for Voice Attribute Comparison** specify categorical agreement or the direction and magnitude of a numerical difference. Gender and regional background compare category agreement; numerical differences are *similar*, *somewhat different*, or *substantially different*. *Similar* denotes a small-difference interval, not equality; direction is stated outside it. Comparisons use underlying numerical values rather than single-attribute label ranks, avoiding exaggerated differences near class boundaries. Age, SNR, and RT60 use absolute differences, while F0 and spectral centroid use relative differences. Sentence-form responses state the reference intervals and units, not exact predicted differences. Appendix A specifies the labels, comparison intervals, and acoustic boundary derivation.

**Targets for Voice Relation Description** follow a five-part structure for full responses: a *profile summary*, a *voice-characteristic summary*, a *recording-condition description*, an *overall-relation summary*, and a *speaker-identity verdict*, in that order. The first two summarize five attribute comparisons—gender, age, regional background, pitch, and timbral brightness—by identifying shared characteristics and differing attributes with their directions and magnitudes. To make the overall-relation summary consistent and traceable to these comparisons, we fix an explicit aggregation policy for this study. It applies attribute-specific weights directly to their categorical agreements and difference magnitudes, rather than to the category-summary text. Only the overall-relation summary takes one of three labels: **largely similar characteristics**, **similarities as well as differences**, or **a clear contrast**. These labels express the adopted description policy, not a universal perceptual similarity scale. Recording conditions are described separately because noise and reverberation can obscure voice cues, and speaker identity is judged independently from source speaker IDs; neither enters the overall-relation aggregate. Appendix B specifies the aggregation and response construction.

**Data for learning and evaluation.** The task requires both profile and identity references and measurable acoustic variation. The VoxCeleb domain supplies enriched profile metadata and speaker identities (Nagrani et al., 2017; Hechmi et al., 2021), while the LibriTTS-R domain, aligned with LibriTTS-P metadata (Koizumi et al., 2023; Kawamura et al., 2024), supports acoustic measurements and controlled recording conditions. These complementary sources support training and **VoiceDiff-Bench** under the same attribute and relation definitions. Appendix C details data construction and reference coverage; Section 4.1 specifies evaluation sets and metrics.

## 3 METHOD

VoiceDiff answers attribute and relation questions about one or two speech recordings, providing a reference model for studying the contribution of different speaker representations.

### 3.1 HIERARCHICAL SPEAKER TOKENIZER

Representations optimized for identity need not support every voice attribute equally. Our hierarchical speaker tokenizer combines two outputs of the same frozen speaker encoder: the *speaker embedding* after temporal pooling and final projection, and *frame-level features* before pooling. Two trainable branches map these aggregated speaker cues and temporal acoustic details to language-model tokens (Figure 2).

For a recording $x$, denote the speaker embedding by $e \in \mathbb{R}^{d_e}$ and features over $L$ frames by $H \in \mathbb{R}^{L \times d_h}$. An MLP expands $e$ into $N$ soft prefix tokens. A Q-Former (Li et al., 2023) uses learnable queries that cross-attend to $H$ to produce $M$ audio tokens. Both branches project to the language model's embedding dimension $d$, and their outputs are concatenated:

$$S^{(emb)} = \mathrm{MLP}(e), \qquad S^{(seq)} = \mathrm{QFormer}(H),$$
$$S = [S^{(emb)}; S^{(seq)}] \in \mathbb{R}^{(N+M)\times d}. \tag{1}$$

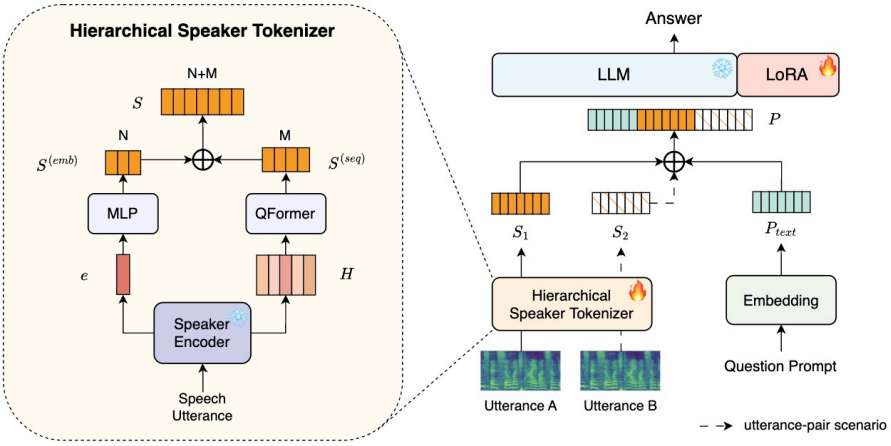



Figure 2: VoiceDiff architecture. The hierarchical speaker tokenizer combines aggregated speaker cues with temporal acoustic detail. An MLP and a Q-Former map the speaker embedding and frame-level features from the same frozen speaker encoder to token sequences, which are concatenated for the language model. The dashed path supports paired inputs.

We refer to both branches' continuous soft tokens in the language model's embedding space as audio tokens. Each voice's tokens accompany the question separately, without assigning attributes to either branch. Section 4.2 and Appendix F.1 compare these trainable mappings with the encoder fixed, including matched-token comparisons.

### 3.2 Training Objective

Let $P_b$ contain the audio tokens and question, and $y_b$ denote the target response. The token loss is $\ell_{b,t} = -\log p_\theta(y_{b,t} \mid P_b, y_{b,<t})$. Targets follow Section 2; metadata and measurements are not provided at inference.

Annotation coverage differs across corpora, so a recording may lack labels for some requested attributes. Rather than restrict multi-attribute questions to fully annotated examples, we retain complete response formats by sampling placeholder values within valid ranges for unlabeled attributes. Individual attribute and comparison spans dependent on these placeholders are masked from the loss. Available facts and fixed wording remain supervised; actual values are learned from labeled examples. Appendix D.2 gives the masking details.

**Stage 1: Voice Difference Understanding.** To connect audio representations to the requested judgments before adapting their verbal expression, we divide Stage 1 into two steps. Stage 1-1 trains attribute prediction, pairwise comparison, and identity judgment with short, label-centered targets, updating only the speaker tokenizer's trainable parameters. Stage 1-2 uses sentence-form targets for the same questions and also updates language-model LoRA parameters (Hu et al., 2022).

Questions request one, several, or all attributes. Multi-attribute targets combine individual answers; single-attribute questions use available labels (Appendix D.1). With $m_{b,t} = 1$ for supervised tokens and 0 otherwise,

$$\mathcal{L}_{\mathrm{S1}} = \frac{\sum_{b,t} m_{b,t}\ell_{b,t}}{\sum_{b,t} m_{b,t}}. \tag{2}$$

**Stage 2: Voice Relation Description.** Stage 2 fine-tunes the Stage 1 model to generate category-level and full descriptions directly from both voices' continuous audio tokens, without separately generating Stage 1 predictions as inputs. Training mixes these descriptions with replay of Stage 1-2-format questions and responses; full responses retain an independently supervised identity verdict.

To train full descriptions on partially annotated pairs, we supplement VoxCeleb metadata with F0 and spectral-centroid measurements. Existing noise and reverberation make these weak acoustic labels. When profile or voice-characteristic labels remain missing, sampled values complete the comparisons

used to construct the summary; their individual expressions remain masked. Unmeasured recording conditions receive a generic caveat, not sampled intensity labels. We downweight the weak acoustic labels and summaries derived from sampled values as $\mathrm{weak}$ supervision. Primary labels and their derived comparisons form $\mathrm{fact}$, and fixed wording forms $\mathrm{form}$. Each group's token losses are averaged separately so lengthy repeated wording does not dominate by token count. Over a batch of size $B$, we minimize

$$\mathcal{L}_{\mathrm{S2}} = \frac{1}{B}\sum_{b=1}^{B}\left(\bar{\ell}_b^{\mathrm{fact}} + 0.1\bar{\ell}_b^{\mathrm{weak}} + 0.2\bar{\ell}_b^{\mathrm{form}} + 2.0\ell_b^{\mathrm{id}}\right) \tag{3}$$

Here, $\bar{\ell}_b^{g}$ is the mean token loss for group $g$ in response $b$. $\ell_b^{\mathrm{id}}$ gives additional weight to the short identity judgment in full responses.

For full descriptions, only LoRA is updated: the entire response loss is blocked from the speaker tokenizer, while both voices' audio tokens remain inputs. This prevents summaries based on sampled values from directly updating the speaker tokenizer. Category descriptions and replay also update its trainable parameters. The speaker encoder and base language-model weights remain frozen. Appendix D details supervision and losses; Section 4.1 and Appendix E give training settings.

# 4 Experiments

## 4.1 Experimental Settings

**Data.** Training uses VoxCeleb1 development and LibriTTS-R `train-clean-100/360`. **VoiceDiff-Bench** contains 1,000 selected VoxCeleb1-O trials, 2,000 LibriTTS-R `test-clean` pairs, and 1,500 noise–reverberation conditions derived from those LibriTTS-R pairs, shared across stages. VoxCeleb metadata and speaker IDs support profile and identity evaluation; LibriTTS-R supports acoustic evaluation, with added MUSAN noise and SLR28 reverberation (Snyder et al., 2015; Ko et al., 2017). Pitch and brightness targets retain pre-augmentation values; recording-condition labels describe added degradation. Attribute-balanced scoring lists use available labels (Appendix C). We additionally evaluate compatible gender, age, and pitch tasks from the synthetic-speech benchmark VoxParadox (Pang et al., 2026), where spoken claims conflict with vocal evidence. No model uses VoxParadox for task adaptation.

**Implementation Details.** VoiceDiff combines frozen ReDimNet-B2 (Yakovlev et al., 2024) and Qwen2.5-1.5B-Instruct (Qwen et al., 2025). Its speaker tokenizer supplies 16 tokens from the speaker embedding and 32 from frame-level features of the same encoder per voice. Stage 1-1 trains only the tokenizer's trainable parameters; Stage 1-2 and Stage 2 also adapt the language model with LoRA (Hu et al., 2022). Main understanding and external results use Stage 1-1; relation descriptions use Stage 2. Appendix E reports optimization settings and checkpoint assignments; Appendix F.1 compares staged and direct joint training.

**Baselines.** We evaluate Qwen2.5-Omni-7B (Xu et al., 2025a), Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b), Kimi-Audio-7B-Instruct (KimiTeam et al., 2025), and Audio Flamingo 3 (AF3) (Ghosh et al., 2025) zero-shot. To test whether our supervision benefits general Audio-LLMs, we also adapt Qwen2.5-Omni, Kimi-Audio, and AF3 on our task data, updating their native audio-to-language connectors and language-model LoRA parameters. All general models receive the same questions, label definitions, and answer options while retaining model-native audio preprocessing; AF3 uses a two-input extension for paired voices. CoLMbo (Baali et al., 2025) is evaluated with its official gender and age prompts. SA-TinyLLaMA (Thebaud et al., 2026) is not included because we could not obtain a usable released checkpoint at evaluation time. Appendix E.3 gives the prompts and input handling.

**Evaluation.** On **VoiceDiff-Bench**, we score attribute prediction (Pred.), direction or *similar* (Dir.), and difference magnitude (Mag.); gender and region use same/different agreement (Match). Relation metrics cover category summaries and, on VoxCeleb pairs with all five comparison labels, overall accuracy, joint correctness with identity, and exact or relaxed comparison agreement. Policy consistency separately checks conclusions against the model's own comparisons. Because the task requires

Table 1: Voice Difference Understanding on VoiceDiff-Bench (accuracy, %↑). Pred.: attribute prediction; Match: same/different; Dir.: direction or *similar*; Mag.: difference magnitude; Acc.: identity verdict. Fine-tuning uses our data; tokenizer variants use Stage 1-1. Dashes denote unevaluated tasks. Bold/underlined: best/second-best displayed scores, including ties.

| Model | Gender | | Age | | | Region | | Pitch | | | Timbral Bright. | | | Noise | | | Reverb. | | | Identity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pred. | Match | Pred. | Dir. | Mag. | Pred. | Match | Pred. | Dir. | Mag. | Pred. | Dir. | Mag. | Pred. | Dir. | Mag. | Pred. | Dir. | Mag. | Acc. |
| **General Audio-LLMs** | | | | | | | | | | | | | | | | | | | | |
| *Zero-shot* | | | | | | | | | | | | | | | | | | | | |
| Qwen2.5-Omni-7B | 92.4 | 52.8 | 29.0 | 40.3 | 40.0 | 33.6 | 50.0 | 18.0 | 38.5 | 40.0 | 19.0 | 40.1 | 40.0 | 28.0 | 40.0 | 40.0 | 14.5 | 39.5 | 40.0 | 50.0 |
| Qwen3-Omni-30B | 96.7 | 95.4 | 44.7 | 45.7 | 49.7 | 59.0 | 74.8 | 22.0 | 47.7 | 40.1 | 24.9 | 43.2 | 39.6 | 39.8 | 42.0 | 39.8 | 32.1 | 33.7 | 30.1 | 54.7 |
| Kimi-Audio-7B | 93.5 | 66.4 | 27.3 | 31.7 | 38.0 | 50.8 | 65.6 | 24.7 | 35.5 | 36.3 | 21.7 | 34.4 | 34.4 | 22.4 | 38.1 | 40.0 | 25.6 | 27.3 | 28.5 | 72.6 |
| Audio Flamingo 3 | 86.0 | 51.4 | 26.3 | 20.7 | 21.3 | 31.5 | 50.2 | 22.9 | 23.9 | 23.3 | 20.3 | 36.3 | 36.4 | 20.6 | 39.2 | 38.9 | 20.0 | 29.0 | 29.4 | 50.6 |
| *Fine-tuning* | | | | | | | | | | | | | | | | | | | | |
| Qwen2.5-Omni-7B | 99.2 | 95.0 | 50.3 | **64.0** | 43.3 | 70.2 | 77.0 | 28.1 | 71.7 | 46.8 | 27.5 | 55.3 | 37.7 | 38.6 | 61.3 | 40.0 | 35.7 | 55.1 | 39.7 | 88.8 |
| Kimi-Audio-7B | **99.4** | 99.6 | **68.7** | 63.0 | 51.3 | **88.1** | 82.8 | 65.5 | 62.3 | 51.1 | 59.6 | 62.7 | 42.5 | **74.8** | **70.8** | **56.3** | 63.7 | 63.4 | 58.7 | 93.6 |
| Audio Flamingo 3 | 99.0 | 98.4 | 31.3 | 38.0 | 43.7 | 70.8 | 76.8 | 18.7 | 43.6 | 40.9 | 20.5 | 39.9 | 40.0 | 20.0 | 15.2 | 40.0 | 21.5 | 0.4 | 40.0 | 82.9 |
| **Speaker-specialized LLMs** | | | | | | | | | | | | | | | | | | | | |
| CoLMbo | 87.9 | – | 30.0 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| **VoiceDiff** | | | | | | | | | | | | | | | | | | | | |
| *Speaker tokenizer type (E: speaker embedding; F: frame-level features)* | | | | | | | | | | | | | | | | | | | | |
| Linear (E) | 94.3 | 45.6 | 41.3 | 12.7 | 12.3 | 47.1 | 44.0 | 25.9 | 28.1 | 17.3 | 24.0 | 19.9 | 19.9 | 26.0 | 21.5 | 21.9 | 2.4 | 27.9 | 27.6 | 44.2 |
| MLP (E) | 98.7 | **100.0** | 51.3 | 58.0 | 45.0 | 60.7 | 83.2 | 53.5 | 61.1 | 44.8 | 44.1 | 44.3 | 28.8 | 39.2 | 52.6 | 41.1 | 34.4 | 48.3 | 40.0 | 97.2 |
| Q-Former (F) | 99.1 | 99.8 | 63.0 | 61.7 | 46.3 | 67.8 | 86.2 | 78.3 | 85.3 | 69.9 | 67.5 | 76.3 | 50.4 | 72.2 | 68.2 | 54.8 | 66.3 | 68.7 | 57.5 | 94.9 |
| **Ours (E+F)** | 99.3 | **100.0** | 62.3 | 62.0 | **55.3** | 68.5 | **87.4** | **81.5** | **85.6** | **70.1** | **70.3** | **76.5** | **50.7** | 72.8 | 68.5 | 55.1 | **67.7** | **68.9** | **60.3** | **98.4** |

a same/different verdict in the response, we score that verdict's accuracy rather than continuous-score discrimination at the equal-error operating point (EER). Missing required answers count as errors (Appendix E.4).

On **VoxParadox**, we report acoustic-label accuracy (GT) and agreement with conflicting spoken claims (ALA). Gender and age use attribute questions; pitch uses three pairwise comparisons to reconstruct a three-segment order, reporting unresolved orders separately. This differs from the original whole-clip multiple-choice protocol (Appendix E.5); Appendix F.3 analyzes the errors.

### 4.2 Results

**Why Train for Voice Difference Understanding?** Task-specific supervision improves general Audio-LLMs across multiple voice attributes and comparisons (Table 1). Zero-shot models often identify gender but struggle with detailed acoustic attributes and difference direction and magnitude: recognizing a familiar attribute does not establish the ability to characterize differences.

Task adaptation improves Qwen2.5-Omni and Kimi-Audio on multiple prediction and comparison measures. Kimi-Audio is strong on age, regional background, and noise; VoiceDiff is stronger on pitch, timbral brightness, reverberation, and identity. Although gains vary by model and attribute, they extend beyond VoiceDiff: direct supervision supports judgments not reliably learned through general audio–language training alone.

**Which Representation Levels Capture Different Voice Attributes?** We compare speaker-tokenizer adapters at 48 tokens per voice—MLP-48, Q-Former-48, and hierarchical $16+32$—with the encoder, language model, and Stage 1-1 protocol fixed. The embedding-based MLP favors identity; the frame-based Q-Former improves pitch and brightness judgments with lower identity accuracy. Combining both improves pitch, brightness, reverberation prediction, and identity over the displayed single-level variants (Table 1). Strong identity performance alone therefore does not establish that a representation supports detailed acoustic comparisons.

The hierarchical speaker tokenizer also leads on joint direction-and-magnitude accuracy across the five ordered attributes, albeit by small margins in several cases. Increasing Q-Former tokens from

Table 2: Voice Relation Description on VoiceDiff-Bench (accuracy, %). (a) Profile and Voice Char.: profile and voice-characteristic summaries; Rec.: recording-condition description; Identity: speaker-identity verdict. Stage 1 + Rule composes Stage 1-2 attribute-wise outputs using the fixed policy. (b) Stage 2 on the VoxCeleb subset with all five comparison labels, including weak acoustic labels. Overall: overall-relation accuracy; Both: correct Overall and Identity. Complete generated examples follow.

**(a) Category summaries and identity**

| Response | Profile | Voice Char. | Rec. | Identity |
|---|---|---|---|---|
| Stage 1 + Rule | 62.90 | 66.07 | 53.40 | 98.30 |
| Stage 2 | 67.30 | 64.81 | 53.07 | 98.40 |

**(b) Stage 2 joint judgments**

| | |
|---|---|
| Overall | 71.18 |
| Identity | 98.49 |
| Both | 70.48 |

**Same speaker despite contrasting characteristics**

Their profiles are similar: both voices sound male and have a North American background. Both voices sound like middle-aged adults and are similar in age (age difference: at most 5 years). Their voice characteristics differ in pitch and timbral brightness: voice 1 has a substantially lower pitch than voice 2 (male-range low vs. male-range very high; F0 difference: at least 40%). voice 1's timbral brightness is substantially darker than voice 2's (very dark vs. very bright; spectral-centroid difference: at least 50%). Noise and reverberation are limited in both voices. Overall, the two voices **show a clear contrast**, **but** the overall voice comparison indicates that the two voices are from **the same speaker**.

**Different speakers despite similar characteristics**

Their profiles are similar: both voices sound female and have a North American background. Both voices sound like young adults and are similar in age (age difference: at most 5 years). Their voice characteristics are similar: both voices have female-range normal pitch, with similar levels (F0 difference: below 20%). Both voices have neutral timbral brightness, with similar levels (spectral-centroid difference: below 20%). Noise and reverberation are limited in both voices. Overall, the two voices **share largely similar characteristics**, **but** the overall voice comparison indicates that the two voices are from **different speakers**.

32 to 48 does not consistently improve comparisons (Appendix F.1): complementary representation levels, not simply more tokens, support acoustic and identity judgments together.

**How Accurately Does VoiceDiff Describe Voice Relations?** Stage 2 generates comparisons, summaries, and identity in one response. Against **Stage 1 + Rule**, which applies the fixed policy to attribute-wise outputs, profile-summary accuracy improves; voice-characteristic and recording-condition summaries and identity remain numerically close (Table 2a). The complete outputs illustrate similar characteristics across different speakers and contrasting characteristics within one speaker, preserving the distinction between similarity and identity.

Overall conclusions are evaluated on VoxCeleb pairs with labels for all five comparisons (Appendix C.3). Applying the fixed policy to profile metadata and acoustic measurements constructs overall-relation targets independently of model outputs; pitch and brightness remain weak labels because existing noise and reverberation can affect their measurement. Stage 2 achieves 71.18% overall-relation accuracy, and both the overall relation and identity are correct in 70.48% of responses (Table 2b). An accurate overall conclusion need not imply that every constituent comparison is correct.

Table 3: Stage 2 comparison agreement (%). Exact: direction and magnitude; Direction: direction or similarity; Adjacent: within one signed-category step. Gender/region require exact matches. Rows are cumulative.

| Matches | Exact | Direction | Adjacent |
|---|---|---|---|
| At least 3 | 95.18 | 97.49 | 99.60 |
| At least 4 | 71.18 | 82.43 | 95.08 |
| All 5 | 25.40 | 37.45 | 66.27 |

In the same responses, at least four of the five comparisons are exactly correct in 71.18%, but all five are correct in 25.40% (Table 3). All-five direction-or-similarity agreement is 37.45%, so magnitude errors alone do not explain the gap. Allowing one adjacent signed-category step raises all-five agreement to 66.27%, with at least four comparisons matching in 95.08%. This tolerance diagnoses error proximity, not strict directional correctness: it also permits confusion between *similar* and a small difference. Integrated descriptions thus contain several correct comparisons, but an account with all five correct requires more precise attribute judgments.

Table 4: Selected VoxParadox tasks (%). GT: acoustic-label accuracy; ALA: agreement with conflicting spoken claims; Unres.: unresolved pitch orders, counted as errors. Q2.5/Q3.0: Qwen2.5-/Qwen3-Omni; AF3: Audio Flamingo 3; VDIFF: VoiceDiff (Stage 1-1). FT uses our data; pitch uses pairwise reconstruction.

| Attribute | Metric | Zero-shot | | | | Task FT | | | VDIFF |
|---|---|---|---|---|---|---|---|---|---|
| | | Q2.5 | Q3.0 | Kimi | AF3 | Q2.5 | Kimi | AF3 | |
| Gender | GT ↑ | 0.5 | 63.5 | 72.9 | 1.0 | <u>99.0</u> | 96.4 | **100.0** | **100.0** |
| | ALA ↓ | 99.5 | 36.5 | 27.1 | 99.0 | 1.0 | 3.6 | **0.0** | **0.0** |
| Age | GT ↑ | 0.0 | 21.6 | 18.6 | 6.0 | **63.8** | 44.2 | 38.7 | <u>48.7</u> |
| | ALA ↓ | 80.9 | 61.3 | 30.2 | 88.9 | 29.1 | <u>5.5</u> | 27.6 | **5.0** |
| Pitch | GT ↑ | 0.0 | 0.0 | 0.0 | 0.0 | <u>41.5</u> | 7.0 | 2.0 | **82.0** |
| | ALA ↓ | 2.5 | 4.0 | 0.5 | 4.5 | 0.5 | **0.0** | 5.5 | **0.0** |
| | Unres. | 97.5 | 96.0 | 99.5 | 95.5 | 55.0 | 93.0 | 92.0 | 18.0 |

Applying the policy to generated comparisons reproduces the conclusion in 99.80% of cases. Almost all incorrect overall conclusions also follow the policy but contain incorrect comparisons (Appendix F.2). This output-level analysis supports Stage 2's intended organization of stated judgments: improving relation descriptions primarily requires more accurate comparisons rather than more consistent aggregation.

**Do Models Follow the Voice or the Words?** Voice-centered supervision improves judgments even when claims in spoken content conflict with vocal evidence. On VoxParadox (Pang et al., 2026), task adaptation raises gender accuracy from 0.5% to 99.0% for Qwen2.5-Omni, 72.9% to 96.4% for Kimi-Audio, and 1.0% to 100.0% for AF3; age accuracy also improves for all three (Table 4). VoiceDiff (Stage 1-1) reaches 100.0% on gender and reconstructs 82.0% of pitch orders, compared with 41.5% for task-adapted Qwen2.5-Omni. Its remaining 18.0% contain a *similar* comparison that prevents a strict order. The pitch-order advantage over task-adapted Qwen2.5-Omni also persists on clips passing the pre-inference segment-boundary check (Appendix F.3). Transfer is attribute-dependent: Qwen2.5-Omni FT exceeds VoiceDiff on age accuracy, despite VoiceDiff's lower agreement with conflicting spoken claims. Appendix F.3 analyzes these distinct errors. Under the derived evaluation, the benefit of voice-centered training thus extends beyond the training-data domains to previously unused conflict cases, supporting accurate vocal judgments rather than mere repetition of spoken claims.

**Additional Analyses.** On natural VoxCeleb recordings without pitch or brightness labels during understanding-stage training, Stage 1-2 predictions correlate with acoustic measurements (pitch $\rho = 0.902/0.876$ for male/female references; brightness $\rho = 0.740$; Appendix F.5). These correlations measure trends, not label accuracy, but suggest cross-corpus transfer despite missing attribute supervision in VoxCeleb. Stage 2 responds to $\pm$3-semitone changes with 78.5% joint pitch-comparison accuracy, versus 44.5/45.5% if its old response is retained (Appendix F.2). Strong combined noise and reverberation reduce Stage 1-2 pitch and brightness prediction accuracy by 6.5 and 8.0 percentage points, while gender and regional-background predictions remain comparatively stable (Appendix F.4). This attribute-dependent sensitivity motivates reporting recording conditions as context for voice judgments.

## 5 Conclusion

Voice Difference Explanation connects individual voice judgments to integrated descriptions while distinguishing characteristic similarity from identity. Voice-centered supervision benefits general Audio-LLMs and our reference model, whose speaker embeddings and frame-level features have complementary strengths. Relation-description training organizes learned comparisons into coherent accounts with category-summary accuracy close to rule-based composition. Most incorrect overall conclusions nevertheless follow the policy applied to incorrect comparisons, identifying the accuracy of constituent comparisons as a key direction for improvement. The evidence remains tied to corpus-specific labels, including weak acoustic labels for natural VoxCeleb recordings, and a fixed description policy rather than a validated perceptual scale. Broader joint annotations, user evaluation, and uncertainty-aware comparisons under diverse recording conditions are important next steps.

### AI USE STATEMENT

We used generative AI tools to support translation, language editing, figure preparation, and code development for research implementation and manuscript preparation, including debugging and LaTeX formatting. The authors reviewed the resulting materials, checked code through relevant tests and manuscript content against sources and experimental records, and take responsibility for the final text, claims, and research artifacts.

### ETHICS STATEMENT

VoiceDiff uses publicly available speech corpora to study voice attributes and relations. Age, gender, and regional-background targets follow corpus annotations rather than independent perceptual judgments; regional categories should not be interpreted as ethnicity or definitive nationality. Voice analysis raises privacy concerns and can facilitate profiling or surveillance. Our evaluation does not establish demographic fairness or suitability for consequential identity decisions, and model outputs should not serve as sole evidence for decisions about individuals.

### REPRODUCIBILITY STATEMENT

Section 3 describes the architecture and training objectives, and Section 4.1 summarizes the experimental setup. Appendices A and B specify attribute and relation labeling; Appendix C documents data sources, supervision coverage, and benchmark construction. Stage-specific prompts, targets, and missing-label training are provided in Appendix D, with model configurations, optimization settings, baseline adaptation, and evaluation procedures in Appendix E. Appendix F reports detailed ablations and supplementary analyses.

## REFERENCES

Massa Baali, Shuo Han, Syed Abdul Hannan, Purusottam Samal, Karanveer Singh, Soham Deshmukh, Rita Singh, and Bhiksha Raj. CoLMbo: Speaker language model for descriptive profiling. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2025. doi: 10.1109/ASRU65441.2025.11434804. URL https://ieeexplore.ieee.org/document/11434804/.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Interspeech 2018*, pp. 1086–1090, 2018. doi: 10.21437/Interspeech.2018-1929. URL https://www.isca-archive.org/interspeech_2018/chung18b_interspeech.html.

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. *arXiv preprint arXiv:2410.00037*, 2024. URL https://arxiv.org/abs/2410.00037.

Soham Deshmukh, Shuo Han, Rita Singh, and Bhiksha Raj. ADIFF: Explaining audio difference using natural language. In *International Conference on Learning Representations*, volume 2025, pp. 11332–11365, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1e337afbbb9fc2de993b88df0db6161e-Abstract-Conference.html.

Dongjie Fu, Xize Cheng, Linjun Li, Xiaoda Yang, Lujia Yang, and Tao Jin. PACHAT: Persona-aware speech assistant for multi-party dialogue. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29325–29342, 2025. doi: 10.18653/v1/2025.emnlp-main.1492. URL https://aclanthology.org/2025.emnlp-main.1492/.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41819–41886, 2025. doi: 10.52202/085713-1396. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3babb6b453cb59d87cb58a1219ef914b-Abstract-Conference.html.

Khaled Hechmi, Trung Ngo Trong, Ville Hautamäki, and Tomi Kinnunen. VoxCeleb enrichment for age and gender recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 687–693. IEEE, 2021. doi: 10.1109/ASRU51503.2021.9688085. URL https://ieeexplore.ieee.org/document/9688085.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Masaya Kawamura, Ryuichi Yamamoto, Yuma Shirahata, Takuya Hasumi, and Kentaro Tachibana. LibriTTS-P: A corpus with speaking style and speaker identity prompts for text-to-speech and style captioning. In *Interspeech 2024*, pp. 1850–1854, 2024. doi: 10.21437/Interspeech.2024-692. URL https://www.isca-archive.org/interspeech_2024/kawamura24_interspeech.html.

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. *arXiv preprint arXiv:2504.18425*, 2025. doi: 10.48550/arXiv.2504.18425. URL https://arxiv.org/abs/2504.18425.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224. IEEE, 2017. doi: 10.1109/ICASSP.2017.7953152. URL https://doi.org/10.1109/ICASSP.2017.7953152.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. LibriTTS-R: A restored multi-speaker text-to-speech corpus. In *Interspeech 2023*, pp. 5496–5500, 2023. doi: 10.21437/Interspeech.2023-1584. URL https://www.isca-archive.org/interspeech_2023/koizumi23_interspeech.html.

Sonal Kumar, Sreyan Ghosh, Yueqian Lin, S Sakshi, Ashish Seth, Yiran Chen, Ramani Duraiswami, and Dinesh Manocha. PolyAudio: Advancing multi-audio reasoning in large audio language models with interleaved multi-audio contexts. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42335–42353, 2026. doi: 10.18653/v1/2026.findings-acl.2101. URL https://aclanthology.org/2026.findings-acl.2101/.

Nadine Lavan, Luke F. K. Burston, and Lúcia Garrido. How many voices did you hear? natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, 110 (3):576–593, 2019. doi: 10.1111/bjop.12348. URL https://doi.org/10.1111/bjop.12348.

Jaejun Lee and Kyogu Lee. Vo-Ve: An explainable voice-vector for speaker identity evaluation. In *Interspeech 2025*, pp. 3988–3992, 2025. doi: 10.21437/Interspeech.2025-591. URL https://www.isca-archive.org/interspeech_2025/lee25e_interspeech.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Potsawee Manakul, Woody Haosheng Gan, Michael J Ryan, Ali Sartaz Khan, Warit Sirichotedumrong, Kunat Pipatanakul, William Barr Held, and Diyi Yang. AudioJudge: Understanding what works in large audio model based speech evaluation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3644–3663, 2026. doi: 10.18653/v1/2026.eacl-long.168. URL https://aclanthology.org/2026.eacl-long.168/.

Matthias Mauch and Simon Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 659–663. IEEE, 2014. doi: 10.1109/ICASSP.2014.6853678. URL https://doi.org/10.1109/ICASSP.2014.6853678.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pp. 2616–2620, 2017. doi: 10.21437/Interspeech.2017-950. URL https://www.isca-archive.org/interspeech_2017/nagrani17_interspeech.html.

Prabhat Pandey, Rupak Vignesh Swaminathan, K V Vijay Girish, Arunasish Sen, Jian Xie, Grant P. Strimel, and Andreas Schwarz. SIFT-50M: A large-scale multilingual dataset for speech instruction fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13921–13942, 2025. doi: 10.18653/v1/2025.acl-long.681. URL https://aclanthology.org/2025.acl-long.681/.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio LLMs listen or read? analyzing and mitigating paralinguistic failures with VoxParadox. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=v7rYbRR9Zw.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2025. URL https://arxiv.org/abs/2412.15115v2.

Emery Schubert and Joe Wolfe. Does timbral brightness scale with frequency and spectral centroid? *Acta Acustica united with Acustica*, 92:820–825, 2006. URL https://www.phys.unsw.edu.au/~jw/reprints/SchubertWolfe06.pdf.

David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015. URL https://arxiv.org/abs/1510.08484.

Thomas Thebaud, Yuzhe Wang, Laureano Moro-Velázquez, Jesús Villalba-Lopez, and Najim Dehak. Speaker verification with speech-aware LLMs: Evaluation and augmentation. In *Interspeech 2026*, pp. 4576–4581, 2026. doi: 10.21437/Interspeech.2026-2670. URL https://www.isca-archive.org/interspeech_2026/thebaud26_interspeech.html.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. arXiv preprint arXiv:2503.20215, 2025a. URL https://arxiv.org/abs/2503.20215.

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. arXiv preprint arXiv:2509.17765, 2025b. URL https://arxiv.org/abs/2509.17765.

Ivan Yakovlev, Rostislav Makarov, Andrei Balykin, Pavel Malov, Anton Okhotnikov, and Nikita Torgashov. Reshape dimensions network for speaker recognition. In *Interspeech 2024*, pp. 3235–3239, 2024. doi: 10.21437/Interspeech.2024-2116. URL https://www.isca-archive.org/interspeech_2024/yakovlev24_interspeech.html.

Wei Zhao, Pengxiang Ding, Min Zhang, Zhefei Gong, Shuanghao Bai, Han Zhao, and Donglin Wang. VLAS: Vision-language-action model with speech instructions for customized robot manipulation. In *The Thirteenth International Conference on Learning Representations*, pp. 51676–51693, 2025. URL https://openreview.net/forum?id=K4FAFNRpko.

# A ATTRIBUTE DEFINITIONS

This appendix explains how we assign labels to voice attributes and their differences. It presents the label tables, pitch and timbral-brightness measurements, difference calculations, and cutoff selection in that order. These criteria connect metadata and measurements to target expressions; stage-specific questions and actual response forms appear in Appendix D.1.

## A.1 ATTRIBUTE LABELS

Table 5 lists the labels used in our responses. Gender and regional background are taken from metadata. Age, pitch, and timbral brightness are labeled by numerical intervals; noise and reverberation use signal-to-noise ratio (SNR) and reverberation time (RT60), respectively. Square brackets include an endpoint, whereas parentheses exclude it.

Table 5: Attribute labels and criteria. F0 and spectral centroid are representative values for the selected audio segment. *Normal* and *neutral* denote the central pitch and timbral-brightness intervals, respectively.

**(a) Speaker profile**

| **Attribute** | **Labels and criteria** |
|---|---|
| Gender | male and female from metadata |
| Age | child: $[1, 10)$ years; adolescent: $[10, 20)$ years; young adult: $[20, 36)$ years; middle-aged adult: $[36, 56)$ years; older adult: at least 56 years |
| Regional background | North American; European; British / Irish; Latin / Hispanic; Oceanian; South Asian; East / Southeast Asian; Middle Eastern / North African; African |

**(b) Voice characteristics: labels and intervals in Hz**

| **Male-range pitch** | **Female-range pitch** | **Timbral brightness** |
|---|---|---|
| very low: below 94 | very low: below 161 | very dark: below 238 |
| low: $[94, 108)$ | low: $[161, 185)$ | dark: $[238, 290)$ |
| normal: $[108, 135)$ | normal: $[185, 221)$ | neutral: $[290, 376)$ |
| high: $[135, 164)$ | high: $[221, 253)$ | bright: $[376, 453)$ |
| very high: at least 164 | very high: at least 253 | very bright: at least 453 |

**(c) Recording conditions**

| **Noise** | **SNR (dB)** | **Reverberation** | **RT60 (s)** |
|---|---|---|---|
| clean | at least 20 | minimal | $[0, 0.3]$ |
| mild | $[10, 20)$ | slight | $(0.3, 0.6]$ |
| moderate | $[5, 10)$ | moderate | $(0.6, 1.0]$ |
| severe | $[0, 5)$ | heavy | $(1.0, 1.5]$ |
| extreme | below 0 | extreme | above 1.5 |

We map country metadata to the nine regional-background labels using fixed lookup tables in the data-construction code. These tables also standardize regional names. The labels are categories used in this task, not labels for spoken language or ethnicity. Countries without a mapping remain unlabeled.

Assigning a male-range or female-range pitch label requires both a valid F0 value and gender metadata. If gender is missing, we do not assign either range as ground truth. We can still compare two F0 values without gender metadata. Invalid or unknown values remain missing in the annotations; Appendix D.2 specifies how training responses retain their format while excluding unsupported individual values and comparisons from direct loss.

### A.2 Pitch and Timbral Brightness Measurement

When a pitch or timbral-brightness measurement is required, we compute it from the selected audio segment. The measured value determines the label in Table 5. We convert the segment to mono 16 kHz and measure it before padding to a common batch length.

Pitch is measured as fundamental frequency (F0). We use the median of valid positive pYIN F0 estimates (Mauch & Dixon, 2014), with a search range of 65–500 Hz, a frame length of 1,024 samples (64 ms), and a hop of 256 samples (16 ms). Frames are centered using constant padding.

For timbral brightness, we use the power-weighted spectral centroid: the average frequency weighted by spectral power. This choice follows the relationship between spectral centroid and perceived brightness (Schubert & Wolfe, 2006). We compute the centroid for each frame using an STFT with a Hann window, then take the median over valid voiced frames:

$$c_t = \frac{\sum_k f_k |X_{t,k}|^2}{\sum_k |X_{t,k}|^2}, \qquad c = \mathrm{median}_{t \in \mathcal{V}}\, c_t. \tag{4}$$

Here, $f_k$ is frequency, $X_{t,k}$ is an STFT coefficient, and $\mathcal{V}$ contains the valid voiced frames. We use the same frame spacing as for pitch and frequencies from 0 to 8 kHz. At least 20 frames must be voiced, and voiced frames must account for at least 5% of all frames. Centroid calculation uses only valid frames with positive energy. A measurement that fails these checks remains missing. Spectral centroid summarizes the spectral distribution in a way related to brightness; it neither captures every aspect of timbre nor removes the effects of recording conditions.

For LibriTTS-R, we measure the original segment before adding noise or reverberation and retain these measurements as the pitch and brightness targets after augmentation. We measure VoxCeleb segments with the same procedure; their existing noise and reverberation can affect the resulting references. These measurements are treated as weak references for full relation-description training, joint relation evaluation, and diagnostic analysis, rather than as ordinary acoustic ground truth for understanding-stage training. Appendix C gives the exact scope of use.

### A.3 Attribute Difference Labeling

For numerical attributes, we first calculate the difference between the two values and assign a magnitude label: *similar*, *somewhat different*, or *substantially different*. Outside the *similar* interval, we also state which voice has the higher or lower value. Thus, *similar* means a difference within the specified interval, not necessarily identical values.

For pitch and timbral brightness, we express how much larger one value is than the other as a percentage of the smaller value. For positive measurements $x_1, x_2$, this is

$$\delta(x_1, x_2) = 100 \left( \frac{\max(x_1, x_2)}{\min(x_1, x_2)} - 1 \right). \tag{5}$$

For age, SNR, and RT60, we use the absolute difference $\delta = |x_1 - x_2|$ in years, dB, and seconds, respectively. Table 6 gives the intervals for the three magnitude labels. Both short-form and sentence-form responses use these labels; sentence-form responses also state the unit and interval.

Table 6: Attribute-specific difference magnitudes. No direction is assigned within the *similar* interval. Otherwise, the response describes both direction and magnitude.

| **Attribute and unit** | **Similar** | **Somewhat different** | **Substantially different** |
|---|---|---|---|
| Age (years) | $\delta \leq 5$ | $5 < \delta < 15$ | $\delta \geq 15$ |
| Pitch (%) | $\delta < 20$ | $20 \leq \delta < 40$ | $\delta \geq 40$ |
| Timbral brightness (%) | $\delta < 20$ | $20 \leq \delta < 50$ | $\delta \geq 50$ |
| Noise (dB SNR) | $\delta \leq 3$ | $3 < \delta < 5$ | $\delta \geq 5$ |
| Reverberation (s RT60) | $\delta \leq 0.1$ | $0.1 < \delta < 0.2$ | $\delta \geq 0.2$ |

We describe direction from the voice specified in the question, or from voice 1 if none is specified. The expressions are younger/older, lower/higher, darker/brighter, cleaner/noisier, and less/more

reverberation. Lower SNR means more noise, whereas larger RT60 means more reverberation. Reversing the input order reverses direction but leaves the magnitude unchanged. For gender and regional background, we instead label whether the categories are the same or different. Speaker identity is determined separately from the original speaker IDs.

**Why we compare values rather than label names.** Two nearby values may fall on opposite sides of a labeling cutoff. For example, male-range F0 values of 100/108 Hz are labeled low/normal, but their 8% difference is *similar*. Conversely, 110/134 Hz are both normal, but their approximately 21.8% difference is *somewhat different*. Age comparisons likewise use numerical ages rather than the distance between age-group labels. This avoids treating every label change as a large difference or every shared label as similarity.

### A.4 Labeling Cutoff Selection

The five pitch and timbral-brightness labels each need four numerical cutoffs. We estimate these cutoffs from corpus measurements, then use the spacing between them to set the cutoffs for difference magnitudes.

**Measurements used to set the cutoffs.** Pitch cutoffs use training recordings from VoxCeleb and LibriTTS-R; timbral brightness uses LibriTTS-R training recordings because its measured distributions differ between the corpora. Held-out validation and test recordings are excluded from boundary estimation; the VoxCeleb development partition is used as training data. Duplicate utterances and invalid measurements are excluded. The targets for individual examples are measured from their selected input segments as described in Appendix A.2.

**Cutoffs for five-level attribute labels.** To keep speakers with many recordings from dominating the cutoffs, we give each speaker equal total weight within VoxCeleb1/2 (Nagrani et al., 2017; Chung et al., 2018) and within LibriTTS-R. We divide that weight equally among the speaker's sessions, then equally among the valid utterances within each session. Sessions are identified by VoxCeleb video IDs or LibriTTS-R chapter IDs. Weights are calculated separately for each attribute using its valid measurements.

We select the weighted 10th, 30th, 70th, and 90th percentiles, denoted Q10, Q30, Q70, and Q90. For each sorted measurement, its percentile position lies at the midpoint of its weight in the cumulative total; we linearly interpolate between these positions. Pitch percentiles are calculated by gender, separately for VoxCeleb1/2 and LibriTTS-R. We then average the corresponding percentiles in log space with equal weight. Brightness uses the LibriTTS-R percentiles alone. Table 7(a) reports the resulting cutoffs before rounding to integer Hz. We round these values to integer Hz to obtain the cutoffs in Table 5.

Table 7: Pitch and timbral-brightness labeling cutoffs. Panel (a) gives the four cutoffs before rounding to integer Hz. Panel (b) converts their average spacing into percentage differences and reports the final rounded cutoffs.

**(a) Single-attribute cutoffs before rounding (Hz)**

| Measured attribute | Q10 | Q30 | Q70 | Q90 |
|---|---|---|---|---|
| Male F0, VoxCeleb1/2 + LibriTTS-R | 93.937 | 108.061 | 134.584 | 164.263 |
| Female F0, VoxCeleb1/2 + LibriTTS-R | 160.976 | 184.913 | 220.536 | 253.329 |
| Spectral centroid, LibriTTS-R | 238.259 | 290.166 | 375.931 | 452.517 |

**(b) Difference cutoffs (%)**

| Attribute | One interval | Two intervals | Final cutoffs |
|---|---|---|---|
| Pitch | 18.38 | 40.13 | 20 / 40 |
| Timbral brightness | 23.84 | 53.36 | 20 / 50 |

**Cutoffs for attribute differences.** We use the average spacing between the four full-precision cutoffs $c_1, \ldots, c_4$, before integer-Hz rounding, as one unit of difference. Spacing is measured in log space so that equal frequency ratios have equal distances:

$$u = \frac{1}{3}\sum_{j=1}^{3} \ln \frac{c_{j+1}}{c_j} = \frac{1}{3} \ln \frac{c_4}{c_1}. \tag{6}$$

For pitch, we average the male and female values of $u$ with equal weight; for timbral brightness, we use the LibriTTS-R value of $u$. We convert one and two intervals to percentage differences using $100(e^u - 1)$ and $100(e^{2u} - 1)$, respectively. We then round *each cutoff independently to the nearest 10 percentage points*, giving 20%/40% for pitch and 20%/50% for brightness (Table 7(b)). The same rounded cutoffs are used to assign labels and to state the ranges in responses. Two intervals double the log distance, not necessarily the percentage difference.

These are the labeling rules adopted for this task, not measured human perceptual thresholds or cutoffs optimized to distinguish speakers. Age, SNR, and RT60 use the separately specified difference intervals in Table 6.

# B Relation Descriptions

A full response for Voice Relation Description presents five components in order: profile summary, voice-characteristic summary, recording-condition description, overall-relation summary, and speaker-identity verdict. The fixed policy below makes the overall-relation summary traceable to the individual comparisons retained in the response, while keeping characteristic similarity separate from identity. Figure 1 shows a complete target example.

## B.1 Summarizing Voice Similarities and Differences

The profile and voice-characteristic descriptions first name the attributes that differ, then describe the shared values and differences. If none differ within a category, the response states that the category is similar. This opening helps the reader locate the main differences; it does not assign a separate numerical score to the category. Each comparison uses the directions, magnitudes, and numerical intervals defined in Appendix A.

To choose the overall-relation summary phrase, we combine the differences in gender, age, regional background, pitch, and timbral brightness directly, rather than using the preceding summary text:

$$Q = 4g + a + r + p + b. \tag{7}$$

The terms $g$ and $r$ indicate whether the gender and regional-background labels differ: each is 0 when the labels match and 1 when they differ. The terms $a, p, b$ describe age, pitch, and timbral-brightness differences, respectively. Each is 0 for *similar*, 1 for *somewhat different*, and 2 for *substantially different*. Here, $r$ denotes regional difference, not RT60. Recording conditions and speaker identity do not contribute to $Q$.

Table 8: Overall-relation summary phrases selected by $Q$. Each expression follows "the two voices." The score summarizes attribute differences; it is not a same-speaker probability.

| Score | Summary | Response wording |
|---|---|---|
| $Q \leq 1$ | Largely similar characteristics | share largely similar characteristics |
| $2 \leq Q \leq 3$ | Similarities and differences | share similarities as well as differences |
| $Q \geq 4$ | Clear contrast | show a clear contrast |

A gender-category difference contributes four points and therefore produces a clear contrast on its own. A broad regional difference contributes one point and cannot do so alone. Age, pitch, and timbral-brightness differences contribute according to their magnitudes. The overall phrase can therefore describe largely similar characteristics despite one small difference, while the preceding category description still states that difference. These weights and score ranges are rules adopted for this task, not estimates of how people weigh the importance of each attribute.

**Missing attribute comparisons.** We require all five attribute comparisons to calculate a reference overall relation. A missing comparison is not treated as a match or assigned zero points. Appendix D.2 explains training with incomplete labels. The joint-reference subset and the distinction between reference agreement and consistency with the model's own comparisons are specified in Appendices C.3 and E.4, respectively.

### B.2 Recording Conditions and Speaker Identity in Descriptions

**Recording conditions.** The response describes noise and reverberation separately to tell the reader which conditions may interfere with the voice comparison. These conditions do not contribute to $Q$. Their labels determine how strongly the response warns about their possible effect: clean/mild noise and minimal/slight reverberation form the low group, moderate labels the intermediate group, and severe/extreme noise and heavy/extreme reverberation the high group.

When all four labels—noise and reverberation for each voice—are known, the highest group determines the warning. If all four are low, the response summarizes noise and reverberation as limited. Otherwise, it names the conditions at moderate or higher levels and identifies the affected voice. Matching labels for both voices are combined into one sentence.

For the intermediate group, the response adds "These conditions may affect the accuracy of this comparison." For the high group, it adds "These conditions may obscure or distort the voice cues used in this comparison, making it less reliable." When only some conditions are known, the response describes those conditions but does not assign an overall warning level. It instead adds "Noise and reverberation can affect the voice cues used in this comparison." When all conditions are unknown, it uses only this sentence. These warnings identify conditions that may make voice cues harder to discern. They do not estimate an individual response's error probability or determine the cause of an actual error.

**Speaker identity.** The final sentence states both the overall similarity or difference and the speaker-identity decision. It uses *but* when the voices have *largely similar characteristics but belong to different speakers*, or *show a clear contrast but belong to the same speaker*. It uses *and* for every other combination:

> Overall, the two voices {overall phrase}, {connector} the overall voice comparison indicates that the two voices are from {identity}.

The speaker-identity label comes from the original speaker IDs, which are not provided as model inputs. The connector expresses how the two judgments relate. It does not claim that the preceding attribute descriptions causally explain the model's internal speaker-identity decision.

Stage-specific questions and representative short-form and sentence-form targets are collected in Appendix D.1.

## C Data and Benchmark Construction

This appendix explains which recordings and targets are used for training and how we construct the fixed evaluation inputs in VoiceDiff-Bench. Labeling criteria are given in Appendices A and B; scoring rules are given in Appendix E.4. Inputs and comparison conditions for additional validation studies are described separately in Appendix E.5.

### C.1 Data Sources and Annotations

Training uses VoxCeleb1 development and LibriTTS-R `train-clean-100/360`. VoxCeleb supplies enriched age and gender metadata and country information mapped to regional background (Nagrani et al., 2017; Hechmi et al., 2021). We align LibriTTS-P metadata with the corresponding LibriTTS-R recordings (Koizumi et al., 2023; Kawamura et al., 2024); LibriTTS-P is not an additional audio corpus.

Table 9 counts recordings of at least one second, eligible as either primary inputs or comparison partners. These are source recordings, not generated questions or pairs. Speaker counts use corpus-specific IDs and do not establish person-level disjointness across corpora.

Table 9: Training recordings and primary training labels. A label is used only when its metadata or measurement is valid.

| Source | Recordings | Speaker IDs | Primary training labels |
|---|---|---|---|
| VoxCeleb1 development | 148,642 | 1,211 | Gender, age, regional background, and speaker identity. |
| LibriTTS-R train-clean-100/360 | 142,692 | 1,147 | Gender and speaker identity; measured pitch and timbral brightness; added noise and reverberation. |
| Total | 291,334 | – | |

**Attribute targets.** LibriTTS-R pitch and timbral brightness are measured on the original audio crop, before adding noise or reverberation. Their targets remain unchanged after augmentation and describe the original segment, not a speaker-invariant property. Noise and reverberation labels follow the added conditions and the intervals in Appendix A, rather than measurements of the conditions already present. Assigning *clean/minimal* when no noise or reverberation is added does not mean that the original recording is entirely noise- or reverberation-free. The representative SNR assigned to *clean* in this case is a labeling convention, not a measured source SNR.

**Supplementary VoxCeleb measurements.** When training Stage 2 full relation descriptions, we measure F0 and spectral centroid on the selected VoxCeleb crops without additional augmentation. Because the original recording conditions affect these measurements, labels derived from them are treated as weak acoustic labels rather than ground truth for ordinary attribute questions, category-specific questions, or Stage 1-2 replay. Unmeasured VoxCeleb noise and reverberation remain unlabeled.

Missing metadata and failed measurements remain unavailable; they are not assigned a class or treated as *similar*. Values used to fill training responses are kept separate from actual annotations. Section 3 explains their treatment during learning.

### C.2 Training Example Construction

**Selecting voices and questions.** Questions, comparison partners, and responses are generated online. The question type and requested attributes guide partner selection to cover the target classes and relations. Requested crops span 1–15 seconds; shorter sources are not repeated to extend them. Whole-recording measurements may guide sampling, but acoustic targets use valid measurements of the selected crops. Speaker identity follows speaker-ID agreement. Ordinary same-speaker pairs use different recordings; recording-condition-only comparisons may use two copies of one crop with different augmentations, without supervising speaker identity on those examples.

**Measuring and augmenting the audio.** We convert crops to mono, 16 kHz and measure acoustic attributes only when requested. The order is *original-crop measurement* → *reverberation* → *noise addition* → *batch padding*. Targets are not remeasured after augmentation. A failed measurement removes only that attribute's target, not other valid targets. We apply no additional peak, RMS, or loudness normalization. Appendix A.2 gives the measurement procedures and validity criteria.

Noise uses non-speech MUSAN recordings at a specified SNR; reverberation uses SLR28 room impulse responses (RIRs) with precomputed RT60 (Snyder et al., 2015; Ko et al., 2017). For single-attribute condition questions, we first select a severity class. For comparisons, we select conditions realizing the requested direction and magnitude. This reduces bias from unequal file counts across severity classes. Appendix E.2 reports the sampling and augmentation probabilities.

### C.3 VOICEDIFF-BENCH CONSTRUCTION

VoiceDiff-Bench uses VoxCeleb1-O and LibriTTS-R `test-clean`. Table 10 separates original pairs from augmented versions of 1,500 LibriTTS-R pairs. The total counts 4,500 pair-input conditions, not independent pairs. Both stages use the same fixed inputs.

Table 10: VoiceDiff-Bench pair inputs. Same and Different refer to speaker identity. Augmented inputs reuse original LibriTTS-R pairs.

| Input condition | Same | Different | Total |
|---|---|---|---|
| Selected VoxCeleb1-O trials | 500 | 500 | 1,000 |
| LibriTTS-R original pairs | 1,000 | 1,000 | 2,000 |
| LibriTTS-R augmented pairs | 750 | 750 | 1,500 |
| Total input conditions | 2,250 | 2,250 | 4,500 |

**Pair selection and augmentation.** The source lists contain 2,000 VoxCeleb1-O recordings from 40 speakers and 2,573 LibriTTS-R recordings from 38 speakers. VoxCeleb1-O trials are selected using metadata alone, not model outputs or scores (seed 4221), retaining their input order and identity labels. LibriTTS-R same-speaker pairs use different utterances; its 1,000 different-speaker original pairs include 800 gender-matched and 200 cross-gender pairs.

Each of 1,500 distinct LibriTTS-R original pairs receives one additional condition: noise only, reverberation only, or both, with 500 examples each. The recording grades R0/R1/R2 (low/intermediate/high) also contain 500 examples each. Grades follow the strongest condition across both voices (Appendix B). These are separate balances, not balanced type–grade combinations, individual attribute classes, or attribute differences. Reused recordings, speakers, and original/augmented pairs make examples statistically dependent.

**Fixed input preparation.** Recordings must be at least two seconds long. We retain the full recording up to fifteen seconds and the central fifteen seconds otherwise. Crops, input order, and augmentation settings are fixed. Audio is mono, 16 kHz; reverberation precedes noise, with no final peak normalization or clipping. Source recordings, within-corpus speaker IDs, and noise/RIR files are disjoint from the task-training lists. This file separation does not establish unseen rooms or out-of-domain recording environments. VoiceDiff-Bench applies fixed questions, comparison pairs, and description criteria for this task to existing public evaluation corpora.

**Attribute questions.** Table 11 lists the fixed questions used for attribute prediction, comparison, and identity evaluation: 5,595 single-voice and 5,800 pair questions, all using the canonical prompt. VoxCeleb1-O supplies profile and identity targets. Its acoustic measurements are excluded from these primary attribute-question scores and used separately as weak references in the joint relation-description analysis. LibriTTS-R supplies gender, identity, acoustic, and recording-condition targets, although reported identity evaluation uses only the selected VoxCeleb1-O trials. Acoustic targets for augmented LibriTTS-R inputs still use the original crop measurements from Appendix C.1. Missing targets are not filled for evaluation.

Prediction questions balance gender, the four age classes and five regions present in the evaluation data, and the five levels of each acoustic attribute; pitch is balanced within gender. Comparison questions balance *same/different*, or the five combinations of *similar* and *somewhat/substantially different* in either direction. Label balance does not imply demographic representativeness: age and region retain metadata uncertainty and speaker concentration.

**Category-summary evaluation.** Category-summary evaluation uses 1,000 VoxCeleb1-O pairs for profiles, 1,998 LibriTTS-R original pairs with complete acoustic targets for voice characteristics, and the 1,500 augmented pairs for recording conditions. Each category target is constructed from the attributes required for that category. These category accuracies are not pooled into an overall-relation accuracy. Appendix E.4 specifies scoring.

Table 11: VoiceDiff-Bench questions by evaluation item and source. Single denotes attribute prediction; Pair denotes comparison or identity judgment. A dash indicates no questions in the reported evaluation.

| Evaluation item | VoxCeleb1-O | | LibriTTS-R | |
|---|---|---|---|---|
| | Single | Pair | Single | Pair |
| Gender | 1,000 | 500 | 1,000 | – |
| Age | 300 | 300 | – | – |
| Regional background | 295 | 500 | – | – |
| Pitch | – | – | 750 | 750 |
| Timbral brightness | – | – | 1,000 | 750 |
| Noise | – | – | 500 | 1,000 |
| Reverberation | – | – | 750 | 1,000 |
| Speaker identity | – | 1,000 | – | – |
| Total | 1,595 | 2,300 | 4,000 | 3,500 |

**Joint references for relation descriptions.** We separately analyze the same VoxCeleb1-O evaluation inputs using profile metadata and acoustic measurement references together. Of the 1,000 pairs, 996 have gender, age, and region metadata and valid pitch and brightness references for both voices. Four pairs lack required acoustic references and are excluded independently of model correctness; missing or contradictory model answers remain errors. We apply the fixed policy in Appendix B to these references to obtain the overall-relation target. Pitch and brightness are measured on the actual input segments using Appendix A.2. Existing noise and reverberation can affect these measurements, so they remain weak references rather than listener-annotated or clean-signal ground truth. Recording conditions are not part of this five-attribute joint evaluation. Table 2(b) and Table 3 analyze the same saved Stage 2 responses, not two independently sampled test sets.

# D Training Targets and Supervision

This appendix details the stage-wise prompts and targets, learning with missing labels, and loss calculation introduced in Section 3. Attribute labels and difference criteria are defined in Appendix A, relation composition in Appendix B, and data construction in Appendix C. Input lengths, padding, token counts, and other model and training settings appear in Appendix E.

## D.1 Prompts and Targets

Stage 1-1 and Stage 1-2 use the same questions with short-form and sentence-form targets, respectively. Stage 2 uses sentence-form targets for category-level and full relational descriptions. The examples below are produced by the actual target generator, not by a trained model. The model receives audio and a question; the metadata used to construct these examples are not part of its input.

**Stage 1: Voice Difference Understanding.** We use the constructed metadata example from Figure 1. The voices have different speaker IDs and are both male with European regional backgrounds. For voices 1/2, the ages are 28/40 years, F0 values 120/125 Hz, spectral centroids 320/340 Hz, SNRs 25/7.5 dB, and RT60 values 0.2/0.8 s. These values instantiate the templates; they are not measurements of recorded audio.

The pitch example above has a small difference. For F0 values of 100/130 Hz, the same question, "Compare the pitch of the two voices.", instead yields the short target "Pitch: somewhat different; voice 1 is lower than voice 2" and the sentence target "voice 1's pitch is somewhat lower than voice 2's (F0 difference: at least 20% and below 40%)."

**Multiple requested attributes.** As described in Section 3.2, a question may request one attribute, a subset, or all attributes. Stage 1-1 lists short answers for the requested attributes; Stage 1-2

Table 12: Single-attribute prompts and targets. Each question is shared across the two response forms.

| **Stage 1-1: short target** | **Stage 1-2: sentence target** |
|---|---|
| *Question:* Describe the gender of this voice. | |
| Gender: male | This is a male voice. |
| *Question:* Describe the age of this voice. | |
| Age group: young adult | This voice sounds like a young adult's voice (ages 20–35). |
| *Question:* Describe the regional background of this voice. | |
| Regional background: European | This voice sounds like it has a European regional background. |
| *Question:* Describe the pitch of this voice. | |
| Pitch: normal (male range) | This voice has a normal male-range pitch (F0 range: at least 108 Hz and below 135 Hz). |
| *Question:* Describe this voice's timbral brightness. | |
| Timbral brightness: neutral | This voice's timbral brightness is neutral (spectral-centroid range: at least 290 Hz and below 376 Hz). |
| *Question:* Describe the noise level in this voice. | |
| Noise: clean | This voice sounds clean, with minimal background noise (SNR range: at least 20 dB). |
| *Question:* Describe the reverberation in this voice. | |
| Reverberation: minimal | This voice has minimal reverberation (RT60 range: at most 0.3 s). |

Table 13: Attribute-comparison and identity prompts and targets. Parenthesized intervals specify the difference-magnitude criteria.

| **Stage 1-1: short target** | **Stage 1-2: sentence target** |
|---|---|
| *Question:* Compare the genders of the two voices. | |
| Gender: same | voice 1 and voice 2 sound like they are the same gender. |
| *Question:* Compare the ages of the two voices. | |
| Age group: somewhat different; voice 1 is younger than voice 2 | voice 1 sounds somewhat younger than voice 2 (age difference: above 5 years and below 15 years). |
| *Question:* Compare the regional backgrounds of the two voices. | |
| Regional background: same | voice 1 and voice 2 sound like they have the same regional background. |
| *Question:* Compare the pitch of the two voices. | |
| Pitch: similar | voice 1's pitch is similar to voice 2's (F0 difference: below 20%). |
| *Question:* Compare the timbral brightness of the two voices. | |
| Timbral brightness: similar | voice 1's timbral brightness is similar to voice 2's (spectral-centroid difference: below 20%). |
| *Question:* Compare the noise levels of the two voices. | |
| Noise: substantially different; voice 1 is cleaner than voice 2 | voice 1 sounds substantially cleaner than voice 2 (SNR difference: at least 5 dB). |
| *Question:* Compare the reverberation in the two voices. | |
| Reverberation: substantially different; voice 1 has less reverberation than voice 2 | voice 1 has substantially less reverberation than voice 2 (RT60 difference: at least 0.2 s). |
| *Question:* Are voice 1 and voice 2 from the same speaker? | |
| Speaker identity: different | voice 1 and voice 2 are from different speakers. |

concatenates their sentence-form answers in the prescribed order. The following examples use the same metadata as the single-attribute tables, preserving the list markers returned by the generator.

Table 14: Subset questions and their two target forms, for single-voice description and pairwise comparison.

| **Stage 1-1: short target** | **Stage 1-2: sentence target** |
|---|---|
| *Question:* Describe this voice's pitch and noise level. | |
| - Pitch: normal (male range)<br>- Noise: clean | This voice has a normal male-range pitch (F0 range: at least 108 Hz and below 135 Hz).<br>This voice sounds clean, with minimal background noise (SNR range: at least 20 dB). |
| *Question:* Compare the pitch and noise levels of the two voices. | |
| - Pitch: similar<br>- Noise: substantially different; voice 1 is cleaner than voice 2 | voice 1's pitch is similar to voice 2's (F0 difference: below 20%).<br>voice 1 sounds substantially cleaner than voice 2 (SNR difference: at least 5 dB). |

Table 15: All-attribute description of one voice, with the same question in both substages.

| **Stage 1-1: short target** | **Stage 1-2: sentence target** |
|---|---|
| *Question:* Describe this voice. | |
| - Gender: male<br>- Age group: young adult<br>- Regional background: European<br>- Pitch: normal (male range)<br>- Timbral brightness: neutral<br>- Noise: clean<br>- Reverberation: minimal | This is a male voice.<br>This voice sounds like a young adult's voice (ages 20–35).<br>This voice sounds like it has a European regional background.<br>This voice has a normal male-range pitch (F0 range: at least 108 Hz and below 135 Hz).<br>This voice's timbral brightness is neutral (spectral-centroid range: at least 290 Hz and below 376 Hz).<br>This voice sounds clean, with minimal background noise (SNR range: at least 20 dB).<br>This voice has minimal reverberation (RT60 range: at most 0.3 s). |

Table 16: All-attribute comparison of two voices. Speaker identity is included last in both target forms.

| **Stage 1-1: short target** | **Stage 1-2: sentence target** |
|---|---|
| *Question:* Compare the two voices. | |
| - Gender: same | voice 1 and voice 2 sound like they are the same gender. |
| - Age group: somewhat different; voice 1 is younger than voice 2 | voice 1 sounds somewhat younger than voice 2 (age difference: above 5 years and below 15 years). |
| - Regional background: same | voice 1 and voice 2 sound like they have the same regional background. |
| - Pitch: similar | voice 1's pitch is similar to voice 2's (F0 difference: below 20%). |
| - Timbral brightness: similar | voice 1's timbral brightness is similar to voice 2's (spectral-centroid difference: below 20%). |
| - Noise: substantially different; voice 1 is cleaner than voice 2 | voice 1 sounds substantially cleaner than voice 2 (SNR difference: at least 5 dB). |
| - Reverberation: substantially different; voice 1 has less reverberation than voice 2 | voice 1 has substantially less reverberation than voice 2 (RT60 difference: at least 0.2 s). |
| - Speaker identity: different | voice 1 and voice 2 are from different speakers. |

An all-attribute Stage 1 response combines individual judgments, whereas a Stage 2 target organizes the same comparisons into category summaries and an overall-relation summary. All requested labels are available in these examples. Appendix D.2 explains how supervision changes when the response format is retained with missing labels.

**Stage 2: Voice Relation Description.** Stage 2 organizes comparisons around shared and differing characteristics. Recording conditions are described separately as factors that may affect the comparison, rather than being added to the overall similarity of voice attributes.

Table 17: Category-level relation prompts and targets.

| Category and question | Target |
|---|---|
| Profile<br>Compare the profiles of the two voices. | Their profiles differ in age: both voices sound male and have a European background; voice 1 sounds somewhat younger than voice 2 (young adult vs. middle-aged adult; age difference: above 5 years and below 15 years). |
| Voice characteristics<br>Compare the voice characteristics of the two voices. | Their voice characteristics are similar: both voices have male-range normal pitch, with similar levels (F0 difference: below 20%). Both voices have neutral timbral brightness, with similar levels (spectral-centroid difference: below 20%). |
| Recording conditions<br>Describe the noise and reverberation in the two voices. | voice 2 has moderate noise (SNR range: at least 5 dB and below 10 dB). voice 2 has moderate reverberation (RT60 range: above 0.6 s and at most 1.0 s). These conditions may affect the accuracy of this comparison. |

A full response brings these observations together and concludes with both the overall-relation summary and speaker-identity verdict. In this example, largely similar characteristics do not change the different-speaker target.

Table 18: Full relation-description prompt and target for the same constructed pair.

| Prompt and target |
|---|
| Explain the relationship between the two voices.<br>Their profiles differ in age: both voices sound male and have a European background; voice 1 sounds somewhat younger than voice 2 (young adult vs. middle-aged adult; age difference: above 5 years and below 15 years). Their voice characteristics are similar: both voices have male-range normal pitch, with similar levels (F0 difference: below 20%). Both voices have neutral timbral brightness, with similar levels (spectral-centroid difference: below 20%). voice 2 has moderate noise (SNR range: at least 5 dB and below 10 dB). voice 2 has moderate reverberation (RT60 range: above 0.6 s and at most 1.0 s). These conditions may affect the accuracy of this comparison. Overall, the two voices share largely similar characteristics, but the overall voice comparison indicates that the two voices are from different speakers. |

This example has complete labels. In training, the same response format can contain expressions with different levels of supervision, as specified below.

### D.2 Training with Missing Labels

**Responses with missing attribute labels.** The corpora provide different annotation coverage: profile metadata, acoustic measurements, and controlled recording-condition labels are not jointly available for every recording. Individual metadata or measurements can also be missing. To train complete multi-attribute response formats on these examples, we distinguish available labels from sampled completions. Single-attribute questions use attributes with available labels. A multi-attribute question may also request attributes without labels; completion values are sampled within valid ranges in a separate context and are not added to the original metadata as new ground truth.

For example, when gender is male but age is missing, the generator can construct the following response. The age is sampled only to complete its format.

Table 19: Missing-age example. The sampled age is not a factual target.

| Question | Target |
|---|---|
| Describe this voice's gender and age. | This is a male voice.<br>This voice sounds like a middle-aged adult's voice (ages 36–55). |

Table 20: Supervision within the missing-age response.

| Response span | Treatment |
|---|---|
| a male | Supervised using the available gender label |
| This voice sounds like | Supervised as fixed response form |
| a middle-aged adult's voice (ages 36–55) | Excluded from direct loss because it depends on a sampled completion |

Loss masking covers completion-dependent labels, comparison directions, difference magnitudes, and numerical intervals, not just numbers. The remaining fixed wording is still supervised. The generator records character spans for the excluded expressions and masks the direct loss on overlapping tokens, including any token that also contains punctuation. Masking removes these tokens from the loss, not from the target sequence used as the preceding context during training.

These examples therefore train the available facts and response form. Attribute values themselves are learned from examples with the corresponding labels.

For two recordings of the same speaker, a missing gender follows the other voice's available gender or a jointly sampled value; gender-conditioned pitch ranges are applied consistently. This does not overwrite observed metadata. Completion values are not selected to make the overall similarity agree with speaker identity.

**Supervision sources in Stage 2.** A full response combines profile and acoustic comparisons even when a recording lacks some of their labels. For VoxCeleb full-description training, F0 and spectral-centroid measurements supplement profile metadata; noise and reverberation already present in these recordings can affect the measurements, so they receive weak supervision. Any remaining sampled attribute values support construction of the summary, not factual supervision of those individual values. Table 21 separates these cases from primary targets and fixed wording. Group-wise averaging prevents lengthy repeated wording from outweighing predicted values simply through token count.

Table 21: Stage 2 supervision by target source. Coefficients are specified in Section 3.2.

| Source of the expression | Loss treatment |
|---|---|
| Valid metadata, primary acoustic targets, controlled recording conditions, speaker IDs, and comparisons derived from them | fact |
| F0 and spectral-centroid measurements used in VoxCeleb full-description training | weak |
| Arbitrary completion values and individual comparisons that depend on them | Excluded from direct loss |
| Overall relation and identity connective computed after completing a context with missing labels | weak |
| Fixed response wording and a generic caveat for unknown recording conditions | form |

Primary acoustic targets for LibriTTS-R are measured before additional augmentation. In contrast, the weaker VoxCeleb measurements are used only for full-description training, not as acoustic targets for Stage 1 or category-level descriptions.

The overall relation and identity connective belong to the fact group only when all five profile and voice-characteristic comparisons have actual reference labels. If any is missing, these expressions are supervised conditionally on the completed context, even when the possible overall relations happen to reduce to a single outcome. Learning a policy-consistent summary is therefore distinct from verifying every comparison against the audio.

Speaker identity is obtained from actual speaker IDs and is not changed to match completed attributes. Unmeasured noise and reverberation receive no invented intensity labels. If both conditions are unknown, the response uses only the generic sentence:

Noise and reverberation can affect the voice cues used in this comparison.

If only some conditions are known, their available labels are described and the generic caveat is appended, without inventing an overall caution grade.

### D.3 Loss Calculation and Parameter Updates

Stage 1 sums the losses over all unmasked target tokens in a batch and divides by their total count; it does not first average each response. Question tokens, audio positions, and padding receive no target loss.

Stage 2 first averages token losses within each fact, weak, and form group of a response. Increasing the length of fixed wording therefore does not increase the form group's weight in proportion to its token count. For the token set $T_b^g$ of group $g$ in response $b$,

$$\bar{\ell}_b^g = \begin{cases} |T_b^g|^{-1} \sum_{t \in T_b^g} \ell_{b,t}, & |T_b^g| > 0, \\ 0, & |T_b^g| = 0. \end{cases}$$

These group means receive the weights in Section 3.2, and the resulting response losses are averaged over the batch. There is no further normalization by the sum of weights or the number of nonempty groups.

The additional identity term is full-vocabulary cross-entropy on the first token of the identity label (`same` or `different`) in a full response. It is not applied to the connective and uses no separate binary classification head. This token also belongs to the fact group; the additional term increases its weight. The term is zero for category descriptions and Stage 1-2 replay. The end-of-sequence token is an ordinary supervised token in Stage 1 and belongs to the form group in Stage 2.

Table 22: Parameter updates by training phase and question type. "Tokenizer" denotes only its trainable token-mapping parameters. The speaker encoder and base language-model weights are always frozen.

| **Phase / question type** | **Target form** | **Updated parameters** |
|---|---|---|
| Stage 1-1 | Short attribute, comparison, and identity answers | Tokenizer |
| Stage 1-2 | Sentence answers to the same questions | Tokenizer + LoRA |
| Stage 2: category description | Comparisons and summary within a fully labeled category | Tokenizer + LoRA |
| Stage 2: full description | Category comparisons, recording-condition description, overall-relation summary, and speaker-identity verdict | LoRA only |
| Stage 2: Stage 1-2 replay | Sentence-form attribute and comparison answers, including identity | Tokenizer + LoRA |

Full descriptions retain actual audio tokens, but detach them from the speaker tokenizer's computation graph. The entire response loss is blocked from updating the tokenizer, including losses on verified factual expressions and the actual identity target. This keeps conditional-summary supervision from directly modifying the trainable tokenizer parameters while preserving audio as input to the language model.

Category descriptions require all labels for the selected category. If segment-level measurement makes a target invalid, the example switches to another supported category or an available individual comparison, falling back to actual speaker identity if necessary. Missing category labels are not filled arbitrarily and treated as factual targets.

# E EXPERIMENTAL SETUP

This appendix specifies the VoiceDiff training recipe, baseline fine-tuning and prompts, and scoring rules used in Section 4. Data construction and supervision coverage are described in Appendix C; the model and objectives are defined in Section 3, with targets and supervision detailed in Appendix D.

## E.1 VOICEDIFF CONFIGURATION AND TRAINING

VoiceDiff uses frozen ReDimNet-B2 (Yakovlev et al., 2024) and Qwen2.5-1.5B-Instruct (Qwen et al., 2025). We use the official VoxCeleb2 speaker-encoder checkpoint `b2-vox2-ft_lm.pt`, obtained after large-margin fine-tuning. The encoder produces a 192-dimensional speaker embedding and 1,152-dimensional frame-level features. The embedding branch uses an MLP, $192 \to 256 \to 16{\times}512$, followed by reshaping and projection to the language model's 1,536-dimensional space. The frame branch projects features to 512 dimensions and uses a four-layer Q-Former with 32 queries, eight attention heads, a 2,048-dimensional feed-forward layer, and dropout 0.1. Cross-attention is applied in every layer. Both branches use output projection and layer normalization; their 16 and 32 tokens are concatenated in that order. Audio is mono, 16 kHz, and 1–15 seconds long. Valid lengths exclude batch padding and are passed to the encoder and frame attention. Audio-token groups are inserted at their designated voice positions in the question context; generation accesses these initially computed tokens without re-encoding the audio.

All stages use AdamW, weight decay 0.01, BF16, and one update per distributed batch. Learning rates increase linearly during warmup and then follow cosine decay over the configured schedule. Model weights carry over between stages, while the optimizer and scheduler are reinitialized. LoRA (Hu et al., 2022) is applied to the attention q/k/v/o projections with rank 16, $\alpha = 32$, and dropout 0.05. The random seed is 4221.

Table 23: VoiceDiff training settings. Learning rates are shown as minimum / maximum; reported epochs are distinguished from schedule horizons. Tokenizer updates apply only to its trainable parameters; the speaker encoder remains frozen.

| Setting | Stage 1-1 | Stage 1-2 | Stage 2 |
|---|---|---|---|
| Response | Short | Sentence | Relational |
| Trainable modules | Tokenizer | Tokenizer + LoRA | Tokenizer + LoRA |
| Global batch size | 128 | 128 | 96 |
| Schedule horizon (epochs) | 20 | 10 | 10 |
| Reported model (epoch) | 15 | 10 | 5 |
| Warmup (updates) | 2,276 | 1,138 | 304 |
| Tokenizer LR | $5{\times}10^{-5}$ / $10^{-4}$ | $2{\times}10^{-6}$ / $4{\times}10^{-6}$ | $4{\times}10^{-7}$ / $2{\times}10^{-6}$ |
| LoRA LR | – | $2{\times}10^{-6}$ / $4{\times}10^{-6}$ | $10^{-5}$ / $5{\times}10^{-5}$ |
| Gradient clipping | None | None | 1.0 |

In Stage 2, full descriptions update LoRA only; category descriptions and Stage 1-2 replay update both the tokenizer and LoRA. The gradient paths are specified in Appendix D.3.

The understanding and tokenizer-ablation results in Table 1 and the external results in Table 4 use Stage 1-1. Rule-composition inputs, single-attribute evaluation under recording degradation, and natural-VoxCeleb pitch and brightness analysis use the final Stage 1-2 model. Direct relational responses in Table 2 and the input-intervention analysis use the five-epoch Stage 2 model above.

## E.2 QUESTION SAMPLING AND AUGMENTATION

Stage 1-1 and Stage 1-2 share the question sampler and differ in response form and trainable modules. Questions and comparison partners are generated online from the combined corpus list; the corpora are not separately resampled to a fixed 1:1 ratio. Table 24 reports requested probabilities, not guaranteed batch counts. Partner availability and valid acoustic measurements can change the realized distribution. Age-question exposure is adjusted across age bins, speakers, and sessions to reduce the influence of prolific sources.

Table 24: Question sampling. Each row describes a separate choice.

| Choice | Probabilities or range |
|---|---|
| *Stage 1-1 and Stage 1-2* | |
| Input | Single voice 0.60; pair 0.40 |
| Scope | One attribute 0.75; subset 0.15; all attributes 0.10 |
| Subset size | 2–4 attributes |
| Question wording | Canonical 0.40; each of two paraphrases 0.30 |
| Explicit reference voice | 0.25 for applicable pair questions |
| Numerical relation | Similar 0.30; somewhat different 0.50; substantially different 0.20 |
| Categorical / identity relation | Same 0.50; different 0.50 |
| *Stage 2* | |
| Training path | Category description 0.40; full description 0.40; Stage 1-2 replay 0.20 |

The default augmentation probabilities are 0.30 for noise, 0.30 for reverberation, 0.30 for both, and 0.10 for neither. Recording-condition questions instead construct conditions for the requested severity or comparison. VoxCeleb full descriptions in Stage 2 receive no additional augmentation. Target measurement and augmentation order follow Appendix C.2.

Loss coefficients are given in Section 3.2; exact masking and aggregation follow Appendix D.3.

### E.3 Baseline Fine-Tuning and Evaluation Prompts

We evaluate Qwen2.5-Omni-7B, Qwen3-Omni-30B-A3B-Instruct, Kimi-Audio-7B-Instruct, and Audio Flamingo 3 (`nvidia/audio-flamingo-3-hf`). The zero-shot and fine-tuned versions share the same base checkpoint for each model. Qwen3-Omni is evaluated zero-shot only. Qwen2.5-Omni, Kimi-Audio, and Audio Flamingo 3 are task-fine-tuned with questions and targets from the same Stage 1 generator. Adaptation updates each model's native audio-encoder–language-model connector; it does not replace the connectors with a shared architecture. We apply LoRA to the language model's attention query, key, value, and output projections, using rank 16, $\alpha = 32$, and dropout 0.05 for all three models.

During evaluation, both zero-shot and fine-tuned models receive the canonical question, shared task definitions, and all permitted answers. These instructions specify the label space but contain no example-specific labels, measurements, or speaker metadata. Table 25 shows actual question and option text; the audio is supplied separately through each model's input interface.

Table 25: Representative baseline prompts. Training uses the Stage 1 question–target templates; evaluation adds task definitions and answer options.

| Question | Permitted evaluation answers |
|---|---|
| Describe the gender of this voice. | Gender: male<br>Gender: female |
| Describe the age of this voice. | Age group: child / adolescent / young adult / middle-aged adult / older adult |
| Compare the pitch of the two voices. | Pitch: similar<br>Pitch: somewhat different; voice 1 is lower than voice 2<br>Pitch: somewhat different; voice 1 is higher than voice 2<br>Pitch: substantially different; voice 1 is lower than voice 2<br>Pitch: substantially different; voice 1 is higher than voice 2 |
| Are voice 1 and voice 2 from the same speaker? | Speaker identity: same<br>Speaker identity: different |

For example, the pitch definitions specify the segment-level F0 measure and the below-20%, 20–40%, and at-least-40% difference intervals. The answer instruction is: *"Return exactly one answer line from the list, without its option letter. Do not repeat the list or add an explanation."* The age alternatives

above are abbreviated for display; evaluation presents each as a separate answer line. Full attribute definitions follow the attribute-label definitions and interval criteria in Appendix A.

Native audio preprocessing is retained. For Audio Flamingo 3, a two-input extension processes the waveforms separately and inserts their features at the corresponding voice positions, rather than concatenating the waveforms. CoLMbo instead uses its official gender and age questions, without our answer options. Its predicted integer-age intervals are mapped by maximum overlap to the task's age groups; ties are unparsed. Its other attributes and pair tasks are outside the reported protocol.

### E.4 Scoring Rules

All models are scored on the same fixed attribute-specific question sets of VoiceDiff-Bench (Table 11). Additional questions on the shared audio pairs do not enter these denominators. Unavailable reference labels are excluded from factual scoring rather than filled in. A deterministic parser normalizes supported paraphrases and reference-voice inversions. Missing, conflicting, and unparseable responses remain errors in the fixed denominator.

**Attribute and difference accuracy.** Prediction requires the requested attribute label. Categorical Match requires the correct same/different relation. Direction requires either the correct higher/lower relation or *similar*, while Magnitude requires the correct difference class. Joint accuracy requires both direction and magnitude for the same answer; a similar reference requires a similar answer. Numerical intervals printed in parentheses are not additional requirements for these label accuracies. Thus, a magnitude-only answer may receive Magnitude credit but cannot receive Direction or Joint credit. Where macro accuracy is reported, it assigns equal weight to the seven attributes and excludes speaker identity.

**Summary accuracy.** Profile and voice-characteristic summaries require the correct constituent relations, their directions, and the set of differing attributes. Recording-condition descriptions require the recording grade and salient per-voice conditions. Each accuracy is evaluated only where its complete primary reference is available. Missing information required by a summary counts as an error. For overall relations, we use the separate mixed-reference subset described below rather than pooling category accuracies.

**Overall relations and constituent comparisons.** On the VoxCeleb pairs with all five comparison labels (Appendix C.3), overall-relation accuracy compares the generated conclusion with the fixed policy applied to metadata and weak acoustic labels, independently of the model's own comparisons. Joint relation–identity accuracy requires both the overall conclusion and the identity verdict to match their labels. A separate policy-consistency diagnostic applies the same rule to the model's generated comparisons and checks whether it reproduces the generated conclusion; it does not substitute for label accuracy.

Within the same responses, exact comparison agreement requires same/different agreement for gender and region, and both direction and magnitude for age, pitch, and brightness. Direction/similarity agreement keeps the categorical criteria but does not require the magnitude of non-similar differences; a similar reference still requires a similar answer. We report the fraction of responses with at least three, at least four, or all five comparisons matching. These are overlapping cumulative criteria. Missing or contradictory comparisons count as errors in the fixed denominator. These measures score the comparison categories, not every numerical-interval string or every sentence in the full response. Appendix F.2 adds an adjacent-category diagnostic without replacing the exact criteria.

**Identity verdict.** Speaker-identity verdict accuracy compares a generated same/different answer with speaker identity. Table 1 uses a standalone identity question. Table 2(a) compares the rule baseline's standalone verdict with the verdict extracted from Stage 2's full description on the same 1,000 VoxCeleb pairs. Table 2(b) uses the 996-pair subset with all five comparison labels. This evaluation scores the verdict actually returned to the user, rather than the threshold trade-off of a continuous similarity score measured by EER; the two metrics are not converted into one another.

### E.5 Supplementary Evaluation Protocols

**VoxParadox task selection and scoring.** We use the gender, age, and pitch-order tasks whose labels map to the supported attribute and comparison responses (Table 26). Models use the same reported versions as in Table 1, without adaptation on VoxParadox. General Audio-LLMs receive the questions, definitions, and options in Appendix E.3; VoiceDiff receives its native questions. Age uses the five-class VoiceDiff vocabulary. Mapping the source's young-adult and elderly-adult labels to young adult and older adult is semantic, not an assertion that numerical age boundaries or chronological ages match.

Table 26: VoxParadox task selection and label correspondence. Clip counts are fixed before scoring.

| Task | Clips | Evaluation label and input |
|---|---|---|
| Gender | 192 | Synthetic voice metadata; eight of 200 clips excluded by the two-second input minimum. |
| Age | 199 | Young-adult / elderly-adult labels mapped to young adult / older adult; one of 200 clips below the input minimum. |
| Pitch order | 200 | Original low–medium–high order, reconstructed from three pairwise comparisons per clip. All source clips retained. |

Each case contributes to one outcome: correct acoustic-label agreement (GT), agreement with a source-provided adversarial label (ALA), another interpretable error, or an unresolved output. All outcomes use the same fixed denominator; multiple adversarial candidates do not duplicate cases. Unresolved outputs count as errors. Low ALA need not imply high GT, since a response can match neither label. Agreement under conflict is not a matched comparison against non-conflicting speech.

**Pitch-order reconstruction.** We split each of the 200 source clips into three equal-duration segments and ask "Compare the pitch of the two voices." for pairs (1,2), (1,3), and (2,3), giving 600 questions per model. Three strict, mutually consistent directions determine an order; a similar answer, missing direction, or cycle makes it unresolved. The resulting order is compared with the source's reference and adversarial orders. Measured F0 does not repair predictions, and no magnitude label is introduced. This differs from the source benchmark's whole-clip multiple-choice question. In particular, our comparison task permits similar below a 20% F0 difference, whereas order reconstruction requires a strict direction.

Exact segment timestamps are unavailable. A pre-inference energy-envelope check, requiring correlation at least 0.95 for every segment pair, supports equal-third segmentation for 158 clips. All 200 clips remain in the main denominator; Appendix F.3 also reports the fixed 158-clip sensitivity subset. Twenty-one segments receive at most 0.246 s of trailing silence to meet the two-second input minimum. No pitch transformation is applied.

**Attribute-wise responses and direct descriptions.** On the same fixed audio inputs, rules compose Stage 1-2 attribute-wise predictions; Stage 2 directly answers "Explain the relationship between the two voices." No baseline fills missing predictions with reference labels. Both use the same profile, voice-characteristic, recording-condition, and identity reference lists and scoring rules. The attribute-wise baseline requires separate questions and is not an inference-cost-matched comparison.

**Audio replacement and pitch shifts.** We use 200 VoxCeleb and 200 LibriTTS-R pairs selected from VoiceDiff-Bench without inspecting predictions. Replacement retains the first voice and replaces the second, constructing 100 same-to-different and 100 different-to-same cases per corpus. References are recomputed from the replacement audio and actual speaker IDs. Pitch shifts of $\pm 3$ semitones affect the second voice of the same LibriTTS-R segments; the existing extractor remeasures pitch and brightness to update their references. We do not assume unchanged identity judgments or invariance of every other attribute after pitch manipulation. Stage 2 receives the same broad question without hints about the intervention or its references. Both the original and newly generated responses are scored against the changed references, separating response changes from correct updates. An exact-waveform replay control is also retained.

**Attribute accuracy under recording degradation.** Profile and voice-characteristic evaluations each use the same 200 source inputs across conditions. Noise uses the same file and offset at SNRs of 20/10/5/0 dB. Reverberation uses RIRs with RT60 ranges of 0.17–0.40/0.60–0.89/1.00–2.03 s; an additional condition combines 0 dB noise with strong reverberation. Original attribute references

remain fixed, rather than being replaced by post-augmentation acoustic measurements. Reverberation conditions use different RIR files and therefore do not isolate RT60 alone. We report Stage 1-2 attribute accuracy and its change from the corresponding original inputs, with speaker-cluster bootstrap 95% intervals for changes.

**Acoustic judgments on natural VoxCeleb speech.** We query the Stage 1-2 model for pitch and timbral brightness on 400 fixed natural VoxCeleb clips from 40 speakers, ten per speaker, without added augmentation. During understanding-stage training, these attributes were supervised in another corpus but not in the VoxCeleb domain. The model and extractor use identical segments. Spearman's $\rho$ measures association between predicted grades and pYIN F0 or power spectral centroid. Pitch correlations are separated by the male/female reference system used for its grades. One clip fails measurement validity checks and is excluded from the corresponding correlations independently of prediction correctness. Rank correlation is not absolute-label accuracy, and this is not transfer to an entirely unseen corpus.

# F Additional Results

## F.1 Speaker-Tokenizer Ablation

The detailed comparison tests whether the two representation sources support different judgments, beyond increasing token count. Table 27 extends Table 1 with 32-token MLP and Q-Former variants. All variants use the same speaker encoder, language backbone, Stage 1-1 settings, epochs, and training exposure. MLP-48 and Q-Former-48 each produce 48 audio tokens per voice, matching the hierarchical tokenizer's 16 tokens from the speaker embedding and 32 from frame-level features. Linear-1 and the 32-token variants use different budgets.

**Attribute prediction and identity.** Panel A reveals complementary strengths within the same speaker encoder. MLP variants using the speaker embedding preserve high identity-verdict accuracy but are weaker on acoustic attributes. Frame-level features substantially improve pitch, brightness, noise, and reverberation prediction, while yielding lower identity accuracy. The hierarchical model combines the highest verdict accuracy with the highest pitch, brightness, and reverberation prediction accuracy among these variants. This is not uniform dominance: Q-Former-48 is slightly better on age prediction, and Q-Former-32 matches the hierarchical model on noise prediction. The benefit is supporting a broader range of judgments with the two sources together.

**Attribute comparisons and token count.** Panel B requires direction and magnitude to be correct together through Joint accuracy. Moving from 32 to 48 Q-Former tokens improves joint pitch, brightness, and noise accuracy but reduces age and reverberation accuracy. The hierarchical tokenizer leads all five Joint measures; relative to the stronger Q-Former configuration for each attribute, its margins are 0.20–1.33 percentage points. Several gains are small and do not establish large or statistically significant improvements. The stronger evidence is the combined support for acoustic prediction, comparison, and identity across both panels. Matching the total of 48 audio tokens rules out a larger token budget as the sole explanation. It does not match every adapter's parameter count or computational cost.

**Staged Training.** We compare direct Stage 1-2 training, which optimizes a new tokenizer and LoRA together from the outset, with the final Stage 1-2 model preceded by Stage 1-1. Both use the same balanced evaluation lists. Direct training runs for 20 epochs; the results below use one seed. Attribute averages weight the seven attributes equally, and direction/magnitude averages weight the five ordered attributes equally. Identity is evaluated separately on 1,000 VoxCeleb pairs.

Staged training yields higher aggregate scores, but not higher accuracy for every attribute. Direct training improves brightness and age comparison by 4.40 and 2.00 percentage points, respectively. This comparison supports the selected training recipe without separately identifying effects of total exposure, initialization, and training order or establishing multi-seed significance.

Table 27: Detailed tokenizer ablation on VoiceDiff-Bench (accuracy, %↑). (A) Attribute prediction and identity; (B) attribute comparison. Audio-token counts are per voice; Joint requires both magnitude and the applicable direction. Bright. denotes timbral brightness; – indicates an inapplicable component. Bold marks the best result for each metric.

**A. Attribute prediction and identity**

| Adapter | Speaker repr. | | Tokens | Profile | | | Voice | | Recording | | Identity |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Embed. | Frames | | Gender | Age | Region | Pitch | Bright. | Noise | Reverb. | |
| Linear | ✓ | | 1 | 94.25 | 41.33 | 47.12 | 25.87 | 24.00 | 26.00 | 2.40 | 44.20 |
| MLP | ✓ | | 32 | 98.50 | 59.00 | 63.05 | 51.73 | 50.50 | 41.00 | 32.27 | 97.40 |
| MLP | ✓ | | 48 | 98.70 | 51.33 | 60.68 | 53.47 | 44.10 | 39.20 | 34.40 | 97.20 |
| Q-Former | | ✓ | 32 | 99.00 | 62.33 | 67.80 | 72.27 | 63.40 | **72.80** | 66.53 | 94.20 |
| Q-Former | | ✓ | 48 | 99.10 | **63.00** | 67.80 | 78.27 | 67.50 | 72.20 | 66.27 | 94.90 |
| Ours | ✓ | ✓ | 16 + 32 | **99.25** | 62.33 | **68.47** | **81.47** | **70.30** | **72.80** | **67.73** | **98.40** |

**B. Attribute difference understanding**

| Adapter | Tokens | Metric | Gender | Region | Age | Pitch | Bright. | Noise | Reverb. |
|---|---|---|---|---|---|---|---|---|---|
| Linear | 1 | Dir./Match | 45.60 | 44.00 | 12.67 | 28.13 | 19.87 | 21.50 | 27.90 |
| | | Mag. | – | – | 12.33 | 17.33 | 19.87 | 21.90 | 27.60 |
| | | Joint | – | – | 12.00 | 12.67 | 19.73 | 12.70 | 14.90 |
| MLP | 32 | Dir./Match | 99.80 | 77.60 | 51.67 | 48.80 | 38.80 | 53.10 | 49.60 |
| | | Mag. | – | – | 41.33 | 38.40 | 29.87 | 39.90 | 39.70 |
| | | Joint | – | – | 37.00 | 38.00 | 29.73 | 24.60 | 23.30 |
| MLP | 48 | Dir./Match | **100.00** | 83.20 | 58.00 | 61.07 | 44.27 | 52.60 | 48.30 |
| | | Mag. | – | – | 45.00 | 44.80 | 28.80 | 41.10 | 40.00 |
| | | Joint | – | – | 40.33 | 43.87 | 28.53 | 25.90 | 22.80 |
| Q-Former | 32 | Dir./Match | 99.60 | 85.00 | 56.33 | 85.07 | 76.00 | 67.90 | 67.10 |
| | | Mag. | – | – | 48.00 | 69.60 | 50.13 | 54.50 | 57.70 |
| | | Joint | – | – | 44.67 | 69.60 | 50.00 | 47.00 | 52.70 |
| Q-Former | 48 | Dir./Match | 99.80 | 86.20 | 61.67 | 85.33 | 76.27 | 68.20 | 68.70 |
| | | Mag. | – | – | 46.33 | 69.87 | 50.40 | 54.80 | 57.50 |
| | | Joint | – | – | 43.00 | 69.87 | 50.27 | 47.30 | 51.70 |
| Ours | 16 + 32 | Dir./Match | **100.00** | **87.40** | **62.00** | **85.60** | **76.53** | **68.50** | **68.90** |
| | | Mag. | – | – | **55.33** | **70.13** | **50.67** | **55.10** | **60.30** |
| | | Joint | – | – | **46.00** | **70.13** | **50.53** | **47.60** | **52.90** |

Table 28: Direct versus staged training (accuracy, %).

| Measure | Direct Stage 1-2 | Stage 1-1 → Stage 1-2 |
|---|---|---|
| Mean attribute prediction | 71.49 | 74.02 |
| Mean attribute comparison | 68.12 | 68.32 |
| Mean direction/similarity | 73.97 | 75.79 |
| Mean difference magnitude | 62.37 | 63.08 |
| Identity verdict | 97.80 | 98.30 |

### F.2 Voice Relation Description

Compared with Stage 1 + Rule, Stage 2 improves profile-summary accuracy, while voice-characteristic and recording-condition summaries and identity remain numerically close (Table 2(a)). The comparison evaluates profile summaries on 1,000 VoxCeleb pairs, voice-characteristic summaries on 1,998 original LibriTTS-R pairs, recording-condition descriptions on 1,500 augmented LibriTTS-R conditions, and identity on the 1,000 VoxCeleb pairs. The shared-input protocol is specified in Appendix E.5. The rule baseline uses additional attribute questions, so this is not an inference-call-

matched comparison. Additional training and response format change together; the results establish integrated output performance, not the isolated causal effect of format.

**Overall conclusions and their constituent comparisons.** Table 2(b) and Table 3 use the same 996 VoxCeleb pairs with all five comparison labels available (Appendix C.3). Stage 2 correctly predicts the overall relation in 709 cases, identity in 981, and both in 702. The 709 responses with at least four exact comparisons and the 709 correct overall conclusions are different success events that happen to have the same count. An overall conclusion can be correct despite errors in some constituent comparisons.

**Consistency with the stated comparisons.** Applying the fixed overall-relation policy to the model's own five comparisons reproduces its generated conclusion in 994 of 996 responses (99.80%). Of the 287 incorrect overall conclusions, 286 remain consistent with those generated comparisons. The output therefore organizes its stated judgments consistently even when those judgments disagree with the references. This identifies constituent-comparison agreement, rather than aggregation consistency, as the main limitation in this evaluation. It does not reveal the model's internal reasoning path or the training stage at which an error originated. The diagnostic concerns the five profile and voice-characteristic comparisons and the overall conclusion, not every category summary, recording condition, or identity verdict.

Table 29: Stage 2 comparison agreement on the joint VoxCeleb subset (%). Exact agreement requires both direction and magnitude for age, pitch, and brightness. Gender and region require same/different agreement under both criteria.

| Attribute | Direction/similarity | Exact agreement |
|---|---|---|
| Gender | 100.00 | 100.00 |
| Age | 76.71 | 70.98 |
| Region | 91.87 | 91.87 |
| Pitch | 88.25 | 82.83 |
| Timbral brightness | 60.54 | 45.58 |

Exactly one comparison is wrong in 456 responses: brightness accounts for 303 (66.45%), age for 87, pitch for 48, and region for 18. Brightness therefore contributes most to the gap between matching at least four comparisons and matching all five. Because brightness is a weak measurement reference, these disagreements do not separate model error from measurement instability.

**Agreement with adjacent difference categories.** As a post-hoc error-severity diagnostic, we allow predictions within one position of the reference on the signed sequence *substantially lower, somewhat lower, similar, somewhat higher, substantially higher*. Age uses younger/older and brightness uses darker/brighter. Gender and region still require exact categorical agreement. This criterion permits similar versus a small difference and is therefore distinct from strict direction accuracy; it is not top-2 accuracy over ranked candidates.

Table 3 reports this diagnostic alongside exact and direction/similarity agreement.

An input-independent diagnostic that always predicts same gender and region and similar age, pitch, and brightness matches all five references in 18.47% of cases exactly and 40.06% under the adjacent criterion. Relaxed agreement alone is therefore not evidence of precise acoustic observation. Main results retain exact comparison and direction criteria.

**Complete generated descriptions.** Table 2 presents two complete Stage 2 outputs. They illustrate how a characteristic relation and an identity verdict can be expressed together, including contrasting characteristics with the same identity and similar characteristics with different identities. They illustrate response structure, not the prevalence of these cases or verified correctness of every claim.

**Updating Descriptions after Input Changes.** Table 30 evaluates Stage 2 under audio replacement and pitch shifts. Both columns are scored against the **post-intervention references**, testing whether a new response reflects the changed facts better than retaining the original answer. Inputs and sampling

follow Appendix E.5. Replacement inputs need not have the same difficulty as the originals, so these results are not accuracy gains over the unmodified task.

Table 30: Agreement with post-intervention references (%). Pitch requires both direction and magnitude.

| Intervention and measure | N | Retain original answer | Generate updated answer |
|---|---|---|---|
| Audio replacement: profile summary | 200 | 4.5 | 48.5 |
| Audio replacement: voice-characteristic summary | 200 | 1.0 | 55.0 |
| Audio replacement: identity verdict | 400 | 4.5 | 95.5 |
| Pitch +3 semitones: pitch comparison | 200 | 44.5 | 78.5 |
| Pitch -3 semitones: pitch comparison | 200 | 45.5 | 78.5 |

Updated responses follow the changed references more closely for all five measures. This supports responsiveness of integrated descriptions to replaced or manipulated audio rather than repetition of fixed text. Summary accuracy remains limited and other attributes can also change; the result does not establish robustness to every transformation or complete faithfulness of internal reasoning. The 400 replacement identity cases pool both diagnostic corpora and do not replace the VoxCeleb-only 1,000-pair metric in Table 2.

### F.3 Judgments under Conflicting Spoken Claims

Table 4 shows that gains from task supervision extend to external cases with conflicting spoken claims, with different error patterns across attributes. The task selection, label correspondence, and pairwise pitch protocol are specified in Appendix E.5. Here we examine what the residual errors imply about this transfer.

**Reference accuracy and adversarial agreement.** Gender and age outputs are all parseable, so their errors are not failures to recover an answer. For binary gender, every error necessarily agrees with the opposing claim; improved GT and reduced ALA describe the same change. Age admits other errors, including middle-aged responses. VoiceDiff's low age ALA therefore does not make it more accurate than Qwen2.5-Omni FT. The task-adaptation gains support better judgments under conflict, but without a matched non-conflicting control they do not isolate a causal reduction in sensitivity to spoken content.

**Pitch-order errors.** VoiceDiff's 36 unresolved orders each contain a *similar* comparison, with no missing directions or cycles. Most zero-shot pitch answers are also *similar*, whereas task-adapted AF3 lacks a recoverable direction in 498 of 600 comparisons. These are distinct output failures under the derived ranking task, not equivalent acoustic judgments. VoiceDiff permits *similar* below a 20% F0 difference, whereas the source reference specifies a strict order; a tie can therefore block reconstruction without establishing a reversed acoustic judgment. Pairwise reconstruction also changes the spoken context and answer space from the source whole-clip question, limiting the conclusion to the evaluated protocol.

**Sensitivity to segment boundaries.** On the fixed 158-clip subset supported by the pre-inference boundary check (Appendix E.5), VoiceDiff reconstructs 144 orders correctly (91.14%), compared with 70 (44.30%) for Qwen2.5-Omni FT. The remaining 42 clips stay in the main denominator, and these scores do not replace the all-200 results. The relative advantage persists on clips whose segmentation is better supported, although this does not establish invariance to other linguistic or recording changes.

**One observed gender conflict.** In stored responses to the same source audio, the acoustic reference is male and the claim in the spoken content suggests female. AF3 zero-shot returns `B. Gender: female`; VoiceDiff returns `Gender: male`. Under the fixed prompts described in Appendix E.3, this example illustrates a judgment that follows vocal evidence despite a conflicting spoken claim. Together with the aggregate GT and ALA results, it supports the utility of voice-centered supervision beyond the training-data domains, while the age and pitch diagnostics delimit that transfer.

### F.4 Attribute Prediction under Recording Degradation

The degradation analysis tests which judgments become harder when recording conditions obscure vocal evidence. Table 31 reports Stage 1-2 accuracy under the matched-input protocol in Appendix E.5: every row uses the same 200 inputs per attribute and retains the pre-augmentation references.

Table 31: Attribute accuracy under added recording conditions (%); parentheses show changes from the same originals in percentage points. Original gender/age/region/pitch/brightness accuracies are 100.0/72.0/81.5/88.5/73.0.

| Added condition | Gender | Age | Region | Pitch | Timbral bright. |
|---|---|---|---|---|---|
| Noise: SNR 20 dB | 100.0 (0.0) | 70.5 (-1.5) | 80.5 (-1.0) | 88.5 (0.0) | 70.5 (-2.5) |
| Noise: SNR 10 dB | 100.0 (0.0) | 71.5 (-0.5) | 82.0 (+0.5) | 86.5 (-2.0) | 73.5 (+0.5) |
| Noise: SNR 5 dB | 100.0 (0.0) | 69.5 (-2.5) | 81.0 (-0.5) | 86.0 (-2.5) | 72.5 (-0.5) |
| Noise: SNR 0 dB | 100.0 (0.0) | 67.0 (-5.0) | 82.0 (+0.5) | 84.5 (-4.0) | 69.0 (-4.0) |
| Reverb.: RT60 0.17–0.40 s | 99.5 (-0.5) | 70.5 (-1.5) | 80.5 (-1.0) | 89.5 (+1.0) | 71.0 (-2.0) |
| Reverb.: RT60 0.60–0.89 s | 100.0 (0.0) | 64.0 (-8.0) | 81.0 (-0.5) | 87.0 (-1.5) | 68.0 (-5.0) |
| Reverb.: RT60 1.00–2.03 s | 100.0 (0.0) | 69.5 (-2.5) | 83.0 (+1.5) | 85.5 (-3.0) | 71.5 (-1.5) |
| Noise 0 dB + reverb. 1.00–2.03 s | 99.5 (-0.5) | 68.0 (-4.0) | 80.5 (-1.0) | 82.0 (-6.5) | 65.0 (-8.0) |

The combined condition reduces pitch and brightness by 6.5 and 8.0 percentage points, with speaker-cluster bootstrap 95% intervals of [-12.09, -0.95] and [-14.93, -1.39]. The -8.0-point age change under moderate reverberation also has an interval excluding zero, [-14.63, -1.81]. Gender and region remain relatively stable, and some conditions improve accuracy. Different RIRs across levels prevent attributing changes to RT60 alone or claiming monotonic degradation. The analysis motivates reporting recording conditions as separate context; it does not establish calibrated confidence or a benefit to user judgments from supplying that context.

### F.5 Pitch and Brightness on Natural VoxCeleb

Do judgments learned with acoustic labels from another corpus follow acoustic variation in natural VoxCeleb recordings? Table 32 reports Stage 1-2 results without further adaptation, using the same-segment protocol in Appendix E.5. VoxCeleb pitch and brightness references were absent during understanding-stage training. Among 400 clips from 40 speakers, one fails measurement-validity checks and is excluded; all 400 model responses are parseable.

Table 32: Spearman rank correlation between predicted grades and measurements on natural VoxCeleb. Values are not accuracy percentages.

| Attribute | Measurement reference | Valid N | Spearman $\rho$ |
|---|---|---|---|
| Pitch – male reference | pYIN F0 | 249 | 0.902 |
| Pitch – female reference | pYIN F0 | 150 | 0.876 |
| Timbral brightness | Power spectral centroid | 399 | 0.740 |

The positive rank correlations show that learned pitch and brightness judgments track measurement trends even without supervision for these attributes in the VoxCeleb domain. VoxCeleb itself was not excluded from training, and rank association does not establish absolute-grade accuracy. Spectral brightness also reflects the recording environment, so this result is neither recovery of intrinsic timbre nor validation against independent perceptual ground truth. The evidence concerns pitch and brightness under the stated extraction rules, not all recording-condition attributes.