

# SpeakerSleuth: Can Large Audio-Language Models Judge Speaker Consistency across Multi-turn Dialogues?

Jonggeun Lee, Junseong Pyo<sup>§</sup>, Gyuhyeon Seo, Yohan Jo<sup>†</sup>  
Graduate School of Data Science, Seoul National University  
{jonggeun.lee, yohan.jo}@snu.ac.kr

## Abstract

Large Audio-Language Models (LALMs) as judges have emerged as a prominent approach for evaluating speech generation quality, yet their ability to assess speaker consistency across multi-turn dialogues remains unexplored. We present **SpeakerSleuth**, a benchmark evaluating whether LALMs can reliably judge speaker consistency across multi-turn dialogues through three tasks reflecting real-world requirements. We construct 1,818 human-verified evaluation instances across four diverse datasets spanning synthetic and real speech, with controlled acoustic difficulty. Evaluating twelve widely-used LALMs, we find that models struggle to reliably detect acoustic inconsistencies. For instance, given audio samples of the same speaker’s turns, some models overpredict inconsistency, whereas others are overly lenient. Models further struggle to identify the exact turns that are problematic. When other interlocutors’ turns are provided as textual context, performance degrades dramatically as models prioritize textual coherence over acoustic cues, failing to detect even obvious gender switches for a speaker. On the other hand, models perform substantially better in comparing and ranking acoustic variants, demonstrating inherent acoustic discrimination capabilities. These findings expose a significant bias in LALMs: they tend to prioritize text over acoustics, revealing fundamental modality imbalances that need to be addressed to build reliable audio-language judges. Our code and data are available at <https://github.com/holi-lab/SpeakerSleuth>.

## 1 Introduction

Recent advances in speech synthesis have enabled systems that produce natural, human-like speech (Du et al., 2024; Zhang et al., 2024b; Défossez

<sup>†</sup>Corresponding author.

<sup>§</sup>Visiting intern from Hanyang University.

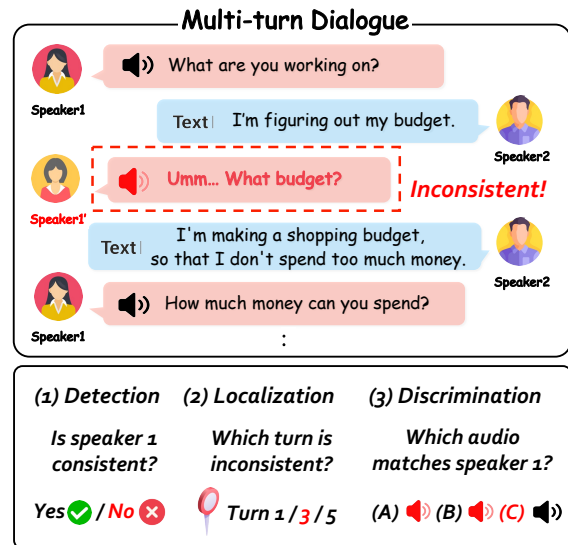


Figure 1: Overview of SpeakerSleuth.

et al., 2024; Lee et al., 2026). These technologies enable diverse applications including voice assistants (Apple, 2024), voice-overs in podcast generation (Google, 2024) and movies (Danell, 2025), and conversational agents (OpenAI, 2024). A fundamental requirement for these systems is maintaining consistent speaker identity (Mullenix and Pisoni, 1990), that is, preserving acoustic characteristics such as timbre, pitch, and voice quality across all utterances in a multi-turn dialogue. This is particularly important in the speech synthesis of multiple dialogue participants, such as voice-overs in movies. However, achieving this consistency across long-form, multi-turn dialogues remains challenging (Xie et al., 2025). Even recent models suffer from speaker confusion (Borsos et al., 2023; Zhang et al., 2024b), timbre drift (Ju et al., 2024), and voice quality variations (Park et al., 2025). These failures are particularly difficult to catch because they emerge only across turns — a single generated utterance may sound natural in isolation, yet be clearly inconsistent when heard in the context of the full dialogue. This necessi-

tates reliable verification methods that can assess speaker consistency at the dialogue level.

Most approaches (Zhang et al., 2024b; Lee et al., 2025) evaluate speaker consistency by computing acoustic similarity between utterances using embedding models (Desplanques et al., 2020; Chen et al., 2022). However, these methods face fundamental limitations when applied to dialogue evaluation: they operate on pairwise comparisons between two utterances, require manually-set thresholds for binary decisions, and cannot assess consistency holistically across entire dialogues.

Recently, Large Audio-Language Models (LALMs) have emerged as potential alternatives for evaluating speech generation quality (Wang et al., 2025a,c). Unlike embedding-based methods that compute pairwise similarities, LALMs can process an entire dialogue at once, receiving both text and audio, and directly outputting a judgment about speaker consistency.

However, two critical gaps remain: First, no unified benchmark exists to systematically evaluate and compare embedding methods and LALMs for multi-turn speaker consistency assessment. Second, whether LALMs possess the acoustic discrimination capabilities necessary for reliable speaker consistency judgment remains unexplored.

To address these questions, we present **SpeakerSleuth**, a benchmark for evaluating both LALMs and embedding-based methods on speaker consistency in multi-turn dialogues. We design our benchmark around three tasks that mirror real-world application requirements (Figure 1). These tasks are *Detection* (identifying whether dialogues contain inconsistencies), *Localization* (pinpointing which specific turns are problematic), and *Discrimination* (comparing and ranking multiple acoustic variants). These capabilities are essential for practical speech generation systems. When dialogue speech is generated, systems must first detect any inconsistencies for each speaker, then localize problematic turns for targeted correction, and finally select optimal outputs from regenerated alternatives. We construct our benchmark from four diverse datasets spanning synthetic and real speech across various conversational settings, comprising 1,818 evaluation instances from 197 speakers, all verified through human annotation.

Our evaluation of 12 state-of-the-art LALMs and 6 embedding methods reveals critical insights into their capabilities. We find that models struggle to reliably detect acoustic inconsistencies due to un-

stable internal thresholds. This leads to inconsistent decisions where some models are too strict while others are too lenient. Moreover, they struggle with fine-grained turn-level acoustic analysis, as evidenced by their inability to localize specific problematic turns even when detecting overall inconsistency. When other interlocutors’ turns are available as dialogue context, LALMs overwhelmingly tend to prioritize textual coherence over acoustic features. They fail to detect even obvious inconsistencies like gender switches for a speaker within coherent dialogue. In contrast, embedding methods achieve stronger detection performance, although they exhibit consistent model-specific biases.

Our contributions are summarized as follows:

- We present SpeakerSleuth, the first benchmark for multi-turn speaker consistency evaluation with 1,818 human-verified instances.
- We comprehensively evaluate 12 LALMs and 6 embedding methods, revealing that models struggle with detection and localization.
- We identify modality imbalances where LALMs prioritize textual context over acoustic discrimination capabilities, providing insights toward reliable audio-language judges.

## 2 Related Work

### 2.1 Speech Synthesis

Speech synthesis has evolved from early end-to-end systems (Wang et al., 2017) to sophisticated controllable generation approaches. Speaker cloning methods (Wang et al., 2023; Li et al., 2025) enabled generating speech in a target speaker’s voice from reference audio, but lacked intuitive control mechanisms. Recent text instruction-guided models (Guo et al., 2023; Du et al., 2024) allow natural language control, significantly improving usability for multi-speaker dialogue generation (Zhang et al., 2024b, 2025; Lee et al., 2025).

However, maintaining consistent speaker identity across long-form, multi-turn dialogues remains challenging (Xie et al., 2025). Models exhibit speaker confusion (Borsos et al., 2023; Zhang et al., 2024b), timbre drift (Ju et al., 2024), and voice quality variations (Park et al., 2025), particularly in zero-shot settings where limited reference audio must generalize to extended conversations. These challenges necessitate robust evaluation methods that can reliably assess both speech quality and speaker consistency across multi-turn dialogues.

## 2.2 Speech Quality Evaluation

Speech quality evaluation systematically assesses speech across dimensions such as naturalness and intelligibility (Loizou, 2011). Traditional approaches rely on objective metrics such as Mel-Cepstral Distortion (MCD) (Kubichek, 1993) and PESQ (Rix et al., 2001), or subjective human assessment through Mean Opinion Score (MOS) (Ribeiro et al., 2011). While neural approaches have enabled automated MOS prediction (Saeki et al., 2022), they typically focus on single quality dimensions. More recently, the LALM-as-a-Judge paradigm (Wang et al., 2025a,c; Chen et al., 2025; Wang et al., 2025d) has emerged, leveraging LALMs trained on joint audio-text data for multi-dimensional quality analysis with natural language reasoning. This enables LALMs to potentially integrate acoustic features with conversational context, making them promising candidates for evaluating speaker consistency in dialogue settings.

## 2.3 Speaker Consistency Evaluation

Beyond assessing individual utterance quality, speaker consistency evaluation measures whether a speaker’s identity remains stable across multiple utterances in a dialogue. Existing approaches include embedding-based methods (Snyder et al., 2017; Khoma et al., 2023) using speaker verification models (Desplanques et al., 2020; Chen et al., 2022) to compute similarity scores across dialogue turns (Zhang et al., 2024b, 2025; Ju et al., 2025; Lee et al., 2025), and human evaluation (Zhang et al., 2024b), which incurs high costs. Given the recent success of LALMs in speech quality evaluation, a natural question arises: can they reliably assess speaker consistency? Their acoustic perception capabilities for this task remain unexplored.

## 3 Task Formulation

To thoroughly evaluate whether LALMs can reliably distinguish speakers based on acoustic features, we propose an evaluation framework. Rather than relying on a single metric, we decompose speaker consistency into three capabilities: *Detection* (identifying whether all turns are consistent), *Localization* (pinpointing which turn is inconsistent), and *Discrimination* (comparing and ranking acoustic variants by their similarity to a target speaker).

Formally, we define a multi-turn dialogue as  $\mathcal{D} = \{(t_1, a_1), (t_2, a_2), \dots, (t_N, a_N)\}$ , where  $t_i$

represents the transcript and  $a_i$  denotes the audio waveform of the  $i$ -th turn. Let  $I \subset \{1, \dots, N\}$  denote the indices of turns belonging to a specific target speaker  $S$ . We denote the audio turns of the target speaker as  $\mathcal{A}_S = \{a_i\}_{i \in I}$ . In our primary evaluation, models receive  $\mathcal{A}_S$  and a reference audio sample of the target speaker to isolate acoustic features from textual cues, with the effect of adding textual context examined separately in Section 7.

### 3.1 Task 1: Detection

In Text-to-Speech (TTS) and voice cloning, ensuring speaker consistency across generated outputs is critical for quality control. The most fundamental requirement is to detect whether all audio turns belong to the same speaker. This requires *absolute judgment capability*, where the model must rely on its internal threshold to determine consistency.

Given  $\mathcal{A}_S$ , the model predicts whether all audio turns maintain the identity of speaker  $S$  (Consistent) or not (Inconsistent). Success on this task demonstrates that the model possesses an appropriate internal threshold to reliably judge speaker identity based on acoustic features. The prompts are provided in Figures 8 and 9.

### 3.2 Task 2: Localization

When speaker inconsistencies are detected in a multi-turn dialogue, identifying the exact problematic turn is essential for efficient correction and regeneration. Merely detecting an anomaly is insufficient; a robust judge must pinpoint where the inconsistency occurs to enable targeted fixes.

Given  $\mathcal{A}_S$ , the model identifies which turn disrupts speaker identity, or predicts None if all turns are consistent. Success on this task demonstrates that the model can distinguish acoustic speaker characteristics at a fine-grained level, rather than relying on dialogue-level patterns. The prompts are provided in Figures 10 and 11.

### 3.3 Task 3: Discrimination

When an inconsistent turn is identified (Task 2), TTS systems typically regenerate multiple candidate outputs and must select the one that best matches the target speaker. This requires *relative judgment capability*, the ability to compare and rank audio samples by their acoustic similarity to a reference speaker, rather than making absolute binary decisions as in Task 1.

Given  $\mathcal{A}_S$  where a target turn is masked, the model is presented with three candidates  $\mathcal{C} =$

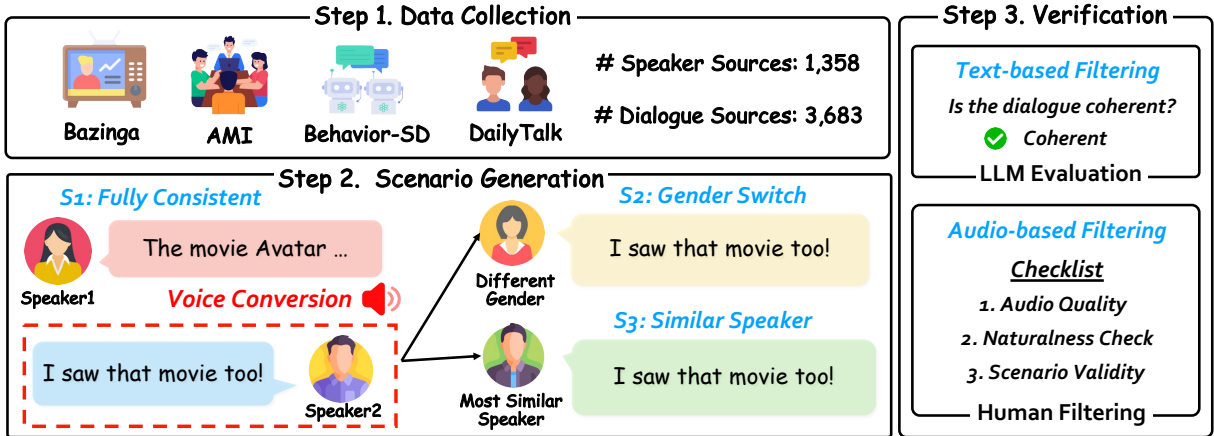


Figure 2: SpeakerSleuth Construction Pipeline.

$\{c_1, c_2, c_3\}$  representing varying levels of acoustic similarity to the original speaker. The order of candidates is randomized to avoid positional bias (Pezeshkpour and Hruschka, 2024). We evaluate two formulations: **classification**, where the model selects the best-matching candidate, and **ranking**, where the model orders all three candidates by acoustic similarity, which poses a strictly harder objective. Success on this task indicates that the model can discriminate speakers acoustically. The prompts for both formulations are provided in Figures 12 and 13.

### 3.4 Distinction from Traditional Speaker Recognition

It is important to note that our tasks differ fundamentally from traditional speaker recognition (Furui, 1996; Gish and Schmidt, 1994), which consists of two main tasks: Speaker Identification determines *who is speaking* from known speakers, while Speaker Diarization segments continuous audio streams to determine *who spoke when*. In contrast, we evaluate *speaker consistency*: given turns that are assumed or claimed to belong to the same speaker, we assess whether the model can verify that they actually maintain acoustic coherence.

## 4 SpeakerSleuth

To systematically evaluate the capabilities defined in Section 3, we introduce SpeakerSleuth, a benchmark composed of multi-turn dialogues with rigorous acoustic and contextual controls. Figure 2 illustrates our benchmark construction pipeline.

### 4.1 Step 1. Data Collection

We collect dialogues with audio and transcripts from four datasets to ensure diversity in con-

versational domains and styles (Figure 2-1): *Bazinga* (Lerner et al., 2022) contains multi-party dialogues from TV shows and movies (e.g., Friends), testing the model’s ability to track speakers in dynamic, scripted interactions. *AMI* (Carletta et al., 2005) consists of spontaneous business meetings, assessing performance in formal, overlapping, and noisy environments. *Behavior-SD* (Lee et al., 2025) provides synthesized dialogues with controlled speech behaviors (e.g., fillers, backchannels), enabling evaluation of generated speech. *DailyTalk* (Lee et al., 2023) captures high-quality everyday conversations, serving as a baseline for casual social interaction.

From these sources, we construct an initial pool of 3,683 dialogues spanning 1,358 unique speakers.

### 4.2 Step 2. Scenario Generation

**Dialogue Extraction.** From the collected pool, we select dialogues and extract segments where a target speaker appears multiple times across the conversation. Each segment contains exactly 5 target speaker turns, with the total number of turns across all speakers capped at 20. This cap is necessary because in multi-party dialogues, the target speaker’s turns are interspersed with those of other participants; without the constraint, the gap between target turns could grow excessively large. We also sample a reference audio of the target speaker from outside these segments.

**Scenario Generation.** As illustrated in Figure 2-2, we create three scenarios per dialogue to systematically test acoustic discrimination capabilities: Original conversations serve as positive samples (**S1: Fully Consistent**), establishing baseline performance when acoustic and textual cues are natu-

Dataset	Instances	Speakers	Avg Turns <sup>†</sup>	Total Duration
Bazinga	636	109	9.9 ± 1.7	3.7 hrs
AMI	138	34	7.9 ± 3.4	0.9 hrs
Behavior-SD	477	52	7.9 ± 1.2	3.0 hrs
DailyTalk	567	2	9.0 ± 0.0	2.6 hrs
<b>Total</b>	<b>1,818</b>	<b>197</b>	<b>8.9 ± 1.7</b>	<b>10.2 hrs</b>

<sup>†</sup>Target speaker: 5 audio turns; other speakers: text.

Table 1: Statistics of SpeakerSleuth.

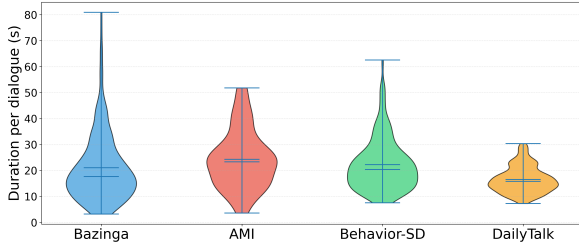


Figure 3: Distribution of per-instance audio duration.

rally aligned. To create inconsistent scenarios, we randomly select one turn and apply voice conversion, transforming acoustic timbre while preserving linguistic and prosodic content. For **S2 (Gender Switch)**, we convert the turn to an opposite-gender voice sampled from the pool, creating clear acoustic deviations. For **S3 (Similar Speaker)**, we convert the turn to an acoustically similar speaker, selected by computing ECAPA-TDNN (Desplanques et al., 2020) embeddings and choosing the one with highest cosine similarity (excluding the target speaker). This requires fine-grained discrimination of subtle timbre differences. By using identical dialogue content across all scenarios, we isolate acoustic features from confounding factors.

For the primary benchmark, we use FreeVC (Li et al., 2023) for voice conversion. We also construct an extended benchmark with CosyVoice3 (Du et al., 2025), OpenVoice (Qin et al., 2023), and YourTTS (Casanova et al., 2022) to verify robustness across voice conversion models (Appendix D.1).

### 4.3 Step 3. Verification

As illustrated in Figure 2-3, we validate dialogue segments through automated text-based filtering and manual audio-based verification. For text-based filtering, we use Qwen3-32B (Yang et al., 2025) to filter out segments that lack natural conversational flow when isolated from their original context (Zhang et al., 2024a). For audio-based verification, expert annotators verify audio quality (clarity, absence of noise or artifacts) and naturalness of voice-converted turns. Annotators also

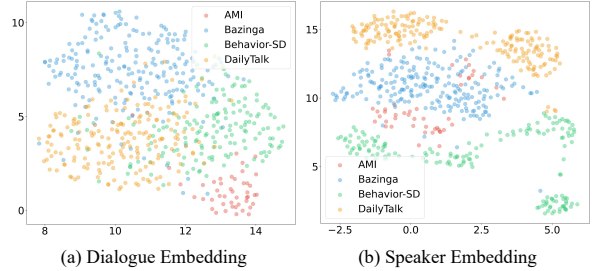


Figure 4: UMAP visualization of dialogue and speaker embeddings.

confirm that each scenario exhibits its intended acoustic characteristics. Only instances meeting all criteria are retained. Details are provided in Appendix A.2.

## 4.4 Benchmark Composition

Our final benchmark contains 606 unique dialogues, each contributing three scenarios, yielding 1,818 total evaluation instances. The benchmark comprises 10.2 hours of audio from 197 speakers across the four source datasets. Table 1 summarizes the dataset statistics, and Figure 3 shows the distribution of sample durations. By holding dialogue content constant across scenarios, performance differences between S1, S2, and S3 directly reflect the model’s ability to detect varying degrees of acoustic deviation. UMAP (McInnes et al., 2018) visualization of dialogue and speaker embeddings confirms substantial diversity in both dialogue content and acoustic characteristics (Figure 4; details in Appendix A.3).

## 5 Experimental Setup

### 5.1 Models

**LALM Judges.** We assess twelve widely-used LALMs: GPT-4o-audio (OpenAI, 2024), Gemini-2.5-Pro (Kavukcuoglu, 2025), Gemini-2.5-Flash/Flash-Lite (Basu Mallick et al., 2025), Qwen2.5-Omni-3B/7B (Xu et al., 2025a), Qwen3-Omni-30B-A3B (Xu et al., 2025b), MiniCPM-o-2.6 (OpenBMB, 2025), Gemma-3n-E4B (Team et al., 2025), Phi-4-multimodal (Microsoft et al., 2025), Omnivinci (Ye et al., 2025), and Audio-Flamingo-3 (Ghosh et al., 2025). These models vary in architecture and scale, enabling analysis of how model capacity affects consistency evaluation.

**Speaker Embedding Methods.** We evaluate three speaker embedding methods, each with two backbone models: WavLM (Chen et al., 2022) and ECAPA-TDNN (Desplanques et al., 2020). All

Model / Method	Detection				Localization								Discrimination							
	S1	S2	S3	Bal	S1	S2			S3			Bal	Classification					Ranking		
	Acc				F1	P	R	F1	P	R	F1	F1	Baz	AMI	B-SD	Daily	Avg	N@1	N@2	EM
<i>Large Audio-Language Models</i>																				
GPT-4o-audio	72.9	32.8	29.5	52.0	71.5	12.6	25.9	15.0	8.7	19.2	11.1	42.3	50.2	30.4	38.4	34.4	40.7	45.4	60.1	19.8
Gemini-2.5-Pro	73.9	71.6	39.3	<b>64.7</b>	65.2	59.3	70.3	62.5	44.0	56.6	47.5	<b>60.1</b>	77.8	76.1	81.1	87.2	<b>81.5</b>	<b>88.8</b>	<b>92.6</b>	<b>71.5</b>
Gemini-2.5-Flash	97.4	45.5	12.0	<u>63.1</u>	64.4	55.6	90.1	62.3	38.6	72.4	45.0	<u>59.0</u>	70.8	73.9	69.2	86.8	<u>75.6</u>	<u>83.6</u>	<u>88.3</u>	<u>61.6</u>
Gemini-2.5-Flash-Lite	40.8	70.3	69.3	55.3	0.3	25.8	94.7	36.4	24.5	94.1	35.4	18.1	47.6	45.7	47.8	44.4	46.5	53.5	64.6	23.1
Qwen2.5-Omni-3B	54.5	49.3	47.8	51.5	21.8	45.6	91.6	47.7	28.9	81.6	36.0	31.8	42.9	39.1	41.5	38.1	40.8	51.7	63.1	22.8
Qwen2.5-Omni-7B	32.0	72.3	69.8	51.5	38.0	14.1	61.8	22.6	13.2	57.5	21.0	29.9	57.5	30.4	39.6	31.7	42.7	52.4	64.2	22.9
Qwen3-Omni-30B-A3B	88.1	29.9	14.0	55.0	0.2	15.1	57.4	23.3	15.2	56.0	23.5	11.8	82.1	69.6	43.4	48.1	60.4	66.9	73.2	36.8
MiniCPM-o-2.6	85.3	0.7	0.0	42.8	66.9	29.8	54.5	35.4	12.3	25.1	17.2	46.6	62.7	45.7	42.8	41.3	49.5	52.2	65.4	23.6
Gemma-3n-E4B	51.2	50.6	48.4	50.4	0.0	19.1	95.1	32.0	19.1	94.4	32.0	16.0	36.8	37.0	33.3	40.7	37.1	47.5	62.1	20.1
Phi-4-multimodal	78.8	24.9	24.3	51.7	99.7	0.0	0.1	0.1	0.0	0.1	0.1	49.9	37.3	39.1	39.0	43.9	39.9	49.5	63.3	21.3
Omnivinci	81.6	46.1	20.9	57.5	48.3	14.9	61.2	23.6	8.8	40.0	15.2	33.8	51.4	39.1	42.1	42.9	45.4	57.0	67.3	25.4
Audio-Flamingo-3	99.2	1.3	1.2	50.2	0.0	17.1	38.4	23.6	17.4	38.8	23.7	11.8	33.0	28.3	37.1	41.3	36.3	44.9	60.7	18.3
<i>Speaker Embedding Methods</i>																				
Pairwise (ECAPA)	36.0	88.4	86.3	61.7	36.0	44.1	92.2	46.5	38.3	81.8	43.5	40.5	97.8	99.0	100.0	100.0	<b>99.2</b>	<b>99.6</b>	<b>92.7</b>	<b>58.6</b>
Pairwise (WavLM)	91.8	38.4	37.7	<b>64.9</b>	91.8	36.0	37.8	31.8	36.8	38.5	31.3	<b>61.7</b>	96.7	95.5	79.9	100.0	<u>93.2</u>	<u>94.4</u>	<u>91.0</u>	<u>55.0</u>
Centroid (ECAPA)	47.3	75.7	73.7	61.0	47.3	54.1	90.7	46.3	47.9	80.8	43.3	46.0	97.8	99.0	100.0	100.0	<b>99.2</b>	<b>99.6</b>	<b>92.7</b>	<b>58.6</b>
Centroid (WavLM)	87.5	38.2	37.7	<u>62.7</u>	87.5	49.0	51.3	31.6	52.1	54.3	31.0	<u>59.4</u>	96.7	95.5	79.9	100.0	<u>93.2</u>	<u>94.4</u>	<u>91.0</u>	<u>55.0</u>
Reference (ECAPA)	8.5	95.5	94.5	51.8	8.5	33.7	89.8	39.0	29.3	81.2	34.6	22.6	97.8	92.0	99.4	76.2	91.0	92.3	87.2	52.7
Reference (WavLM)	79.3	32.2	32.6	55.9	79.3	23.8	25.7	16.9	23.0	24.6	17.1	48.3	100.0	94.3	73.0	68.8	82.8	88.3	84.3	51.4

Table 2: Main results (%). Detection reports per-scenario and Balanced Accuracy (Bal). Localization reports F1 for S1, Precision (P), Recall (R), and F1 for S2/S3, and Balanced F1. Discrimination reports per-dataset and average Classification Accuracy, and Ranking metrics (N@1: NDCG@1, N@2: NDCG@2, EM: Exact Match). **Bold** indicates the highest and underline the second-highest per group for Bal Acc, Bal F1, Avg, N@1, N@2, and EM.

methods flag turns as inconsistent when their similarity (or equivalently distance) crosses a threshold  $\tau$ , flag all such turns for localization, and rank candidates via their similarity metric for discrimination. **Pairwise Similarity** computes each turn’s average cosine similarity to all other turns ( $\tau_{\text{pair}} = 0.4$ ). **Centroid Distance** computes each turn’s distance from the centroid of all turns ( $\tau_{\text{cent}} = 0.3$ ). **Reference Comparison** computes similarity between each embedding and a reference speaker audio ( $\tau_{\text{ref}} = 0.4$ ). Thresholds are set based on preliminary validation. Detailed algorithms and threshold sensitivity analysis are in Appendix B.2.

## 5.2 Evaluation Protocol

As described in Section 3, LALMs receive all target speaker turns at once and judge consistency across the full dialogue. Our primary evaluation uses audio-only input with a reference sample of at least 3 seconds (Wang et al., 2023) from the target speaker. This setting reflects realistic TTS scenarios where target speaker samples guide generation (Li et al., 2025), and enables fair comparison with embedding-based methods that also operate on audio-only input. To isolate the contribution of different factors, we also examine per-turn comparison against the reference (same as Reference comparison in Speaker Embedding Methods; full results in Appendix C), removal of the reference audio, and addition of textual context (Section 7).

**Evaluation Metrics.** For Detection, we report per-scenario accuracy and Balanced Accuracy:

$$\text{Bal} = \frac{1}{2} \left( \text{Acc}_{S1} + \frac{\text{Acc}_{S2} + \text{Acc}_{S3}}{2} \right)$$

which equally weights consistent (S1) and inconsistent (S2/S3) scenarios. This balancing is necessary because S1 and S2/S3 penalize opposite model behaviors: a model that always predicts consistent achieves 100% on S1 but 0% on S2/S3, and vice versa. For Localization, we compute Precision, Recall, and F1-score per dialogue instance and report their macro-averages across all instances for each scenario, along with Balanced F1 following the same balancing scheme. For Discrimination, the classification reports per-dataset accuracy and overall accuracy across all samples, while the ranking reports NDCG@1, NDCG@2, and Exact Match. Detailed definitions are in Appendix B.3.

## 6 Main Results

### 6.1 Detection and Localization Performance

**LALMs.** As shown in Table 2 and Figure 5, detection results reveal that LALMs lack balanced internal thresholds for speaker consistency judgment. Models cluster along the anti-diagonal (i.e., high S1 with low S2/S3, or vice versa): those like MiniCPM-o-2.6 and Audio-Flamingo-3 overwhelmingly predict consistent, failing to detect

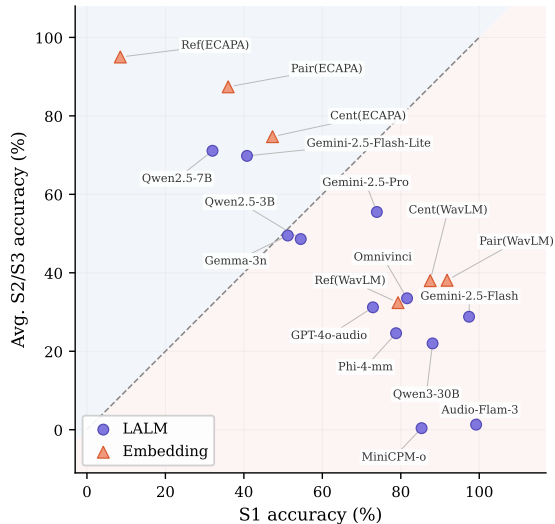


Figure 5: **Detection threshold bias across models.** Each point shows a model’s S1 accuracy vs. average S2/S3 accuracy.

even obvious speaker changes, while Gemini-2.5-Flash-Lite and Qwen2.5-Omni-7B exhibit the opposite bias. This instability results in poor balanced accuracy, with most models scoring below 60%. The best model, Gemini-2.5-Pro, achieves 64.7% but remains notably weak on S3 (39.3%), indicating it can detect gender switches but struggles when the substituted speaker is acoustically similar.

Localization results further reveal the limitations of current LALMs. Most models exhibit extreme behavior: some default to marking no turns as inconsistent (e.g., Phi-4-multimodal with near-zero F1 across S2/S3), while others flag nearly all turns indiscriminately (e.g., Gemma-3n-E4B with 95% recall but 19% precision, at the chance level for 5 turns). Critically, models in the latter group show near-identical scores across S2 and S3, confirming that they cannot distinguish gender switches from subtle timbre differences. These models flag turns based on a fixed bias rather than actual acoustic content. Only Gemini-2.5-Pro (S2/S3 F1: 62.5%/47.5%) and Gemini-2.5-Flash (62.3%/45.0%) maintain meaningful precision alongside high recall, and notably show a drop from S2 to S3, indicating that these models do respond to acoustic difficulty.

**Speaker Embedding Methods.** Speaker embedding methods achieve comparable detection performance, with Pairwise (WavLM) reaching 64.9% balanced accuracy, but exhibit the same systematic biases: ECAPA-TDNN-based methods over-detect changes while under-performing on S1, and

WavLM-based methods show the opposite pattern. For localization, even methods with strong detection do not proportionally improve at pinpointing the inconsistent turn (best: 61.7% balanced F1), suggesting that the ability to detect inconsistency does not transfer to pinpointing where it occurs.

## 6.2 Discrimination Performance

**LALMs.** Table 2 presents discrimination results. Compared to detection, discrimination performance improves substantially for stronger models: Gemini-2.5-Pro achieves 81.5% classification accuracy with 92.6% NDCG@2 and 71.5% Exact Match, followed by Gemini-2.5-Flash (75.6%, 88.3%, 61.6%) and Qwen3-Omni-30B-A3B (60.4%, 73.2%, 36.8%). This dissociation between detection and discrimination validates our task design: models that struggle with absolute binary judgments due to unstable thresholds can still perceive acoustic differences when comparing candidates.

**Speaker Embedding Methods.** Pairwise and Centroid methods achieve near-perfect classification (93–99%) and high NDCG@2 (91–93%), but Exact Match drops to 55–59%: they reliably identify the best match but cannot order the remaining candidates. Gemini-2.5-Pro shows the opposite tradeoff, with lower classification accuracy (81.5%) but substantially higher Exact Match (71.5%). Embedding methods and LALMs thus exhibit complementary strengths on discrimination: embeddings excel at pinpointing the closest match, while Gemini-2.5-Pro better captures the relative ordering among candidates.

## 7 Further Analyses of LALMs

### 7.1 Impact of Textual Context

Our main results evaluate models using only the target speaker’s audio turns. A natural hypothesis is that providing the interlocutors’ turns as text would help models better focus on the target speaker’s voice by anchoring the conversational flow, allowing them to allocate more attention to acoustic features of the target. To test this, we provide the full dialogue to LALM judges, with non-target interlocutors’ turns in text form while the target speaker’s turns remain as audio. By construction, all dialogues are textually coherent (Section 4.3), so the text itself offers no signal of inconsistency, allowing us to test whether textual context helps models focus on acoustic features.

Model	Impact of Textual Context									Impact of Reference Audio								
	S1			S2			S3			S1			S2			S3		
	Audio	+C	Δ	Audio	+C	Δ	Audio	+C	Δ	w/ Ref	w/o	Δ	w/ Ref	w/o	Δ	w/ Ref	w/o	Δ
GPT-4o-audio	72.9	93.4	<b>+20.5</b>	32.8	6.3	<b>-26.5</b>	29.5	5.0	<b>-24.5</b>	72.9	80.5	<b>+7.6</b>	32.8	16.2	<b>-16.6</b>	29.5	13.9	<b>-15.6</b>
Gemini-2.5-Pro	73.9	34.5	<b>-39.4</b>	71.6	46.8	<b>-24.8</b>	39.3	32.8	<b>-6.5</b>	73.9	47.4	<b>-26.5</b>	71.6	38.4	<b>-33.2</b>	39.3	16.7	<b>-22.6</b>
Gemini-2.5-Flash	97.4	91.9	<b>-5.5</b>	45.5	16.7	<b>-28.8</b>	12.0	10.7	<b>-1.3</b>	97.4	97.0	<b>-0.4</b>	45.5	41.3	<b>-4.2</b>	12.0	15.5	<b>+3.5</b>
Gemini-2.5-Flash-Lite	40.8	93.4	<b>+52.6</b>	70.3	3.3	<b>-67.0</b>	69.3	3.3	<b>-66.0</b>	40.8	92.6	<b>+51.8</b>	70.3	13.9	<b>-56.4</b>	69.3	12.5	<b>-56.8</b>
Qwen2.5-Omni-3B	54.5	85.3	<b>+30.8</b>	49.3	15.8	<b>-33.5</b>	47.8	12.7	<b>-35.1</b>	54.5	92.7	<b>+38.2</b>	49.3	10.6	<b>-38.7</b>	47.8	9.0	<b>-38.8</b>
Qwen2.5-Omni-7B	32.0	59.8	<b>+27.8</b>	72.3	41.7	<b>-30.6</b>	69.8	42.5	<b>-27.3</b>	32.0	98.5	<b>+66.5</b>	72.3	2.8	<b>-69.5</b>	69.8	2.0	<b>-67.8</b>
Qwen3-Omni-30B-A3B	88.1	93.1	<b>+5.0</b>	29.9	8.6	<b>-21.3</b>	14.0	6.9	<b>-7.1</b>	88.1	99.8	<b>+11.7</b>	29.9	11.7	<b>-18.2</b>	14.0	1.2	<b>-12.8</b>
MiniCPM-o-2.6	85.3	85.7	<b>+0.4</b>	0.7	0.0	<b>-0.7</b>	0.0	0.0	0.0	85.3	86.5	<b>+1.2</b>	0.7	3.1	<b>+2.4</b>	0.0	0.3	<b>+0.3</b>
Gemma-3n-E4B	51.2	93.9	<b>+42.7</b>	50.6	5.8	<b>-44.8</b>	48.4	5.8	<b>-42.6</b>	51.2	36.4	<b>-14.8</b>	50.6	66.9	<b>+16.3</b>	48.4	65.6	<b>+17.2</b>
Phi-4-multimodal	78.8	61.5	<b>-17.3</b>	24.9	39.2	<b>+14.3</b>	24.3	38.2	<b>+13.9</b>	78.8	92.5	<b>+13.7</b>	24.9	8.1	<b>-16.8</b>	24.3	7.3	<b>-17.0</b>
Omnivinci	81.6	98.7	<b>+17.1</b>	46.1	2.3	<b>-43.8</b>	20.9	1.5	<b>-19.4</b>	81.6	88.1	<b>+6.5</b>	46.1	61.5	<b>+15.4</b>	20.9	16.3	<b>-4.6</b>
Audio-Flamingo-3	99.2	97.5	<b>-1.7</b>	1.3	1.8	<b>+0.5</b>	1.2	1.7	<b>+0.5</b>	99.2	95.7	<b>-3.5</b>	1.3	4.3	<b>+3.0</b>	1.2	3.8	<b>+2.6</b>

Table 3: **Impact of Textual Context and Reference Audio on Detection Accuracy (%)**. Left: performance change when adding textual context (+C) vs. audio-only (Audio). Right: performance change when removing reference audio (w/o) vs. with reference (w/ Ref). Δ denotes the difference.

Table 3 reveals the opposite: adding textual context degrades rather than improves acoustic judgment. For most models, it sharply improves S1 accuracy while collapsing S2/S3 accuracy, with the exception of Gemini-2.5-Pro, which degrades across all scenarios, and Phi-4-multimodal, whose audio-only baseline already collapses, inverting the pattern. This asymmetric pattern indicates that models default to judging speakers as consistent whenever the dialogue text flows naturally, regardless of acoustic evidence, even failing to detect obvious gender switches. Rather than helping models focus on the target speaker’s voice, textual context shifts their judgment toward text-based reasoning, revealing a modality imbalance that may reflect disproportionate attention to text over audio tokens in LALMs (Wang et al., 2025b). Localization results exhibit a similar pattern (Table 12). These findings reveal that improving LALMs as speaker consistency judges requires not just better acoustic representations, but mechanisms to balance attention allocation during multi-modal fusion. Developing methods to better leverage dialogue context while maintaining acoustic sensitivity would be a promising future direction.

## 7.2 Impact of Reference Audio

In our main evaluation, we provide models with a single reference audio sample from the target speaker to establish a comparison baseline. We investigate what happens when this reference is absent, requiring models to judge consistency solely from the dialogue turns themselves.

Table 3 shows that for several models, removing reference audio leads them to default to *Consistent* judgments: S1 accuracy increases sharply while S2/S3 accuracy drops substantially. Without an

explicit comparison anchor, these models adopt lenient thresholds, failing to detect even obvious inconsistencies such as gender switches. Localization results exhibit the same pattern (Table 12). This reveals that many LALMs fail to establish appropriate decision boundaries without explicit references. Rather than developing robust internal speaker representations from dialogue context alone, they rely heavily on explicit reference audio to anchor their judgments. This pattern connects to our earlier finding from the Discrimination task (Table 2), where models achieved better performance through relative judgment. These findings have important practical implications: in real-world applications such as TTS validation, providing reference audio is essential for reliable speaker consistency judgment.

## 7.3 Effect of Speaker Turns and Clip Duration

To examine whether our findings generalize across dialogue lengths, we construct a 10-turn dataset comprising 759 instances from 253 unique dialogues using the same pipeline as the primary benchmark. As shown in Table 4, the 10-turn setting is slightly more challenging on average (−1.2% Detection, −4.2% Localization, −3.1% Discrimination), though individual trends vary. Model rankings are largely preserved, indicating our findings are not specific to the 5-turn setting (full results in Table 11; analysis in Appendix D.2).

We further analyze how clip duration affects performance on inconsistent scenarios (S2/S3). Figure 6 shows results for the top three LALMs by Discrimination accuracy (Gemini-2.5-Pro, Gemini-2.5-Flash, Qwen3-Omni-30B-A3B), grouped by duration quartile. Both settings exhibit a clear monotonic trend: longer clips yield higher Detec-

Model	Det Bal (%)			Loc Bal F1 (%)			Disc Acc (%)		
	5t	10t	$\Delta$	5t	10t	$\Delta$	5t	10t	$\Delta$
GPT-4o-audio	52.0	51.3	-0.7	42.3	38.8	-3.5	40.7	39.2	-1.5
Gemini-2.5-Pro	<b>64.7</b>	<b>69.3</b>	<b>+4.6</b>	<b>60.1</b>	55.4	-4.7	<b>81.5</b>	80.1	-1.4
Gemini-2.5-Flash	63.1	56.3	-6.8	59.0	<b>62.4</b>	<b>+3.4</b>	75.6	<b>82.2</b>	<b>+6.6</b>
Gemini-2.5-Flash-Lite	55.3	48.1	-7.2	18.1	15.1	-3.0	46.5	46.6	+0.1
Qwen2.5-Omni-3B	51.5	51.8	+0.3	31.8	19.8	-12.0	40.8	40.7	-0.1
Qwen2.5-Omni-7B	51.5	51.1	-0.4	29.9	35.6	+5.7	42.7	38.7	-4.0
Qwen3-Omni-30B-A3B	55.0	50.9	-4.1	11.8	15.4	+3.6	60.4	54.9	-5.5
MiniCPM-o-2.6	42.8	50.2	+7.4	46.6	36.3	-10.3	49.5	37.9	-11.6
Gemma-3n-E4B	50.4	51.5	+1.1	16.0	8.2	-7.8	37.1	35.2	-1.9
Phi-4-multimodal	51.7	50.2	-1.5	49.9	50.0	+0.1	39.9	31.6	-8.3
Omnivinci	57.5	52.0	-5.5	33.8	16.4	-17.4	45.4	40.3	-5.1
Audio-Flamingo-3	50.2	48.5	-1.7	11.8	7.7	-4.1	36.3	32.0	-4.3

Table 4: 5-turn vs. 10-turn comparison. Det Bal: Balanced Detection Accuracy. Loc Bal F1: Balanced Localization F1. Disc Acc: Discrimination classification accuracy.  $\Delta = 10t - 5t$ .

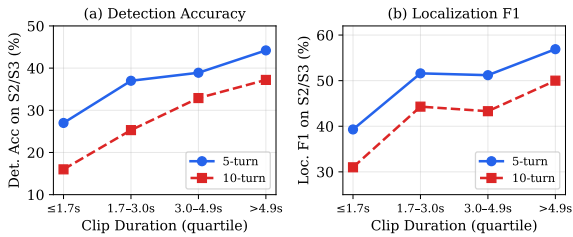


Figure 6: Performance on inconsistent scenarios (S2/S3) by clip duration quartile for the top three LALMs.

tion Accuracy and Localization F1, confirming that models require sufficient acoustic evidence within each clip for reliable speaker judgment.

#### 7.4 LALMs as Voice Cloning Evaluators

Beyond judging speaker consistency across dialogues, a natural downstream application is evaluating voice cloning systems: automatically ranking outputs from different models by their acoustic similarity to a reference speaker. We investigate whether LALMs can serve this role reliably, producing rankings that agree with human judgment.

To explore this, we construct a VC Quality Ranking task. From the benchmark dialogues, we randomly sample a pair of audio clips per dialogue: one serving as the reference speaker audio and the other as the source. We extract the transcript from the source clip and use it along with the reference audio to generate cloned outputs via three models — OpenVoice (Qin et al., 2023), YourTTS (Casanova et al., 2022), and CosyVoice3 (Du et al., 2025) — yielding four candidates per sample (source, OpenVoice, YourTTS, CosyVoice3). Three human annotators independently ranked 485 samples, with 50 shared samples for inter-annotator reliability (Kendall’s  $W = 0.860$ ). Human rankings serve as ground truth. Models are asked to rank all four candidates by

Model	Acc	NDCG@1	NDCG@2	EM
GPT-4o-audio	7.3	10.5	12.7	1.7
Gemini-2.5-Pro	<b>48.0</b>	<b>63.0</b>	<b>75.1</b>	<b>16.8</b>
Gemini-2.5-Flash	37.1	53.8	67.4	8.9
Gemini-2.5-Flash-Lite	27.0	43.8	55.2	5.6
Qwen2.5-Omni-3B	23.1	39.2	51.2	4.9
Qwen2.5-Omni-7B	25.8	41.9	53.6	5.4
Qwen3-Omni-30B-A3B	37.1	52.9	62.7	11.5
MiniCPM-o-2.6	32.0	45.4	57.1	6.6
Gemma-3n-E4B	26.0	40.0	49.3	3.5
Phi-4-multimodal	27.6	42.2	51.3	5.4
Omnivinci	27.6	41.2	51.3	5.8
Audio-Flamingo-3	30.3	43.0	53.4	4.3
Random	25.0	39.3	50.4	4.2

Table 5: VC Quality Ranking results.

acoustic similarity to the reference speaker. Evaluation protocol is the same as the Discrimination task (§ 5.2). The prompt is provided in Figure 16.

As shown in Table 5, Gemini-2.5-Pro achieves the strongest performance (Acc: 48.0%, NDCG@2: 75.1%), followed by Gemini-2.5-Flash and Qwen3-Omni-30B-A3B. However, most models perform modestly above the random baseline, and Exact Match remains low even for the best model (16.8%). Unlike the Discrimination task, where candidates originate from distinct speakers, all candidates here are generated from the same source audio targeting the same speaker identity, requiring models to perceive much finer-grained acoustic variations. These results indicate that while some LALMs show promise as automatic voice cloning evaluators, reliable quality ranking aligned with human judgment remains a challenging open problem.

## 8 Conclusion

We present SpeakerSleuth, a benchmark for evaluating whether LALMs can reliably judge speaker consistency across multi-turn dialogues. Built around three complementary tasks that mirror real-world application requirements (Detection, Localization, and Discrimination), SpeakerSleuth comprises 1,818 human-verified instances across four diverse datasets. Our evaluation reveals fundamental limitations: models lack stable internal thresholds, struggle with fine-grained turn-level analysis, and prioritize textual coherence over acoustic features. At the same time, they show a clear dissociation between detection and discrimination, indicating that acoustic discrimination capability is present but not effectively integrated into consistency judgment. These findings point to calibration, fine-grained reasoning, and modality integration as core challenges for building reliable audio-language judges.

## Limitations

SpeakerSleuth has several limitations. First, our benchmark covers only English dialogues. The synthetic generation pipeline itself is language-agnostic and can be extended to other languages. Second, our four datasets span diverse acoustic conditions (TV shows, meetings, synthesized speech, studio recordings), but we do not isolate individual acoustic factors such as background noise, reverberation, or recording quality. While our pipeline readily supports adding controlled perturbations via standard audio augmentation, systematic analysis of each factor is left to future work. Finally, we do not analyze how model performance varies across speaker demographics such as accents or age groups. Whether LALMs exhibit demographic biases in speaker consistency judgment remains an open question. Despite these limitations, SpeakerSleuth provides a systematic framework for probing how LALMs reason about speaker identity in multi-turn dialogues, and the pipeline naturally accommodates future extensions along each of the axes above.

## Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants (RS-2024-00333484 and RS-2024-00414981) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (RS-2025-02215122, Development and Demonstration of Lightweight AI Model for Smart Homes), all funded by the Korean government (MSIT).

## References

Apple. 2024. Apple intelligence: Ai for the rest of us. <https://www.apple.com/apple-intelligence/>. Accessed 2026-04-19.

Shrestha Basu Mallick, Sid Lall, Zach Gleicher, and Kate Olszewska. 2025. Continuing to bring you our latest models, with an improved gemini 2.5 flash and flash-lite release. <https://developers.googleblog.com/en/continuing-to-bring-you-our-latest-models-with-an-improved-gemini-2-5-flash-and-flash-lite-release/>. Google Developers Blog; Accessed 2026-04-19.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audioldm: a language modeling approach to audio generation.

*IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.
- Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pages 2709–2720. PMLR.
- Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhe-huai Chen, Chao Zhang, Chao-Han Huck Yang, and EngSiong Chng. 2025. Audio large language models can be descriptive speech quality evaluators. In *The Thirteenth International Conference on Learning Representations*.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Victor Danell. 2025. Watch the skies. <https://www.imdb.com/title/tt14807348/>. Accessed 2026-04-19.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuyck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*. ISCA.
- Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.
- Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.
- Sadaoki Furui. 1996. An overview of speaker recognition technology. *Automatic Speech and Speaker Recognition: Advanced Topics*, pages 31–56.

- 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. 2025. [Audio flamingo 3: Advancing audio intelligence with fully open large audio language models](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Herbert Gish and Michael Schmidt. 1994. Text-independent speaker identification. *IEEE signal processing magazine*, 11(4):18–32.
- Google. 2024. NotebookLM now lets you listen to a conversation about your sources. <https://blog.google/technology/ai/notebooklm-audio-overviews/>. Accessed 2026-04-19.
- Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. [Prompttts: Controllable text-to-speech with text descriptions](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. 2024. [Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models](#). In *International Conference on Machine Learning*, pages 22605–22623. PMLR.
- Zeqian Ju, Dongchao Yang, Kai Shen, Yichong Leng, Zhengtao Wang, Songxiang Liu, Xinyu Zhou, Tao Qin, Xiangyang Li, Jianwei Yu, and Xu Tan. 2025. [Mooncast: High-quality zero-shot podcast generation](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Koray Kavukcuoglu. 2025. Gemini 2.5: Our most intelligent ai model. <https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-thinking-updates-march-2025/>. Google Developers Blog; Accessed 2026-04-19.
- Volodymyr Khoma, Yuriy Khoma, Vitalii Brydinskyi, and Alexander Kononov. 2023. [Development of supervised speaker diarization system based on the pyannote audio processing library](#). *Sensors*, 23(4).
- R. Kubichek. 1993. [Mel-cepstral distance measure for objective speech quality assessment](#). In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128 vol.1.
- Jonggeun Lee, Junseong Pyo, Jeongmin Park, and Yohan Jo. 2026. [Spokenus: A spoken user simulator for task-oriented dialogue](#). *arXiv preprint arXiv:2603.16783*.
- Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. [Dailytalk: Spoken dialogue dataset for conversational text-to-speech](#). In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Sehun Lee, Kang-wook Kim, and Gunhee Kim. 2025. [Behavior-SD: Behaviorally aware spoken dialogue generation with large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9574–9593, Albuquerque, New Mexico. Association for Computational Linguistics.
- Paul Lerner, Juliette Bergoënd, Camille Guinaudeau, Hervé Bredin, Benjamin Maurice, Sharleyne Lefevre, Martin Bouteiller, Aman Berhe, Léo Galmant, Ruiqing Yin, and Claude Barras. 2022. [Bazinga! a dataset for multi-party dialogues structuring](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3434–3441, Marseille, France. European Language Resources Association.
- Jingyi Li, Weiping Tu, and Li Xiao. 2023. [Freevc: Towards high-quality text-free one-shot voice conversion](#). In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yinghao Aaron Li, Cong Han, and Nima Mesgarani. 2025. [Styletts: A style-based generative model for natural and diverse text-to-speech synthesis](#). *IEEE Journal of Selected Topics in Signal Processing*.
- Philipos C. Loizou. 2011. *Speech Quality Assessment*, pages 623–654. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).
- Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. [Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras](#). *Preprint*, arXiv:2503.01743.
- John W. Mullennix and David B. Pisoni. 1990. [Stimulus variability and processing dependencies in speech](#)

- perception. *Perception & Psychophysics*, 47:379–390.
- OpenAI. 2024. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>. Accessed 2026-04-19.
- OpenBMB. 2025. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone. <https://openbmb.vercel.app/minicpm-o-2-6-en>. Accessed 2026-04-19.
- Se Jin Park, Julian Salazar, Aren Jansen, Keisuke Kinoshita, Yong Man Ro, and RJ Skerry-Ryan. 2025. Long-form speech generation with spoken language models. In *Forty-second International Conference on Machine Learning*.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.
- Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer. 2011. Crowdmos: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2416–2419.
- A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2.
- Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Interspeech 2022*. ISCA.
- David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. 2017. Deep neural network embeddings for text-independent speaker verification. In *Interspeech 2017*, pages 999–1003.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffroy Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Keanealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesh Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petriani, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szepktor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. *Gemma 3 technical report*. *Preprint*,

- arXiv:2503.19786.
- Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.
- Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, Yan Lu, and Yong Qin. 2025a. [Speechllm-as-judges: Towards general and interpretable speech quality evaluation](#). *Preprint*, arXiv:2510.14664.
- Junyu Wang, Ziyang Ma, Zhengding Luo, Tianrui Wang, Meng Ge, Xiaobao Wang, and Longbiao Wang. 2025b. Pay more attention to audio: Mitigating imbalance of cross-modal attention in large audio language models. *arXiv preprint arXiv:2509.18816*.
- Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. 2025c. [QualiSpeech: A speech quality assessment dataset with natural language reasoning and descriptions](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23588–23609, Vienna, Austria. Association for Computational Linguistics.
- Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, et al. 2025d. Enabling auditory large language models for automatic speech quality evaluation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech 2017*, pages 4006–4010.
- Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. 2025. [Towards controllable speech synthesis in the era of large language models: A systematic survey](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 764–791, Suzhou, China. Association for Computational Linguistics.
- 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. 2025a. [Qwen2.5-omni technical report](#). *Preprint*, arXiv: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. 2025b. [Qwen3-omni technical report](#). *Preprint*, arXiv:2509.17765.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chuji Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. 2025. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*.
- Chen Zhang, Luis Fernando D’Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024a. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19515–19524.
- Leying Zhang, Yao Qian, Xiaofei Wang, Manthan Thakker, Dongmei Wang, Jianwei Yu, Haibin Wu, Yuxuan Hu, Jinyu Li, Yanmin Qian, and sheng zhao. 2025. [Covomix2: Advancing zero-shot dialogue generation with fully non-autoregressive flow matching](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Leying Zhang, Yao Qian, Long Zhou, Shujie Liu, Dongmei Wang, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Lei He, et al. 2024b. Covomix: Advancing zero-shot speech generation for human-like multi-talker conversations. *Advances in Neural Information Processing Systems*, 37:100291–100317.

## A SpeakerSleuth Details

Table 6: Breakdown of TV series and episodes used (all from Season 1) from Bazinga dataset in SpeakerSleuth.

TV Series/Movie	Episodes
24	24
Battlestar Galactica	13
Breaking Bad	7
Buffy The Vampire Slayer	12
ER	25
Friends	24
Game of Thrones	10
Homeland	12
Lost	25
Six Feet Under	13
Star Wars	7
The Big Bang Theory	17
The Office	6
The Walking Dead	6
<b>Total</b>	<b>201 episodes</b>

### A.1 Dataset Details

We use four datasets for SpeakerSleuth. This section describes the specific subset used from each and its licensing. Table 1 summarizes the resulting statistics.

**Bazinga.** Bazinga (Lerner et al., 2022) is a multi-party dialogue dataset from TV series and movies. We use 14 series spanning 201 episodes total, covering comedy, drama, and documentary formats to capture diverse speaking styles. Table 6 provides the per-series breakdown.

**AMI Meeting Corpus.** The AMI Meeting Corpus (Carletta et al., 2005) consists of 100 hours of meeting recordings across three rooms, predominantly featuring non-native speakers. We use the evaluation set of 16 meetings, which provides challenging real-world acoustic conditions: spontaneous disfluencies, overlapping speech, variable room acoustics, and non-native accents.

**Behavior-SD.** Behavior-SD (Lee et al., 2025) is a large-scale dataset of synthesized dialogues (100K+ dialogues, 2,164 hours) with annotations for conversational behaviors such as fillers, backchannels, and interruptions. We use the test set of 925 dialogues. Behavior-SD serves as a control condition with clean, synthesized speech against the more challenging real-world recordings from other datasets.

**DailyTalk.** DailyTalk (Lee et al., 2023) is a high-quality conversational TTS dataset derived from

DailyDialog, containing 2,541 studio-quality dialogues between two participants (one male, one female). We use all 2,541 dialogues. While the speaker diversity is limited, the consistent studio conditions enable evaluation of within-speaker consistency across varied conversational scenarios.

**Licensing.** All datasets are used under their respective licenses: AMI Corpus and Behavior-SD under CC BY 4.0, Bazinga under CC BY-NC 4.0, and DailyTalk under CC BY-SA 4.0. Our use for benchmark evaluation is consistent with their intended academic research purposes.

### A.2 Verification Details

This section details the verification pipeline outlined in Section 4.3. For text-based filtering, we use Qwen3-32B (Yang et al., 2025) with the prompt shown in Figure 17. For audio-based verification, three annotators from our research team, all with expertise in speech processing and audio evaluation, evaluated all samples based on the criteria below. Annotators also confirmed that each reference audio shares consistent speaker identity and acoustic environment with the target speaker’s turns in the dialogue, ensuring it serves as a reliable anchor for acoustic comparison.

**Audio Quality.** Annotators checked for excessive background noise interfering with speaker characteristics, confirmed clear human speech in all utterances, and flagged excessive clipping or silence.

**Naturalness.** Voice-converted turns were checked for robotic artifacts, and annotators verified that pitch, timbre, tone, and emotion flowed naturally throughout each turn.

**Scenario Validity.** For S1, annotators confirmed consistent speaker identity and acoustic environment across all turns. For S2, they verified clear gender distinction between the original and converted speaker. For S3, they ensured the substituted speaker was acoustically distinguishable from the target, despite the intended similarity.

### A.3 Dialogue and Speaker Embedding Visualization Details

The visualization in Figure 4 was generated using the UMAP (McInnes et al., 2018) algorithm with `n_neighbors` set to 15. For dialogue embeddings, we employed the all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) text embedding model, while

speaker embeddings were extracted using ECAPA-TDNN (Desplanques et al., 2020).

## B Experimental Setup Details

### B.1 LALM Judges

Table 7 summarizes the twelve LALMs we evaluate, along with their parameter counts. All inference is conducted with temperature 0 on NVIDIA A100 80GB GPUs using CUDA 12.4.

Model	Parameters
<i>Proprietary</i>	
GPT-4o-audio	N/A
Gemini-2.5-Pro	N/A
Gemini-2.5-Flash	N/A
Gemini-2.5-Flash-Lite	N/A
<i>Open-source</i>	
Qwen2.5-Omni-3B	3B
Qwen2.5-Omni-7B	7B
Qwen3-Omni-30B-A3B	30B total, 3B active
MiniCPM-o-2.6	8B
Gemma-3n-E4B	8B total, 4B effective
Phi-4-multimodal	5.6B
OmniVinci	9B
Audio-Flamingo-3	7B

Table 7: LALM judges evaluated in our benchmark. “N/A” indicates parameter counts not publicly disclosed.

### B.2 Speaker Embedding Methods

We describe the three speaker embedding methods introduced in Section 5.1. For each target speaker audio set  $\mathcal{A}_S = \{a_i\}_{i \in I}$ , we first extract per-turn speaker embeddings  $e_i$  from  $a_i$  using either WavLM (Chen et al., 2022) or ECAPA-TDNN (Desplanques et al., 2020). The three methods differ in how they aggregate these embeddings to produce a consistency score per turn and a similarity score per candidate.

**Pairwise Similarity.** The per-turn consistency score is the mean pairwise cosine similarity:

$$s_i = \frac{1}{|I|-1} \sum_{j \neq i} \text{sim}(e_i, e_j).$$

Turn  $i$  is flagged if  $s_i < \tau_{\text{pair}}$ . For *Discrimination*, each candidate  $o_j$  is scored as

$$q_j = \frac{1}{|I|} \sum_{i \in I} \text{sim}(o_j, e_i).$$

**Centroid Distance.** With context centroid  $c = \frac{1}{|I|} \sum_{i \in I} e_i$  and  $\text{dist}(x, y) = 1 - \text{sim}(x, y)$ , the per-turn score is

$$s_i = \text{dist}(e_i, c).$$

Turn  $i$  is flagged if  $s_i > \tau_{\text{cent}}$ . For *Discrimination*, each candidate is scored as

$$q_j = \text{dist}(o_j, c).$$

**Reference Comparison.** Given a reference embedding  $r$  extracted from the target speaker’s reference audio, the per-turn score is

$$s_i = \text{sim}(e_i, r).$$

Turn  $i$  is flagged if  $s_i < \tau_{\text{ref}}$ . For *Discrimination*, each candidate is scored as

$$q_j = \text{sim}(o_j, r).$$

**Task Application.** Given these per-method scores, we apply the three tasks uniformly. *Detection* classifies  $\mathcal{A}_S$  as inconsistent if any turn is flagged. *Localization* outputs the flagged turn(s). *Discrimination* returns  $\arg \max_j q_j$  (or  $\arg \min$  for Centroid Distance) for the classification formulation, and  $\{q_j\}$  sorted in descending order (ascending for Centroid) for the ranking formulation.

**Threshold Sensitivity Analysis** We sweep  $\tau$  from 0.1 to 0.7 in 0.1 increments for each of the six method-extractor combinations and evaluate detection accuracy and localization F1-score. As Figure 7 shows, optimal thresholds vary substantially across configurations. Our chosen values ( $\tau_{\text{pair}} = \tau_{\text{ref}} = 0.4$ ,  $\tau_{\text{cent}} = 0.3$ ) provide reasonable performance across both backbones.

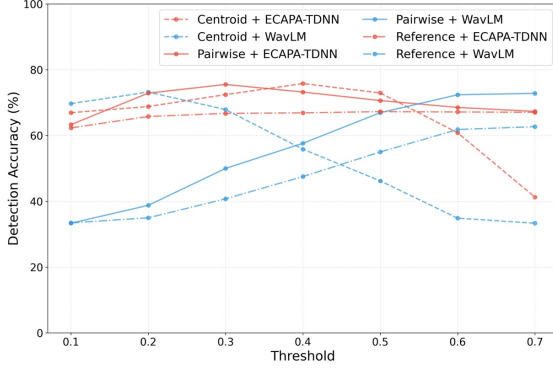
### B.3 Evaluation Metrics

This section defines the evaluation metrics used for the three tasks.

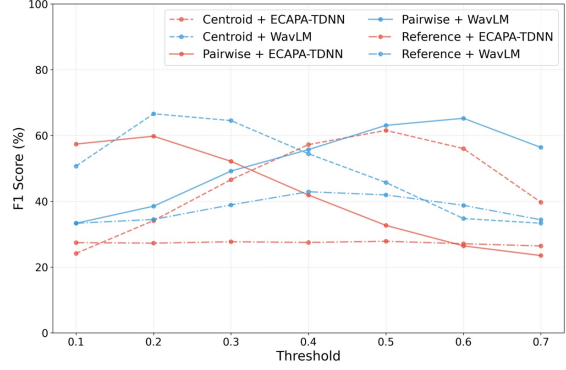
**Task 1: Detection.** Detection is a binary classification problem where the model predicts whether a dialogue is consistent or inconsistent. Since Scenario 1 (S1) contains only fully consistent dialogues and Scenarios 2 and 3 (S2, S3) contain only dialogues with an inconsistent turn, per-scenario accuracy reduces to the proportion of dialogues correctly classified within each scenario:

$$\text{Acc}_S = \frac{1}{N_S} \sum_{d \in S} \mathbb{1}[\hat{y}_d = y_d],$$

where  $N_S$  is the number of dialogues in scenario  $S \in \{S1, S2, S3\}$ ,  $\hat{y}_d$  is the prediction, and  $y_d$  is the ground-truth label. Reporting accuracy per scenario reveals threshold calibration patterns that would be hidden under a single aggregate score.



(a) Detection Accuracy



(b) Localization F1

Figure 7: Detection and Localization Performance with varying thresholds  $\tau$ .

**Task 2: Localization.** Given all target speaker turns at once, the model outputs the set of turns it judges to be inconsistent (or None). We evaluate this output as a multi-label classification: for each sample, let  $\hat{G}$  be the predicted set and  $G$  the ground-truth set of inconsistent turns. Per-sample precision and recall are  $P = |\hat{G} \cap G|/|\hat{G}|$  and  $R = |\hat{G} \cap G|/|G|$ , with F1 defined as

$$F1 = \frac{2P \cdot R}{P + R}.$$

We use the convention  $P = 0$  when  $|\hat{G}| = 0$ ,  $R = 0$  when  $|G| = 0$ , and  $F1 = 1$  when  $|\hat{G}| = |G| = 0$ . All metrics are macro-averaged across samples.

**Task 3: Discrimination.** Discrimination presents three candidates, one from each scenario (S1, S2, S3), in a randomly shuffled order to avoid positional bias. We evaluate two formulations.

*Classification.* The model selects the single best-matching candidate (i.e., S1, the original). We report accuracy over  $N$  samples:

$$\text{Acc} = \frac{1}{N} \sum_{j=1}^N \mathbb{1}[\hat{c}_j = c_j],$$

where  $\hat{c}_j$  is the predicted candidate and  $c_j$  is the ground-truth candidate (i.e., S1).

*Ranking.* The model orders all three candidates by acoustic similarity to the target speaker. The ground-truth ordering is  $S1 \succ S3 \succ S2$ : S1 is the original consistent clip; S3 is voice-converted from the most acoustically similar speaker; and S2 is voice-converted from a speaker of a different gender, the most dissimilar option. Relevance scores are assigned as 2, 1, and 0 for ranks 1, 2, and 3,

respectively. Per sample,  $\text{NDCG}@k$  is computed as

$$\text{NDCG}@k = \frac{\text{DCG}@k}{\text{IDCG}@k},$$

$$\text{DCG}@k = \sum_{i=1}^k \frac{2^{\text{rel}_i} - 1}{\log_2(i + 1)}$$

where  $\text{rel}_i$  is the relevance score of the candidate at rank  $i$  and  $\text{IDCG}@k$  is the DCG of the ideal ranking.  $\text{NDCG}@k$  is averaged across samples. We additionally report Exact Match, the proportion of samples where the full predicted ranking matches the ground truth.

## C Per-Turn Reference Comparison

Our main evaluation presents all target speaker turns simultaneously, allowing models to reason over the collective acoustic pattern across turns. Here we evaluate an alternative *per-turn* setting that mirrors the Reference Comparison in speaker embedding methods (Appendix B.2): each turn is independently compared to the reference audio with the prompt “Is this clip from the same speaker as the reference?” (Figure 14). For Detection, a dialogue is marked inconsistent if the model answers “No” on any turn; for Localization, all turns answered “No” are returned as the predicted inconsistent set. For Discrimination, each candidate is presented together with only the reference audio, omitting the surrounding consistent turns available in the main setting (Figure 15).

Table 8 compares this setting to our main evaluation, revealing structural limitations across all three tasks. For Detection, most models collapse to flagging every dialogue: any single “No” response suffices to mark the dialogue inconsistent,

Model	Detection				Localization								Discrimination				
	S1	S2	S3	Bal	S1	S2		S3		Bal	Classification						
	Acc				F1	P	R	F1	P	R	F1	F1	Baz	AMI	B-SD	Daily	Avg
<i>Per-Turn Pairwise</i>																	
GPT-4o-audio	0.2	99.8	100.0	50.0	0.2	20.9	98.6	34.3	20.4	97.9	33.7	17.1	55.9	43.5	40.3	30.7	43.0
Gemini-2.5-Pro	18.0	95.0	91.0	55.5	18.0	38.2	82.4	48.8	27.8	67.5	36.7	30.4	41.7	69.6	54.4	51.9	50.3
Gemini-2.5-Flash	11.6	96.4	92.9	53.1	11.6	38.6	89.6	50.4	28.3	72.2	38.2	28.0	47.6	80.4	76.1	74.1	<b>65.8</b>
Gemini-2.5-Flash-Lite	4.1	97.9	98.6	51.2	4.1	26.9	86.0	39.1	24.4	80.5	35.9	20.8	30.7	34.8	47.2	57.7	43.7
Qwen2.5-Omni-3B	0.0	100.0	100.0	50.0	0.0	20.4	99.3	33.8	20.3	98.4	33.5	16.8	53.8	47.8	45.9	45.5	48.7
Qwen2.5-Omni-7B	0.7	100.0	100.0	50.3	0.7	21.8	98.1	35.3	21.6	96.0	34.7	17.9	44.8	30.4	34.0	31.2	36.6
Qwen3-Omni-30B-A3B	2.9	99.5	98.6	51.0	2.9	25.7	96.0	39.1	22.0	87.6	34.3	19.8	79.2	76.1	46.5	44.4	59.6
MiniCPM-o-2.6	85.3	35.6	17.8	<b>56.0</b>	85.3	25.2	28.3	26.1	6.4	7.6	6.7	<b>50.9</b>	36.3	30.4	40.9	42.3	38.9
Gemma-3n-E4B	33.9	68.2	67.2	50.8	33.9	13.3	30.9	17.4	13.6	31.1	17.7	25.7	33.0	34.8	39.6	42.3	37.8
Phi-4-multimodal	11.4	90.2	87.9	50.2	11.4	21.9	59.1	30.3	19.6	54.9	27.4	20.1	31.6	30.4	37.1	41.3	36.0
Omnivinci	0.3	100.0	99.5	50.0	0.3	23.2	99.3	37.1	22.0	94.5	35.1	18.2	41.5	34.8	38.4	41.3	40.1
Audio-Flamingo-3	75.3	27.7	23.2	50.4	75.3	5.6	17.3	7.9	4.4	14.9	6.4	41.2	33.0	30.4	36.5	41.3	36.3
<i>Multi-Turn (from Table 2)</i>																	
GPT-4o-audio	72.9	32.8	29.5	52.0	71.5	12.6	25.9	15.0	8.7	19.2	11.1	42.3	50.2	30.4	38.4	34.4	40.7
Gemini-2.5-Pro	73.9	71.6	39.3	<b>64.7</b>	65.2	59.3	70.3	62.5	44.0	56.6	47.5	<b>60.1</b>	77.8	76.1	81.1	87.2	<b>81.5</b>
Gemini-2.5-Flash	97.4	45.5	12.0	63.1	64.4	55.6	90.1	62.3	38.6	72.4	45.0	59.0	70.8	73.9	69.2	86.8	75.6
Gemini-2.5-Flash-Lite	40.8	70.3	69.3	55.3	0.3	25.8	94.7	36.4	24.5	94.1	35.4	18.1	47.6	45.7	47.8	44.4	46.5
Qwen2.5-Omni-3B	54.5	49.3	47.8	51.5	21.8	45.6	91.6	47.7	28.9	81.6	36.0	31.8	42.9	39.1	41.5	38.1	40.8
Qwen2.5-Omni-7B	32.0	72.3	69.8	51.5	38.0	14.1	61.8	22.6	13.2	57.5	21.0	29.9	57.5	30.4	39.6	31.7	42.7
Qwen3-Omni-30B-A3B	88.1	29.9	14.0	55.0	0.2	15.1	57.4	23.3	15.2	56.0	23.5	11.8	82.1	69.6	43.4	48.1	60.4
MiniCPM-o-2.6	85.3	0.7	0.0	42.8	66.9	29.8	54.5	35.4	12.3	25.1	17.2	46.6	62.7	45.7	42.8	41.3	49.5
Gemma-3n-E4B	51.2	50.6	48.4	50.4	0.0	19.1	95.1	32.0	19.1	94.4	32.0	16.0	36.8	37.0	33.3	40.7	37.1
Phi-4-multimodal	78.8	24.9	24.3	51.7	99.7	0.0	0.1	0.1	0.0	0.1	0.1	49.9	37.3	39.1	39.0	43.9	39.9
Omnivinci	81.6	46.1	20.9	57.5	48.3	14.9	61.2	23.6	8.8	40.0	15.2	33.8	51.4	39.1	42.1	42.9	45.4
Audio-Flamingo-3	99.2	1.3	1.2	50.2	0.0	17.1	38.4	23.3	17.4	38.8	23.7	11.8	33.0	28.3	37.1	41.3	36.3

Table 8: Per-turn reference comparison vs. main results (multi-turn). In the per-turn setting, each turn is independently compared against the reference audio; the main setting presents all turns simultaneously. Detection reports per-scenario and Balanced Accuracy (Bal). Localization reports F1 for S1, Precision (P), Recall (R), and F1 for S2/S3, and Balanced F1. Discrimination reports per-dataset and average Classification Accuracy.

yielding near-zero accuracy on S1 and near-perfect on S2/S3. Balanced Detection Accuracy thus converges to  $\sim 50\%$  across almost all models, compared to the 42.8–64.7% spread in the main setting. Localization exhibits the same over-flagging: Recall exceeds 95% for most models while Precision stays around 20%. Discrimination also degrades—Gemini-2.5-Pro drops from 81.5% to 50.3% and Gemini-2.5-Flash from 75.6% to 65.8%—because a single reference clip provides a weaker acoustic anchor than the full set of consistent turns available in the main setting. These limitations underscore that multi-turn context is not merely a design complexity but a necessary condition for meaningful speaker consistency evaluation.

## D Extended Benchmarks

### D.1 Multiple Voice Conversion Models

To verify that our benchmark findings are not specific to a single voice conversion (VC) model, we construct an extended benchmark by applying three additional VC models—CosyVoice3 (Du et al., 2025), OpenVoice (Qin et al., 2023), and YourTTS (Casanova et al., 2022)—alongside the original FreeVC (Li et al., 2023). The extended benchmark shares exactly the same set of multi-

turn dialogue instances as the original: the same dialogues, target speakers, and scenario structure, with the only difference being which VC model generates the inconsistent audio. This design enables a direct, controlled comparison isolating the effect of VC model variability from all other factors. The resulting inconsistent clips are distributed as follows: CosyVoice3 (44%), OpenVoice (24%), FreeVC (23%), and YourTTS (9%).

Table 9 reports the complete evaluation results on the extended benchmark, following the same format as our main results (Table 2). Table 10 summarizes the comparison against the original FreeVC-only benchmark using the three headline metrics: Balanced Detection Accuracy, Balanced Localization F1-score, and Discrimination Accuracy. The performance differences are small for most models: the average absolute  $\Delta$  is 1.9% for Detection, 2.0% for Localization, and 2.8% for Discrimination, with no systematic direction of change. Model rankings are generally preserved, and the key qualitative patterns from Section 6 hold under the extended benchmark: the anti-diagonal clustering indicating threshold instability, the Precision/Recall divergence in Localization, and the detection–discrimination dissociation.

Model / Method	Detection				Localization								Discrimination							
	S1	S2	S3	Bal	S1	S2			S3			Bal	Classification					Ranking		
	Acc				F1	P	R	F1	P	R	F1	F1	Baz	AMI	B-SD	Daily	Avg	N@1	N@2	EM
<i>Large Audio-Language Models</i>																				
GPT-4o-audio	77.2	33.9	32.0	55.1	68.9	14.7	26.8	17.4	9.9	21.4	12.4	41.9	42.6	30.4	28.5	35.4	35.7	47.2	63.5	19.1
Gemini-2.5-Pro	66.2	65.1	43.9	<b>60.3</b>	67.2	57.0	65.8	59.6	46.7	58.3	50.0	<b>61.0</b>	86.7	82.6	82.9	91.5	<b>86.9</b>	<b>93.5</b>	<b>93.9</b>	<b>72.8</b>
Gemini-2.5-Flash	97.2	53.6	17.7	<b>66.4</b>	64.4	59.6	90.1	66.3	45.4	78.1	51.7	<b>61.7</b>	73.8	82.6	76.6	88.4	79.8	<u>86.3</u>	<u>90.1</u>	<u>62.9</u>
Gemini-2.5-Flash-Lite	32.7	75.4	74.3	53.8	0.7	23.6	94.7	36.2	23.2	92.5	35.4	18.2	25.7	50.0	45.0	51.1	40.6	56.5	68.0	25.0
Qwen2.5-Omni-3B	54.8	51.6	48.5	52.4	21.6	16.1	59.7	24.7	15.4	56.8	23.6	22.9	35.2	41.3	42.4	38.6	38.6	50.3	63.9	21.9
Qwen2.5-Omni-7B	32.8	70.9	70.5	51.8	38.3	14.5	64.0	23.2	13.3	58.6	21.2	30.3	44.3	37.0	34.2	40.7	40.0	50.6	64.0	22.2
Qwen3-Omni-30B-A3B	88.1	33.2	14.6	56.0	0.2	40.5	93.4	50.1	26.5	85.6	37.0	21.9	73.3	69.6	38.6	52.4	57.4	67.6	77.0	40.0
MiniCPM-o-2.6	99.8	1.7	0.2	50.4	62.5	35.4	62.6	43.1	15.5	32.6	19.7	47.0	59.5	50.0	48.1	51.3	53.2	57.0	69.6	29.4
Gemma-3n-E4B	51.2	48.4	48.0	49.7	0.0	19.2	96.0	31.9	19.2	95.2	31.9	15.9	37.1	43.5	32.9	39.2	37.1	46.5	61.0	19.7
Phi-4-multimodal	79.0	24.1	23.3	51.4	99.7	0.0	0.0	0.0	0.1	0.3	0.1	49.9	36.2	39.1	41.1	45.0	40.5	47.6	62.8	20.6
Omnivinci	81.7	46.0	21.2	57.6	48.3	14.5	63.3	23.3	9.5	43.9	15.5	33.9	41.9	50.0	41.8	50.8	45.3	56.2	67.7	27.4
Audio-Flamingo-3	99.2	1.3	1.0	50.2	0.0	17.2	38.7	23.5	17.3	38.7	23.6	11.8	32.9	28.3	36.1	42.9	36.5	44.3	61.2	16.9
<i>Speaker Embedding Methods</i>																				
Pairwise (ECAPA)	26.9	98.2	94.2	61.5	26.9	42.8	92.9	53.8	37.5	82.1	47.0	38.7	99.5	97.8	99.4	100.0	<b>99.5</b>	<b>99.6</b>	<b>93.2</b>	<b>61.2</b>
Pairwise (WavLM)	96.7	35.7	36.3	<b>66.3</b>	96.7	32.6	33.6	32.9	33.5	34.3	33.8	<b>65.0</b>	92.9	97.8	81.0	98.9	<b>92.0</b>	<u>94.2</u>	<u>91.5</u>	<u>55.7</u>
Centroid (ECAPA)	21.6	99.5	96.0	59.7	21.6	52.6	91.1	62.8	46.7	81.1	55.8	40.5	99.5	97.8	99.4	100.0	<b>99.5</b>	<b>99.6</b>	<b>93.2</b>	<b>61.2</b>
Centroid (WavLM)	91.4	53.4	56.3	<b>73.1</b>	91.4	48.5	49.9	49.0	51.0	52.3	51.4	<b>70.8</b>	92.9	97.8	81.0	98.9	<u>92.0</u>	<u>94.2</u>	<u>91.5</u>	<u>55.7</u>
Reference (ECAPA)	5.8	99.0	96.5	51.8	5.8	29.7	90.1	41.9	25.0	81.1	36.1	22.4	93.3	97.8	99.4	78.4	90.6	91.3	86.4	53.1
Reference (WavLM)	91.1	27.6	27.5	59.3	91.1	20.4	21.5	20.7	19.1	20.0	19.4	55.7	92.4	97.8	74.1	70.8	81.3	86.5	81.2	49.5

Table 9: Full evaluation results on the extended benchmark, which applies four voice conversion models to the original dialogues (FreeVC, CosyVoice3, OpenVoice, YourTTS) (%). Detection reports per-scenario and Balanced Accuracy (Bal). Localization reports F1 for S1, Precision (P), Recall (R), and F1 for S2/S3, and Balanced F1. Discrimination reports per-dataset and average Classification Accuracy, and Ranking metrics (N@1: NDCG@1, N@2: NDCG@2, EM: Exact Match).

Model	Detection			Localization			Discrimination		
	FVC	4VC	$\Delta$	FVC	4VC	$\Delta$	FVC	4VC	$\Delta$
GPT-4o-audio	52.0	55.1	<b>+3.1</b>	42.3	41.9	<b>-0.4</b>	40.7	35.7	<b>-5.0</b>
Gemini-2.5-Pro	<b>64.7</b>	60.3	<b>-4.4</b>	<b>60.1</b>	61.0	<b>+0.9</b>	<b>81.5</b>	<b>86.9</b>	<b>+5.4</b>
Gemini-2.5-Flash	63.1	<b>66.4</b>	<b>+3.3</b>	59.0	<b>61.7</b>	<b>+2.7</b>	75.6	79.8	<b>+4.2</b>
Gemini-2.5-Flash-Lite	55.3	53.8	<b>-1.5</b>	18.1	18.2	<b>+0.1</b>	46.5	40.6	<b>-5.9</b>
Qwen2.5-Omni-3B	51.5	52.4	<b>+0.9</b>	31.8	22.9	<b>-8.9</b>	40.8	38.6	<b>-2.2</b>
Qwen2.5-Omni-7B	51.5	51.8	<b>+0.3</b>	29.9	30.3	<b>+0.4</b>	42.7	40.0	<b>-2.7</b>
Qwen3-Omni-30B-A3B	55.0	56.0	<b>+1.0</b>	11.8	21.9	<b>+10.1</b>	60.4	57.4	<b>-3.0</b>
MiniCPM-o-2.6	42.8	50.4	<b>+7.6</b>	46.6	47.0	<b>+0.4</b>	49.5	53.2	<b>+3.7</b>
Gemma-3n-E4B	50.4	49.7	<b>-0.7</b>	16.0	15.9	<b>-0.1</b>	37.1	37.1	0.0
Phi-4-multimodal	51.7	51.4	<b>-0.3</b>	49.9	49.9	0.0	39.9	40.5	<b>+0.6</b>
Omnivinci	57.5	57.6	<b>+0.1</b>	33.8	33.9	<b>+0.1</b>	45.4	45.3	<b>-0.1</b>
Audio-Flamingo-3	50.2	50.2	0.0	11.8	11.8	0.0	36.3	36.5	<b>+0.2</b>

Table 10: Comparison of model performance on the original benchmark (FVC: FreeVC only) versus the extended benchmark (4VC: FreeVC + CosyVoice3 + OpenVoice + YourTTS). Det: Balanced Accuracy. Loc: Balanced F1. Disc: Avg Classification Accuracy.

## D.2 Longer Dialogues

To examine whether our findings generalize across dialogue lengths, we construct a 10-turn benchmark comprising 759 instances from 253 unique dialogues using the same pipeline as our primary 5-turn benchmark (Section 4). This extension preserves all benchmark construction components and changes only the number of target speaker turns.

Table 11 reports the full per-scenario Detection, Localization, and Discrimination results on the 10-turn benchmark, following the same format as our main results (Table 2). Section 7.3 summarizes these results using the three headline metrics (Table 4) and discusses the overall trend: the 10-turn setting is slightly more challenging

on average ( $-1.2\%$  Detection,  $-4.2\%$  Localization,  $-3.1\%$  Discrimination), while model rankings are largely preserved. The full per-scenario breakdowns here confirm that the key qualitative patterns observed in the 5-turn setting remain intact. The anti-diagonal clustering persists (e.g., Gemini-2.5-Flash reaches 100% on S1 but only 22.5%/2.8% on S2/S3, while Gemini-2.5-Flash-Lite shows the opposite bias), Precision/Recall divergence in Localization continues (Gemma-3n-E4B:  $P \approx 9\%$ ,  $R \approx 95\%$  on S2), and Gemini-2.5-Pro remains strong on Discrimination (80.1% Avg, 63.1% Exact Match) despite its modest Detection performance (69.3% Balanced Accuracy), confirming the detection–discrimination dissociation.

## E Impact of Textual Context and Reference Audio on Localization

Table 12 reports the Localization F1 results for the same ablation conditions analyzed for Detection in Section 7: adding textual context (+C vs. Audio) and removing reference audio (w/o vs. w/ Ref).

Both conditions reproduce the asymmetric behavior observed for Detection. Adding textual context sharply improves S1 F1 (e.g., Qwen2.5-Omni-7B: 38.0  $\rightarrow$  97.8) while collapsing S2 and S3 F1 (e.g., Gemini-2.5-Pro drops by  $-54.6\%$  on S2 and  $-41.2\%$  on S3), confirming that textual context biases models toward declaring all target-

Model / Method	Detection				Localization								Discrimination							
	S1	S2	S3	Bal	S1	S2			S3			Bal	Classification				Ranking			
	Acc				F1	P	R	F1	P	R	F1	F1	Baz	AMI	B-SD	Daily	Avg	N@1	N@2	EM
<i>Large Audio-Language Models</i>																				
GPT-4o-audio	85.4	18.6	15.8	51.3	70.4	7.0	13.4	8.8	4.3	10.7	5.8	38.8	60.0	41.4	35.4	52.4	39.2	45.6	60.5	12.0
Gemini-2.5-Pro	73.5	75.9	54.2	<b>69.3</b>	63.6	46.3	60.1	49.9	40.2	56.1	44.4	<b>55.4</b>	80.0	71.2	81.4	95.2	<b>80.1</b>	<b>87.7</b>	<b>90.0</b>	<b>63.1</b>
Gemini-2.5-Flash	100.0	22.5	2.8	<b>56.3</b>	74.7	59.1	83.4	63.4	33.1	62.8	36.8	<b>62.4</b>	72.7	81.7	81.4	95.2	<b>82.2</b>	85.5	88.8	61.3
Gemini-2.5-Flash-Lite	43.1	51.4	54.9	48.1	12.6	10.9	74.7	17.1	11.1	84.2	18.1	15.1	41.1	55.6	47.4	52.9	46.6	49.9	59.4	19.4
Qwen2.5-Omni-3B	76.3	28.1	26.5	51.8	29.2	6.2	52.2	10.6	5.9	52.6	10.2	19.8	54.5	43.3	39.8	33.3	40.7	46.4	60.2	19.4
Qwen2.5-Omni-7B	48.2	54.2	53.8	51.1	64.0	5.1	29.6	8.0	3.8	25.7	6.3	35.6	81.8	48.3	31.7	42.9	38.7	48.1	61.5	20.2
Qwen3-Omni-30B-A3B	97.6	5.5	2.8	50.9	1.2	30.4	84.6	36.1	16.7	77.9	23.1	15.4	72.7	68.3	48.4	57.1	54.9	59.4	68.0	30.0
MiniCPM-o-2.6	100.0	0.4	0.4	50.2	51.0	22.7	45.8	29.2	10.8	24.5	14.2	36.3	54.5	51.7	30.4	47.6	37.9	48.5	63.8	20.2
Gemma-3n-E4B	72.7	30.8	29.6	51.5	0.0	8.9	94.5	16.3	9.0	95.7	16.5	8.2	18.2	36.7	36.6	28.6	35.2	44.9	59.9	18.6
Phi-4-multimodal	95.3	5.9	4.3	50.2	100.0	0.0	0.0	0.0	0.0	0.0	0.0	50.0	54.5	28.3	29.2	47.6	31.6	44.3	59.7	18.6
Omnivinci	79.8	27.3	20.9	52.0	22.5	6.4	66.4	11.4	5.0	58.1	9.2	16.4	36.4	41.7	39.8	42.9	40.3	51.4	64.5	21.3
Audio-Flamingo-3	92.9	4.7	3.6	48.5	0.4	9.5	56.5	15.2	9.3	58.1	15.0	7.7	54.5	28.3	29.2	52.4	32.0	43.7	60.2	17.0

Table 11: Full evaluation results on the 10-turn benchmark (%). Detection reports per-scenario and Balanced Accuracy (Bal). Localization reports F1 for S1, Precision (P), Recall (R), and F1 for S2/S3, and Balanced F1. Discrimination reports per-dataset and average Classification Accuracy, and Ranking metrics (N@1: NDCG@1, N@2: NDCG@2, EM: Exact Match).

Model	Impact of Textual Context									Impact of Reference Audio								
	S1			S2			S3			S1			S2			S3		
	Audio	+C	Δ	Audio	+C	Δ	Audio	+C	Δ	w/ Ref	w/o	Δ	w/ Ref	w/o	Δ	w/ Ref	w/o	Δ
GPT-4o-audio	71.5	88.6	<b>+17.1</b>	15.0	3.6	<b>-11.4</b>	11.1	3.7	<b>-7.4</b>	71.5	79.9	<b>+8.4</b>	15.0	11.4	<b>-3.6</b>	11.1	7.2	<b>-3.9</b>
Gemini-2.5-Pro	65.2	97.3	<b>+32.1</b>	62.5	7.9	<b>-54.6</b>	47.5	6.3	<b>-41.2</b>	65.2	87.3	<b>+22.1</b>	62.5	44.0	<b>-18.5</b>	47.5	28.2	<b>-19.3</b>
Gemini-2.5-Flash	64.4	98.8	<b>+34.4</b>	62.3	30.3	<b>-32.0</b>	45.0	14.3	<b>-30.7</b>	64.4	94.1	<b>+29.7</b>	62.3	57.9	<b>-4.4</b>	45.0	32.4	<b>-12.6</b>
Gemini-2.5-Flash-Lite	0.3	15.0	<b>+14.7</b>	36.4	29.6	<b>-6.8</b>	35.4	31.0	<b>-4.4</b>	0.3	6.6	<b>+6.3</b>	36.4	42.5	<b>+6.1</b>	35.4	41.0	<b>+5.6</b>
Qwen2.5-Omni-3B	21.8	76.0	<b>+54.2</b>	47.7	3.7	<b>-44.0</b>	36.0	4.1	<b>-31.9</b>	21.8	82.8	<b>+61.0</b>	47.7	4.7	<b>-43.0</b>	36.0	4.2	<b>-31.8</b>
Qwen2.5-Omni-7B	38.0	97.8	<b>+59.8</b>	22.6	0.6	<b>-22.0</b>	21.0	0.3	<b>-20.7</b>	38.0	88.3	<b>+50.3</b>	22.6	3.2	<b>-19.4</b>	21.0	2.8	<b>-18.2</b>
Qwen3-Omni-30B-A3B	0.2	26.4	<b>+26.2</b>	23.3	36.3	<b>+13.0</b>	23.5	26.0	<b>+2.5</b>	0.2	1.3	<b>+1.1</b>	23.3	52.6	<b>+29.3</b>	23.5	30.4	<b>+6.9</b>
MiniCPM-o-2.6	66.9	59.0	<b>-7.9</b>	35.4	23.9	<b>-11.5</b>	17.2	14.9	<b>-2.3</b>	66.9	62.8	<b>-4.1</b>	35.4	38.9	<b>+3.5</b>	17.2	19.7	<b>+2.5</b>
Gemma-3n-E4B	0.0	1.3	<b>+1.3</b>	32.0	28.0	<b>-4.0</b>	32.0	27.9	<b>-4.1</b>	0.0	0.0	0.0	32.0	30.7	<b>-1.3</b>	32.0	30.2	<b>-1.8</b>
Phi-4-multimodal	99.7	100.0	<b>+0.3</b>	0.1	0.0	<b>-0.1</b>	0.1	0.0	<b>-0.1</b>	99.7	100.0	<b>+0.3</b>	0.1	0.0	<b>-0.1</b>	0.1	0.0	<b>-0.1</b>
Omnivinci	48.3	17.8	<b>-30.5</b>	23.6	10.7	<b>-12.9</b>	15.2	9.8	<b>-5.4</b>	48.3	90.0	<b>+41.7</b>	23.6	8.9	<b>-14.7</b>	15.2	2.7	<b>-12.5</b>
Audio-Flamingo-3	0.0	3.5	<b>+3.5</b>	23.3	17.6	<b>-5.7</b>	23.7	17.8	<b>-5.9</b>	0.0	0.7	<b>+0.7</b>	23.3	23.3	0.0	23.7	22.7	<b>-1.0</b>

Table 12: **Impact of Textual Context and Reference Audio on Localization F1 (%)**. Left: performance change when adding textual context (+C) vs. audio-only (Audio). Right: performance change when removing reference audio (w/o) vs. with reference (w/ Ref). Δ denotes the difference.

speaker turns as consistent, even when an inconsistent turn is present. Removing reference audio produces an analogous pattern for several models (e.g., Qwen2.5-Omni-3B S2: 47.7 → 4.7, S3: 36.0 → 4.2), though the direction and magnitude of change are more variable across models than in the textual context setting. These results mirror the Detection findings in Section 7.1 and 7.2, demonstrating that the observed modality imbalances are not specific to the Detection formulation but extend to fine-grained turn-level localization.

## F Potential Risks

Despite positive applications, we acknowledge several risks. Advanced speaker consistency evaluation could be exploited to create more convincing synthetic voices for impersonation, fraud, or misinformation campaigns. As synthetic speech becomes increasingly indistinguishable from human speech, public trust in audio evidence and voice-

based authentication may diminish. Additionally, technologies that analyze speaker characteristics could be misappropriated for unauthorized surveillance or profiling.

## G AI Assistants in Research or Writing

We used AI assistants to refine and proofread the text, and assist with coding experiments. However, all core ideas, experimental design, analysis, and scientific contributions are entirely the work of the authors.

## H Prompt Templates

### Prompt Template for Detection

You are an expert at speaker recognition. You can determine if audio turns are consistent with a target speaker based on voice characteristics alone. Focus on the main speaker's voice characteristics. Ignore backchannels, background noise, or short interjections.

First, listen to this reference audio clip from the target speaker:

[Audio: {reference\_audio}]

Now listen to the following turns from the conversation. Focus ONLY on the voice identity.

Turn 1: [Audio: {turn\_1}]

Turn 2: [Audio: {turn\_2}]

...

Turn  $N$ : [Audio: {turn\_ $N$ }]

Question: Is the voice consistent across all these turns?

Answer with ONLY one word: YES or NO.

Figure 8: Prompt template for Detection: determining speaker consistency across dialogue turns using audio only.

### Prompt Template for Detection (with Textual Context)

You are an expert at speaker recognition. You can determine if audio turns provided for a specific speaker in a conversation are consistent. Focus on the main speaker's voice characteristics. Ignore backchannels, background noise, or short interjections.

First, listen to this reference audio clip from the target speaker ({target\_speaker}):

[Audio: {reference\_audio}]

Now listen to the following conversation. Focus on the turns spoken by {target\_speaker}.

Turn 1 ({speaker\_1}): [Audio: {turn\_1}]

Turn 2 ({speaker\_2}): {text\_2}

...

Turn  $N$  ({speaker\_ $N$ ): [Audio: {turn\_ $N$ }]

Question: Is the voice of {target\_speaker} consistent across all their turns in this conversation?

Answer with ONLY one word: YES or NO.

Figure 9: Prompt template for Detection (with Textual Context): determining speaker consistency when textual transcripts of other speakers' turns are also provided.

### Prompt Template for Localization

You are an expert at speaker recognition. You can identify if audio turns are from the same speaker by analyzing voice characteristics alone. Focus on the main speaker's voice characteristics. Ignore backchannels, background noise, or short interjections.

First, listen to this reference audio clip from the target speaker:

[Audio: {reference\_audio}]

Now listen to the following turns. Identify which turns (if any) are likely spoken by a different speaker than the target speaker.

Turn 1: [Audio: {turn\_1}]

Turn 2: [Audio: {turn\_2}]

...

Turn *N*: [Audio: {turn\_N}]

Question: Which turns are inconsistent with the target speaker? List the turn numbers (e.g., "Turn 3, Turn 5"). If all turns are consistent, answer "None".

Figure 10: Prompt template for Localization: identifying which turns are inconsistent with the target speaker using audio only.

### Prompt Template for Localization (with Textual Context)

You are an expert at speaker recognition. You can determine if audio turns provided for a specific speaker in a conversation are consistent. Focus on the main speaker's voice characteristics. Ignore backchannels, background noise, or short interjections.

First, listen to this reference audio clip from the target speaker ({target\_speaker}):

[Audio: {reference\_audio}]

Now listen to the following conversation involving {target\_speaker}. Identify which turns (if any) attributed to {target\_speaker} are likely spoken by a different speaker than the target speaker.

Turn 1 ({speaker\_1}): [Audio: {turn\_1}]

Turn 2 ({speaker\_2}): {text\_2}

...

Turn *N* ({speaker\_N}): [Audio: {turn\_N}]

Question: Which turns are inconsistent with the target speaker? List the turn numbers (e.g., "Turn 3, Turn 5"). If all turns are consistent, answer "None".

Figure 11: Prompt template for Localization (with Textual Context): identifying which turns are inconsistent when textual transcripts of other speakers' turns are also provided.

### Prompt Template for Discrimination (Classification)

You are an expert at speaker recognition. You can identify speaker consistency by analyzing voice characteristics alone. Focus on the main speaker's voice characteristics. Ignore backchannels, background noise, or short interjections.

You will hear a sequence of turns. One position in the sequence is MISSING. You will be given three options for what should go in that position. Your task is to select the option that makes the entire sequence most consistent with the target speaker.

Listen to the turn sequence:

Turn 1: [Audio: {turn\_1}]

...

Turn  $k$ : [MISSING – Choose the correct option below]

...

Turn  $N$ : [Audio: {turn\_N}]

Here are the options for the missing position (Turn  $k$ ):

Option A: [Audio: {option\_A}]

Option B: [Audio: {option\_B}]

Option C: [Audio: {option\_C}]

Question: Which option (A, B, or C) makes the entire sequence most consistent with the target speaker?

Answer with ONLY one letter: A, B, or C.

Figure 12: Prompt template for Discrimination (Classification): selecting the option that best fits the missing position to maximize speaker consistency.

### Prompt Template for Discrimination (Ranking)

You are an expert at speaker recognition. You can identify speaker consistency by analyzing voice characteristics alone. Focus on the main speaker's voice characteristics. Ignore backchannels, background noise, or short interjections.

You will hear a sequence of turns. One position in the sequence is MISSING. You will be given three options for what should go in that position. Your task is to rank all options by how well they match the target speaker's voice.

Listen to the turn sequence:

Turn 1: [Audio: {turn\_1}]

...

Turn  $k$ : [MISSING – Rank the options below]

...

Turn  $N$ : [Audio: {turn\_N}]

Here are the options for the missing position (Turn  $k$ ):

Option A: [Audio: {option\_A}]

Option B: [Audio: {option\_B}]

Option C: [Audio: {option\_C}]

Question: Rank all options (A, B, C) from most consistent to least consistent with the target speaker.

Answer with ONLY the letters in order, separated by ">". For example: A > B > C.

Figure 13: Prompt template for Discrimination (Ranking): ranking candidates by acoustic consistency with the target speaker given the surrounding dialogue context.

### Prompt Template for Per-Turn Reference Comparison (Detection and Localization)

You are an expert at speaker recognition. You can determine if two audio clips are from the same speaker by analyzing voice characteristics alone. Focus on the main speaker's voice characteristics. Ignore backchannels, background noise, or short interjections.

First, listen to this reference audio clip from the target speaker:

[Audio: {reference\_audio}]

Now listen to this audio clip:

[Audio: {test\_audio}]

Question: Is this clip from the same speaker as the reference?

Answer with ONLY one word: YES or NO.

Figure 14: Prompt template for Per-Turn Reference Comparison (Detection and Localization): determining whether each target speaker turn is from the same speaker as the reference.

### Prompt Template for Per-Turn Reference Comparison (Discrimination)

You are an expert at speaker recognition. You can determine the speaker's identity in audio recordings by analyzing their voice characteristics. Focus on the main speaker's voice characteristics. Ignore backchannels, background noise, or short interjections.

First, listen to this reference audio clip from the target speaker:

[Audio: {reference\_audio}]

You will hear three versions of the same utterance. Your task is to select the option that is most consistent with the target speaker.

Option A: [Audio: {option\_A}]

Option B: [Audio: {option\_B}]

Option C: [Audio: {option\_C}]

Question: Which option (A, B, or C) is most consistent with the target speaker?

Answer with ONLY one letter: A, B, or C.

Figure 15: Prompt template for Per-Turn Reference Comparison (Discrimination): selecting the original target speaker audio among three candidates compared against the reference.

### Prompt Template for VC Quality Ranking

You are an expert at speaker recognition. You can identify speakers by analyzing voice characteristics such as pitch, tone, speaking style, and timbre. Focus on the main speaker's voice characteristics. Ignore backchannels, background noise, or short interjections.

First, listen to this reference audio clip from the target speaker:

[Audio: {reference\_audio}]

Now, you will hear several audio clips. Your task is to rank all clips by how similar they sound to the reference speaker's voice.

Option A: [Audio: {option\_A}]

Option B: [Audio: {option\_B}]

Option C: [Audio: {option\_C}]

Option D: [Audio: {option\_D}]

Question: Rank all options (A, B, C, D) from most similar to least similar to the reference speaker's voice.

Answer with ONLY the letters in order, separated by ">". For example: A > B > C > D.

Figure 16: Prompt template for VC Quality Ranking: ranking voice-cloned candidates by acoustic similarity to the reference speaker.

### Prompt Template for Dialogue Coherence Filtering

You are an expert dialogue analyzer. Read the following conversation and determine if it flows naturally and coherently as a whole.

Conversation:

{speaker\_1}: {text\_1}

{speaker\_2}: {text\_2}

...

{speaker\_N}: {text\_N}

Task:

If the conversation flows naturally with logical connections between turns, output “coherent”.

If there are sudden disruptions, random insertions, contradictions, or parts that don’t make sense in context (at any point in the dialogue), output “incoherent”.

Return ONLY one word: “coherent” or “incoherent”.

Figure 17: Prompt template for Dialogue Coherence Filtering: evaluating conversational naturalness and flow for benchmark construction.