CoLMbo-SV: A Grounded Speaker Language Model for Explainable Voice Comparison

Anonymous ACL submission

Abstract

Speaker verification systems achieve strong accuracy but offer no explanation for their decisions, limiting their use in forensic and other high-stakes settings. Audio language models can describe speech but are not trained for pairwise verification and do not guaranty that their numerical claims reflect the underlying signal. We present **CoLMbo-SV**, an explainable comparative speaker verification system that ties each verdict to inspectable acoustic evidence. To train and evaluate the system, we construct **VoxReason**, a dataset of speaker pairs annotated with structured acoustic-phonetic evidence and forensic-style chain-of-thought reports, filtered by a deterministic numerical validator and an LLM-as-judge that discard reports whose claims drift from the underlying measurements. Verdicts are accompanied by a Fisher-combined confidence label computed over per-feature z-scores against gender-conditional within-speaker variability, giving each decision a likelihood-ratio-style audit trail. In the TIMIT pairs test set, CoLMbo-SV reaches 77.1% verdict accuracy versus 24.3% for a Qwen3-Omni-30B baseline, with explanation faithfulness 0.70 vs. 0.23 and a 0% unknown-verdict rate vs. 50%. The system's emitted recording-condition measurements empirically predict its own reliability, establishing a direction for auditable speaker verification with explicit statistical confidence.

1 Introduction

In Daubert hearings on the admissibility of voice comparison evidence in U.S. federal courts (Morrison and Thompson, 2017), forensic examiners are routinely asked a question that automatic speaker verification systems cannot answer: which acoustic properties drove the match, what sources of uncertainty were present, and on what statistical basis the conclusion rests. Modern systems return a similarity score, but neither the court nor the defense can obtain a principled account of what the score means or how it was reached. This is not an incidental gap, but a structural one: speaker verification systems are optimized to produce decisions, not to justify them (Morrison, 2009). In forensic, security, and legal settings, the basis of a decision carries the same evidentiary weight as the decision itself, which renders a similarity score without explanation insufficient as evidence.

Forensic speaker comparison also imposes a stronger evidentiary requirement than ordinary speaker verification: the task is not only to decide whether two recordings are similar, but to evaluate the strength of the observed speech evidence under competing hypotheses, i.e., whether the questioned recording was produced by the known speaker or by another speaker drawn from a relevant population subgroup (Morrison and Thompson, 2017; Drygajlo et al., 2015). This requires reasoning over substantial inter- and intra-class variability across lexical content, speaking style, dialect, health, emotion, vocal effort, and time, together with extrinsic variability introduced by microphones, codecs, channels, background noise, reverberation, and duration. A forensic system must therefore assess not only whether two recordings are similar, but also whether the observed similarities and differences are expected under same-speaker variability or are more typical of different-speaker comparisons.

Modern automatic speaker verification systems are poorly aligned with this requirement. State-of-the-art models based on x-vectors (Snyder et al., 2018), ECAPA-TDNN (Desplanques et al., 2020) and self-supervised backbones like WavLM (Chen et al., 2022) compress the identity of the speaker into a fixed-dimensional embedding and reduce the comparison to a cosine similarity. The representation is powerful, but opaque. It does not encode a separable notion of the acoustic categories on which human examiners and forensic phoneticians rely, such as fundamental frequency, voice quality, formant structure, or articulation timing. Recent au-

dio language models (Chu et al., 2023; Tang et al., 2024; Ghosh et al., 2024) offer a complementary but incomplete capability. They can produce fluent natural-language descriptions of speech, but are not trained for calibrated pairwise speaker comparison, do not structure their reasoning around forensically relevant acoustic evidence, and offer no guaranty that their stated observations reflect what the underlying signal actually contains. For example, a model may correctly identify a voice as "high-pitched" while reporting an F0 value that differs by tens of Hertz from the measured signal. The field is split between two paradigms: systems that **decide without explaining** and systems that **describe without deciding**. Neither meets the standard demanded by high-stakes deployment, where a verification outcome must be both accurate and traceable to inspectable acoustic evidence.

We argue that explainability in speaker verification cannot be achieved by post-hoc narration of a black-box decision; it must be built into the decision process itself. We ground a language model's reasoning in a structured pipeline of forensically relevant acoustic features and require it to cite specific signal-level observations before producing a decision. Critically, each feature comparison is contextualized by typical within-speaker variability: we compute per-feature z-scores against gender-conditional intra-speaker standard deviations, combine evidence across features via Fisher's method, and emit a final verdict with an explicit high/medium confidence label on an ENFSI-style verbal scale (Drygajlo et al., 2015). The verification decision thus emerges from a structured analysis that cites inspectable measurements, rather than from an embedding similarity that cannot be interrogated.

We introduce CoLMbo-SV, an explainable comparative speaker verification system that integrates a forensically grounded evidence pipeline with a language-model backbone. It contains a pretrained W2V-BERT 2.0 speaker encoder, a learnable audio mapper that projects pairs of utterances into the LLM embedding space, and a frozen Gemma backbone with LoRA adapters, trained jointly with a verification head so that the same/different decision and the textual analysis are produced from a shared representation. To train and evaluate this system, we also construct VoxReason, a dataset of paired recordings annotated with structured acoustic-phonetic evidence and forensic-style chain-of-thought comparison reports, filtered by an LLM-as-judge stage that discards any report whose numerical claims drift from the underlying measurements.

2 VoxReason

The task of explainable comparative speaker verification consists of taking two recordings, extracting acoustic-to-linguistic evidence from each, and producing a natural-language report that justifies a same-speaker / different-speaker decision. To support this task, we construct **VoxReason**, a dataset of speaker pairs annotated with structured grounding evidence and forensic-style chain-of-thought comparison reports. To our knowledge, no existing speaker verification dataset provides reasoning supervision of this kind: standard SV benchmarks (Nagrani et al., 2017) provide labels and recordings, and audio-language datasets (Kim et al., 2019; Drossos et al., 2020) provide free-form descriptions, but neither pairs the two with the structured phonetic evidence that grounded reasoning requires. The full data construction pipeline is shown in Figure 1.

2.1 Speech recordings

We source speech recordings from TIMIT (Garofolo et al., 1993) and LibriSpeech (Panayotov et al., 2015). TIMIT contributes phonetically rich, hand-transcribed read speech across eight U.S. dialect regions, which is uniquely suited to dialect-sensitive forensic analysis. LibriSpeech contributes a substantially larger pool of speakers and longer utterances under cleaner recording conditions. Together they cover the two regimes that matter for our task: dense phonetic-dialectal annotation and broad speaker-population coverage. For every utterance, we extract a structured feature set spanning the acoustic-to-linguistic hierarchy recording conditions (SNR, bandwidth, pause profile), prosody and rhythm (F0 statistics, energy dynamics, syllable rate, rhythm class), phonation quality (H1–H2, spectral slope, HNR, jitter, shimmer), segmental acoustics (VOT, per-vowel formants, rhotic F3, lateral F2, flap duration), and contextual-phonotactic features (rhoticity, flapping, dialect classification). The resulting per-utterance JSON record is the grounding evidence that anchors all downstream reasoning, ensuring that every claim a model makes about an utterance can be traced to a measured acoustic correlate.

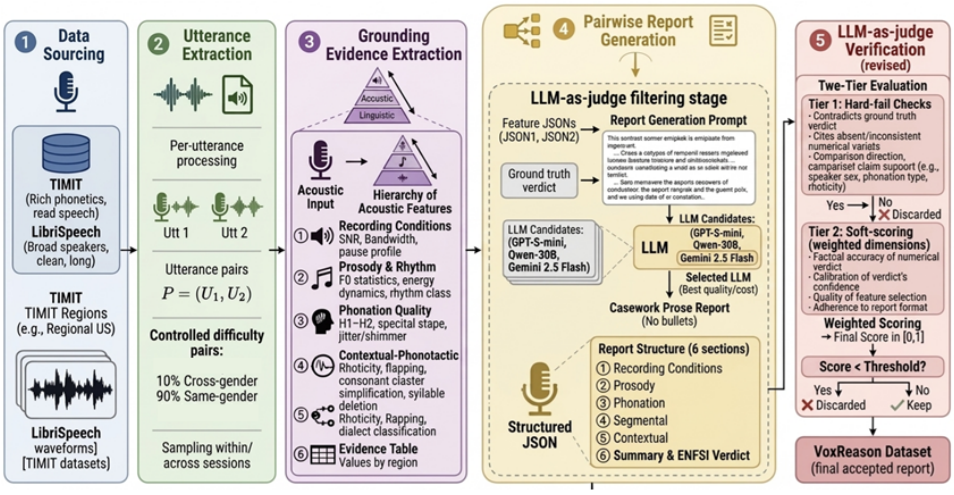

Figure 1: VoxReason Data Construction Pipeline

2.2 Comparison reports

For each utterance pair, we generate a forensic-style speaker comparison report grounded in the two feature JSONs. First, we instruct the LLM to produce continuous casework prose rather than bulleted summaries, with numerical evidence embedded in analytical sentences and accompanied by phonetic interpretation what each value means articulatorily, what the difference would sound like, and what could plausibly cause it. Second, we structure the report into the six sections that mirror forensic phonetic practice (recording conditions, prosody and rhythm, phonation quality, segmental acoustics, contextual-phonotactic features, and a weight-of-evidence summary), terminated by a compact evidence table and an ENFSI (Drygajlo et al., 2015) verbal-scale verdict. To select a generator, we benchmarked three candidates GPT-5-mini, Qwen-30B, and Gemini 2.5 Flash on a development set of 100 pairs and found Gemini 2.5 Flash to provide the best quality-to-cost tradeoff at the scale required to annotate the full corpus.

2.3 Statistical confidence grounding

Free-form LLM reports tend to attach verbal confidence (*"strongly suggests"*, *"high confidence"*) to verdicts inconsistently, with no principled tie to the underlying acoustic distances between the two utterances. To replace these subjective hedges with a measurable signal, we attach an explicit statistical confidence label to every report using the following procedure.

For each discriminative scalar feature f (mean F0, jitter, shimmer, HNR, formants F1/F2/F3, voicing ratio, speaking rate, voiced and voiceless VOT, and the voice-quality descriptors), we precompute a gender-conditional within-speaker standard deviation $\sigma_w(f, g)$ from the training-side speakers — i.e., how much the feature naturally varies across utterances of the same speaker of gender g. Given a pair (a, b) with measured values $v_a^{(f)}, v_b^{(f)}$ and inferred gender g, we score each feature with a z-statistic against this within-speaker baseline:

$$z_f = \frac{|v_a^{(f)} - v_b^{(f)}|}{\sigma_w(f, g)}, \qquad p_f = 2\big(1 - \Phi(z_f)\big) \quad (1)$$

where Φ is the standard normal CDF. A small p_f means the observed difference is too large to be explained by within-speaker variation alone, evidence in favor of *different speaker*; a large p_f means the observed difference is comfortably within typical intra-speaker variation, evidence in favor of *same speaker*.

Per-feature p-values are then combined across all k available features via Fisher's method (Fisher, 1970):

$$\chi^2 = -2 \sum_{f=1}^{k} \ln p_f, \qquad \chi^2 \sim \chi^2_{2k} \quad (2)$$

yielding a single combined p-value p_{Fisher}. We discretize this into a four-level confidence label using fixed thresholds: $p_{\text{Fisher}} > 0.40$ (same, high), $0.25 < p_{\text{Fisher}} \leq 0.40$ (same, medium), $0.05 < p_{\text{Fisher}} \leq 0.25$ (different, medium), and $p_{\text{Fisher}} \leq 0.05$ (different, high). Recording-condition features (SNR, noise floor, speech ratio) are not speaker-discriminative and are excluded from the combination.

The resulting confidence label is spliced into each report as a closing sentence reporting the top three discriminating features and their z-statistics, followed by the rewritten **Conclusion** line. Reports therefore deliver three layers of evidence: a qualitative analysis grounded in measured features, a per-feature statistical strength, and a calibrated verdict aligned with the likelihood-ratio framework standard in forensic phonetics (Morrison, 2009; Rose, 2002).

2.4 Deterministic numerical validation

LLM judges, including the one described in Section 2.5, evaluate reports holistically and can be inconsistent on quantitative claims — a known weakness of LLM-based evaluation (Liu et al., 2023; Zheng et al., 2023). We complement the LLM judge with a deterministic validator that performs lossless numerical verification of every candidate report before it reaches the judge.

The validator (i) tokenizes the report and extracts every numerical literal via regular expression; (ii) strips spurious matches that are not feature claims — markdown section numbers (`## 1.`), formant labels (`F1`, `F2`, `F3`), utterance and speaker indices (`utterance 2`, `speaker 1`), and the computed-statistic artifacts introduced by Section 2.3 (`z = 1.40`, `p = 0.235`, χ^2 `= 7.98`, `df = 10`); and (iii) checks each remaining number against the union of values in the two feature JSONs and the comparison block, accepting a match when the cited number is within an absolute tolerance of 0.5 or a relative tolerance of 5% of any ground-truth value. A report is rejected outright if more than 20% of its numerical claims fail to match.

The validator runs as a cheap pre-filter before the LLM-as-judge stage, catching the bulk of numerical errors without an LLM call. The remaining samples are passed to the LLM judge for the qualitative checks that can't be reduced to regex matching.

2.5 LLM-as-judge

Even with an explicit instruction to ground every claim in the provided feature JSON, LLM generators systematically produce *post-hoc rationalizations* when the ground-truth verdict is supplied in the prompt: rather than letting the evidence drive the verdict, they retrofit a plausible-sounding justification to the given label, often citing numerical values that do not appear in the input JSON. We observe this empirically on a development sample: when given the correct verdict and the two feature JSONs, the generator cites at least one numerical value absent from or inconsistent with the JSON, even when explicitly prompted to use only the provided values. This failure mode is well-attested in the LLM evaluation literature (Turpin et al., 2023; Lyu et al., 2023) but is especially damaging here, where the numerical claims *are* the evidence. The combination of the deterministic validator (Section 2.4) and the LLM judge described below is designed specifically to catch and discard these rationalized reports. Generated reports are not used as supervision until they pass an LLM-as-judge filtering stage designed to remove samples that drift from the underlying numerical evidence. The judge is given the two feature JSONs, the ground-truth verdict, and the candidate report, and applies a two-tier evaluation: a set of hard-fail checks that discard any report whose verdict contradicts ground truth, cites numerical values absent from or inconsistent with the feature JSONs, inverts the direction of a key A-versus-B comparison, or asserts categorical claims unsupported by the features (dialect, speaker sex, phonation type, rhoticity); and a weighted soft-scoring stage along five dimensions, factual accuracy of numerical claims, calibration of the verdict's confidence to the evidence, quality of feature selection, robustness reasoning under adverse recording conditions, and adherence to the prescribed report format, yielding a final score in [0,1]. Reports that trigger any hard-fail or score below threshold are discarded. This filter is the single most important guardrail in VoxReason: without it, the dataset would inherit the LLM's tendency to invent precise-sounding numbers and to ignore recording-condition confounds, which are exactly the failure modes an explainable speaker verification system must avoid.

3 CoLMbo-SV

CoLMbo-SV combines a large-scale pretrained speaker encoder, a learnable audio mapper, and a frozen language-model backbone with LoRA adapters. The system takes two recordings and a user prompt, projects each utterance into the LLM embedding space through the bridge, and produces both a same/different verification logit and a textual response. The full architecture is shown in Figure 2.

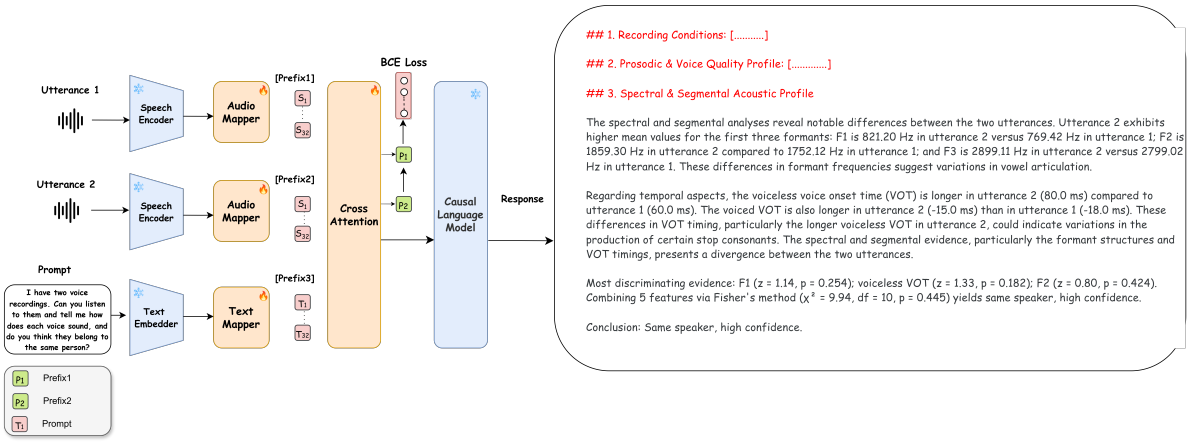

Figure 2: **CoLMbo-SV pipeline.** Two utterances are independently encoded and projected into the language-model embedding space via a Q-Former with learnable queries. The utterance tokens are concatenated with the user prompt and passed to a frozen causal language model, which generates a structured comparison response.

3.1 Speaker Encoder

We use a W2V-BERT 2.0 backbone pretrained for speaker verification with curriculum ranking loss on a large-scale speaker corpus (Baali et al., 2026), which exposes both frame-level hidden states $\mathbf{H} \in \mathbb{R}^{T \times D_e}$ and a pooled speaker embedding $\mathbf{e} \in \mathbb{R}^{D_p}$ per utterance. The encoder is frozen during stage-1 training (the language-model adaptation stage) and optionally unfrozen for stage-2 fine-tuning with a reduced learning rate. We use the pooled embeddings $\mathbf{e}_a, \mathbf{e}_b$ for the verification head and the bridge prefix.

3.2 Audio Mapper

The audio mapper maps the two speaker embeddings into the LLM embedding space, producing a sequence of soft tokens that prepend the language model's input. We adopt a transformer-mapper design (Mokady et al., 2021): each utterance embedding $\mathbf{e}_u \in \mathbb{R}^{D_p}$ is first projected and split into $K = 10$ prefix tokens via a two-layer transformer projector, producing $\mathbf{S}_u \in \mathbb{R}^{K \times D_l}$ where D_l is the LLM hidden dimension. The user prompt is tokenised and embedded by the LLM's input embedding layer, then passed through a third transformer projector that produces a prompt-conditioned prefix $\mathbf{S}_p \in \mathbb{R}^{K \times D_l}$. The three prefixes are concatenated with a learned separator token $\mathbf{s}_{\text{sep}}$:

$$\mathbf{X} = [\mathbf{S}_a; \mathbf{s}_{\text{sep}}; \mathbf{S}_b; \mathbf{s}_{\text{sep}}; \mathbf{S}_p] \quad \in \mathbb{R}^{(3K+2) \times D_l}.$$

A 4-layer cross-attention block lets the prompt-prefix tokens attend to the two audio prefixes before they enter the LLM, producing the bridge output $\tilde{\mathbf{X}}$.

Verification head. In parallel, a small classifier head reads the two pooled embeddings $\mathbf{e}_a, \mathbf{e}_b$ and outputs a single same/different logit $\ell \in \mathbb{R}$. The head is a two-layer MLP applied to a fused representation $[\mathbf{e}_a; \mathbf{e}_b; |\mathbf{e}_a - \mathbf{e}_b|; \mathbf{e}_a \odot \mathbf{e}_b]$, the standard concatenation used in speaker verification.

3.3 Language Model

The mapper's output $\tilde{\mathbf{X}}$ is treated as a soft-token prefix for a frozen Gemma 4 (Team et al., 2025) backbone, with LoRA adapters (Hu et al., 2022) applied to all attention and MLP projection layers. The LLM is queried in 4-bit NF4 quantisation (Dettmers et al., 2023) with bfloat16 compute for memory efficiency. Generation proceeds autoregressively from the prefix, producing the textual analysis and verdict.

3.4 Training Objective

CoLMbo-SV is trained with a joint objective: a binary cross-entropy loss on the verification head, and a token-level language-modelling loss on the textual response.

Language-modelling loss. Given the mapper output $\tilde{\mathbf{X}}$ and the target response token sequence $y_1, \ldots, y_T$ (which includes the forensic analysis, statistical confidence, and final verdict line), we minimise the standard next-token cross-entropy over the response tokens only:

$$\mathcal{L}_{\text{LM}} = -\frac{1}{T} \sum_{t=1}^{T} \log P_\theta\big(y_t \mid \tilde{\mathbf{X}}, y_{<t}\big).$$

The audio and prompt prefix positions are masked from the loss; the gradient only flows from re-

sponse tokens.

Verification loss. The verification head's logit ℓ is supervised against the ground-truth label $y_{\text{sv}} \in \{0, 1\}$ (1 = same speaker, 0 = different) with binary cross-entropy:

$$\mathcal{L}_{\text{BCE}} = -\big[y_{\text{sv}} \log \sigma(\ell) + (1 - y_{\text{sv}}) \log(1 - \sigma(\ell))\big].$$

Total loss. The two terms are combined with fixed weights:

$$\mathcal{L} = \lambda_{\text{LM}}\, \mathcal{L}_{\text{LM}} + \lambda_{\text{BCE}}\, \mathcal{L}_{\text{BCE}}.$$

We set $\lambda_{\text{LM}} = 1.0$ and $\lambda_{\text{BCE}} = 0.5$ in all experiments.

The verification head and the language model thus train jointly on every batch: the head provides a direct speaker-discrimination signal to the encoder and bridge, while the LM loss shapes the bridge to produce a representation that supports faithful textual analysis.

Training-time feature injection. A subset of training prompts includes the per-utterance feature JSON inline ("Acoustic features extracted from the recordings: F0 = 120 Hz, jitter = 0.52, ..."). This serves as auxiliary grounding that helps the model learn to not invent specific numerical values into analytical prose and to follow the prescribed report structure. At inference time, no features are provided in the prompt: the model must produce the same form of report conditioned on the two audio prefixes and the user's question alone.

3.5 Inference

At test time the two utterances are encoded, the mapper produces the soft-token prefix $\tilde{\mathbf{X}}$ from the audio embeddings and the user prompt, and the language model generates the textual response autoregressively. Decoding is greedy with a repetition penalty of 1.15 and a no-repeat n-gram size of 3, up to a maximum of 768 new tokens. The verification head is computed in parallel from the pooled audio embeddings and yields the continuous score $\sigma(\ell) \in [0, 1]$ used for ROC/EER evaluation; the textual response carries the final verdict and confidence level used for verdict accuracy. Crucially, no acoustic features are passed to the model at inference; all reported numerical claims in the response are produced from the audio alone.

Table 1: Comparison of CoLMbo-SV against Qwen3-Omni on the TIMIT pairs test split (N=1917). Best in bold.

	Qwen3-Omni	CoLMbo-SV
Strict verdict (%) ↑	24.31	**77.10**
Answered verdict (%) ↑	48.54	**77.10**
Unknown rate (%) ↓	49.92	**0.00**
Faithfulness ↑	0.23	**0.70**
Perfectly grounded (%) ↑	9.12	**47.47**
$\geq 50\%$ grounded (%) ↑	23.20	**84.42**
Relevance (%) ↑	71.05	**95.36**
Repetition (%) ↓	92.07	**1.62**

Table 2: Mapper prefix shift from a neutral baseline under feature-specific prompts, averaged over 100 pairs. The 42% spread across probes (max/mean − min/mean for Δ_F) confirms feature-specific routing.

Probe	Δ_F	Cos. dist.
Breathiness	0.146	0.010
F0	0.180	0.015
Jitter	0.189	0.017
HNR	0.197	0.018
Formants	0.228	0.025
VOT	0.228	0.024

4 Results

We evaluate CoLMbo-SV on the TIMIT pairs test split against Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025). Our analysis covers verdict accuracy, explanation faithfulness, and interpretability properties relevant to forensic deployment.

4.1 Comparison with audio language models

Table 1 compares CoLMbo-SV and Qwen3-Omni on a matched 1917-pair subset (identical pair_id and section across the two systems). CoLMbo-SV correctly classifies 77.10% of pairs versus Qwen3-Omni's 24.31% (strict). Notably, Qwen3-Omni fails to produce a parseable verdict in 49.92% of responses — its outputs frequently end without committing to a same/different judgment — whereas CoLMbo-SV achieves 0.00% unknown rate. Among the numerical claims each model makes, 70% of CoLMbo-SV's are recoverable from the underlying signal evidence (within tolerance), compared to 23% for Qwen3-Omni. The 47.47% perfectly-grounded rate (every cited number verifiable) versus Qwen3-Omni's 9.12% reflects the central design contribution: CoLMbo-SV's responses are built around inspectable acoustic measurements rather than free-form description.

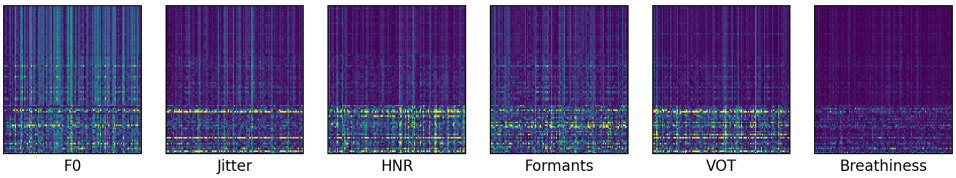

Figure 3: CoLMbo-SV activation maps showing distinct patterns for acoustic-feature queries.

Calibrated evidence reporting. Beyond producing a same/different verdict, CoLMbo-SV reports the statistical strength of its decision: a Fisher-combined p-value computed from per-feature z-scores against gender-conditional within-speaker variability, expressed in a high/medium confidence label. This mirrors the likelihood-ratio framework used in forensic phonetics (Morrison, 2009; Rose, 2002), where a verification result is grounded in calibrated evidence rather than an uninterpretable score. The mechanism is illustrated by the segmental example in Table 4 in the (Appendix): five weak per-feature observations (F_1, F_2, F_3, voiced VOT, voiceless VOT, individually $z < 2$) combine via Fisher's method into a single high-confidence same-speaker conclusion, with χ^2=7.98 (df=10, p=0.635).

4.2 Interpretability

Visualizing Mapper activations A central design goal of CoLMbo-SV is that the audio–language mapper condition its output on the user's prompt, so that queries about specific acoustic features draw on the corresponding regions of the speaker representation rather than producing a single prompt-invariant prefix. To verify this empirically, we run the bridge on 100 test pairs under six feature-specific prompts (F0, jitter, HNR, formants, VOT, breathiness) and a neutral pairwise prompt as baseline. For each feature-specific prompt we compute the relative Frobenius shift of the bridge's prefix representation from the neutral baseline. Table 2 reports the per-probe shifts: queries about formants and VOT produce the largest shifts (22.8% Frobenius), queries about breathiness the smallest (14.6%), with a 42% spread across probes. Figure 3 visualises this directly, plotting $|h_{\text{probe}} - h_{\text{neutral}}|$ over the top-120 most variable feature dimensions: distinct activation regions are recruited by different feature queries, and the breathiness panel is visibly dimmer than the others, matching its smaller Frobenius shift. The model does not produce a prompt-invariant audio summary; it routes queries to different parts of its learned representation, exactly as the design intends.

Recording-condition reliability. A forensic deployment must expose, rather than mask, uncertainty arising from poor recording quality. Stratifying CoLMbo-SV's predictions by the minimum SNR within each test pair (Table 3) reveals a monotone relationship between recording quality and verdict accuracy: pairs above 35 dB SNR achieve 79.0% accuracy, whereas the small set of pairs below 15 dB SNR fall to 25%. Mean numerical faithfulness shows the same direction — recoverable numerical claims rise from 0.44 on the noisiest pairs to 0.72 on the cleanest — though the effect concentrates between the extreme low-SNR bucket and the rest of the test set rather than progressing smoothly. The same monotone trend holds for the other two recording-condition features extracted by the pipeline (noise floor: 69.4% → 78.0% accuracy; speech ratio: 64.5% → 79.6%; full breakdown in the appendix). Because the model emits these measurements explicitly as the first section of every response, the user receives both a decision and an audit trail of the conditions under which that decision was made — an interpretability property that single-score embedding systems cannot offer.

5 Conclusion

We introduced **CoLMbo-SV**, an explainable comparative speaker verification system that ties each verdict to inspectable acoustic evidence, and **VoxReason**, a dataset of speaker pairs annotated with forensically structured reasoning supervision. CoLMbo-SV combines a pretrained speaker encoder, a transformer-mapper bridge, and a language-model backbone trained jointly under a verification head and a token-level language-modelling objective. Verdicts are accompanied by a Fisher-combined statistical confidence label grounded in gender-conditional within-speaker variability, providing a likelihood-ratio-style audit trail that single-score embedding systems cannot offer.

Table 3: Verdict accuracy and faithfulness on TIMIT pairs stratified by minimum SNR (N=3000). Best in bold.

SNR bucket	n	Mean SNR (dB)	Verdict (%)	Faithfulness	Perfect (%)
≤ 15 dB†	4	13.06	25.00	0.44	0.00
15–25 dB	142	21.75	71.83	0.71	48.23
25–35 dB	1672	30.96	73.15	0.73	48.85
> 35 dB	1182	37.83	**79.02**	0.72	44.59

† Sample size is small; the value should be read as suggestive.

On the TIMIT pairs benchmark, CoLMbo-SV substantially outperforms a zero-shot audio-language baseline on verdict accuracy (77.1% vs. 24.3%), explanation faithfulness (0.70 vs. 0.23), and unknown verdict rate (0% vs. 50%), while emitting recording-condition measurements that empirically predict its own reliability. The system trades absolute verification accuracy against dedicated SV baselines for explanation faithfulness and calibrated confidence — a deliberate design choice for high-stakes settings where auditability matters as much as raw discrimination.

Limitations

CoLMbo-SV is trained on read speech (TIMIT, LibriSpeech) under controlled recording conditions; performance on conversational speech and field-recorded forensic audio remains to be evaluated. Per-feature routing analysis indicates that the bridge concentrates its capacity on coarse-grained discriminators (F0, voicing, VOT) more strongly than on fine-grained spectral detail, leaving room for improvement on same-gender, same-dialect pairs where dedicated SV systems still hold an advantage. Future work includes scaling VoxReason to conversational corpora, integrating richer phonetic features into the bridge, and exploring confidence-calibrated likelihood-ratio outputs aligned with forensic admissibility standards.

Ethics

Speaker verification has both legitimate and potentially harmful applications. Legitimate use includes secure authentication, forensic phonetic analysis under court-admissible standards, and accessibility applications. Potential harms include surveillance of vulnerable populations and over-reliance on machine output in legal proceedings. The structured analysis trace this work introduces is intended to support the legitimate forensic use case by making the model's reasoning auditable. We caution against deploying the system in legal contexts without the trace explicitly inspected by qualified phoneticians, and against use cases for which speaker verification was not designed.

References

Imen Ben Amor and Jean-François Bonastre. 2022. Balr: Binary-attribute-based likelihood ratio estimation for forensic voice comparison. In *2022 International workshop on biometrics and forensics (IWBF)*, pages 1–6. IEEE.

Massa Baali, Sarthak Bisht, Rita Singh, and Bhiksha Raj. 2026. What and when to learn: Curriculum ranking loss for large-scale speaker verification. *arXiv preprint arXiv:2603.24432*.

Massa Baali, Shuo Han, Syed Abdul Hannan, Purusottam Samal, Karanveer Singh, Soham Deshmukh, Rita Singh, and Bhiksha Raj. 2025. Colmbo: Speaker language model for descriptive profiling. *arXiv preprint arXiv:2506.09375*.

Imen Ben-Amor, Jean-François Bonastre, and Salima Mdhaffar. 2024. Extraction of interpretable and shared speaker-specific speech attributes through binary auto-encoder. In *Proc. Interspeech*, volume 2024, pages 3230–3234.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 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.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Soham Deshmukh, Shuo Han, Rita Singh, and Bhiksha Raj. 2025. Adiff: Explaining audio difference using natural language. *arXiv preprint arXiv:2502.04476*.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *interspeech*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Andrzej Drygajlo, Michael Jessen, Stefan Gfroerer, Isolde Wagner, Jos Vermeulen, Tuija Niemi, and 1 others. 2015. *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*. Verlag für Polizeiwissenschaft Frankfurt.

Ronald Aylmer Fisher. 1970. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, and Nancy L Dahlgren. 1993. Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93:27403.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Vincent Hughes, Chenzi Xu, Paul Foulkes, Philip Harrison, Poppy Welch, Jessica Hazel Wormald, Finnian Kelly, and David van der Vloed. 2024. Exploring individual speaker behaviour within a forensic automatic speaker recognition system. In *Proceedings of Odyssey: The Speaker and Language Workshop*. York.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329.

Yi Ma, Shuai Wang, Tianchi Liu, and Haizhou Li. 2025. Expo: Explainable phonetic trait-oriented network for speaker verification. *IEEE Signal Processing Letters*, 32:731–735.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Geoffrey Stewart Morrison. 2009. Forensic voice comparison and the paradigm shift. *Science & Justice*, 49(4):298–308.

Geoffrey Stewart Morrison and William C. Thompson. 2017. Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review*, 18:326–434.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. *Interspeech 2017*, page 2616.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Philip Rose. 2002. *Forensic Speaker Identification*. Taylor and Francis.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust DNN embeddings for speaker recognition. In *ICASSP*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pages 16607–16629.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Appendix

A.1 Related Work

Forensic voice comparison concerns the comparison of known-speaker and questioned-speaker recordings to assist a court or other decision-maker in evaluating speaker identity. Morrison and Thompson (2017) distinguish between the technical approach used to extract information from speech, the concrete system used to process a case, and the reasoning framework used to report the strength of evidence. Under the likelihood-ratio framework, the practitioner evaluates the probability of the observations under competing hypotheses, i.e. whether the questioned recording was produced by the known speaker or by some other speaker from a relevant population subgroup, which requires consideration of both similarity to the known speaker and typicality with respect to the population subgroup. Hence, similarity alone is not a sufficient strength-of-evidence statement. The ENFSI guidelines (Drygajlo et al., 2015) similarly frame automatic and semiautomatic forensic speaker recognition as an evaluative task embedded within a Bayesian interpretation framework, focusing on case assessment, net duration of speech, technical quality, mismatched recording conditions, relevant population, validation, calibration, and reporting.

Modern speaker verification systems represent each utterance as a dense speaker embedding and compare embeddings using cosine similarity, PLDA, or a learned backend. The x-vector family popularized neural embedding-based speaker recognition (Snyder et al., 2018), and ECAPA-TDNN, with its channel-attentive temporal aggregation, remains a strong supervised baseline (Desplanques et al., 2020). Self-supervised speech encoders such as wav2vec 2.0, HuBERT, WavLM, and W2V-BERT have further improved speaker representations when adapted to speaker-discriminative objectives (Chen et al., 2022), achieving EERs below 1% in VoxCeleb1. These systems are accurate but interpretable only in aggregate: the speaker embedding entangles pitch, voice quality, formant structure, segmental timing, and channel conditions into a single vector, and no per-feature reasoning can be recovered from a similarity score.

Recent forensic automatic speaker recognition research has emphasized that global performance metrics can hide substantial variability between speakers and samples. Hughes et al. (2024) examine individual-speaker behavior in a forensic x-vector system and show that, despite strong overall performance, a minority of speakers produce substantially worse calibrated log-likelihood-ratio cost values. Their phonetic-content ablation further shows that vowels generally improve performance, while stops and fricatives can degrade it for many speakers, and that the contribution of phonetic classes varies by speaker. They also report that long-term laryngeal features, such as fundamental frequency and voice quality, are not well correlated with ASR performance, suggesting that such features may provide complementary information for problematic speakers. These findings motivate comparison-specific explanations that identify which phonetic and acoustic observations are informative, unreliable, or missing from the embedding-based decision.

A separate line of work seeks to make the speaker recognition score itself interpretable. The BA-LR framework represents each speech extract as a binary vector of shared speaker attributes, where each coefficient indicates the presence or absence of a latent voice attribute (Amor and Bonastre, 2022). Each attribute is associated with explicit behavioral parameters, and the final likelihood ratio is computed as a product of attribute-level likelihood ratios under an independence assumption. Ben-Amor et al. (2024) extend this approach with a binary autoencoder that natively produces binary attribute representations under an attribute-oriented loss that encourages dimensions to behave like shared speaker-specific attributes. This work is closely related to CoLMbo-SV in that it treats speaker recognition as decomposable evidence rather than opaque embedding comparison. CoLMbo-SV differs in that its evidence is grounded in named acoustic-phonetic measurements (F0, formants, jitter, shimmer, HNR, VOT, voice-quality descriptors) and is verbalized as a structured forensic-style report rather than a vector of latent binary attributes.

Phonetic-trait models offer another route toward interpretable speaker verification. ExPO (Ma et al., 2025) introduces phonetic trait layers

into an ECAPA-TDNN-style architecture: a pretrained phone recognizer supplies phone boundaries, frame-level embeddings are averaged within phone segments, and the resulting phone-indexed trait embeddings are pooled into an utterance-level speaker embedding. At inference time, ExPO outputs both a final verification score and a phone-wise phonetic-trait similarity vector, allowing users to inspect which phonetic traits support a trial. This provides a human-interpretable view of neural speaker verification, but its explanations are learned phone-level similarity scores rather than named acoustic measurements.

Recent multimodal systems pair a speech encoder with a language model to perform audio captioning, speech understanding, and instruction following. Qwen-Audio and Qwen3-Omni (Chu et al., 2023) caption and answer questions about speech; SALMONN (Tang et al., 2024) integrates a Whisper encoder with an LLM via Q-Former layers; LTU and GAMA (Ghosh et al., 2024) train on instruction-following over audio content. These models can describe spectral and prosodic properties, but they are trained on caption-style data, not on pairwise speaker verification, and their numerical claims about a recording are typically not verified against the signal. In high-stakes speaker comparison, numerical claims must correspond to measured acoustic quantities, and the final decision must be traceable to the evidence cited in the report.

A small but growing body of work targets interpretable speech analysis with LLMs. The original CoLMbo system (Baali et al., 2025) produces natural-language descriptions of single utterances along dimensions such as accent, emotion, and voice quality, but it does not compare two recordings or render a verification decision. ADIFF (Deshmukh et al., 2025) performs a comparative voice-quality description for audio LLMs but operates only on perceived voice attributes; it does not extract acoustic measurements, does not ground its statements in intra-speaker variability, and does not emit a calibrated verification verdict. Recent work on explainable speaker recognition has also explored attention-based saliency and prototype methods, but these explanations are post-hoc visualizations of a fixed-embedding decision rather than structural reformulations of the decision itself. To our knowledge, CoLMbo-SV is the first system that unifies LLM-based comparative speaker analysis with the calibrated, likelihood-ratio-style evidence framework standard in forensic phonetics and that evaluates its generated explanations for numerical faithfulness as well as verification accuracy.

Table 4: Representative CoLMbo-SV outputs across different prompt types. Numerical citations match the underlying signal evidence (Faithfulness ≥ 0.9 in each case shown). Confidence levels are derived from Fisher-combined p-values over per-feature z-scores.

Prompt	Generated response
Overall assessment. *"Give an overall assessment of whether these two recordings come from the same speaker."*	**## 1. Recording Conditions.** *"Utterance 1 exhibits a higher signal-to-noise ratio (SNR) of 35.49 dB compared to utterance 2's 31.21 dB. ... While utterance 1 benefits from a cleaner recording environment, the differences in SNR and noise floor are not so substantial as to preclude a reliable acoustic comparison."* **## 2. Prosodic & Voice Quality.** *"Mean F0: 113.24 vs 116.56 Hz; jitter higher in utterance 1 (1.99 vs 1.69); HNR 10.50 vs 11.90 dB."* **## 3. Spectral.** *"Utterance 1 F1=800 Hz, F2=1800 Hz, F3=2830 Hz; Utterance 2 F1=750, F2=1750, F3=2750."*
Segmental analysis. *"Analyze the segmental acoustic features F1, F2, F3 means, voiced and voiceless VOT and finish with a Conclusion line on speaker identity."*	*"Utterance 2 mean F1 827.08 Hz vs utterance 1's 756.44 Hz; utterance 1 F2=1859 Hz, F3=2919 Hz vs utterance 2's F2=1759, F3=2811. Voiceless VOT 80 ms (utt 2) vs 70 ms (utt 1); voiced VOT* -15 *vs* -18 *ms."* *"Most discriminating evidence: F1 (z=1.16, p=0.245); F3 (z=0.78, p=0.437); F2 (z=0.80, p=0.424). Combining 5 features via Fisher's method (χ^2=7.98, df=10, p=0.635) yields same speaker, high confidence."* **Conclusion:** *Same speaker, high confidence.*
Single-feature query (HNR). *"Examine HNR differences and explain how vocal clarity contributes to speaker comparison."*	*"Utterance 1 has higher HNR of 12.95 dB compared to utterance 2's 10.54 dB, indicating cleaner vocal tone."* *"Typical within-speaker variation for HNR is 1.50 (utterance 1 12.95 dB, utterance 2 10.54 dB; z=1.64, p=0.100), indicating different speaker, medium confidence."* **Conclusion:** *Different speakers, medium confidence.*
Single-feature query (breathiness). *"Analyze breathiness in both utterances and explain whether it supports or weakens speaker differentiation."*	*"Breathiness index slightly higher in utterance 2 (0.55) vs utterance 1 (0.51)."* *"Most discriminating evidence: breathiness index (z=0.52, p=0.605); voice_quality.breathiness (z=0.52, p=0.605); voice_quality.whisper (z=0.12, p=0.903). Combining 3 features via Fisher's method (χ^2=2.33, df=6, p=0.886) yields same speaker, high confidence."* **Conclusion:** *Same speaker, high confidence.*