A Multilingual Voice Analytics Module for Contact-Center Hiring

Wagner W. Ávila Bombardelli[1] **Vanessa Marquiafavel Serrani**[2] **Edgard Kuboo**[1]
Erica C. Marins Missão[1] **Marcelo Noronha**[2]
[1]Atento Co. / Data Science, São Paulo, Brazil
[2]Atento Co. / Language User Interface, São Paulo, Brazil
{wagner.bombardelli, vanessa.serrani, edgard.kuboo, erica.missao, marcelo.noronha}@atento.com

Abstract

Contact-center operations often face significant challenges in identifying candidates whose vocal performance aligns with high-quality customer interactions. Existing speech analytics tools typically assess only content, providing limited insight into how candidates speak. To address this gap, we introduce a module to Smart Recruiter Voice (SR-Voice), a multilingual speech analytics module designed to support call-center hiring. SR-Voice extends a previous text-only auditor by integrating segment-level, audio-native analysis capable of generating judgments, concise evidence-based rationales, and 0–10 scores across three dimensions: Emotion, Communication, and Rhythm. Our two-stage architecture first applies an audio-native model to propose a label, which is then reassessed by a lightweight auditor that combines transcript cues with acoustic and timing indicators grounded in phonetic and prosodic theory. We evaluate SR-Voice on a production-like volunteer dataset, reporting strong agreement and calibration performance reaching Macro-F1 = 0.83; Expected Calibration Error (ECE) = 0.053. The hybrid system achieves state-of-the-art calibration without post-hoc adjustment, with the audio-only variant attaining the lowest Negative Log-Likelihood (NLL) = 0.472). Designed for operational practicality, SR-Voice emphasizes traceability, short rationales, and well-calibrated probabilities suitable for threshold-based decisions and human-in-the-loop triage. We also discuss privacy-preserving storage and the prospective masking of Personally Identifiable Information (PII) for archival data.

1 Introduction

This product began as a text-centric auditor for hiring simulations: candidates interacted with scripted scenarios and were scored on clarity, completeness, and cordiality based on *what* they wrote. However, it missed prosodic and voice-quality cues critical for contact-center roles.

Why a dedicated voice module now. In real call-center audio, prosody and voice quality convey information the transcript cannot (Picard, 2010; Jurafsky and Martin, 2024). Production artifacts—overlap, background noise, accents, multilinguality—also degrade conventional Speech Emotion Recognition (SER) and rigid checklists (Swain et al., 2018; Busso et al., 2008). For PT-BR specifically, acted/lab corpora underrepresent crosstalk and channel variability (Torres Neto et al., 2018; Costantini et al., 2014). These gaps motivate SR-Voice (Smart Recruiter-Voice module), which scores *how* candidates speak, not only *what* they say.

Scope of this paper. SR-Voice is a multilingual (EN-US, ES, PT-BR) module that takes raw speech as input, derives acoustic and temporal features, and outputs calibrated, sequence-aligned segment-level judgments alongside an overall performance score for candidate assessments. In the present study, we restrict our experiments to Brazilian Portuguese (PT-BR). This choice is motivated by (i) corpus availability, (ii) the applied focus on Brazilian contact-center assessment scenarios, and (iii) the need for controlled experimental conditions when analyzing acoustic–prosodic indicators. Low-level preprocessing (e.g., segmentation or diarization) is considered an auxiliary front-end signal for session-level integrity and remains out of scope. The system combines different models to compose the final score.

Problem and setting. In contact-center hiring scenarios, evaluators must determine whether candidates demonstrate stable and intelligible voice quality, as well as an appropriate speaking pace and fluency—attributes that directly affect customer experience.

Designing automated support for this task is challenging. Publicly available speech resources remain limited for several languages, particularly

Brazilian Portuguese, and widely used speech emotion recognition (SER) benchmarks (e.g., IEMOCAP, RAVDESS) only partially reflect real call-center conditions, as they are typically collected in controlled or acted environments. As a result, models trained on such data may not generalize to spontaneous service interactions.

Beyond data mismatch, operational Quality Assurance (QA) settings impose additional requirements: systems must provide time-stamped, traceable justifications for their assessments and remain robust to diarization errors and Automatic Speech Recognition (ASR) uncertainty.

Approach. We adopt a dual-stage architecture combining an audio-native model with a lightweight LLM auditor. Additional features help to improve the quality. Voice Activity Detector (VAD) removes non-speech, *reference*-based diarization supports continuity checks, and low-support depreciation stabilizes scores. Results follow a unified JSON schema for dashboards and review.

Contributions.

(1) A production-ready, multilingual speech module for contact-center hiring that provides time-stamped, evidence-based evaluations of vocal delivery.

(2) A novel two-stage design combining an audio-language model with a lightweight LLM auditor, enhanced by integrity mechanisms (e.g. low-support depreciation) for reliable operation in noisy environments.

(3) A practical assessment protocol reporting both agreement with human raters (Macro-F1, Cohen's κ) and critical probability-quality metrics (Expected Calibration Error, Brier score) on a production-like dataset, with a focus on Brazilian Portuguese.

2 Related Work

Text-only assessment frameworks capture task completion, discourse structure, and politeness strategies, but systematically miss prosodic and voice-quality cues that shape intelligibility, affect, and perceived professionalism in call-center speech (Picard, 2010; Jurafsky and Martin, 2024). While transcript-based auditing benefits from advances in large-vocabulary ASR, lexical evidence alone cannot account for phonatory stability, rhythmic control, or temporal fluency—dimensions that strongly influence communication quality in hiring scenarios.

Prior work has proposed a wide range of acoustic metrics to quantify voice quality and prosodic behavior; however, reported normative values vary substantially across tools, recording conditions, and corpora. As a result, published norms are better treated as soft priors rather than fixed thresholds. Acceptable ranges are typically tool- and corpus-dependent (de Felippe et al., 2006; Wertzner et al., 2005; Leme et al., 2016), motivating speaker- or session-level normalization strategies.

Recent work in multimodal speech analytics integrates acoustic embeddings and language models to improve robustness in real-world evaluation tasks. Such systems combine speech representations with textual features derived from ASR, aiming to capture both semantic and paralinguistic information. However, many approaches rely on high-dimensional neural embeddings and prioritize predictive performance over interpretability, leaving open questions regarding explainability and auditability in professional decision-making contexts.

2.1 Acoustic and Timing Cues

Classical phonetics and clinical acoustics formalize reproducible extraction of acoustic indicators and provide theoretical grounding for voice-quality analysis (Titze, 2000; Boersma and Weenink, 2001; Lerch, 2022).

Pitch statistics (F_0 mean, variance, median) track register and emphasis, contributing to perceived engagement and pragmatic structuring.

Jitter and *shimmer* capture cycle-level instability linked to roughness, breathiness, and perceived clarity (Maryn and Roy, 2009; Titze, 2000; Teixeira et al., 2013).

HNR and *GNE* summarize harmonic organization versus turbulent noise, providing compact measures of phonatory regularity (Yumoto et al., 1982; Michaelis et al., 1998, 1997).

Prosodic fluency—operationalized via speech rate and pause ratio—quantifies pacing and hesitation. These temporal dimensions exhibit well-documented variation across speaking styles and communicative contexts (Tauroza and Allison, 1990; Yuan and Liberman, 2010; Campione and Véronis, 2002; Goldman-Eisler, 1968). In applied evaluation settings, temporal control directly influences perceptions of confidence, clarity, and communicative effectiveness.

Large-vocabulary ASR systems provide segment-indexed lexical evidence (e.g., repetitions, truncations, quotations) as well as temporal alignment that supports prosodic summarization. Nevertheless, recognition errors and partial coverage motivate cross-checks against acoustic cues and conservative evidence rules (Jurafsky and Martin, 2024; Radford et al., 2023). In hybrid auditing settings, transcripts typically serve as structured evidence rather than acting as standalone classifiers.

2.2 Agreement and Calibration

Perceptual judgments of speech are inherently subjective. When segments lack salient lexical or prosodic cues, annotators may diverge in their evaluations. Consequently, evaluation frameworks often anchor thresholds to human gold standards (e.g., double annotation followed by adjudication), quantify agreement using Cohen's κ—including weighted variants for ordinal or ternary labels—and report macro-F1 to capture class-balanced performance (McHugh, 2012).

Beyond accuracy, probability quality and calibration are increasingly recognized as critical for decision-support systems. Reliability diagrams, Expected Calibration Error (ECE), and the multiclass Brier score provide complementary views of probabilistic consistency (Guo et al., 2017; Brier, 1950). In operational settings, calibrated outputs are particularly relevant when model predictions inform human decisions.

2.3 Linguistic Foundations: Prosody, Rhythm, and Pragmatic Function

Acoustic and timing cues used in speech assessment are grounded in established theories of prosody. Autosegmental–Metrical frameworks conceptualize intonational patterns as configurations of pitch accents and boundary tones associated with information structure and pragmatic function (e.g., focus marking, turn-taking). These prosodic dimensions directly affect perceived communication quality and emotional expressivity.

Rhythm metrics (e.g., pairwise variability indices) and measures of syllable-timed versus stress-timed behavior quantify temporal alternation patterns that shape fluency and naturalness in spontaneous dialogue. By linking extracted acoustic indicators to linguistic constructs, feature-level representations remain directly traceable to phonetic theory. Unlike end-to-end neural embeddings, such indicators facilitate transparent auditor rationales and enable targeted diagnostics (e.g., distinguishing local hesitation from global tempo instability).

Despite advances in acoustic modeling, ASR-based auditing, and multimodal fusion, relatively few systems explicitly combine theoretically grounded acoustic indicators, transcript-derived linguistic evidence, and calibrated probabilistic outputs aligned with human agreement in hiring scenarios. This gap is particularly salient for contact-center speech in Portuguese, where interpretability, auditability, and linguistic grounding are central requirements.

3 Methods

We operationalize SR-Voice as the composition of a speech model and an auditor (see §1). Figure 1 outlines the flow; here we specify inputs, feature extraction, evidence rules, and evaluation metrics. The results are reported at the segment level.

We run [1] SR-Voice in Cloud-1 (managed cloud) , orchestrated with K8s-Managed (managed Kubernetes) . Automatic speech recognition (ASR) uses ASR-Cloud (large-vocabulary ASR) to speech-to-text (STT) The language auditor is LLM-Based , and the audio-native model is ALM-Based served as a containerized service in the same Kubernetes cluster. Kubernetes node pools are backed by C-GPU-80G (data-center GPU, with at least 80 GB VRAM) and all calls, model versions, prompts, and features are logged for auditability. To improve robustness and latency, long recordings are split into segments of at most 30 s, preferring pause-aligned boundaries when available. This window size matches the reliability we observe on short utterances and prevents degradation on very long clips.

Optional front-end modules (e.g., VAD, neural diarization with reference enrollment, and overlap detection) are used solely to gate candidate-attributed segments and to emit session-level integrity flags. These signals are logged for traceability but are intentionally excluded from segment-level evaluation and scoring. This design isolates the contribution of acoustic–prosodic and linguis-

[1] Certain implementation details (including specific commercial providers, proprietary configurations, model variants, and cloud infrastructure identifiers) have been anonymized due to industrial confidentiality constraints. The abstractions reported here preserve the technical characteristics relevant for scientific evaluation (model class, deployment paradigm, hardware profile, and logging strategy) while omitting commercially sensitive information.

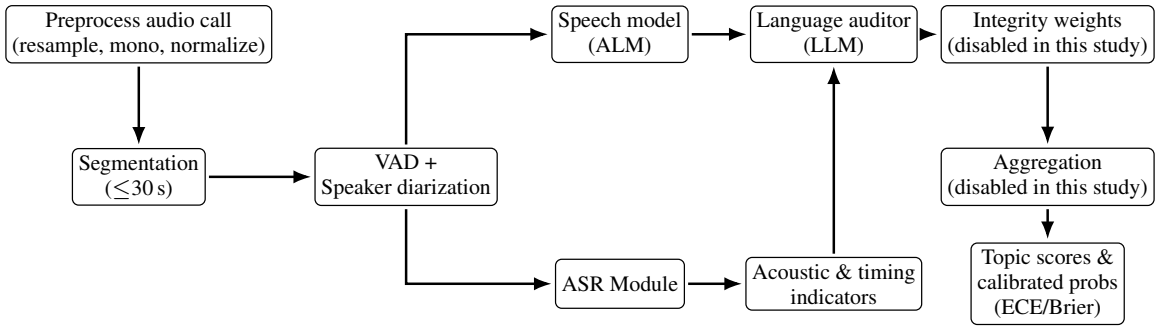

Figure 1: SR-Voice pipeline overview.

tic features, avoids propagating front-end segmentation errors into the evaluation metrics, and ensures that reported results reflect assessment quality rather than preprocessing performance.

For each candidate-attributed segment, the audio-only speech model emits per-topic judgments with confidences; when available, ASR word-level timestamps are used to derive timing features (e.g., WPM, pause ratio). The auditor LLM consumes (i) the speech model's hypothesis, (ii) acoustic and timing statistics (e.g., F_0 summary; jitter/shimmer; HNR/GNE; WPM; pause ratio), and (iii) segment-indexed transcript cues (e.g., repetitions, truncations, filled pauses). It outputs an audited verdict per topic and a mapped score $s_t \in [0, 10]$ with a concise justification; scores are a monotonic mapping of confidences to a fixed 0–10 scale (no post-hoc probability calibration).

Auditor ranges and indicators. We pass explicit units and soft priors in the prompts to reduce tool-dependent ambiguity. The numeric ranges and indicators used are summarized in the Tables below (Tables 1, 2).

3.1 Internal dataset and human annotation

Test calls were collected in a production-like sandbox with consented internal volunteers acting as candidates across realistic support scenarios. Audio was stored per caller and segmented per the policy in §3 ($\leq$ 30 s, pause-aligned when available).

Corpus description The internal evaluation corpus comprises speech recordings from 16 consented volunteer speakers (11 male, 5 female) collected in a production-like sandbox simulating realistic support-call scenarios. The segmented audio yielded 94 candidate-attributed speech segments. Each segment constitutes an independent evaluation unit and is assessed across three topics—Emotion, Communication, and Rhythm—resulting in 282 segment–topic instances. This design enables topic-level analysis while preserving segment independence in both modeling and evaluation.

Annotation unit and schema. The basic annotation unit is a *candidate-attributed speech segment*. Each segment is evaluated independently for each topic in $\{\text{Emotion}, \text{Communication}, \text{Rhythm}\}$.

For every segment–topic pair, we record:

- **File identifier**;
- **Hybrid system verdict** $\hat{y} \in \{\text{Yes}, \text{No}, \text{Undetermined}\}$;
- **Mapped score** $s_t \in [0, 10]$, accompanied by a brief rationale.

For human adjudication, we additionally collect:

- **Agreement flag** (binary: *agree* vs. *disagree*);
- **Human score** $h_t \in [0, 10]$;
- **Short textual justification**.

This schema enables direct comparison between automated predictions and human evaluations at the topic level.

Rater guidance and consensus. Scores are "higher is better": calm/professional voice (*Emotion*); clear, intelligible speech with minimal disfluency (*Communication*); fluent rhythm with natural pace and pauses (*Rhythm*). Raters mark `agree` only when the hybrid verdict *and* its cited evidence (lexical quote or prosodic cue) align with the perceived segment. Each item is triple-annotated by independent raters; disagreements are resolved via consensus (with *Undetermined* preserved when evidence is insufficient). Inter-rater agreement and calibration diagnostics are reported in §2.1 and Results.

Table 1: Two-column summary of acoustic and timing metrics used by the auditor (LLM-only) model.

Metric	Unit	Guidance (soft priors)	Role in system
Jitter (local)	%	∼0.3–0.6% indicates stable periodicity; higher ↑ = instability/strain (supportive, never decisive).	Emotion/quality support
Shimmer (local)	dB	∼0.3–0.5 dB "safe"; higher ↑ = amplitude instability.	Emotion/quality support
HNR	dB	>12 dB clean voicing; 9–12 moderate; ≪9 hoarse/noisy.	Supportive cue (never decisive)
GNE	–	∼0.1–0.3 clearer phonation (tool-dependent); use with HNR, not alone.	Quality support
F_0 (mean)	Hz	M: 85–180; F: 165–255; moderate dispersion; out-of-range cues arousal/strain.	Emotion support
WPM	w/min	110–145 fluent; >160 fast; <110 slow (cross-check with pauses/syll/s).	*Rhythm*
Syllables/s	syll/s	∼3–5 conversational; useful when segments are short.	*Rhythm* (backup)
Pause ratio	% time	10–20% typical; high values can signal hesitation; avoid sole use on short clips.	*Rhythm*/*Communication*
Energy dynamics	dB (rel.)	Abrupt changes/peaks indicate emphasis or instability; helps segmentation.	Emphasis/segmentation
Formants (F1–F3)	Hz	Vowel/speaker dependent; used qualitatively for plausibility/clarity.	Quality sanity check
Integrity signals (VAD/overlap)	–	Filter/flag non-speech or crosstalk; *logged only; not used for evaluation*.	Integrity (logged)
Diarization (DER)	%	Continuity vs. *reference* enrollment; *logged only; not used for evaluation*.	Integrity (logged)

Notes / refs. Ranges are supportive heuristics, not rules; decisions combine acoustics with transcript evidence (literal quotes + timestamps). Acoustic/voice-quality: (de Felippe et al., 2006; Leme et al., 2016; Maryn and Roy, 2009; Titze, 2000; Boersma and Weenink, 2001; Yumoto et al., 1982; Michaelis et al., 1997, 1998). Rate/timing: (Tauroza and Allison, 1990; Yuan and Liberman, 2010; Campione and Véronis, 2002; Goldman-Eisler, 1968; Lerch, 2022). Diarization/overlap: (Bredin et al., 2019).

Table 2: Two-column summary of transcript-derived indicators used by the auditor and for evidence rendering.

Indicator	Type/Unit	Guidance (soft priors)	Role in system
Adjacent word repetitions	count / seg.	Immediate token repeats in a clause (e.g., "eu, eu…"); case/diacritic–insensitive; deduplicate ASR alts.	*Communication*: dysfluency cue.
Truncated / partial words	boolean / count	Use ASR cut-off/hesitation markers (e.g., "co–", "con…") to flag broken onsets/endings.	*Communication*: articulation/flow.
Filled pauses (PT-BR)	count / seg.	Language-specific fillers ("ah", "é…", "hum"); ignore evaluator backchannels.	*Communication*/*Rhythm*.
Lexical repetition flag (n-gram loops)	boolean (+ count)	Flag repeated bi/tri-grams over a sliding window (20–30 s) to capture looping/word search.	*Communication*: cohesion/fluency.
Mean utterance length (candidate)	tokens / turn	Average tokens per candidate turn; robust to micro-pauses and minor timing noise.	*Communication*: verbosity
Fraction of call time aligned to word timestamps; low coverage down-weights transcript cues.	Confidence modulator.		
Text anomalies	boolean / tags	Detect nonsensical fragments or obvious ASR artifacts; if no literal support, use "prosodic inference" wording.	Justification hygiene.

Notes / refs. Transcript indicators are supportive; final decisions combine acoustics and text per the auditor prompt. Timing–based fluency metrics (e.g., *WPM*, text-timed pause ratio, syllables/s) appear in the rhythm/interaction table. Background on discourse pauses and text evidence: (Jurafsky and Martin, 2024; Goldman-Eisler, 1968; Campione and Véronis, 2002; Lerch, 2022).

3.2 Validation metrics

Cohen's κ for ternary judgments. Following (McHugh, 2012; Landis and Koch, 1977), let p_o be the observed agreement and p_e the chance agreement induced by the label marginals over $\mathcal{Y} = \{\text{Yes}, \text{No}, \text{Undet.}\}$. Cohen's κ is

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \tag{1}$$

For ordinal variants (e.g., penalizing *Yes* vs. *No* more than disagreements involving *Undet.*), weighted κ uses a cost matrix w_{ij} and reweights the

observed/expected proportions accordingly. In SR-Voice, κ captures criterion-level agreement beyond prevalence, aligning with human-facing audits.

Macro-F1 for class balance. For each class $c \in \mathcal{Y}$ with precision P_c and recall R_c, define $F1_c = 2P_cR_c/(P_c + R_c)$. Macro-F1 averages $F1_c$ over classes, reducing sensitivity to class imbalance—relevant since *Undet.* is expected under noisy or short segments; see (Jurafsky and Martin, 2024) for a general overview.

Probabilistic calibration (ECE and reliability). As in (Guo et al., 2017), we partition predictions into M confidence bins $\{B_m\}_{m=1}^{M}$ based on the top-class confidence and compute the Expected Calibration Error:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|, \tag{2}$$

where $\mathrm{acc}(B_m)$ is the empirical accuracy and $\mathrm{conf}(B_m)$ the mean predicted confidence in bin B_m. Reliability diagrams visualize acc vs. conf across bins (Guo et al., 2017).

Multiclass Brier and NLL. The multiclass Brier score (Brier, 1950) complements ECE by assessing the full probability vector:

$$\mathrm{Brier} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k \in \mathcal{Y}} \left(\hat{p}_{ik} - o_{ik} \right)^2, \tag{3}$$

where o_{ik} is one-hot (1 if $y_i k = k$, else 0) and $\hat{p}_{ik}$ is the predicted probability for class k. We also report the negative log-likelihood (log-loss),

$$\mathrm{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \log \hat{p}_{i,y_i}, \tag{4}$$

which is especially sensitive to overconfident errors.

Why these metrics fit SR-Voice. (i) κ discounts chance agreement and is robust to prevalence shifts—crucial when *Undet.* varies with noise/overlap. (ii) **Macro-F1** prevents dominance by majority classes when *No* (normal) or *Yes* (issue) outweighs *Undet.*. (iii) **ECE/Brier/NLL**evaluate probability quality used in dashboards and thresholding, supporting human-in-the-loop triage and safe deployment.

Table 3: Overall performance on the filtered evaluation set (*Undetermined* excluded). Best per column in bold; ties in bold

Model	Macro-F1	κ	ECE_{10}	Brier	NLL
Hybrid	**0.861**	0.536	**0.053**	**0.109**	0.630
ALM	**0.861**	**0.607**	0.091	0.189	**0.472**
LLM	0.760	0.446	0.076	0.158	0.602

Note on post-hoc calibration. We did not apply post-hoc calibration in this study; when miscalibration is observed, methods such as Bayesian binning (BBQ) and beta calibration can be applied without retraining (Naeini et al., 2015; Kull et al., 2017).

3.3 Evaluation protocol

We evaluate at the segment level against a triple-annotated human gold (§3): items with *Undetermined* gold response are excluded from agreement metrics unless stated otherwise. We report (i) Macro-F1 per topic and overall, (ii) chance-corrected agreement via Cohen's κ (weighted for ternary labels when noted), and (iii) probabilistic calibration for the hybrid model's dashboard probabilities using top-class ECE (M=10 bins) and the multiclass Brier score (McHugh, 2012; Guo et al., 2017; Brier, 1950).

4 Results

We evaluate on a filtered set that excludes *Undetermined* gold labels (123 segment–topic pairs: 40 *Communication*, 45 *Emotion*, 38 *Rhythm*). We report Macro-F1, Cohen's κ, and calibration metrics: top-class ECE (10 bins), multiclass Brier, and negative log-likelihood (NLL). Overall performance results are reported in Table 3. Higher is better for Macro-F1/κ; lower is better for ECE/Brier/NLL.

Overall. Hybrid matches the best Macro-F1 while achieving the lowest calibration error and Brier (ECE 0.053; Brier 0.109). ALM attains the lowest NLL (0.472), indicating sharper probability mass when correct, though it is less well-calibrated overall. LLM trails both in accuracy and calibration.

Per-topic summary. *Communication* is essentially solved by Hybrid and ALM (near-perfect agreement); residual ECE underscores that perfect accuracy does not imply perfect calibration. In *Emotion*, Hybrid shows strong agreement but higher NLL driven by a few overconfident errors.

Rhythm is the most variable dimension, consistent with its sensitivity to segment length and ASR timing noise.

Visual diagnostics. Figure 2 shows reliability diagrams for Hybrid, ALM, and LLM (accuracy vs. confidence, $M=10$ bins); Hybrid adheres most closely to the diagonal, matching its lowest ECE/Brier, while ALM is sharper but less calibrated in mid-confidence bins. Figure 3 presents the Hybrid confusion matrix (row-normalized, three classes), with mass concentrated in the *No* column due to class prevalence and a conservative operating point. Both figures use the full three-class set including gold *Undetermined* for visualization; Table 3 reports metrics on the filtered set.

5 Discussion

Agreement. Overall κ values fall in the *moderate–substantial* range (Landis and Koch, 1977; McHugh, 2012): ALM reaches $\kappa=0.607$ (substantial), Hybrid $\kappa=0.536$ (moderate). For *Communication*, both models achieve near-perfect agreement—a ceiling effect of this sample, which contains only one clip with a clear communication failure; most segments are fluent speech.

Calibration. ECE_{10} surfaces a distinct quality axis: Hybrid is best calibrated (0.053), outperforming LLM (0.076) and ALM (0.091). Well-calibrated probabilities matter for thresholding and human-in-the-loop triage (Guo et al., 2017). Across topics, *Rhythm* is the least calibrated on average, consistent with its higher ambiguity and dependence on timing cues.

Error profile and dataset effects. The Hybrid confusion matrix (Fig. 3, visualized on the full three-class set) shows a tendency toward *No*, especially when gold is *Undetermined*. This aligns with corpus composition (volunteers attempting to perform well) and a conservative operating regime. This functions as a continuous safety guardrail within SR-Voice. All segments are automatically screened for potentially harmful content, such as offensive language, hate speech, or sexist intent, and any flagged segment is vetoed before downstream scoring. This ensures that SR-Voice remains within a controlled safety envelope, preserving ethical and operational integrity even under noisy or high-risk inputs

In *Emotion*, a handful of overconfident mistakes inflate NLL despite low ECE/Brier .

Takeaways. Hybrid matches ALM on Macro-F1 while offering markedly better calibration; LLM trails on both. Given the small, clean dataset, these are likely lower-bound estimates. We expect larger gains with more diverse acoustics and adjudicated dysfluency samples; future work will expand the corpus and, if needed, explore post-hoc calibration on held-out data.

Linguistic interpretability and operationalization From a linguistic perspective, the indicators extracted by SR-Voice correspond to three major phonetic–prosodic dimensions. (i) Voice quality measures (jitter, shimmer, GNE) capture phonatory stability and vocal effort, correlating with perceived emotional valence and physiological tension. (ii) F0-based statistics reflect both affective modulation and intonational structure, linking to pragmatic functions such as emphasis, politeness, or engagement. (iii) Temporal organization metrics—speech-rate, pause ratio, and Pairwise Variability Index (PVI)—represent rhythmical properties tied to fluency and conversational turn management. Operationalizing these constructs as soft priors in the auditor's reasoning prompts enables rationales that are acoustically grounded and linguistically interpretable. This connection between model output and theoretical prosody provides transparency and supports targeted feedback for human auditors and trainees.

6 Ethical impact

This work uses internally collected speech from volunteer colleagues in a production-like setting. Participation was informed and voluntary; no minors or sensitive groups were targeted. Audio was processed inside an access-restricted environment with encryption in transit and at rest, auditable access logs, and role-based permissions. For this study, raw audio was analyzed without automated PII redaction to preserve acoustic fidelity during evaluation; before long-term retention or any sharing, we store only de-identified artifacts (e.g., annotations, derived features, and transcripts with manual/automatic redaction where applicable), and we will not release raw audio.

We recognize potential fairness risks (e.g., accent, speaking rate). To mitigate these, we (i) report inter-rater agreement and use a ternary label space to encode residual ambiguity, and (ii) plan stratified analyses and bias audits in future releases. No attempts were made to infer protected attributes,

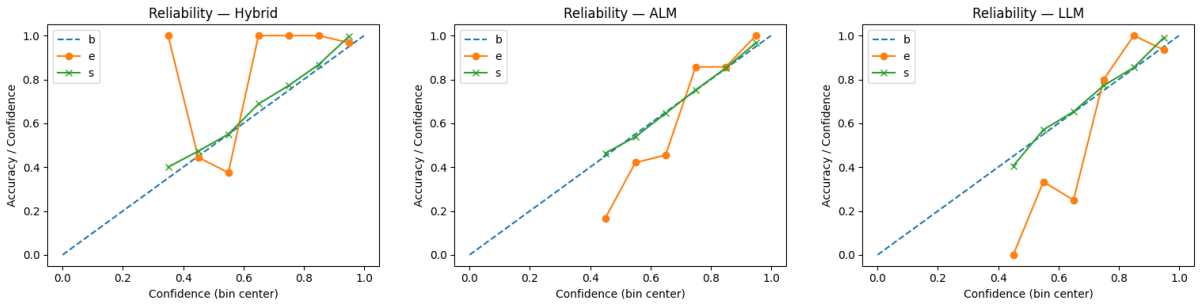

Figure 2: Reliability diagrams (accuracy vs. confidence). Legend: **b** = ideal baseline (dashed diagonal, perfect calibration); **e** = empirical accuracy per confidence bin; **s** = mean predicted confidence per bin. Hybrid tracks the diagonal more closely, consistent with the lowest ECE/Brier.

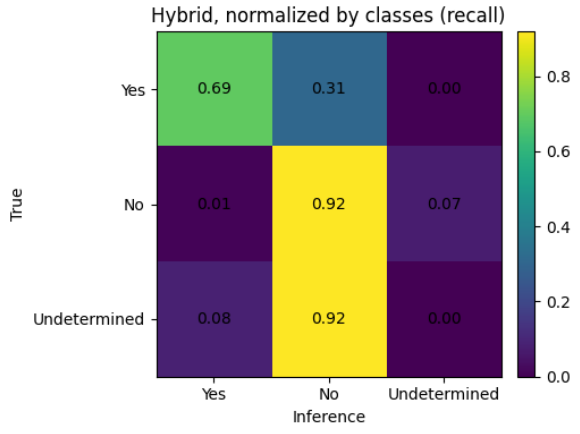

Figure 3: Confusion matrix on the full three-class set (*Undetermined* included for visualization; evaluation metrics are reported on the filtered set). The concentration along the *No* column reflects the conservative operating point and class distribution.

and no harmful-content data were curated. This submission follows the ACL Policy on Publication Ethics regarding authorship, conflicts of interest, privacy, and use of generative assistance.

7 Limitations

Our evaluation uses a small, relatively clean internal dataset with volunteer speakers, which biases class prevalence and under-represents difficult cases (noise, overlap, clinically validated dysfluency). Speech judgments are inherently subjective; we mitigate with double annotation and adjudication, but residual ambiguity remains. We report segment-level results only; low-level preprocessing and session-level aggregation/integrity weighting are treated as replaceable front ends and are out of scope. We did not apply post-hoc calibration in this study. Future work will expand and diversify the corpus, report confidence intervals (e.g., non-parametric bootstrap over segments), and—if warranted—study calibration methods and per-threshold trade-offs on held-out data.

8 Conclusion

The SR-Voice module demonstrates that hybrid audio–text modeling can produce reliable and interpretable assessments of vocal delivery in realistic hiring scenarios.

Our findings highlight that Communication reaches a performance ceiling even for human raters, while Rhythm remains the most challenging criterion to assess, suggesting an inherent perceptual ambiguity that future models must address. By combining phonetic indicators with transparent probability estimates, SR-Voice advances the practical use of speech analytics for recruitment, offering traceability suitable for threshold-based and human-in-the-loop decision pipelines.

Future work will explore multilingual generalization, longer-form conversational data, and refined calibration across acoustic domains, as well as ethical safeguards for privacy and bias. Taken together, these directions move SR-Voice toward a robust, explainable foundation for audio-native evaluation in real-world, multilingual settings.

9 Acknowledgments

The authors would like to thank Atento for their institutional support throughout this project. We are especially grateful to the Innovation Board at Atento for their strategic guidance and continuous encouragement of research and development initiatives. We also sincerely thank our Atento colleagues who contributed to the creation and organization of the corpus. Their collaboration, domain expertise, and careful work were essential to the successful development of this resource.

References

Paul Boersma and David Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Grégory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2019. pyannote.audio: Neural building blocks for speaker diarization. *arXiv preprint arXiv:1911.01255*.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Estelle Campione and Jean Véronis. 2002. A large-scale multilingual study of silent pause duration. In *Proceedings of Speech Prosody*, pages 199–202.

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. Emovo corpus: An italian emotional speech database. In *Proceedings of LREC*, pages 3501–3504, Reykjavik, Iceland.

Ana Clara Naufel de Felippe, Maria Helena Marotti Martelletti Grillo, and Thaís Helena Grechi. 2006. Normatização de medidas acústicas para vozes normais. *Revista Brasileira de Otorrinolaringologia*, 72(5):659–664.

Frieda Goldman-Eisler. 1968. *Psycholinguistics: Experiments in Spontaneous Speech*. Academic Press, London.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330. Introduces reliability diagrams/ECE for modern networks.

Daniel Jurafsky and James H. Martin. 2024. Speech and language processing. Draft third edition.

Meelis Kull, Telmo de Menezes e Silva Filho, and Peter Flach. 2017. Beyond platt scaling: Beta calibration for binary classifiers. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 623–631.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

André Luis Maciel Leme, Márcio Abud Marcelino, and Pedro Paulo Leite do Prado. 2016. Medidas vocais acústicas de mulheres sem queixas de voz e com laringe normal. *CoDAS*, 28(5):610–617.

Alexander Lerch. 2022. *An Introduction to Audio Content Analysis: Music Information Retrieval Tasks and Applications*, 2 edition. Wiley-IEEE Press.

Youri Maryn and Nelson Roy. 2009. Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia: A systematic review. *Journal of Voice*, 23(3):291–301. Discusses reliability and limitations of acoustic measures such as jitter and shimmer.

Mary L. McHugh. 2012. Interrater reliability: the Kappa statistic. *Biochemia Medica*, 22(3):276–282.

Dirk Michaelis, Mattias Fröhlich, and Hans W. Strube. 1998. Selection and combination of acoustic features for the description of pathologic voices. *The Journal of the Acoustical Society of America*, 103(3):1628–1639. Introduces and validates the Glottal-to-Noise Excitation (GNE) ratio.

Dirk Michaelis, Thomas Gramss, and Hans W. Strube. 1997. Glottal-to-noise excitation ratio: A new measure for describing pathological voices. *Journal of Voice*, 11(3):319–331.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2901–2907. Also known as Bayesian Binning into Quantiles (BBQ).

Rosalind W. Picard. 2010. Affective computing: From laughter to ieee. *IEEE Transactions on Affective Computing*, 1(1):11–17.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*, volume 202 of *PMLR*, pages 28492–28518.

Madhusmita Swain, Aurobinda Routray, and Prithwiraj Kabisatpathy. 2018. Databases, features and classifiers for speech emotion recognition: A review. *International Journal of Speech Technology*, 21:93–120.

Steve Tauroza and Desmond Allison. 1990. Speech rates in british english. *Applied Linguistics*, 11(1):90–105.

João Paulo Teixeira, Carla Oliveira, and Carla Lopes. 2013. Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia technology*, 9:1112–1122.

Ingo R. Titze. 2000. *Principles of Voice Production*. Iowa City: National Center for Voice and Speech.

José R. Torres Neto, Geraldo P. Rocha Filho, Leandro Y. Mano, and Jó Ueyama. 2018. Verbo: Voice emotion recognition database in portuguese language. *Journal of Computer Science*, 14(11):1420–1430.

Haydée F. Wertzner, Solange Schreiber, and Luciana Amaro. 2005. Análise da frequência fundamental, jitter, shimmer e intensidade vocal em crianças com transtorno fonológico. *Revista Brasileira de Otorrinolaringologia*, 71(5).

Jiahong Yuan and Mark Liberman. 2010. The loudness and speaking rate of broadcast news speech. In *Proceedings of Interspeech*, pages 1393–1396.

Eiji Yumoto, William J. Gould, and Thomas Baer. 1982. Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, 71(6):1544–1550.