From Naturalness to Norms: Interactional Cultural Competence for SpeechLMs

Santosh T.Y.S.S.
Amazon, Seattle, U.S.A
santoshtyss@gmail.com

Abstract

Spoken language models (SpeechLMs) are increasingly real-time conversational actors. Yet many culturally consequential aspects of spoken interaction are not primarily lexical. Across sociolinguistics, linguistic anthropology, and conversation analysis, meaning emerges through how talk is produced and coordinated—prosody, timing, turn-taking, overlap, backchannels, and repair—within situated speech events. A transcript can be semantically correct yet interactionally inappropriate because many culture-bearing signals are audible and sequential rather than textual. This position paper argues for a speech-first view of cultural competence as interactional competence: the ability of a spoken agent to participate appropriately in event-situated interaction with locally normative conduct, while allowing plural acceptable realizations. Here, *appropriate* does not imply generic human-likeness; in many applications, the desired behavior may instead be constrained, neutral, predictable, or tool-like under an application-specific interaction contract. We synthesize social-science foundations into a theory-derived taxonomy of culture-bearing signals in speech, identify interactional phenomena where transcript correctness fails to predict appropriateness, and ground the agenda in today's SpeechLM stacks and evaluation practice. We propose an evaluation framing that complements WER/MOS and broad capability suites by making speech events and interaction contracts explicit, diagnosing where modern pipelines lose interactional cues, and treating cultural appropriateness as a norm-conditioned target rather than generic "naturalness."

1 Introduction: Culture Lives in Interaction

Spoken interfaces are increasingly deployed as real-time conversational actors: SpeechLMs now operate as *participants* in live interaction, not merely as recognizers or offline generators (Cui et al., 2025b; Peng et al.; Arora et al., 2025). In this setting, success is not only whether the system *understands* or *answers* correctly, but whether it behaves in ways that interlocutors interpret as appropriate for the situation—including when to speak, how long to pause, how to signal stance, and how to manage overlap. This aligns with dialogue accounts of *grounding*, where interlocutors continuously coordinate evidence of understanding under time and attention constraints (Clark and Brennan, 1991).

Transcript-identical minimal pair (speech changes the action).
Utterance: "That's great."
Affiliative (sincere): timely response, smooth onset, moderate pitch range; heard as alignment.
Ironic / face-threatening: delayed onset, flat or exaggerated contour, audible sigh/laughter; heard as sarcasm or reproach.

Figure 1: Why transcript correctness does not determine interactional appropriateness.

A core premise from the social sciences is that culture is not a static inventory of facts that speakers retrieve on demand. Rather, culture is enacted as practice in situated activity (Duranti, 2009). Much of what is culturally consequential in speech does not reside in propositional content alone, but in the coordination of multiple semiotic resources—prosody, timing, participation structure, and sequential organization—that guide interpretation in real time (Gumperz, 1982; Tannen, 2005; Sacks et al., 1974; Schegloff, 2007). This implies a practical point for speech systems: a transcript-first view (treating speech as "text plus noise") can erase precisely the cues that make talk culturally intelligible and normatively evaluable.

This paper advances a speech-first position: **cultural competence in spoken agents should be operationalized as interactional competence.** Figure 1 illustrates the core intuition: the transcript

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31787–31802

can remain fixed while the social action changes because the relevant cues are prosodic, sequential, and participatory rather than purely lexical. This does *not* claim that speech research has ignored prosody or turn-taking. On the contrary, speech and dialogue research has deep literatures on expressive TTS, prosody control, end-of-turn detection, incremental dialogue, and full-duplex interaction (Wang et al., 2018; Skantze, 2021; Ekstedt and Skantze, 2022; Lin et al., 2025; Castillo-López et al., 2025). Our claim is narrower: these phenomena are often framed as generic expressiveness, fluency, or human-likeness, whereas cultural competence requires evaluating them as normatively situated conduct relative to a speech event and its participation norms.

Crucially, *appropriate* should not be read as synonymous with *human-like*. In many deployments, the desired behavior is intentionally constrained: predictable, neutral, minimally affective, and explicitly tool-like. A spoken tutor, interpreter, scheduling assistant, or clinical interface may each require different interactional conduct, including different allowable uses of backchannels, hesitation, laughter, affective stylization, or persona. We therefore treat appropriateness as relative to an *interaction contract*: the application-specific expectations that define what kinds of conduct are desirable, optional, or disallowed in a given event.

Scope and non-goals. We use *culture* in a deliberately practical sense: locally shared, contestable expectations for how to conduct talk in a given *speech event* (e.g., meeting, classroom, service encounter) (Hymes, 1974), including norms for stance display, participation, and repair. These norms may be associated with institutions, roles, genres, and communities of practice, and need not map neatly onto nationality or demographic categories. We do *not* equate culture with speaker identity, and we do not advocate inferring user attributes. Instead, we treat norms as part of the task specification (event, roles, relationship, setting, goals, and interaction contract) while allowing plural acceptable realizations. We also distinguish event norms from individual preferences: personalization may be useful in some applications, but it is not a prerequisite for norm-conditioned evaluation. Finally, we do not argue that interactional competence exhausts cultural competence; semantic and world-knowledge aspects are complementary, but speech makes the interactional layer especially operational and especially easy to under-specify.

This perspective complements recent calls in NLP to treat culture as theoretically grounded and evaluation as intentionally cultural (Zhou et al., 2025; Oh et al., 2025; AlKhamissi et al., 2025), and aligns with socially aware NLP's emphasis that language is inseparable from social factors (Hovy and Yang, 2021; Yang et al., 2025). Our contribution is to translate these insights into the realities of SpeechLM design and evaluation: modern stacks have identifiable bottlenecks (endpointing, diarization, streaming latency, text intermediates, prosody control interfaces) that systematically shape interactional behavior, and therefore encode conversational norms by construction.

Contributions. We contribute (i) a theory-derived taxonomy of culture-bearing signals in spoken interaction (prosody, timing, participation, repair) with links to speech-event conditioning and interaction contracts; (ii) an analysis of where modern SpeechLM pipelines lose or reshape these cues (cascades and end-to-end systems); and (iii) a norm-conditioned evaluation framing built around explicit *speech events*, plural acceptable realizations, and diagnostic probes.

2 Foundations: Why Prosody, Timing, and Participation Are Meaning-Bearing

This section motivates why a speech-first account of cultural competence must treat prosody, timing, and participation as meaning-bearing, not ornamental. Across linguistic anthropology, interactional sociolinguistics, conversation analysis, and pragmatics, social meaning is not exhausted by lexical content but accomplished through audible, sequential, and participation-organized practices whose interpretation depends on the speech event and locally recognizable expectations.

2.1 Linguistic anthropology: meaning as practice and indexicality

Linguistic anthropology emphasizes that meaning emerges through participation in socially organized activity rather than residing solely in abstract linguistic form (Duranti, 2009). A central mechanism is *indexicality*: forms point beyond referential content to social meanings such as authority, familiarity, stance, and alignment (Silverstein, 2003). Indexical meanings are frequently carried by prosody and voice quality (e.g., how certainty, deference, irritation, or playfulness is displayed), and by interactional positioning (e.g., whether a response

comes immediately, with delay, with overlap, or through repair).

2.2 Interactional sociolinguistics: contextualization cues

Interactional sociolinguistics operationalizes this view through *contextualization cues* (Gumperz, 1982). Prosody, rhythm, pausing, and other interactional signals instruct interlocutors how to interpret an utterance (as serious vs joking, cooperative vs hostile, deferential vs blunt). Importantly, misunderstanding can occur even when words and grammar are shared, because interlocutors differ in the cues they treat as salient and in the inferences they draw from them (Gumperz, 1982; Tannen, 2005). These differences need not map neatly onto nationality-level categories; they may be organized by institution, community of practice, genre, generation, or interactional role.

2.3 Conversation analysis: sequential organization and normative accountability

Conversation analysis shows that talk is organized through systematic practices of turn-taking, sequence organization, and repair (Sacks et al., 1974; Schegloff, 2007; Schegloff et al., 1977). Participants treat these practices as normatively accountable: pauses, overlaps, interruptions, and repairs are not neutral timing artifacts but actions that can display respect, dominance, attentiveness, or dispreference. Cross-linguistic work suggests both universals and culturally patterned variation in turn-taking and timing (Stivers et al., 2009). Work on silence in intercultural interaction further illustrates that what counts as an appropriate gap is culturally learned (Nakane, 2007).

2.4 Pragmatics and politeness: culturally variable realizations of speech acts

Pragmatics distinguishes speech acts (requests, refusals, apologies) from their culturally variable realizations (Brown and Levinson, 1987). Large comparative traditions in cross-cultural pragmatics show that directness, mitigation, and sequencing differ systematically, and that prosody and timing are often central to how "politeness" and "stance" are enacted (Blum-Kulka et al., 1989). Thus, there is rarely a single correct realization; appropriateness is negotiated in context and admits plural acceptable forms.

Implication. Across these traditions, a shared conclusion follows: **many culture-bearing signals in spoken interaction are audible and sequential, and cannot be reliably recovered from transcripts alone.** Just as importantly, the relevant norms are local and event-situated rather than reducible to fixed demographic categories. For spoken agents, this means that evaluation should condition on the speech event, participant roles, and interaction contract under which conduct is being judged, while allowing plural acceptable realizations within those constraints. This is the theoretical basis for treating cultural competence in SpeechLMs as interactional competence.

3 Grounding the Agenda in Speech and Dialogue Research

A speech-first account of cultural competence does not begin from the premise that speech researchers have ignored prosody, timing, or interaction. Instead, it reframes what "good" prosody and "good" timing mean when the target is culturally and situationally appropriate conduct under an explicit interaction contract.

3.1 Prosody is a core modeling target - typically as expressiveness or control

TTS research has made major progress on representing and controlling speaking style and prosody, including unsupervised latent style representations such as Global Style Tokens (Wang et al., 2018) and other controllability interfaces. This matters because it establishes an engineering fact: modern systems increasingly have *degrees of freedom* to shape prosodic realization at inference time. The cultural competence question is how to connect these controls to norm-conditioned targets: what prosodic realizations count as appropriate for a given speech event, participant configuration, and interaction contract.

3.2 Turn-taking and timing are central: from incremental SDS to full-duplex agents

Turn-taking has long been recognized as a defining challenge for spoken agents, including end-of-turn detection, interruption handling, and response timing tradeoffs (Skantze, 2021). Recent work reframes turn-taking as predicting future joint voice activity (Voice Activity Projection), enabling turn-shift, overlap, and backchannel prediction without dense labels (Ekstedt and Skantze,

2022). Community surveys document rapid growth in datasets and methods and highlight remaining evaluation challenges (Castillo-López et al., 2025). With full-duplex spoken dialogue models, timing becomes a first-class capability boundary; Full-Duplex-Bench evaluates pause handling, backchanneling, turn-taking, and interruption management under scenario-driven setups (Lin et al., 2025).

3.3 SpeechLM stacks and their evaluation are expanding—but culture-bearing interaction remains underspecified

Surveys of SpeechLMs synthesize architectures, training recipes, and evaluation coverage (Cui et al., 2025b; Peng et al.; Arora et al., 2025; Wu et al., 2024). Recent benchmarks and suites increasingly evaluate more than recognition accuracy, including holistic capability categories, robustness, and fairness (Lee et al., 2025; Cui et al., 2025a). However, these evaluations still typically under-specify the normative context that makes interactional behavior culturally interpretable: a "snappy" response might be praised for low latency in one speech event but judged pushy or disrespectful in another, and a backchannel policy that increases acknowledgments might be supportive in one interactional style and interruptive in another.

The same point extends beyond assistant-style dialogue systems. In speech-to-speech translation, the relevant target may be preserving or adapting politeness, register, and pragmatic force without introducing new floor-management behavior. More broadly, across tutoring, customer support, scheduling, and clinical interfaces, interactional competence depends on which aspects of conduct the application should preserve, adapt, suppress, or make predictable.

Key reframing. Speech research already develops the *models* of prosody and turn-taking; cultural competence requires specifying the *normative conditioning* under which prosody and timing should be judged and optimized. Different applications may therefore have different targets: in some settings the appropriate realization may be affiliative and richly interactive, while in others it may be deliberately neutral, minimal, and tool-like.

4 Where Culture-Bearing Signals Are Lost in Today's SpeechLM Stacks

Modern spoken agents are implemented either as cascades (ASR→LLM→TTS) or as more end-to-end SpeechLMs; in practice, many deployments retain component-like interfaces even when models are integrated (Cui et al., 2025b; Peng et al.; Arora et al., 2025). These interfaces define what information is preserved, what is controllable, and what is treated as "content." As a result, they do not merely transmit interactional meaning; they also shape which forms of conduct a system can realize or suppress, under a given interaction contract.

4.1 Cascades: transcript-first bottlenecks

Cascaded pipelines create a structural bottleneck: ASR outputs a transcript; the LLM plans in text; TTS re-synthesizes speech with default or loosely controlled prosody. Prosody, overlap, hesitation, and voice quality are often discarded at the text interface and may be reintroduced only as generic "naturalness" or style. This encourages optimization for propositional correctness, while interactional competence becomes an emergent side effect rather than a target.

4.2 Component-level loci of interactional meaning (and failure)

Several loci repeatedly matter for culturally interpretable interaction:

- **Endpointing / VAD policies:** silence thresholds define what counts as "the end" of a turn, directly shaping how pauses, hesitation, and floor-yielding are treated.
- **Streaming latency:** response delays become interactional signals (e.g., engagement, deliberation, reluctance), not mere engineering noise.
- **Diarization and overlap handling:** overlap is common in human talk; errors collapse backchannels, interruptions, and multi-party participation frameworks.
- **Text intermediates:** transcript-like representations erase prosody, timing, and paralinguistics unless explicitly preserved.
- **Prosody control interfaces:** even when TTS can express prosody, the system must decide *which* prosody fits the speech event, local norms, and interaction contract.

Turn-taking research already highlights the fragility of silence-threshold policies and the centrality of interruptions and response delays (Skantze, 2021; Castillo-López et al., 2025), and full-duplex benchmarks make these behaviors measurable (Lin et al., 2025). Our contribution is to treat these design

choices as norm-encoding decisions that must be evaluated as culturally situated interactional conduct rather than as purely technical details.

5 A Theory-Derived Taxonomy of Culture-Bearing Signals in Spoken Interaction

This taxonomy synthesizes distinctions repeatedly drawn in linguistic anthropology, interactional sociolinguistics, conversation analysis, and pragmatics (Section 2). The aim is to describe *where cultural meaning is enacted* in speech and *how it becomes measurable* for evaluation and diagnosis.

5.1 Speech-event conditioning (situation and participation)

Before describing signal layers, we foreground a higher-order conditioning variable: the *speech event*. Following the ethnography of communication tradition, culturally appropriate talk depends on participants, roles, goals, setting, genre, and "key" (the socially recognized tone) (Duranti, 2009). For spoken agents, this implies that evaluation is ill-posed if it treats utterances as context-free: the same content can be appropriate or inappropriate depending on roles (e.g., peer vs clinician), institutional constraints, relationship distance, and the system's *interaction contract*. By interaction contract, we mean the application-specific expectations that define what kinds of conduct are desirable, optional, or disallowed in a speech event—for example, whether the system should be highly affiliative, minimally backchanneling, strictly neutral, or explicitly non-personified. This makes clear that appropriateness is not equivalent to generic human-likeness.

5.2 Layer 1: lexical and pragmatic resources (what is said)

Cultural meanings can be indexed by lexical/pragmatic choices: address terms, honorifics, hedges, forms of indirectness, and code-switching. Pragmatics and politeness theory emphasize that acts like refusing or apologizing vary in directness and mitigation (Brown and Levinson, 1987; Blum-Kulka et al., 1989). In speech systems, this layer often dominates evaluation because it is easiest to access via transcripts.

5.3 Layer 2: prosody (how stance and epistemics are displayed)

Prosody (intonation, rhythm, tempo, prominence) shapes stance, epistemic commitment, affect, and discourse structure. Interactional sociolinguistics treats prosody as a contextualization cue that frames interpretation (Gumperz, 1982). Cross-cultural conversational style work shows that differences in pacing and intonational patterns can yield divergent interpretations even with the same words (Tannen, 2005). Speech systems increasingly model prosody for expressiveness (Wang et al., 2018); the cultural competence question is how to evaluate prosody as norm-conditioned stance display rather than generic "nice-sounding speech."

5.4 Layer 3: paralinguistics and voice quality (alignment, persona, social indexing)

Laughter, sighs, breathiness, creakiness, and other voice-quality cues can index affect, irony, affiliation, or discomfort. These cues often interact with prosody and lexical content: a laugh particle can mitigate disagreement; a sigh can display reluctance; voice quality can index social identity and stance. In SpeechLMs, these cues are frequently altered by ASR normalization, codec compression, or TTS voice selection, which can inadvertently "flatten" culturally meaningful variation. They may also be intentionally constrained by the interaction contract when a system is expected to remain neutral, predictable, or minimally personified.

5.5 Layer 4: interactional organization (timing, turn-taking, repair, feedback)

Conversation analysis identifies the organization of turn-taking and repair as foundational to how actions are recognized and normatively evaluated (Sacks et al., 1974; Schegloff et al., 1977; Schegloff, 2007). Timing (gap length), overlap, interruption, backchannels, and repair strategies are culturally patterned and consequential (Stivers et al., 2009; Nakane, 2007). Speech and dialogue research explicitly models many of these phenomena (Skantze, 2021; Ekstedt and Skantze, 2022; Lin et al., 2025); our cultural framing emphasizes that "good" interactional organization is speech-event- and norm-conditioned.

5.6 Layer 5: social factors as conditioning variables

Hovy and Yang argue that social factors (e.g., speaker attributes, audience, setting) structure linguistic variation and should be modeled explicitly (Hovy and Yang, 2021). For SpeechLMs, treating social factors as conditioning variables helps avoid conflating culture with demographics: the same "community norm" can be associated with genres, institutions, and interactional roles rather than speaker identity alone. This category also clarifies what should *not* be inferred: cultural evaluation can be norm-conditioned without requiring demographic profiling.

6 Case Studies: When Transcript Correctness Fails to Predict Appropriateness

We highlight documented interactional phenomena where the "same words" can yield different social actions and cultural interpretations because the meaning-bearing cues are prosodic, sequential, or participatory. These cases illustrate why transcript correctness and generic naturalness are insufficient proxies for event-conditioned appropriateness.

6.1 Refusals and dispreferred responses

Conversation analysis notes that refusals are often *dispreferred* actions managed through delay, mitigation, and accounts; these patterns vary across settings and communities (Schegloff, 2007). Cross-cultural pragmatics shows systematic variation in refusal strategies, including indirectness and sequencing (Blum-Kulka et al., 1989; Brown and Levinson, 1987). In speech, delay (a pause before responding), hesitation, and prosodic softening can function as mitigation, even if the lexical content remains constant. A spoken agent that produces the correct refusal content but responds with an immediate, flat contour may be heard as abrupt or dismissive in events where delay indexes face sensitivity (Goffman, 1967). Conversely, in time-pressured or explicitly efficiency-oriented events, the same delay may be judged unnecessary or evasive. What counts as a "good" refusal is therefore not universal, but conditioned by the speech event and its interaction contract.

6.2 Backchannels

Backchannels (e.g., "mm-hm," "yeah," laughter particles) signal attention and alignment and have been studied since early descriptive work (Yngve, 1970). Their timing and frequency are interactional and culturally patterned: what counts as supportive engagement in one conversational style can be heard as interruptive in another. Full-duplex spoken agents make this salient because backchannels are no longer optional—they are expected in natural interaction and are now benchmarked explicitly (Lin et al., 2025). Cultural competence requires that backchannel policies be evaluated relative to the speech event (e.g., clinician-patient vs peer conversation) and local norms, not only relative to aggregate "human-likeness."

6.3 Turn-transition timing

Cross-linguistic work suggests that while turn-taking has universal structure, cultures vary in preferred gap length and in how silence is interpreted (Stivers et al., 2009; Nakane, 2007). In some events, rapid transition displays engagement; in others, a brief silence displays thoughtfulness or deference. Speech stacks impose response latencies through endpointing and streaming delays; without event-conditioned evaluation, systems may be rewarded for being "fast" even when speed produces culturally inappropriate conduct for the event.

6.4 Repair

Repair practices (self-correction, other-correction, indirect clarification) are systematically organized and normatively consequential (Schegloff et al., 1977). A spoken agent that corrects too directly, too quickly, or too publicly can threaten face in events where indirect repair is preferred. Conversely, excessive hedging can be inefficient in events where direct correction is valued. These are not merely "tone" issues; they are sequential practices shaping whether the interaction proceeds smoothly.

Takeaway. Across these cases, the central pattern is stable: **interactional outcomes depend on prosody, timing, and sequential organization, not only on transcript-level content.** For reproducible evaluation, these phenomena should be judged relative to explicit speech-event constraints rather than a single generic standard of "natural" or "human-like" behavior. The Appendix makes this operational with Speech-Event Cards, transcript-fixed minimal pairs, a worked example, and a stack-aware reporting checklist.

Taxonomic unit	Culture-bearing phenomena	Observable signals	Typical SpeechLM failure / bottleneck
Speech event / interaction contract (conditioning)	Roles, relationship distance, institutional genre, "key"/tone, goals, allowed/disallowed conduct	Metadata; scenario prompts; role labels; application constraints	Evaluation ill-posed if event variables or contract are implicit; norms default to evaluator's community or to generic human-likeness
Lexical/pragmatic	Honorifics, address forms, hedges, indirectness, code-switching	Transcript; lexical markers; discourse markers	Over-optimization for "polite English" defaults; cultural mismatch despite correct semantics
Prosody	Boundary tones, prominence, tempo, pitch range, hesitation contours	F0/energy contours; timing; prominence measures	Text interface discards cues; TTS controls not norm-conditioned; codec compresses stance cues
Paralinguistics / voice	Laughter, sighs, breathiness, creak; affect display; persona indexing	Non-lexical vocalizations; voice quality measures	ASR deletes non-lexicals; TTS normalizes; voice selection collapses meaningful variation or introduces unwanted persona cues
Interactional organization	Turn timing, overlap, backchannels, repair, interruption management	Gap/overlap distributions; VAP-like activity; feedback placement	Endpointing/latency artifacts; diarization collapse; full-duplex overlap mismanagement
Social factors (conditioning)	Audience design; register; style-shifting; community-of-practice norms	Speaker role/audience cues; register features	Conflating culture with demographics; missing situational conditioning

Table 1: A theory-derived taxonomy of culture-bearing signals in spoken interaction.

7 Evaluation Landscape and the Case for Norm-Conditioned Cultural Evaluation

Speech evaluation has historically emphasized intelligibility and perceptual quality (e.g., WER, MOS). SpeechLM evaluation is broadening to include holistic capability suites and speech-native QA benchmarks (Lee et al., 2025; Cui et al., 2025a). Bias-oriented benchmarks further emphasize that audio introduces additional axes (e.g., acoustic vs. content) absent in text-only evaluation (Choi et al., 2025; Wu et al., 2025; Koenecke et al., 2020).

7.1 What existing metrics capture well

WER captures transcript accuracy; **MOS** has historically served as a broad measure of perceived naturalness or perceptual quality; holistic suites like **AHELM** aggregate tasks across perception, reasoning, fairness, safety, and robustness (Lee et al., 2025); speech-native QA benchmarks like **VoxEval** test knowledge and reasoning in speech-in/speech-out interaction (Cui et al., 2025a). These tools are valuable and should remain central. At the same time, especially for modern high-quality systems, MOS is often too coarse and under-diagnostic to localize interactional failures, even before cultural appropriateness is considered.

7.2 What remains under-instrumented for cultural competence

For interactional cultural competence, the missing piece is often *normative situatedness*: specifying the speech event and evaluating prosody, timing, and participation as appropriate conduct under that event's norms. Recent cultural NLP work argues that "culture" must be theory-grounded and that evaluation should be intentionally cultural rather than implicitly normed by annotators' defaults (Zhou et al., 2025; Oh et al., 2025; AlKhamissi et al., 2025). Culture surveys further document that "culture-aware NLP" often collapses into either multilinguality or culture-as-facts unless evaluation targets are made explicit (Pawar et al., 2025; Liu et al., 2025).

For SpeechLMs, an analogous risk is that prosody and turn-taking are optimized as generic expressiveness or low-latency fluency, while the evaluative norms remain implicit. Bias benchmarks already show that audio introduces distinct channels (content vs. acoustics) that can differentially drive outcomes (Choi et al., 2025; Wu et al., 2025). Interactional cultural competence extends this logic beyond bias: it asks whether timing, prosody, and participation are appropriate for an explicit speech event and interaction contract.

Making the framing usable. The Appendix provides a Speech-Event Card template, a plural-acceptability protocol, transcript-fixed minimal pairs, and a reporting checklist that specifies what to log. These artifacts support reproducible comparisons across systems while making normative assumptions inspectable and revisable.

7.3 How this differs from generic "turn-taking quality" and expressive TTS

Prior speech evaluation often treats timing and prosody as either (i) *generic* interaction quality

Common target	What it measures well	What it can miss for cultural competence	Speech-first extension
WER / ASR accuracy	Lexical correctness; recognition robustness	Prosody/voice/overlap deleted; hesitation and non-lexicals erased	Add interaction tags (overlap, backchannels), preserve timing features; evaluate event-conditioned turn segmentation
MOS / naturalness	Global perceptual quality of synthesis	"Natural" $\neq$ norm-appropriate; default conversational style becomes implicit norm	Add event-conditioned appropriateness ratings; compare plural acceptable realizations within event
Speech QA / reasoning (e.g., VoxEval)	Speech-native understanding and knowledge in speech-in/out	Content correctness can ignore stance/timing; interactional breakdown not measured	Add probes holding transcript fixed while varying prosody/timing; measure sequential fit and repair
Holistic suites (e.g., AHELM)	Broad capability coverage; includes fairness/safety/robustness	Dialogue categories may not isolate interactional norms; event norms implicit	Add explicit speech-event schema; evaluate participation management (backchannels, overlap, interruption) by event
Bias benchmarks (e.g., VoiceBBQ; spoken bias evals)	Diagnose content vs acoustic bias in SLMs/SDMs	Often focuses on stereotypes, not broader interactional appropriateness	Use the same "two-channel" insight for culture-bearing interaction: content vs prosody/timing/participation
Turn-taking benchmarks (e.g., Full-Duplex-Bench)	Measures pause handling, interruptions, backchannel behavior	Optimizes "human-like" timing without specifying which norms apply	Condition scenarios on speech-event variables; allow plural normative timing regimes

Table 2: Strengths of current speech evaluation and extensions needed for norm-conditioned interactional evaluation.

(e.g., "good" endpointing / low-latency turn-taking) or (ii) *expressiveness* and "naturalness" in TTS. Our claim is not that these lines of work are wrong, but that they often leave the normative target implicit: *appropriate for whom, in which speech event, under which expected conduct?* Norm-conditioned cultural evaluation makes that target explicit, so the same acoustic behavior can be judged differently across events (and can be acceptable in multiple ways within an event), and so failures can be attributed to specific pipeline choices rather than a vague notion of "naturalness."

7.4 A minimal evaluation framing (overview)

We propose an evaluation framing that is compatible with existing practice while making norms explicit: (1) **Specify the speech event** (roles, relationship, setting, goal, genre, key, interaction contract); (2) **Evaluate interactional conduct** (sequential fit, stance display, participation management, repair); (3) **Allow plural acceptability** within an event, rather than forcing a single "gold" realization; (4) **Diagnose stack bottlenecks** (endpointing, diarization, latency, text interfaces, prosody controls) as part of evaluation, since they encode norms by design. A compact illustration is the repair turn "Could you repeat that?" In a clinic-intake event, a softened self-repair framing with a slightly longer gap may better fit an anxious-client interaction; in a time-pressured service encounter, a faster, more direct clarification may be preferable. WER may treat both identically, and MOS may rate both similarly natural, while event-conditioned appropriateness distinguishes which better fits the interaction.

Scalability and subjectivity. "Event cards" and plural norms need not imply an unbounded catalog of cultures. A practical path is to start with a small, well-scoped set of high-frequency speech events (e.g., customer support, scheduling, tutoring), define norms at the level of roles and interactional contingencies, and report both mean judgments and disagreement. Disagreement is signal: it can reflect plural acceptable realizations or mismatched event framing. In practice, this suggests a tiered approach: structured event/stack reporting for all systems, plus targeted human judgments on event-stratified subsets where interactional appropriateness matters. Where feasible, pair human ratings with diagnostics that localize deviation sources (latency, endpointing, prosody control), so evaluation remains actionable rather than merely subjective. The Appendix provides concrete schemas, rubrics, minimal-pair probes, and a worked example.

8 Implications and Research Directions

Framing cultural competence as interactional competence suggests concrete directions that build on ongoing speech research trends.

8.1 Event-conditioned interaction modeling and evaluation

Turn-taking and prosody modeling can be made norm-conditioned by incorporating speech-event

and interaction-contract variables and by reporting performance stratified by event type. This aligns with capability-aware evaluation for audio-language models while making the normative context explicit.

8.2 Interfaces that preserve culture-bearing structure

Because many culture-bearing signals are lost at interfaces, speech systems should treat timing, prosody, and participation as first-class representations: explicit overlap markers; backchannel channels; prosody tokens; timing features; and controllable synthesis parameters tied to event schemas and interaction contracts rather than generic style prompts. The goal is not maximal expressiveness, but controllable, norm-conditioned conduct.

8.3 From bias-only to broader cultural interactional diagnostics

Bias work in speech shows audio adds channels that influence outcomes (acoustic vs. content) (Choi et al., 2025; Wu et al., 2025; Koenecke et al., 2020). A broader cultural agenda generalizes the same insight: many culturally meaningful interactional outcomes are driven by *how* speech unfolds (timing, stance, participation), not only by *what* is said. Diagnosing *where* these cues are lost in stacks (Section 4) becomes part of cultural evaluation, not just engineering debugging.

8.4 Localization as norm alignment, not demographic inference

Recent cultural NLP work argues against treating culture as trivia and emphasizes theory-grounded constructs and interpretive evaluation (Zhou et al., 2025; AlKhamissi et al., 2025). For SpeechLMs, this supports a localization framing: align interactional norms to events and communities of practice, rather than inferring sensitive demographic attributes. This is compatible with socially aware NLP's focus on social factors as modeling dimensions (Hovy and Yang, 2021; Yang et al., 2025). It also suggests distinguishing event norms from individual preferences: personalization may be useful in some applications, but it is not a prerequisite for norm-conditioned evaluation.

9 Relation to Cultural NLP and Socially Aware NLP

Culture-aware NLP surveys show that much prior work focuses on text and often operationalizes culture as knowledge, preferences, or annotation differences, with growing emphasis on theory-grounded definitions and evaluation design (Pawar et al., 2025; Liu et al., 2025). Position papers argue for sociocultural theory and intentionally cultural evaluation (Zhou et al., 2025; Oh et al., 2025), and anthropological critiques emphasize that benchmarks can hard-code narrow cultural assumptions if norms remain implicit (AlKhamissi et al., 2025). Socially aware NLP emphasizes that language use is shaped by social factors and that modeling should account for these dimensions (Hovy and Yang, 2021; Yang et al., 2025). Our contribution is not to newly claim that culture matters for language technology, but to make these commitments concrete for speech. For spoken agents, the primary culture-bearing phenomena are often prosodic, sequential, and participatory; modern SpeechLM stacks have identifiable interfaces where these signals are discarded or standardized; and evaluation must therefore specify speech events and interaction contracts, and allow plural acceptable realizations rather than treating "naturalness" as a proxy for appropriateness.

10 Conclusion

SpeechLMs are increasingly judged not only on correctness, but on how they behave as conversational participants. Social-science accounts of interaction show that culture is enacted through prosody, timing, participation, and sequential organization; these cues are often audible and norm-conditioned, and transcripts can be insufficient proxies. We suggest that cultural competence in spoken agents is productively framed as interactional competence: norm-conditioned participation in explicit speech events with plural acceptable realizations. Crucially, the relevant target is not generic human-likeness. In many deployments, the appropriate spoken behavior may be intentionally constrained, neutral, predictable, and explicitly tool-like; what matters is that the interaction contract be made explicit and evaluated accordingly. Grounded in today's speech modeling practice (prosody control, turn-taking, full-duplex dialogue) and in the realities of SpeechLM stacks (endpointing, diarization, streaming latency, text interfaces), this agenda motivates evaluation protocols that complement WER, MOS, and broad capability suites by making norms explicit and diagnosing where interactional cues are lost.

Limitations

This paper is conceptual and synthetic: it proposes a framing, taxonomy, and evaluation agenda rather than introducing a new dataset or reporting a full empirical study. While the Appendix includes reusable event cards, probe templates, rubrics, and a worked example intended to support pilot studies, we do not claim validated measurement instruments in this manuscript.

Appropriateness is norm-conditioned by speech events and local participation expectations, but norms are often tacit, contested, and heterogeneous even within a community. Our event schema can improve reproducibility by making assumptions explicit, yet it cannot eliminate interpretive variation or disagreement. Moreover, interaction commonly supports multiple legitimate styles; emphasizing plural acceptability improves realism but complicates standardization.

The taxonomy has intrinsic boundary ambiguity because prosody, paralinguistics, and interactional timing tightly interact (e.g., filled pauses can simultaneously manage turn-holding and stance; silence can mark both prosodic boundaries and dispreference). We treat the taxonomy as a functional guide for measurement targets rather than mutually exclusive bins; overlaps are expected.

Generalization across languages, genres, and modalities is limited. Many examples draw from well-studied languages and interactional settings. While turn-taking and repair have broad cross-linguistic relevance (Stivers et al., 2009), specific norms vary across languages, communities, and speech events. We also focus on speech (audio) because SpeechLMs foreground it, but gesture, gaze, and posture can further condition interactional meaning.

We also note a practical boundary issue: culture, identity perception, and bias can intertwine in real interaction, and norm-conditioning could be misused for profiling if framed in demographic terms. We therefore advocate event- and role-based specifications rather than demographic inference, but acknowledge residual risks in sensitive applications.

Finally, real-time deployments introduce confounds. Streaming latency, endpointing policies, prosody-control interfaces, and diarization errors can create interactional artifacts, making it hard to separate "model behavior" from "system behavior." We argue evaluations should document system settings and, where possible, pair human judgments with diagnostics that attribute deviations to specific pipeline choices. Norm-conditioned evaluation also increases human-judgment burden: recruiting raters with relevant norm knowledge is costly, and deciding whose norms are represented remains a substantive methodological choice.

Ethics Statement

This manuscript proposes a conceptual framing, taxonomy, and evaluation agenda; it does not collect new data, recruit participants, or run human-subject studies. The primary ethical risks therefore arise from how the agenda could be implemented in future work.

Local norms are not a license to reproduce harmful behavior or unsafe content. In deployment and evaluation, safety requirements and platform policies should take precedence when they conflict with situated interactional expectations, and reports should document how such constraints shape what counts as "appropriate" within an event.

Norm-conditioned cultural evaluation can be misapplied to justify stereotyping or to treat a single conversational style as "correct." We therefore argue for plural acceptability and for conditioning on speech events and participation roles rather than inferring sensitive demographic attributes. More generally, richer modeling of prosody, voice quality, and interactional conduct can increase the risk of profiling or unwanted identity inference if used for surveillance or discriminatory personalization. We recommend that implementations avoid demographic classification as a prerequisite for cultural evaluation and instead rely on explicit, task-appropriate event specifications and opt-in localization settings. Finally, benchmarking choices can encode evaluator defaults; to mitigate this, our framing emphasizes making event assumptions explicit and documenting rater instructions and norm sources, consistent with broader critiques of culture benchmarks that warn against implicit normative baselines (AlKhamissi et al., 2025; Zhou et al., 2025; Oh et al., 2025).

Use of AI During the writing process, the authors used ChatGPT to refine the content of the manuscript. Prior to submission, a comprehensive review was carried out and appropriate revisions were made, with the authors assuming complete responsibility for the published content.

References

Mai AlKhamissi, Yunze Xiao, Badr AlKhamissi, and Mona Diab. 2025. Hire your anthropologist! rethinking culture benchmarks through an anthropological lens. *arXiv preprint arXiv:2510.05931*.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.

Shoshana Blum-Kulka, Juliane House, and Gabriele Kasper. 1989. Cross-cultural pragmatics: Requests and apologies. *(No Title)*.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Galo Castillo-López, Gaël de Chalendar, and Nasredine Semmar. 2025. A survey of recent advances on turn-taking modeling in spoken dialogue systems. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 254–271.

Junhyuk Choi, Ro-hoon Oh, Jihwan Seol, and Bugeun Kim. 2025. Voicebbq: Investigating effect of content and acoustics in social bias of spoken language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28713–28724.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. 2025a. Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. *arXiv preprint arXiv:2501.04962*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y Guo, and Irwin King. 2025b. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970.

Alessandro Duranti. 2009. Linguistic anthropology: History, ideas, and issues. *Linguistic anthropology: A reader*, pages 1–60.

Erik Ekstedt and Gabriel Skantze. 2022. Voice activity projection: Self-supervised learning of turn-taking events. *arXiv preprint arXiv:2205.09812*.

Erving Goffman. 1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. Pantheon Books.

John J Gumperz. 1982. *Discourse strategies*. 1. Cambridge University Press.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 588–602.

Dell Hymes. 1974. *Foundations in Sociolinguistics: An Ethnographic Approach*. University of Pennsylvania Press.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.

Tony Lee, Haoqin Tu, Chi Heem Wong, Zijun Wang, Siwei Yang, Yifan Mai, Yuyin Zhou, Cihang Xie, and Percy Liang. 2025. Ahelm: A holistic evaluation of audio-language models. *arXiv preprint arXiv:2508.21376*.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. 2025. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *arXiv preprint arXiv:2503.04721*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Ikuko Nakane. 2007. Silence in intercultural communication.

Juhyun Oh, Inha Cha, Michael Saxon, Hyunseung Lim, Shaily Bhatt, and Alice Haeyun Oh. 2025. Culture is everywhere: A call for intentionally cultural evaluation. In *The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*. Association for Computational Linguistics (ACL).

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, and 1 others. A survey on speech large language models for understanding. *Authorea Preprints*.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Emanuel A Schegloff. 2007. *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge university press.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & communication*, 23(3-4):193–229.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178.

Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, and 1 others. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Deborah Tannen. 2005. *Conversational style: Analyzing talk among friends*. Oxford University Press.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, pages 5180–5189. PMLR.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai-wei Chang, Ho-Lam Chung, Alexander H Liu, and Hung-yi Lee. 2024. Towards audio language modeling–an overview. *arXiv preprint arXiv:2402.13236*.

Yihao Wu, Tianrui Wang, Yizhou Peng, Yi-Wen Chao, Xuyi Zhuang, Xinsheng Wang, Shunshun Yin, and Ziyang Ma. 2025. Evaluating bias in spoken dialogue llms for real-world decisions and recommendations. *arXiv preprint arXiv:2510.02352*.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2025. Socially aware language technologies: Perspectives and practices. *Computational Linguistics*, 51(2):689–703.

Victor H Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, pages 567–578.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

Appendix: Practical Materials for Event-Conditioned Interactional Evaluation

This appendix collects reusable artifacts that make event-conditioned, speech-native evaluation reproducible: (i) a Speech-Event Card template (§A), (ii) a rubric with anchors and a plural-acceptability protocol (§B), (iii) a transcript-fixed minimal-pair library (§C), and (iv) a stack-aware reporting checklist (§D). Where helpful, we also provide a minimal pilot protocol (§E) and a worked example (§F).

Speech-Event Card (Template)

Event name: (e.g., "clinic intake," "advisor meeting," "peer tutoring," "customer support call")
Setting / institutionality: (informal / semi-formal / institutional; constraints, stakes)
Participants & roles: (A: role; B: role; role obligations; entitlements)
Power / status relation: (symmetric / hierarchical; how authority is displayed)
Relationship distance: (strangers / acquaintances / close; history)
Goal / activity type: (inform, request, refuse, negotiate, troubleshoot, comfort, advise)
Genre: (service encounter, interview, tutoring, narrative sharing, deliberation)
Interactional "key": (warm, neutral, authoritative, apologetic, playful, urgent)
Local constraints (norm targets): (e.g., "allow reflective silence," "avoid overlap," "use mitigated refusals")
Plural acceptability note: (list 2–3 acceptable styles/variants within this event)
Disallowed failure patterns: (e.g., "interrupting the client," "overly intimate terms," "performative cheerfulness")
Rater competence assumption: (who is a qualified rater for this event and why)

Figure 2: Reusable Speech-Event Card to surface normative assumptions and condition evaluation.

A Speech-Event Card Template and Schema

Event conditioning is a design requirement: without an explicit interactional situation, "appropriateness" is ill-defined. The Speech-Event Card below makes normative assumptions explicit and supports reproducible judgments.

How to use. (1) Create 3–8 event cards spanning settings and role relations you care about. (2) Provide the card to raters (and optionally to the model/system policy). (3) Evaluate the same transcript under different event cards to expose event dependence.

Compact schema (for benchmarks / rater instructions). If you want a lightweight, checklist-style operationalization of the event card:

Field	Description (examples)
Setting / genre	Institutional (clinic, customer support), semi-institutional (classroom tutoring), informal (friends)
Roles & power	User role vs. agent role; power asymmetry (expert–novice, service provider–customer)
Relationship distance	Stranger, acquaintance, recurring relationship; degree of familiarity
Goal	Solve a task, provide advice, coordinate action, provide emotional support
Key / tone expectation	Neutral-professional, warm-supportive, formal-deferential, playful
Turn-taking norm	Overlap tolerance; expected gap length; barge-in expectations
Politeness/face norm	Directness expectations; preferred mitigation for refusals/repairs
Constraints	Safety constraints; time pressure; noise; channel constraints (phone vs. in-person)

Table 3: Event card schema. The goal is not exhaustive sociological modeling; it is to make the normative context explicit enough for reproducible evaluation.

B Rubric for Event-Conditioned Interactional Appropriateness

Use this rubric for human judgments (researcher annotation, expert panels, or community-competent raters), or as a structured guide for qualitative analysis. The goal is not a single gold output: it is to score appropriateness *relative to the Speech-Event Card* while allowing plural acceptable realizations.

B.1 Plural acceptability protocol (practical)

Plurality is expected in cultural interaction. To operationalize it:

- **Acceptability vs. preference:** ask raters to mark whether an issue is (i) an event-norm violation or (ii) a personal preference among acceptable variants.
- **Record alternatives:** for any response rated acceptable, ask raters to write (or choose) at least one alternative acceptable realization (e.g., different contour, different backchannel timing).
- **Treat disagreement as diagnostic:** if qualified raters split, interpret it as (a) plural norms or (b) underspecified event constraints. Revise the event card rather than forcing consensus.

B.2 Optional scoring formalisms (if you want numbers)

If you need a summary score while respecting plurality, two compatible options are:

- **Constraint violation rate + preference distribution:** define a small set of event constraints (hard failures) and report their violation rate; separately report pairwise preferences among acceptable variants.
- **Pairwise "fits event better" + Bradley–Terry:** collect pairwise judgments across system outputs within each event and fit a Bradley–Terry model for rankings; report uncertainty and event-conditioned differences.

C Transcript-Fixed Minimal-Pair Library

Each template defines (i) a fixed transcript, (ii) a controlled manipulation of prosody/timing/participation, and (iii) an expected interactional effect that depends on the Speech-Event Card. These can be instantiated via TTS/prosody editing, full-duplex policy edits, or curated from corpora.

C.1 A. Refusal / disagreement (mitigation via timing and prosody)

1. **Delay mitigation vs immediate refusal**
   Transcript: "I don't think I can do that."
   A: immediate onset; flat contour; strong prominence on *can't*.
   B: 300–600ms delay; softened prominence; hesitation; lower pitch range.
   Hypothesis: A reads blunt; B reads considerate in face-sensitive events.

2. **Tentative stance via rising contour**
   Transcript: "That might be difficult."
   A: falling terminal; high prominence (certainty).
   B: rising or level terminal; reduced prominence (tentativeness).
   Hypothesis: B better fits events requiring deference or hedging.

3. **Disagreement as affiliation vs confrontation**
   Transcript: "I'm not sure I agree."
   A: short gap; clipped delivery; narrow pitch range.
   B: supportive preface backchannel ("mm"), warm pitch movement, slower tempo.

Dimension	1 (Poor)	3 (Adequate)	5 (Excellent)	Notes / cues
Event Fit	Violates event constraints (tone/role mismatch; wrong formality).	Generally fits event; minor mismatches.	Clearly aligned with setting, roles, and key; feels "right for this situation."	Use Speech-Event Card. Penalize role/power and tone mismatches.
Stance & Face	Prosody/timing conveys wrong stance (too blunt, too certain, too cold).	Stance mostly reasonable; occasional mis-signals.	Prosody and wording jointly convey appropriate stance and facework.	Attend to mitigation, prominence, pitch range, hedges, apology/refusal delivery.
Interactional Flow	Turn timing/backchannels/overlap disrupt flow (interrupts, awkward gaps).	Flow is workable; occasional timing issues.	Smooth turn-taking with event-appropriate backchannels and timing.	Include latency, overlap handling, and backchannel placement.
Repair Conduct	Handles trouble poorly (ignores confusion; wrong repair type; escalates).	Repair works but is suboptimal for event.	Repair strategy is event-appropriate and maintains face.	Score only when repair is triggered. Track self-repair vs other-repair framing.
Plural Acceptability	One rigid style; feels norm-imposing.	Allows some variation.	Supports multiple acceptable realizations within event constraints.	Record alternative acceptable variants; don't treat disagreement as noise by default.

Table 4: Rubric for event-conditioned interactional appropriateness with anchors admitting plural acceptability.

Hypothesis: B preserves affiliation; A may sound adversarial.

C.2 B. Apologies and responsibility (prosodic sincerity markers)

1. **Apology warmth via voice quality and tempo**
   Transcript: "I'm sorry about that."
   A: fast tempo; creaky/flat affect; immediate transition to solution.
   B: slightly slower; softened voice quality; micro-pause before "sorry."
   Hypothesis: B reads sincere/supportive in consoling events; A reads perfunctory.

2. **Apology escalation vs minimization**
   Transcript: "Sorry, I misunderstood."
   A: low prominence on "misunderstood"; quick recovery.
   B: increased prominence; explicit confirmation question after apology.
   Hypothesis: B fits high-stakes events; A may fit low-stakes casual talk.

C.3 C. Backchannels (timing, density, and overlap tolerance)

1. **Dense vs sparse backchannels**
   Transcript (speaker A): long explanation (held fixed).
   Agent behavior: backchannels inserted every 1.5s vs every 4s.
   Hypothesis: Dense may be supportive in some styles, interruptive in others.

2. **Backchannel placement relative to completion points**
   Agent backchannel: placed mid-intonational phrase vs at phrase boundary.
   Hypothesis: Mid-phrase can be heard as interruption in many events.

3. **Acknowledgment token type with constant meaning**
   Transcript token: "yeah" vs "mm-hm" (or language-appropriate equivalents).
   Hold timing constant; vary token.
   Hypothesis: Token choice indexes stance/commitment; may shift perceived alignment.

C.4 D. Turn transitions and latency (gap length as meaning)

1. **Fast vs deliberative response onset**
   Transcript: "That's a good question."
   A: onset $<$200ms after user question.
   B: onset 600–900ms with audible thinking marker (breath/"uh").
   Hypothesis: B can index deliberation/respect; A can index eagerness or pushiness depending on event.

2. **Silence tolerance in sensitive events**
   Transcript: "I can help with that."
   Manipulation: insert 1.2s silence before reply vs 0.2s silence.
   Hypothesis: Long silence may be respectful in consoling events, awkward in service encounters.

C.5 E. Overlap and interruption (full-duplex conduct)

1. **Barge-in policy: allow user interruption vs block interruption**
   Scenario: user begins speaking mid-agent sentence.
   A: agent stops immediately and yields.
   B: agent continues, then responds.
   Hypothesis: Yielding can signal respect; continuing can signal dominance, event-dependent.

2. **Collaborative overlap vs competitive overlap**
   Transcript: agent produces "right" overlap during user's story.
   Manipulation: overlap at narrative climax vs overlap during informational content.
   Hypothesis: Overlap timing changes whether it reads as collaborative enthusiasm.

C.6 F. Repair trajectories (strategy selection)

1. **Direct clarification vs hedged confirmation**
   Transcript (repair): "Did you mean X?" vs "Just to confirm, did you mean X?"
   Manipulation: add hedge, apology, or rationale while holding propositional content.
   Hypothesis: Hedged repair fits face-sensitive events; direct fits time-pressured service.

2. **Self-repair vs other-repair framing**
   Transcript: "I may have misheard—could you repeat that?" (self-repair)
   vs "You said X, right?" (other-repair leaning).
   Hypothesis: Self-repair preserves face; other-repair may sound accusatory in some events.

C.7 G. Prosodic stance minimal pairs (same words, different action)

1. **Sarcasm/irony vs sincerity**
   Transcript: "That's great."
   A: increased pitch range, positive contour.
   B: flat contour, delayed onset, creaky voice quality.
   Hypothesis: B can index irony; event card should specify whether irony is acceptable.

2. **Authority vs solidarity**
   Transcript: "Let's do it this way."
   A: strong prominence, falling terminal (authoritative).
   B: softened prominence, inclusive tone, slight rise (collaborative).
   Hypothesis: A fits hierarchical roles; B fits peer roles.

C.8 H. Acoustic identity control (content held fixed; voice varied)

1. **Same transcript, different voice/acoustic realization**
   Transcript: hold words constant for an advice/refusal/repair turn.
   Manipulation: synthesize/realize with different accents/voices/speaking rates while keeping timing/prosody policy constant (or, alternately, keep voice constant while varying timing/prosody).
   Goal: separate judgments driven by interactional conduct (timing/prosody/repair) from judgments driven by acoustic presentation; report event-conditioned effects rather than collapsing into a single "naturalness" score.

D Suggested Reporting Checklist for Stack-Aware Evaluation

To support reproducibility and interpretability, report the following alongside evaluation results:

- **Interface:** cascaded ASR→LLM→TTS vs end-to-end speech-to-speech.
- **Endpointing / VAD:** thresholds, hangover time, barge-in enabled/disabled.
- **Streaming:** average response onset latency; buffering policy; incremental decoding on/off.
- **Overlap handling:** overlap detection, diarization method, whether backchannels vs interruptions are distinguished.
- **ASR normalization (if cascaded):** disfluency handling; punctuation/segmentation; whether laughter/breath/particles are retained or deleted.
- **Prosody control:** what controls exist (tokens, reference, pitch/energy sketches) and which are used during evaluation.
- **Audio representation (if end-to-end):** codec/tokenizer type, bitrate/quantization, whether paralinguistics are preserved.
- **Event conditioning:** how event cards are provided (prompt, system policy, fine-tuning labels).

- **Plural scoring protocol:** whether multiple acceptable realizations are allowed and how disagreement is recorded.

E Minimal Pilot Protocol

If you choose to include a small empirical "proof-of-need" study, the following protocol is intentionally lightweight:

- **Events:** 4–6 Speech-Event Cards spanning role relations and settings (e.g., peer tutoring; clinic intake; customer support; advisor meeting).
- **Items:** 20–30 base transcripts per event (requests, refusals, apologies, clarifications).
- **Manipulations:** 2–3 minimal-pair manipulations per transcript (latency, prosody contour, backchannel placement, overlap policy).
- **Raters:** define eligibility per event (experience with setting; language variety; role familiarity).
- **Outcomes:** rubric scores + pairwise "fits event better" judgments; report disagreement patterns as findings.
- **Release:** event cards, audio stimuli, rubric instructions, and stack reporting checklist.

F Worked Example

Same transcript, different event cards. *Transcript:* "Could you repeat that?"

Event Card 1: Clinic intake (hierarchical, anxious client). Norm targets: calm pacing; soften directives; allow reflective silence. Expected: hedged self-repair framing ("Sorry—I may have misheard...") + gentle prosody.

Event Card 2: Time-pressured service encounter (queue behind). Norm targets: efficiency; minimal mitigation; quick repair. Expected: direct clarification with short latency and neutral contour.

Interpretation. Under Card 1, a fast/direct contour can read brusque; under Card 2, long hedges can read inefficient. This illustrates why "good repair" is not universal: it is event-conditioned.