

From Output Metrics to Cognitive Process: Disfluency and Lexical Entrainment for Evaluating Real-Time AI-Mediated Communication

Wanqing Psyche He and Susan R. Fussell

Department of Information Science, Cornell University

Ithaca, NY, USA

{wh385, sfussell}@cornell.edu

Abstract

Real-time communication mediated by artificial intelligence (AI) can produce an accurate output while changing how a speaker plans an utterance and what a listener receives. Output measures, such as word error rate, machine translation scores, and task accuracy, leave these process effects unobserved. We propose a dual-axis framework that aligns disfluency measures with suggestion correspondence, including lexical entrainment and semantic similarity. Disfluency type, rate, timing, and duration characterize observable production difficulty. Correspondence measures characterize whether suggestion content appears in the response. Observing both axes within the same episode distinguishes behavioral configurations that either family alone leaves ambiguous and supports explicit hypotheses about planning, uptake, and resistance. An illustrative corpus demonstrates event-detection feasibility and shows that semantic correspondence varies with coded incorporation and exceeds a comparison matched on presentation format. These checks leave cognitive interpretations open to experimental validation. We also derive a prediction concerning listeners: removing natural disfluencies may remove cues that listeners use to anticipate difficult or novel material. The resulting framework and reporting protocol extend standard evaluation to the human processes through which real-time AI-mediated communication succeeds or fails.

1 Introduction

Natural language processing (NLP) evaluation of spoken dialogue systems has long asked whether a system produced an accurate or useful output. Word error rate (WER) evaluates automatic speech recognition (ASR), Bilingual Evaluation Understudy (BLEU) and chrF evaluate machine translation (MT), and task completion evaluates goal-oriented dialogue (Papineni et al., 2002; Post, 2018). These measures have driven progress in

speech recognition, spoken language understanding, and end-to-end (E2E) dialogue modeling (Shriberg, 1994; Jurafsky and Martin, 2024). Their unit of analysis becomes incomplete when a system intervenes during communication itself. Real-time AI-mediated communication (AI-MC) now supplies suggestions, lexical clarifications, and spoken completions while a person is planning and producing speech (Duan et al., 2021; He and Fussell, 2026; Qin et al., 2025; Fang et al., 2025; Xia et al., 2023). Two systems can consequently receive the same output score while differing in planning disruption, response reuse, or preservation of cues that support a listener. Evaluation that overlooks these differences can favor a system that improves the observable product while degrading the interaction through which it was produced.

Speech provides evidence about these otherwise hidden process differences. Speakers produce approximately 2–3 words per second while coordinating lexical selection, phonological encoding, prosodic planning, and articulation (Levelt, 1989). Filled pauses, repetitions, and restarts occur in roughly 6–10% of spoken words in naturalistic conversation (Bortfeld et al., 2001; Fox Tree, 1995), and their distributions vary with planning, retrieval, monitoring, and interaction management (Goldman-Eisler, 1968; Shriberg, 1994; Clark and Tree, 2002; Corley and Stewart, 2008). An event still has several possible sources because its interpretation depends on the speaker, task, conversational position, and linguistic context. Its type and timing can localize observable difficulty, while correspondence between the AI suggestion and the response establishes whether related content appeared. The two forms of evidence answer complementary questions about the same episode.

We therefore advance a framework with two observable axes: *speech measures* and *suggestion correspondence*. Lexical entrainment provides the established word-choice measure within the second

axis, while semantic and human-coded measures capture correspondence beyond shared wording. Aligning the axes yields four behavioral profiles that organize competing explanations without assigning a unique cognitive state. A reporting protocol then specifies the anchors, comparisons, and missing-data distinctions needed to test those explanations. The same process perspective produces an additional prediction for listeners, who may lose useful anticipatory cues when AI systems smooth disfluent speech (Arnold et al., 2003; Fox Tree, 2001; Corley et al., 2007). Our contribution is a way to formulate and evaluate these process questions alongside output quality. We begin by establishing what speech-production theory permits the measures to mean.

2 Theoretical Basis for Process Measures

Using disfluency as a process measure requires a theory that constrains its possible sources. Levelt’s 1989 model organizes speech production into three incrementally operating stages: the *conceptualizer* generates a preverbal communicative intent, the *formulator* translates this intent into a linguistic plan, and the *articulator* executes the motor program as speech (Indefrey and Levelt, 2004; Levelt et al., 1999). A *perceptual loop monitor* compares inner speech and auditory feedback with intent, and detected mismatches can interrupt an ongoing plan (Roelofs, 2020; Nozari et al., 2011; Hartsuiker and Kolk, 2001).

Predictive processing accounts add that speakers and listeners anticipate utterances at multiple linguistic levels (Pickering and Garrod, 2013; Haagoort, 2019). Together, these accounts explain why external content arriving during production may affect planning, retrieval, monitoring, or management of the interaction. They also set an inferential boundary: event type and timing narrow candidate processes without identifying a unique stage.

2.1 Relation to NLP Research on Disfluency

NLP has usually treated disfluency as material to detect and remove. Work on the Switchboard corpus established taxonomies of interruption points, reparanda, and repairs (Shriberg, 1994), and subsequent sequence labelers, pretrained transformers, and E2E ASR rescorers filtered these events before parsing or MT (Wang et al., 2020; Jamshid Lou and Johnson, 2020; Rocholl et al., 2021). This approach supports downstream processing while discarding

evidence about how that output was produced.

A separate literature shows that AI can change human language. Smart-reply and predictive-text systems narrow lexical diversity and shift register in text-based communication (Arnold et al., 2020; Mieczkowski et al., 2021; Hohenstein et al., 2023; Yakura et al., 2024), while work on conversational entrainment measures lexical and syntactic convergence between people (Nenkova et al., 2008; Reitter, 2014; Niederhoffer and Pennebaker, 2002). Real-time systems now connect these previously separate concerns: E2E models such as Moshi (Défossez et al., 2024) and SpeechGPT (Zhang et al., 2023), together with incremental turn-taking systems (Skantze, 2021), supply content during ongoing production. The resulting evaluation problem concerns both the system contribution and the human process of responding to it. This paper uses disfluency to preserve process evidence and suggestion correspondence to determine how that evidence relates to system content.

2.2 Disfluency Types and Production Processes

Disfluency categories become informative through their different distributions across production and interactional contexts. *Silent pauses* at clause and phrase boundaries are associated with conceptual planning and with the complexity or novelty of upcoming material (Goldman-Eisler, 1968; Butterworth, 1980). *Filled pauses* (“uh/um”) occur during conceptual planning and lexical retrieval and can signal an upcoming delay to listeners (Clark and Tree, 2002).

Other categories provide stronger evidence about managing or revising an utterance. *Repetitions* can maintain the conversational floor while formulation continues (Schegloff et al., 1977; Engelhardt et al., 2010). *Self-repairs* provide overt evidence that a speaker halted and revised speech after monitoring (Hartsuiker, 2006; Levelt, 1983), and their onset can follow an error within hundreds of milliseconds (Hartsuiker and Kolk, 2001; Postma, 2000). Their functions overlap with effects of task difficulty, language background, turn management, and the surrounding utterance. Inference therefore depends on contrasts among types and positions. For example, a shift from fillers toward repairs is consistent with increased monitoring and correction, while remaining compatible with other contextual explanations.

2.3 How AI Mediation Perturbs Production

These associations between disfluency type and cognitive process make the production model experimentally useful. **Lexical clarifications** can resolve unfamiliar terms during comprehension and change the information available for response planning (Duan et al., 2021; He and Fussell, 2026). **Content suggestions**, such as topic prompts and factual completions, may change conceptual planning (Arnold et al., 2016; He and Fussell, 2026; Qin et al., 2025). **Lexical and stylistic suggestions**, such as word recommendations and register adjustments, supply candidate words or syntactic frames during formulation (Arnold et al., 2020; Hohenstein et al., 2023). **Timing cues**, such as turn-taking prompts, may change turn initiation and floor management (Skantze, 2021). These mappings specify predictions while allowing one intervention to affect several components.

Evaluating external content while preparing speech can place simultaneous demands on working memory (Oomen and Postma, 2001; Engelhardt et al., 2013). A nearby disfluency is consequently consistent with evaluation of the intervention and with unrelated production difficulty. This ambiguity separates *cognitive fluency*, the efficiency of planning and retrieval, from *utterance fluency*, observable speech production (Segalowitz, 2010). Lexical evidence, matched comparisons, and converging measures are required to connect an observable event to the intervention.

These mechanisms therefore motivate directional hypotheses. Relevant content available before a complex response may reduce pauses associated with planning. Content that arrives during formulation or conflicts with an emerging plan may increase repairs and repetitions. Manipulating production demand should chiefly alter speech event timing and frequency, whereas manipulating suggestion usefulness or attractiveness should chiefly alter suggestion correspondence. Crossing these manipulations tests whether the two measures respond selectively and predicts changes in the four profiles introduced next.

3 A Two-Axis Evaluation Framework

A useful framework must translate the theoretical distinction into observations that can be compared across systems. We organize those observations along two axes: speech measures identify when and how production difficulty becomes observable,

and suggestion correspondence identifies whether system content appears in the response. Low disfluency can accompany smooth uptake, no consultation, silence, or withdrawal, and high correspondence can reflect considered use, automatic reuse, or content the speaker would have produced independently. Aligning the axes within the same episode reduces these complementary ambiguities and defines the evidence needed to distinguish the remaining explanations (Jakesch et al., 2023; Hohenstein et al., 2023).

Comparison also requires a common protocol. Studies should report anchors, windows, counting rules, baselines, and extraction procedures while adapting their values to the task and language. Unmediated baselines within each speaker are preferred when available. Studies without that condition should use matched windows or control trials and model variation among speakers. Raw disfluency values are unsuitable for unadjusted comparisons because individual rates vary substantially (Bortfeld et al., 2001).

3.1 Speech Measures

The speech axis begins with event identity and progressively adds duration, position, and change over time. Each layer restricts the plausible explanation of an event while preserving uncertainty arising from the speaker, task, and conversational context.

Disfluency rate by type. Report counts and rates (per 100 words, per minute, or per elapsed response time) of filled pauses, silent pauses, repetitions, repairs, and false starts. Stating the denominator makes systems comparable, and retaining categories preserves differences among their likely functions.

Disfluency duration. Report the temporal extent of each event because equal counts can conceal different amounts of interruption. Longer pauses and repair phases can accompany greater planning or reprocessing demands, with interpretation determined by event position and context (Hartsuiker, 2006).

Positional distribution. Locate event onsets relative to AI appearance, question completion, response onset, wording related to the suggestion, and response end. A filled pause after appearance is consistent with planning or lexical retrieval difficulty, whereas a repair after wording related to the suggestion provides evidence of revision. Posi-

tion connects an event to a candidate intervention period and operationalizes the *speech event timing* axis in §3.3. Alignment windows should be specified before analysis and tested for sensitivity to plausible boundary changes.

Disfluency trajectory. Report change across a conversation to distinguish initial disruption from adaptation or accumulated difficulty. Record silence, abandoned turns, and missing speech separately because each can otherwise appear as a low event rate.

3.2 Content Correspondence Measures

Once a speech event has been localized, the second axis asks how the response relates to the intervention. We distinguish surface lexical overlap, semantic similarity, and incorporation coded by people because each supports a different inference. *Suggestion correspondence* names the general axis, and *lexical entrainment* refers specifically to convergence in word choice across turns, following the dialogue literature (Nenkova et al., 2008; Niederhoffer and Pennebaker, 2002; Reitter, 2014).

Moving-average type-token ratio (MATTR). This measure summarizes lexical diversity across a session and uses a moving window to reduce the length sensitivity of raw type-token ratio (Covington and McFall, 2010). A decline can accompany repeated use of vocabulary suggested by AI (Hohenstein et al., 2023). It therefore detects change in expressive range across time. Attribution to an individual suggestion requires the correspondence measures that follow.

Suggestion correspondence. Compare the participant’s response text with the displayed suggestion. Use surface measures such as n-gram overlap to assess wording reuse and sentence-embedding cosine similarity or BERTScore to assess semantic correspondence (Reimers and Gurevych, 2019; Zhang et al., 2020). Human coding can establish whether recognizable suggestion content appears. Reporting them separately prevents semantic similarity, lexical reuse, and incorporation from being treated as interchangeable. Together they operationalize response content while leaving endorsement and reliance to additional evidence.

Connotation and register shift. Compare the speaker’s affective, formal, or stylistic profile with the AI output using formality classifiers (Pavlick and Tetreault, 2016), function-word style matching

(Niederhoffer and Pennebaker, 2002), or affective representations. Persistence into later unmediated turns provides a stronger adaptation hypothesis than immediate reuse. Topic, audience, and interactional goals remain competing explanations that matched comparisons must address.

3.3 Combining Speech and Content Measures

The two axes become informative when measured for the same episode. Suggestion correspondence establishes the relation between response and system content, whereas speech event timing locates observable production difficulty around the content’s availability and possible use. Crossing them yields four behavioral profiles (Table 1). The profiles convert two ambiguous signals into a structured set of competing explanations and specify which additional evidence could separate them.

Table 2 draws on a secondary analysis of the published XPLAIN Wizard-of-Oz study (He and Fussell, 2026). In that study, speakers using English as a second language completed a standardized collaborative planning task. The interface provided private, clickable lexical clarifications and presented predefined idea or sentence suggestions at scripted points. The source paper reports the system design and survey and interview findings. Here, timestamped recordings and response annotations serve only to demonstrate event alignment and correspondence measurement. Trained annotators coded disfluency events against intervention timestamps and coded suggestion incorporation from separate records. The excerpts illustrate measurement decisions and carry no validated profile labels.

Reading the excerpts descriptively, Row 1 shows close phrase correspondence in the reconstructed materials and a brief filler at retrieval onset. Row 2 contains a filler at clause onset outside the prespecified analysis window and uses a different lemma. Row 3 contains a pre-uptake *um*, the phonological *pack-/pack-aged/packable* repair, and content uptake of the AI’s term “packable.” Row 4 contains a silent pause and filler followed by lexical substitution within the AI’s syntactic frame. These observations illustrate the component measures and remain compatible with multiple cognitive accounts.

The examples also show why a profile starts an analysis. Frequent Profile A observations could arise from smooth fit, uncritical reuse, or missed event detection, whereas frequent Profile D observations could arise from substitution, rejection, un-

	No event in analysis window	Disfluency or repair in analysis window
Higher suggestion correspondence	<i>Profile A:</i> response content corresponds with the suggestion, and no event occurs in the prespecified window. Consistent with smooth uptake, uncritical reuse, or missed event detection.	<i>Profile B:</i> response content corresponds with the suggestion and an event occurs in the prespecified window. Consistent with effortful uptake or concurrent production difficulty.
Lower suggestion correspondence	<i>Profile C:</i> response content has little correspondence with the suggestion and no event occurs. Consistent with independent production, missed support, or undetected correspondence.	<i>Profile D:</i> response content has little correspondence with the suggestion and an event occurs in the prespecified window. Consistent with substitution, rejection, unrelated planning difficulty, or floor management.

Table 1: Four behavioral profiles defined by suggestion correspondence and speech events in a prespecified analysis window. The cells summarize observable configurations and retain multiple candidate interpretations. Analyses should preserve both axes as continuous measures, and the quadrants are descriptive and support hypothesis generation.

Illustrated relation	Reconstructed illustration	Timing of speech events	Suggestion correspondence
Phrase match	“ <i>um</i> like rolling suitcase”	Brief filled pause at retrieval onset and short response latency. Their relation to the AI event depends on the prespecified window and coding rule.	Surface and content correspondence with the reconstructed phrase suggested by AI.
Different content	“ <i>um</i> I think we can include some basic items”	Filled pause at clause onset outside the prespecified analysis window.	Response contains content that differs from the reconstructed suggestion.
Match with repair	“ <i>um</i> it’s a pack-...packaged...packable back-pack”	Filled pause before wording related to the suggestion, followed by truncation and self-repair.	Content correspondence with the AI’s term (“packable”) and phonological variation in its production.
Frame with substitution	“it’s... <i>um</i> but maybe instead of a wool hat use a beanie”	Silent pause and filled pause before a lexical revision within the prespecified analysis window.	Syntactic frame correspondence with lexical substitution (wool hat → beanie).

Table 2: Four reconstructed illustrations of the component measures in Table 1. Content words were paraphrased into a different topical domain, and the original disfluency forms and positions were preserved. The rows illustrate measurement without validated profile assignments. Two trained annotators independently coded events against intervention timestamps and resolved disagreements through discussion. The source study received institutional review board approval (protocol 0149408). Italicized tokens mark filled pauses, and “...” marks silent pauses.

related planning difficulty, or floor management. A system comparison should use profile frequencies to select follow-up measures and should retain output quality, task performance, and participant reports as complementary outcomes.

Operationalizing the axes. Retain correspondence and speech timing as continuous measures in primary analyses, with the four profiles summarizing regions of their joint distribution. Any categorical cutpoint should be justified for the task, specified before analysis, and examined in a sensitivity analysis. This preserves information that universal high and low thresholds would discard.

Chance baseline for suggestion correspondence. Similar wording may have been likely without exposure to a particular suggestion. Compare each response with unshown suggestions matched on gen-

erator, task, topic, communicative function, length, and presentation format when possible. The displayed score, matched distribution, and their difference estimate how much correspondence is specific to the displayed content. Report unmatched dimensions because they limit the comparison.

Illustrative measurement check. The corpus first tests event detection and describes where events occurred. The source study recruited 38 speakers using English as a second language, and 35 recordings were usable for disfluency analysis. Two independently coded event intervals counted as a match when their temporal intersection divided by their union reached .25, yielding symmetric event-detection $F_1 = .981$. The adjudicated record contained 1,276 events with 2,286 category labels, of which filled pauses were the most fre-

quent category (1,039 labels). At least one event occurred in 62.3% of suggestion response intervals and 44.8% of clarification response intervals. The design included no unassisted condition, and the contexts differed in support function, response content, and observation time. This contrast therefore provides only a descriptive estimate of coverage.

The next check asks whether aggregate disfluency rate is associated with suggestion incorporation in this sample. Incorporation codes were available for 312 responses from two sources: 268 had graded codes and 44 had binary codes. Each source supplied a final code without independent duplicate coding, and the 22 positive binary cases lacked partial or full extent. Within participants, the probability that an incorporated response had a higher disfluency rate was .541 (exact $p = .347$, Holm-adjusted $p = .694$, 31 participants, 277 responses). This inconclusive association does not support using aggregate disfluency rate as a substitute for a separate correspondence measure.

The correspondence checks then ask whether the content measure behaves as intended. Using final corrected response spans, similarity increased with incorporation among 261 responses with graded codes and semantic scores (descriptive Spearman's $\rho = .483$). Across 332 responses, displayed suggestions were more similar than the mean of other suggestions of the same format that were not shown on that trial in 299 cases (90.1%), with mean similarity of .617 for displayed suggestions and .352 for comparisons (difference = .265). Both summaries contain repeated responses from participants and items. The alternative suggestions also concerned different items, so the second summary is a weak baseline that retains topic differences. These results establish event-detection feasibility and a descriptive relation between semantic correspondence and coded incorporation. The corpus has no measure of cognitive load aligned to individual events or independent labels for the profiles, leaving their cognitive interpretations for prospective tests.

3.4 Anticipated Objections

The most direct objection is that WER and surveys already cover system quality and user experience. They answer questions at different temporal scales: WER evaluates transcript accuracy, and surveys summarize retrospective experience. Analysis aligned to individual events asks where production changed around a particular intervention and whether its content entered the response.

The framework is useful when this local process question matters to the system comparison.

A stronger objection concerns the nonspecificity of disfluency. Planning, retrieval, language background, and floor management can produce similar events. Temporal alignment and suggestion correspondence reduce this ambiguity, especially with comparisons within speakers or across matched episodes. Causal and cognitive claims still require the contextual and converging evidence specified in Table 3. This constraint defines the framework's evidential standard and prevents a behavioral profile from becoming a diagnostic label.

4 Application Across AI-MC Contexts

The framework earns practical value only when it changes how systems are compared. The following cases expose different ways that equivalent output scores can conceal process effects, and each case yields a distinct contrast for the two axes.

4.1 Proactive Content Scaffolding

Contextual suggestions and factual completions can improve the final response while imposing different demands during its production (Fang et al., 2025; Xia et al., 2023). Research on text input already documents changes in language use under reply and composition support (Arnold et al., 2020; Mieczkowski et al., 2021). In speech, relevant content available before response planning may reduce pauses at clause boundaries, whereas content that conflicts with an emerging plan may increase fillers, false starts, or repairs. Tests can separate these accounts by varying timing and semantic fit, comparing matched opportunities, and modeling individual baselines. Session trajectories then show whether an initial effect diminishes through adaptation or accumulates across interventions.

4.2 Lexical and Stylistic Adjustment

Word replacements, register adjustments, and tonal reframings intervene closer to formulation by supplying candidate lemmas and syntactic frames. Interactive alignment theory predicts lexical convergence through priming and shared expectations (Pickering and Garrod, 2004, 2013). Computational measures can compare convergence with the human partner and with the AI across turns (Nenkova et al., 2008; Reitter, 2014; Niederhoffer and Pennebaker, 2002). This comparison becomes

Requirement	Reference implementation and reportable choices
Unit and contrast	Analyze each suggestion opportunity and the participant’s first substantive response. Report any unmediated condition, matched control trials, and adjustment for repeated observations within speakers.
Event anchors and windows	Record suggestion appearance, question completion, response onset, and response end. The illustrative corpus used four intervals: 10 seconds before appearance, appearance to question completion, question completion to response onset, and response onset to the earliest competing event, substantive partner turn, page transition, or 30 seconds after question completion.
Disfluency assignment	Assign an event to a window by onset. Report categories, counts, durations, denominator, annotation or extraction method, reliability, and timestamp sensitivity.
Content correspondence	Analyze only the participant’s response text. Report surface overlap, semantic similarity, and incorporation codes separately. The illustrative check used paraphrase-multilingual-mpnet-base-v2. Report the model and segmentation rule.
Chance baseline	Compare the response with unshown suggestions matched on task, topic, communicative function, length, and presentation format. The illustrative corpus matched task and presentation format only. Report unmatched dimensions, the displayed score, comparison mean, and their difference.
Thresholds and missingness	Preserve continuous scores in primary analyses. Mark categorical profiles as exploratory. Record silence, abandoned turns, unavailable speech, and extraction failures separately.
Interpretation	Report language, task, turn position, floor management, and plausible alternative explanations. Use self-report, performance, multimodal measures, or experimental manipulation when making cognitive claims.

Table 3: Minimum reporting protocol. Values can vary across tasks and languages, and the listed information is required for comparison and interpretation.

consequential when accommodation is distributed unequally. Sociolinguistic research predicts greater accommodation from speakers with less positional power (Giles and Ogay, 2007). Higher suggestion correspondence and lower lexical diversity in that group would document an asymmetry requiring explanation. Establishing an AI effect then requires controls for power, topic, item content, and baseline language use.

4.3 Self-Presentation and Emotional Attunement

Tone refinement, empathic phrasing, and politeness support extend the same problem to self-presentation (Goffman, 1959; Walther, 1996). Suggestion correspondence localizes immediate reuse, and MATTR shows whether lexical diversity changes across the session. Comparisons within speakers across mediated and unmediated conditions can therefore test changes in expressive range. Participant and listener judgments remain necessary for assessing genuineness and appropriateness, outcomes that linguistic correspondence alone cannot determine. Across all three contexts, valid inference depends on following the relevant people, turns, and modalities beyond an isolated response.

5 Methodological Extensions

Extending the unit of analysis beyond one speaker and one turn introduces dependencies that the basic two-axis profile must represent explicitly.

5.1 Dyadic and Multiparty Settings

Dyadic AI-MC permits comparison of each speaker’s speech rate, disfluency, and lexical choices across turns (Pickering and Garrod, 2004). Additional participants introduce overlap, floor holding, shifting addressees, and several potential sources of lexical convergence (Clark, 1996). Analyses should therefore model addressee, turn position, overlap, and group role. With these dependencies represented, the framework can test whether access to suggestions, content correspondence, and speech events are distributed asymmetrically across participants.

5.2 Speaker and Listener Roles

The analysis of speakers raises a corresponding question about what mediation changes for listeners. Automatic speech recognition captions, speech-to-speech translation, and speech language models may remove pauses and repairs from the signal (Wang et al., 2020; Rocholl et al., 2021; Défossez et al., 2024; Zhang et al., 2023; Skantze, 2021). Psycholinguistic studies show that filled pauses can direct attention toward novel or complex content and facilitate lexical integration (Arnold et al., 2003; Fox Tree, 2001; Corley et al., 2007), and listeners also adapt to a partner’s disfluency patterns (Brennan and Schober, 2001). Smoothing may therefore improve surface fluency while reducing information used during comprehension.

This literature supports an untested prediction

about AI smoothing: removing natural disfluencies may reduce anticipatory cues and impair comprehension of novel material. A controlled study should manipulate filled pauses, silent pauses, and repairs separately. For fillers and silent pauses, lexical content and voice can remain constant while overall duration is balanced or modeled. Repair studies can compare matched fluent and repaired recordings and model duration differences. Recall, comprehension accuracy, clarification requests, echoed repetitions, and planning time in the listener’s next turn capture consequences at different stages. This design could test whether an apparent improvement in output fluency creates a cost for the listener.

5.3 Multimodal Triangulation

The remaining ambiguity concerns what a localized speech event means. Complementary streams can connect the event to consultation, arousal, or broader acoustic change while preserving the limits of each measure.

Gaze and interaction logs. Gaze toward an AI display and interface logs can show that prompt consultation coincided with a speech event. Their inferential contribution is orientation to a region at the precision supported by the recording and coding method.

Physiological measures. Electrodermal activity (EDA) and heart rate variability (HRV) can provide converging measures of autonomic arousal (Taptiklis et al., 2023). Arousal remains compatible with deliberation, overload, and other explanations.

Acoustic features beyond disfluency. Speech rate, pitch range, and pause duration complement categorical counts. Extraction should use raw audio or a pipeline validated for retaining the target events.

Each stream addresses a different ambiguity, so the useful combination follows the claim being tested. Inferences about cognitive benefit or harm ultimately require performance, experience, or experimental outcomes aligned with the same episode.

6 Design Implications and Conclusion

Two systems can achieve the same WER, task completion, or satisfaction score while creating different production demands and different input for listeners. Disfluency measures locate observable

production difficulty in time, and suggestion correspondence establishes how the response relates to system content. Their joint analysis makes these otherwise hidden differences available for system comparison.

The immediate use case is offline evaluation. Researchers can derive the measures through human coding, forced alignment, or ASR configured to retain disfluencies (Shriberg, 1994; Wang et al., 2020; Jamshid Lou and Johnson, 2020; Rocholl et al., 2021), with explicit validation of event retention, timestamps, and extraction error. Adaptive deployment sets a higher evidential threshold: a measure must prospectively predict benefit or difficulty in the target population, and responding to it must improve an outcome. An isolated disfluency spike or decline in lexical diversity cannot justify intervention.

The framework contributes four behavioral profiles, a minimum reporting protocol, and experiments that can test their interpretations for speakers and listeners. Its broader claim is that real-time AI-MC should be evaluated through the interaction that produces an output as well as through the output itself. This expanded unit of evaluation reveals when apparent system success coincides with observable production difficulty, response reuse, or fewer comprehension cues.

Limitations

The framework’s value depends on disciplined interpretation of its boundaries. The illustrative corpus establishes event-detection and correspondence feasibility. Causal validation and independent labels for the four behavioral profiles require prospective data.

Pipeline dependence. Extracting fluency metrics requires human coding or a pipeline for ASR and disfluency detection. Outputs from many E2E speech models (Whisper, SeamlessM4T) and commercial ASR application programming interfaces (APIs) omit fillers and reparanda. Applying the framework therefore requires systems that preserve disfluencies (Wang et al., 2020; Jamshid Lou and Johnson, 2020; Rocholl et al., 2021), forced alignment over raw audio, or human annotation. Each approach introduces measurement error and cost that implementation work must quantify.

Speaker variability. Disfluency rates vary 5–10× across individuals as a function of age, language background, working memory capacity, and

clinical conditions (Bortfeld et al., 2001; Engelhardt et al., 2010). Contrasts within each speaker are preferred. Designs without an unmediated condition should use matched episodes, model speaker effects, and avoid unadjusted system rankings across speakers.

Limits of profile identification. The two axes are noisy observable measures. Similar response content can arise without suggestion use, and an event in the analysis window can reflect planning, retrieval, turn management, or measurement error. Chance baselines address prior content similarity, and contextual and multimodal evidence narrows the interpretations of speech events. The profiles remain descriptive and can be used to generate hypotheses.

Theoretical anchoring. Levelt’s architecture organizes candidate production processes. Predictive processing and interactional accounts (Pickering and Garrod, 2013; Hagoort, 2019) motivate additional explanations. The proposed relations between disfluency types and production processes are probabilistic and require tests that account for context.

Operationalization. Table 3 specifies minimum reporting requirements. Lemmatizer choice, embedding model, n-gram window, MATTR window, and event windows still depend on language, register, and task. Comparisons across studies require reporting these choices and testing sensitivity to plausible alternatives.

Scope across languages. Our taxonomy and predictions are derived primarily from psycholinguistic literature on English. Filled-pause inventories, repair conventions, and cues at clause boundaries vary across languages, so the framework requires replication in each language before cross-language application.

Adaptive feedback. A system that responds to disfluencies may change how users speak over time. Periodic recalibration, control conditions that withhold intervention, and longitudinal evaluation are required to detect such effects.

Ethical Considerations

Speech behaviors have probabilistic associations with cognitive processes, creating risks in both data collection and interpretation.

Bias across speaker populations. Disfluency baselines vary across populations. Stuttering, aphasia, and Parkinson’s-related dysarthria can produce

disfluency unrelated to AI-MC, and speakers using a second language (L2) can differ from speakers using their first language (L1) (Segalowitz, 2010). Unadjusted application may conflate speaker characteristics with system quality. Designs that compare observations within each speaker, stratified analysis, and explicit baseline adjustment are prerequisites for population comparisons.

Surveillance dual use. Real-time extraction of speech features can support workplace monitoring. These measures should be limited to consented research evaluation. Any field deployment would require informed consent, bystander protection, data minimization protocols, and an explicit prohibition on inference about individual performance.

Privacy and institutional review. Applying the framework requires recording and processing speech, which is highly identifying personal data. Extracted features may retain sensitive information, and the source audio contains identity and content. Institutional review, secure storage, data minimization, and feature extraction on devices where feasible are appropriate safeguards. The recording used for Table 2 was collected under institutional review board protocol 0149408. The illustrations reconstruct the original events with content words paraphrased into a different topical domain.

Power asymmetry and equity. §4.2 motivates testing whether speakers with less positional power correspond more closely with AI suggestions. Such a pattern would raise a deployment concern because AI-MC may affect speakers unequally. Field evaluations should report outcomes by positional power and include pooled estimates only with the corresponding stratified results.

Feedback through training data. Training on speech processed through ASR, speech language models, or translation pipelines may change the distribution of disfluencies represented in future systems. Datasheets and provenance metadata should distinguish naturally produced speech from AI-MC speech.

Use of AI Assistants. We used large language model (LLM) writing assistants, including chat assistants and integrated development environment tools, to refine prose, copyedit text, and format LaTeX. The authors developed the research ideas, framework, analyses, citations, arguments, and bibliography and reviewed all assisted edits.

Materials Availability. Supporting materials can be found on the Open Science Framework at <https://osf.io/yc6dh/>.

References

- Jennifer E Arnold, Maria Fagnano, and Michael K Tanenhaus. 2003. Disfluencies signal thee, um, new information. *Journal of psycholinguistic research*, 32(1):25–36.
- Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI)*, pages 128–138. ACM.
- Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. 2016. [On suggesting phrases vs. predicting words for mobile text composition](#). In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 603–608, Tokyo, Japan. Association for Computing Machinery.
- Heather Bortfeld, Silvia D Leon, Jonathan E Bloom, Michael F Schober, and Susan E Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, 44(2):123–147.
- Susan E Brennan and Michael F Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of memory and language*, 44(2):274–296.
- Brian Butterworth. 1980. Evidences from pauses in speech, language production: Speech and talk, vol. i.
- Herbert H Clark. 1996. *Using language*. Cambridge university press.
- Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.
- Martin Corley, Lucy J MacGregor, and David I Donaldson. 2007. It’s the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, 105(3):658–668.
- Martin Corley and Oliver W Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602.
- Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.
- Wen Duan, Naomi Yamashita, Yoshinari Shirai, and Susan R. Fussell. 2021. [Bridging fluency disparity between native and nonnative speakers in multilingual multiparty collaboration using a clarification agent](#). *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–31.
- Paul E Engelhardt, Martin Corley, Joel T Nigg, and Fernanda Ferreira. 2010. The role of inhibition in the production of disfluencies. *Memory & Cognition*, 38(5):617–628.
- Paul E Engelhardt, Joel T Nigg, and Fernanda Ferreira. 2013. Is the fluency of language outputs related to individual differences in intelligence and executive function? *Acta psychologica*, 144(2):424–432.
- Cathy Mengying Fang, Yasith Samaradivakara, Pattie Maes, and Suranga Nanayakkara. 2025. Mirai: A wearable proactive ai" inner-voice" for contextual nudging. In *Proceedings of the extended abstracts of the CHI conference on human factors in computing systems*, pages 1–9.
- Jean E. Fox Tree. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34(6):709–738.
- Jean E Fox Tree. 2001. Listeners’ uses of um and uh in speech comprehension. *Memory & cognition*, 29(2):320–326.
- Howard Giles and Tania Ogay. 2007. Communication accommodation theory. In Bryan B. Whaley and Wendy Samter, editors, *Explaining Communication: Contemporary Theories and Exemplars*, pages 293–310. Lawrence Erlbaum, Mahwah, NJ.
- Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Anchor Books, New York.
- Frieda Goldman-Eisler. 1968. *Psycholinguistics: Experiments in Spontaneous Speech*. Academic Press, London.
- Peter Hagoort. 2019. The neurobiology of language beyond single-word processing. *Science*, 366(6461):55–58.
- Robert J Hartsuiker. 2006. Are speech error patterns affected by a monitoring bias? *Language and Cognitive Processes*, 21(7-8):856–891.
- Robert J Hartsuiker and Herman HJ Kolk. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive psychology*, 42(2):113–157.
- Wanqing Psyche He and Susan R. Fussell. 2026. [XPLAIN: A proactive scaffold across speech processing stages: Supporting non-native speakers in real-time ai-mediated turn-taking](#). In *Proceedings of the 8th ACM Conference on Conversational User Interfaces*, pages 1–17, New York, NY, USA. Association for Computing Machinery.
- Jess Hohenstein, Rene F Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific reports*, 13(1):5487.

- Peter Indefrey and Willem JM Levelt. 2004. The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144.
- Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. [Human heuristics for ai-generated language are flawed](#). *Proceedings of the National Academy of Sciences*, 120(11).
- Paria Jamshid Lou and Mark Johnson. 2020. Disfluency detection using a noisy channel model and a neural reranker. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5745–5755. ACL.
- Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing*, 3rd edition. Draft chapters available at <https://web.stanford.edu/~jurafsky/slp3/>.
- Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.
- Willem JM Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.
- Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. 1999. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(1):1–38.
- Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-mediated communication: Language use and interpersonal effects in a referential communication task. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–14.
- Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers (ACL-HLT)*, pages 169–172. ACL.
- Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.
- Nazbanou Nozari, Gary S. Dell, and Myrna F. Schwartz. 2011. Is comprehension necessary for error detection? a conflict-based account of monitoring in speech production. *Cognitive psychology*, 63(1):1–33.
- Claudy CE Oomen and Albert Postma. 2001. Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research*, 30(2):163–184.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.
- Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 61–74. MIT Press.
- Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.
- Martin J. Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. ACL.
- Albert Postma. 2000. Detection of errors during speech production: A review of speech monitoring models. *Cognition*, 77(2):97–132.
- Peinuan Qin, Zicheng Zhu, Naomi Yamashita, Yitian Yang, Keita Suga, and Yi-Chieh Lee. 2025. [AI-based speaking assistant: Supporting non-native speakers’ speaking in real-time multilingual communication](#). *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–28.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. ACL.
- David Reitter. 2014. Alignment in web-based dialogue: Studies in big data computational psycholinguistics. *Topics in Cognitive Science*. Earlier formulation: Reitter, Keller & Moore (2011), “A computational cognitive model of syntactic priming,” *Cognitive Science* 35.
- Johann C. Rocholl, Vicky Zayats, Daniel D. Walker, Noah B. Murad, Aaron Schneider, and Daniel J. Liebling. 2021. Disfluency detection with unconstrained in-span operations. In *Proceedings of Inter-speech 2021*.
- Ardi Roelofs. 2020. [Self-monitoring in speaking: In defense of a comprehension-based account](#). *Journal of Cognition*, 3(1):18.
- 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.
- Norman Segalowitz. 2010. *Cognitive bases of second language fluency*. Routledge.
- Elizabeth Ellen Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Phd dissertation, University of California, Berkeley.

- Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.
- Nick Taptiklis, Merina Su, Jennifer H Barnett, Caroline Skirrow, Jasmin Kroll, and Francesca Cormack. 2023. Prediction of mental effort derived from an automated vocal biomarker using machine learning in a large-scale remote sample. *Frontiers in Artificial Intelligence*, 6:1171652.
- Joseph B Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research*, 23(1):3–43.
- Shaolei Wang, Wanxiang Che, Qi Liu, Pengda Qin, Ting Liu, and William Yang Wang. 2020. Disfluency detection with unlabeled data and small annotated data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2162–2172. ACL.
- Haijun Xia, Tony Wang, Aditya Gunturu, Peiling Jiang, William Duan, and Xiaoshuo Yao. 2023. Crosstalk: Intelligent substrates for language-oriented interaction in video-based communication and collaboration. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–16.
- Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2024. Empirical evidence of large language model’s influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.
- Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773. ACL.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.