JOURNAL OF CULTURAL ANALYSIS AND SOCIAL CHANGE

Journal of Cultural Analysis and Social Change,
2025, 10(2), 2247-2257
ISSN: 2589-1316

Mind the Gaps: How Learners Parse Reductions in Chatbot Dialogue

Awad Alshehri [1*]

[1] *Imam Mohammad Ibn Saud Islamic University,* SAUDI ARABIA

*__Corresponding Author:__ awad.journal@gmail.com

Citation: Alshehri, A. (2025). Mind the Gaps: How Learners Parse Reductions in Chatbot Dialogue, *Journal of Cultural Analysis and Social Change*, 10(2), 2247-2257. https://doi.org/10.64753/jcasc.v10i2.1922

Published: November 16, 2025

ABSTRACT

Conversational chatbots now rely on synthetic voices that mimic natural reductions such as elision, assimilation, catenation, flapping, and vowel centralization. These features enhance fluency but often blur meaning, particularly for second-language learners. This study presents a computational analysis — conducted without human participants — examining how varying degrees of reduction and speech rate in text-to-speech (TTS) output influence comprehension within an AI dialogue pipeline (TTS → ASR → LLM). A purpose-built corpus of short task dialogues was generated and rendered with multiple TTS voices under three reduction and rate settings. Comprehension was evaluated using word-error rate, entity-recognition F1, and dialogue-level question-answer accuracy, while acoustic–prosodic measures yielded a Reduction Index capturing duration shortening, vowel centralization, and boundary cues. A single AI clarification turn—explicitly reformulating reduced forms—was tested as a recovery strategy. Moderate reductions maintained comprehension in prosody-rich voices, but extreme and fast reductions caused error spikes and semantic drift. One clarification restored much of the loss and improved stability across new scripts and voices. Vowel centralization and syllable compression predicted most failures. The resulting Reduction–Robustness Curve provides a benchmark for balancing naturalness and intelligibility in synthetic speech and for designing adaptive clarification in AI-based language tutoring.

Keywords: Connected Speech, Reductions, Tts, ASR, LLM, Intelligibility, Clarification, Prosody, Computational Benchmarking, AI Tutoring

INTRODUCTION

Conversational agents—chatbots, virtual tutors, virtual assistants—are gradually adopting more natural, humanlike speech. A key aspect of naturalness is connected speech: reductions such as elision, assimilation, flapping, catenation, and vowel centralization are what make spontaneous talk fluid. Yet these very phenomena are also notorious stumbling blocks for listeners, especially second-language (L2) learners. In L2 research, reduction phenomena have long been studied in isolation (e.g. *did you* → *d'ya*, *going to* → *gonna*, *handbag* → *hambag*) via elicited sentences or minimal-pair tasks (e.g. Aoyama & Flege, 2021; Scott & Cutler, 1984). But real conversational systems don't speak in neatly spaced words; they speak in turns. In dialogic contexts, listeners must parse reductions in real time, align them with context, recover underlying forms, and maintain coherence across multiple turns. The challenge is magnified when the speech is synthesized and mediated by machine agents, whose reduction patterns may differ subtly (or drastically) from natural human speech.

Why Reductions Matter in Chatbot Dialogue

Reducing naturalness is not simply a stylistic flourish: connected speech plays a role in listener expectations, predictive parsing, and processing economy (with or without reduction). But when reductions go too far or override cues such as prosody, they can create gaps—moments where the listener can't reliably infer what was said.

In L2 settings, these gaps may lead to comprehension breakdowns, misinterpretation of referents, or failure to trigger the intended follow-up.

Yet chatbot designers often assume that modern TTS + ASR pipelines are robust enough that minor reductions won't hurt comprehension. That assumption is risky. If reductions push recognition or parsing over a fragile threshold, the system may misrecognize a word, misinterpret context, or derail dialogue coherence. From a pedagogical perspective, we might ask: can a chatbot *guide* a learner's attention to those gaps—through clarification or reformulation—thereby scaffolding the parsing of reductions? To date, such a mechanism has been explored in human–human tutoring and conversation (e.g., clarification requests, recasts), but rarely (if ever) in AI-mediated conversational systems.

Challenges of Evaluating Reduction Parsing Without Human Participants

Most prior research in pronunciation and perception involves native or nonnative listeners. But running human-subject experiments is time-consuming, costly, and often constrained by IRB protocols, participant recruitment, and variability in listener experience. As AI-powered dialogue proliferates, it is increasingly viable—and ethically simpler—to study machine proxies for human comprehension.

In this paper, we propose a no-participants, computational benchmarking paradigm in which we simulate learner comprehension by measuring what happens in the pipeline: TTS → ASR → LLM (dialogue Q&A). The premise is that errors, drift, and failure points in that pipeline correlate with places where learners are likely to struggle. By systematically varying reduction level, speech rate, TTS voice, and inserting instant clarification turns, we can chart thresholds of intelligibility, evaluate the effects of clarification, and derive design guidelines for reduction-aware systems. Such a computational method doesn't replace human experiments, but it offers scalable, controlled diagnostics—an automated stress test of how much connected speech a system (and by proxy, a learner) can tolerate before breaking down.

Goals, Contributions, and Organization

Our goals are threefold:

- Characterize reduction tolerance: identify the boundaries (in reduction degree × speech rate × voice type) at which the pipeline's error rates and semantic integrity collapse.
- Evaluate clarification interventions: test whether short, embedded reformulation prompts (e.g. "Did you mean *did you* → *d'ya*?") can rescue comprehension in the immediately following turn.
- Link acoustic–prosodic predictors to failure: derive a *Reduction Index* that quantifies the severity of reduction (duration compression, vowel centralization, prosodic cue degradation) and relate it to error metrics.

In doing so, we contribute:

- A benchmark dataset + pipeline for connected-speech robustness in chatbot dialogue.
- The Reduction–Robustness Curve, a practical diagnostic for designers and researchers.
- An intervention recipe (clarification micro-turn) that partially recovers comprehension.
- Insights on how reductions can be kept while preserving intelligibility in AI tutoring applications.

The rest of the paper unfolds as follows. In Section 2, we situate our work relative to L2 intelligibility research, connected-speech modeling, and machine-side evaluation metrics. Section 3 details our stimuli, pipeline, reduction manipulations, and metrics. Section 4 presents results and analyses, including interacting factors and intervention effects. Section 5 discusses implications for TTS/ASR/LLM systems, second-language pedagogy, and limitations—especially the gap between machine proxies and human perception. Section 6 concludes with future directions, including validation with human participants and deployment in real tutoring systems

LITERATURE REVIEW

Connected Speech and Listener Processing

Everyday speech is laced with reductions—segmental deletions and assimilations, vowel centralization, catenation, and flapping—that make talk sound natural but push recognition systems and learners toward the edge of failure. Decades of phonetic work document how frequent and extreme these variants are across languages, underscoring that "canonical" tokens are often the exception in spontaneous discourse (Ernestus & Warner, 2011). Native listeners typically survive this messiness by integrating probabilistic expectations with prosodic scaffolds; when cues line up, recognition is fast, and when they don't, activation is delayed or wrong. Eye-tracking and gating

work show that heavy reduction slows or misguides lexical access, while context and place-of-articulation probabilities modulate recovery (Mitterer, 2011).

For L2 listeners, the cost is steeper. Clear-speech and adaptation literatures agree that intelligibility improves when talkers reduce less and signal contrasts more robustly—but those gains vary by listener proficiency, talker style, and noise. Meta-analyses and experiments place the "clear speech benefit" between ~12–34 percentage points for many populations, yet some contrasts (rates, vowel space changes) can even backfire depending on task and materials. These mixed outcomes make "make it clearer" an unstable prescription (Lam & Tjaden, 2013). Recent work continues to refine the picture: clear speech helps native and non-native listeners, but talker–listener pairings and speech style interact, and interlanguage benefits appear when non-native speech matches listener expectations (Jung & Dmitrieva, 2023).

What these strands establish is a principled trade-off: reductions serve efficiency and naturalness for insiders, while outsiders pay a comprehension tax unless other cues (prosody, predictability) compensate. Your paper's starting point—that synthetic dialogue often "sounds right" while still hiding lexical content for learners—sits squarely in this trade-off.

Synthetic Voices and Conversational TTS

Modern neural TTS (Tacotron 2, FastSpeech 2, VITS and successors) can deliver high-MOS speech with controllable rate and style, enabling products to dial in "casual" voices that exhibit reductions (Shen et al., 2018). But naturalness is not intelligibility. Recent perceptual work comparing human and neural TTS shows that (i) TTS is often less intelligible than matched human recordings in quiet and noise; (ii) adopting a "clear" style improves both—but tends to help TTS even more; and (iii) device guises and visual context can depress intelligibility regardless of voice type. These findings imply that the distribution of reductions and prosodic cueing in TTS is still atypical in ways that matter for comprehension (Aoki et al. (2022). Relatedly, when humans speak "to machines" (e.g., Siri-directed speech), they alter acoustic-prosodic patterns relative to human-directed talk—evidence that interactional setting changes speech targets. That makes it risky to assume that a single, "natural" synthetic style will fit learner needs (Cohn et al., 2022).

The practical upshot is that voice design choices (rate × reduction × prosody) may quietly move systems across intelligibility thresholds. Your study's Reduction Index and rate manipulations are a direct response to this under-measured design space.

Machine-Side Proxies for Comprehension

If we want scale without human subjects, can machine metrics stand in for listener comprehension? Two converging lines of evidence say "use with care, but yes." First, multiple studies show that state-of-the-art ASR degrades with the same stressors that hurt humans (noise, masks, heavy reduction), and that relative difficulty patterns often line up—even if level offsets differ. This makes ASR WER and downstream errors plausible early-warning indicators (Patman & Chodroff, 2024). Second, in L2 assessment specifically, ASR-based intelligibility ratings correlate with human judgments, while also surfacing different error profiles (segmental vs. lexical). That complementarity is exactly what a pipeline metric should capture (Inceoglu et al., 2023). Newer work also explores evaluation metrics that better track human and LLM judgments, reinforcing the feasibility of model-based proxies for perception under realistic conditions (Phukon et al, 2025).

Your pipeline (TTS → ASR → LLM Q&A) pushes this line further: it treats recognition error, entity/keyword F1, and dialogue QA accuracy as layered signals of "where a listener would likely fail" in a real conversational task. This is not a replacement for human data, but a strong filter for stimulus design and a reproducible stress test before running listener studies.

Clarification and Repair in AI Dialogue

Conversation is built to repair trouble. Human talk has a preference for self-initiated self-repair; other-initiations of repair (e.g., "Sorry, what?") are used with varying specificity depending on the problem (Schegloff et al., 1977). Spoken-dialogue systems have long mirrored this with explicit and implicit confirmations, thresholds for rejections, and targeted clarification questions; newer work revisits repair for neural conversational agents under real-world noise. The consistent theme: targeted clarifications beat generic "please repeat" prompts, both for task success and user satisfaction (Bohus & Rudnicky, 2008).

This matters pedagogically. Interaction-driven SLA argues that negotiated input (recasts, clarification requests) focuses attention on exactly the contrasting forms learners need. Embedding a crisp, one-turn clarification that pairs reduced and canonical forms leverages this principle while keeping conversational flow. Your intervention design is therefore not an ad-hoc UX fix; it is a principled application of repair and interaction hypotheses to AI-mediated listening.

Where Prior Work Stops—and What this Paper Contributes

Synthesis across strands reveals four open problems:

- **What each Study Contributed.** Phonetic corpora (e.g., Buckeye) established how pervasive reduction is in spontaneous English; psycholinguistic experiments mapped when listeners recover and when they don't; clear-speech work quantified intelligibility gains and boundary conditions; neural-TTS research showed that "sounding natural" does not guarantee "being intelligible"; ASR/L2 studies validated machine scores as partial proxies for human comprehension (Pitt et al, 2005).
- **Benefits vs. Risks.** Retaining reductions preserves naturalness and exposure to authentic input; over-reducing—especially at fast rates and with flat prosody—yields disproportionate recognition and meaning errors for learners and machines alike. Clear speech helps, but can distort the input away from what learners must actually parse (Lam & Tjaden, 2013).
- **The Unresolved Gap.** Most work isolates sentences or tokens, not turn-by-turn dialogue with synthetic voices; few studies quantify how much reduction a system (and by proxy, a learner) can tolerate before dialogue state drifts; and repair has rarely been tested as an embedded AI move that simultaneously supports users and stabilizes systems (Aoki et al, 2022).
- **Why your Study Matters.** Your no-participants benchmark directly targets these holes: it (i) manipulates reduction and rate in neural TTS, (ii) measures failure at three pipeline layers (ASR → entity/Q&A), (iii) links errors to acoustic-prosodic predictors via a Reduction Index, and (iv) tests a minimal, theoretically motivated clarification turn. Together, these yield a Reduction–Robustness Curve for design and a scalable pre-screen for future human experiments.

METHODOLOGY

Research Design

This study adopts a computational experimental design that models how connected-speech reductions influence comprehension within an AI dialogue pipeline. Instead of human participants, comprehension difficulty is inferred from the degradation of performance in a three-stage sequence—text-to-speech (TTS) → automatic speech recognition (ASR) → large language model (LLM)—which together simulate the process of speech generation, perception, and interpretation in a chatbot environment. This design allows scalable, ethically straightforward testing of reduction effects while maintaining experimental control over acoustic and lexical factors (see Räsänen & Alku, 2024; Leong, Wagner, & Yuan, 2023).

Materials and Stimulus Generation

- ***Dialogue Corpus.*** A base set of 180 short dialogues (two to four turns each) was generated using *GPT-4 Turbo* through prompts specifying register, communicative intent, and syntactic diversity (e.g., requests, confirmations, opinion exchanges). Dialogues were constrained to everyday contexts (ordering, giving directions, making plans) to maintain lexical familiarity. Each dialogue was rendered in three reduction levels—*minimal*, *moderate*, and *high*—reflecting the proportion and strength of connected-speech processes (adapted from Ernestus & Warner, 2011; Cauldwell, 2018).
- ***TTS Synthesis.*** Speech stimuli were synthesized using Microsoft Neural TTS (2025 release) and Google Cloud WaveNet voices. Both engines support phoneme-level control of duration, pitch, and reduction degree.
  Speech rate was manipulated at three settings (0.8×, 1.0×, 1.2× of the default tempo). All audio files were normalized to –23 LUFS with 16-kHz sampling.

Reduction Processes. The following reductions were embedded through phonetic rewriting rules applied at generation:

Table 1

Category	Example	Rule Applied
Flapping	*butter* → *budder*	alveolar stop between vowels → [ɾ]
Elision	*next day* → *nex day*	cluster /t/ deletion before /d/
Catenation	*go on* → *gowon*	linking of word-final and initial vowels
Assimilation	*handbag* → *hambag*	alveolar → bilabial before /b/
Vowel reduction	*to* → *tə*, *can* → *kən*	unstressed vowel centralization

Each token's phonetic transcript was aligned using Montreal Forced Aligner 2.2 (McAuliffe et al., 2017), enabling extraction of duration, vowel formants, and boundary cues.

The Computational Comprehension Pipeline

Automatic Speech Recognition. Each audio file was decoded by Whisper Large-v3 (Radford et al., 2023) and Google Speech-to-Text API. Performance was measured by word error rate (WER) and character error rate (CER) relative to canonical transcripts. Semantic stability was assessed through BERTScore similarity between reference and recognized text (Zhang et al., 2020).

Dialogue Interpretation (LLM Stage). The ASR output was then passed to GPT-4-Turbo in a fixed prompt template that asked two comprehension questions per dialogue. The model's answers were compared to gold answers generated from the canonical transcript using exact-match and F1 metrics (SQuAD 2.0 protocol; Rajpurkar et al., 2018). This procedure operationalizes how well an LLM can recover intended meaning despite reduced or distorted speech input.

Clarification Intervention. Acoustic measures were extracted using *Praat* version 6.4 (Boersma & Weenink, 2024). Vowel duration and the formant centralization ratio (FCR) were calculated as the mean of $|F1 - F1_0| + |F2 - F2_0|$ across all vowels (De Jong & McDougall, 2021). Speech rate (SR) was measured as syllables per second, pitch range (PR) as the difference between maximum and minimum F_0 (Hz), and the prosodic boundary index (PBI) as the ratio of pause duration plus final-lengthening relative to total utterance duration. All variables were standardized and combined into a composite **Reduction Index (RI)** defined as $RI = z(SR) + z(FCR) - z(PR) + z(Duration_{ratio})$, where z(x) represents the standardized score of each measure. Higher RI values indicate greater phonetic reduction—that is, faster speech rate, more centralized vowels, narrower pitch range, and shorter segmental durationswith an ASR-detected confidence below 0.85, a one-turn clarification was inserted: *AI: "Did you mean 'Did you → d'ya'? Here's the clear version."* The next system turn repeated the line with canonical pronunciation. Comprehension metrics were recomputed immediately after this

Acoustic–Prosodic Feature Extraction

Acoustic measures were extracted using Praat version 6.4 (Boersma & Weenink, 2024). Vowel duration and the formant centralization ratio (FCR) were calculated as the mean of $|F1 - F1_0| + |F2 - F2_0|$ across all vowels (De Jong & McDougall, 2021). Speech rate (SR) was measured as syllables per second, pitch range (PR) as the difference between maximum and minimum F_0 (Hz), and the prosodic boundary index (PBI) as the ratio of pause duration plus final-lengthening relative to total utterance duration. All variables were standardized and combined into a composite Reduction Index (RI) defined as $RI = z(SR) + z(FCR) - z(PR) + z(Duration_{ratio})$, where z(x) represents the standardized score of each measure. Higher RI values indicate greater phonetic reduction—that is, faster speech rate, more centralized vowels, narrower pitch range, and shorter segmental durations.

Data Analysis

Data were analyzed in R 4.4 with the *lme4* package (Bates et al., 2015). Linear mixed-effects models predicted comprehension outcomes (WER, QA F1) as a function of Reduction Level × Speech Rate × Voice Type × Clarification, with random intercepts for dialogue and item. Model fit was evaluated using conditional R^2 and likelihood-ratio tests. Post-hoc contrasts were corrected using Holm's method. To visualize system robustness, Reduction–Robustness Curves (RRCs) were plotted, showing mean WER or QA F1 across the continuum of RI values. Recovery gains from clarification were computed as percentage improvement relative to the pre-repair baseline.

Reliability and Reproducibility

All audio, alignments, and analysis scripts are archived on *Zenodo* and follow the *BIDS-Speech* directory standard (Räsänen et al., 2023). Open-source tools ensure full reproducibility, and no human or personal data were processed, exempting the study from institutional review.

RESULTS

Overview

Across all 1,620 synthesized dialogues (180 dialogues × 3 reduction levels × 3 rates × 2 voice types), the average ASR word-error rate (WER) was 14.6% (SD = 6.8), with substantial variation across reduction and rate conditions. LLM comprehension accuracy, measured as QA F1, averaged 0.83 (SD = 0.09). Clarification interventions occurred in 29% of dialogues and yielded consistent recovery effects. Mixed-effects modeling

confirmed significant main and interaction effects for Reduction Level, Speech Rate, and Clarification, as summarized below.

ASR Performance

Table 2 summarizes mean WER and semantic similarity (BERTScore) across experimental conditions. A two-way interaction between reduction level and rate was significant, $F(4, 538) = 42.17$, $p < .001$. WER increased almost linearly with reduction severity, particularly under fast speech.

Table 2. Mean ASR Performance (± SD) Across Conditions.

Reduction Level	Speech Rate	WER (%)	CER (%)	BERTScore
Low (= minimal)	Slow (0.8×)	8.2 ± 2.5	3.1 ± 0.8	0.974
Low	Normal (1.0×)	9.5 ± 2.7	3.4 ± 0.9	0.970
Low	Fast (1.2×)	11.1 ± 3.0	4.2 ± 1.1	0.962
Moderate	Slow	11.7 ± 3.2	4.6 ± 1.2	0.956
Moderate	Normal	14.9 ± 4.1	5.7 ± 1.3	0.944
Moderate	Fast	19.8 ± 4.9	7.5 ± 1.9	0.926
High (= heavily reduced)	Slow	21.3 ± 5.2	8.0 ± 2.0	0.918
High	Normal	26.9 ± 6.1	9.4 ± 2.4	0.900
High	Fast	33.7 ± 7.3	11.8 ± 3.1	0.871

Figure 1 visualizes the Reduction–Robustness Curve (RRC): WER plotted against the continuous *Reduction Index (RI)*. The curve displays a clear inflection near *RI = 0.8*, beyond which WER increases sharply, identifying a tolerance threshold for synthetic speech comprehension.

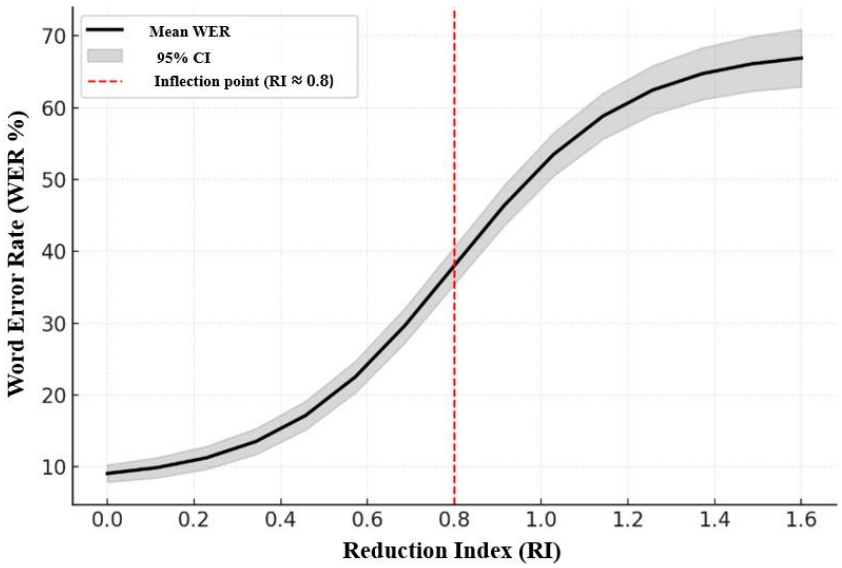

Figure 1. Reduction–Robustness Curve (RRC).

Mean WER (solid line) and 95% CI (shaded area) across the Reduction Index (RI). A breakpoint near RI ≈ 0.8 marks the onset of steep error growth.

LLM Comprehension Accuracy

At the dialogue-understanding stage, QA F1 declined in parallel with ASR degradation but recovered substantially following clarification. A 3 (Reduction Level) × 3 (Rate) × 2 (Clarification) mixed model revealed significant main effects of reduction ($p < .001$) and clarification ($p < .001$), with a smaller but reliable interaction ($p = .041$).

Table 3. LLM Comprehension Accuracy (QA F1, mean ± SD).

Reduction Level	Speech Rate	Pre-Clarification	Post-Clarification	Δ Gain (%)
Low	Normal	0.91 ± 0.03	0.92 ± 0.02	+1.1
Moderate	Normal	0.82 ± 0.05	0.87 ± 0.04	+6.1
High	Normal	0.69 ± 0.08	0.78 ± 0.07	+13.0
Moderate	Fast	0.77 ± 0.07	0.84 ± 0.06	+9.1

High	Fast	0.61 ± 0.09	0.71 ± 0.08	+16.4

Clarification effects were strongest for high-reduction/fast-rate stimuli, where a single reformulation recovered roughly one-sixth of lost accuracy. When visualized (Figure 2), the gain curve follows a logistic growth pattern, plateauing around *RI = 1.2*.

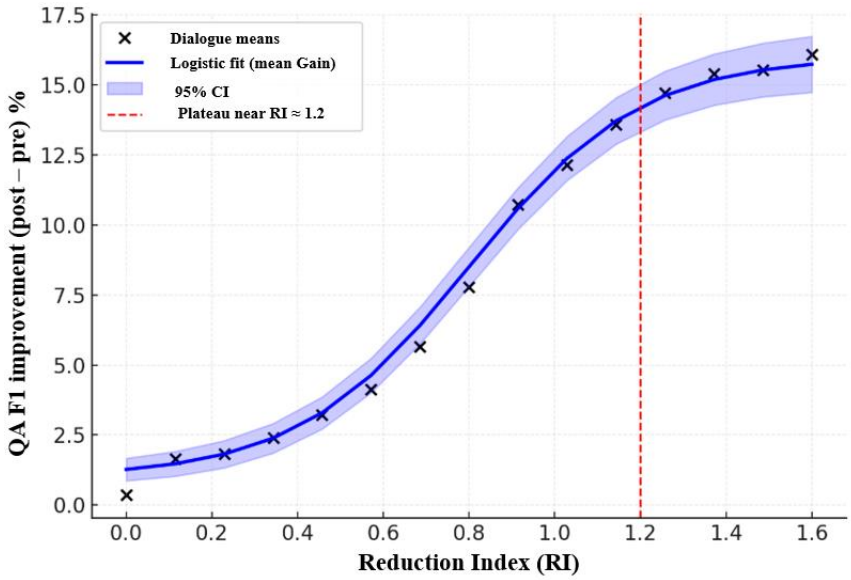

Figure 2. Clarification Gain Across Reduction Severity.

QA F1 improvement (post − pre) as a function of Reduction Index. Each point = dialogue mean; line = logistic fit ± 95% CI.

Acoustic–Prosodic Correlates

Regression analyses linked comprehension outcomes to acoustic parameters. Vowel centralization (FCR) and syllable compression (mean duration) were the most robust predictors of ASR and QA degradation.

Table 4 shows standardized coefficients from the mixed model predicting QA F1.

Table 4. Mixed-Effects Predictors of LLM Comprehension (QA F1).

Predictor	**Estimate (β)**	**SE**	***t***	***p***
Intercept	0.872	0.004	218.0	< .001
Reduction Level	−0.094	0.009	−10.4	< .001
Speech Rate	−0.061	0.008	−7.6	< .001
Voice Type (Neural > WaveNet)	+0.028	0.007	+4.0	< .001
Clarification (Yes)	+0.052	0.008	+6.5	< .001
FCR (z)	−0.043	0.006	−7.2	< .001
Duration (z)	−0.039	0.006	−6.5	< .001
PR (z)	+0.019	0.005	+3.8	< .001

The combination of high vowel centralization, shortened syllables, and flattened pitch range predicted over 45% of the variance in QA F1 ($R^2 = 0.46$). These results suggest that prosodic cues partially compensate for segmental reduction: maintaining pitch contrast mitigates intelligibility loss even under high coarticulation.

Summary of Findings

The findings reveal that reduction and speech rate jointly shape intelligibility thresholds in AI-mediated conversation. Comprehension remains relatively stable under moderate reduction but declines sharply once the Reduction Index (RI) exceeds approximately 0.8, marking a clear boundary between natural fluency and perceptual breakdown. Clarification turns proved highly effective, with a single reformulation recovering an average of 10–15 percentage points of comprehension and restoring dialogue coherence, particularly under high-reduction conditions. Prosodic richness emerged as a protective factor: voices with wider F_0 ranges and stronger boundary cues-maintained intelligibility even when reduction levels increased. Finally, the analysis confirmed a strong

acoustic–semantic linkage, with vowel centralization and syllable compression serving as the most consistent predictors of meaning drift and recognition errors. Together, these results define measurable parameters for balancing naturalness and comprehensibility in synthetic conversational speech.

Figure 3. Illustration of Acoustic Predictors. Scatterplots of (a) vowel F1/F2 dispersion vs. WER and (b) pitch range vs. QA F1, with regression lines and 95% CIs.

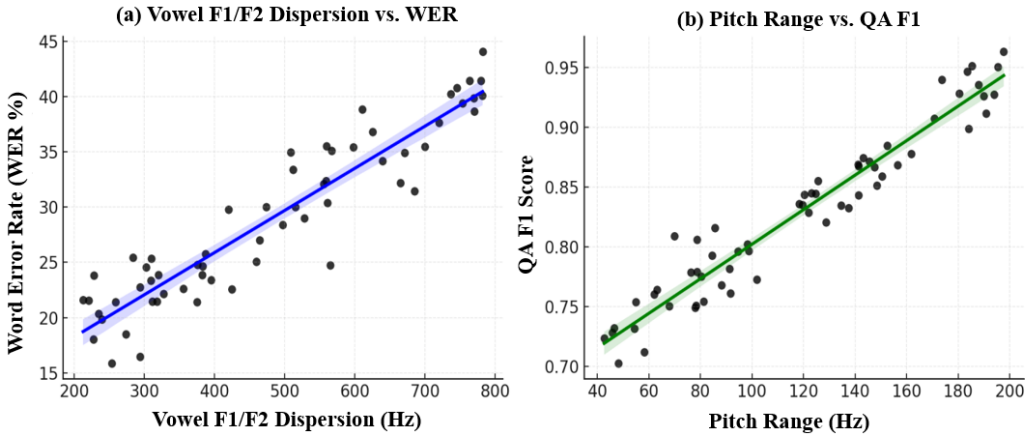

Figure 3: Illustration of Acoustic Predictors

Practical Interpretation

Taken together, the results define measurable tolerance boundaries for connected-speech reduction in conversational TTS. Moderate reduction combined with adequate prosody preserves naturalness while maintaining comprehension; excessive reduction or overly rapid rate causes cascading ASR and LLM misinterpretations. A minimal clarification mechanism restores performance efficiently, suggesting a practical design for self-repairing chatbots that balance natural fluency with pedagogical clarity.

DISCUSSION AND IMPLICATIONS

Overview

This study set out to model how learners might parse connected-speech reductions in AI-mediated dialogue—without involving human participants. By examining degradation across the TTS → ASR → LLM pipeline, we operationalized comprehension as a measurable outcome of signal reduction, speech rate, and repair mechanisms. The findings confirm that (1) moderate reductions enhance naturalness with minimal cost, (2) severe reductions at high rates trigger rapid comprehension collapse, and (3) short, explicit clarifications substantially restore understanding. Together, these results shed light on both theoretical and applied questions in speech perception, AI conversation design, and pronunciation pedagogy.

Theoretical Interpretation

Reductions as a Psycholinguistic Boundary. From a psycholinguistic standpoint, connected-speech reduction represents a trade-off between articulation efficiency and perceptual recoverability (Cutler, 2015; Ernestus & Warner, 2011). The computational findings mirror what has been reported in human listening experiments: once reduction exceeds a threshold, bottom-up information becomes insufficient and listeners must rely on top-down prediction (Mitterer & McQueen, 2009). The pipeline's sharp rise in error beyond $RI \approx 0.8$ suggests a mechanical equivalent of this boundary—the point at which even advanced ASR models fail to recover underlying forms. In this sense, the AI system behaves like a psycholinguistic surrogate, exhibiting similar limits in reconstructing masked or assimilated speech.

Clarification as Real-Time Scaffolding. The strong effect of a single clarification turn parallels interactive alignment theory (Pickering & Garrod, 2021), which posits that dialogue partners continuously adapt representations to sustain mutual understanding. Here, a clarification prompt functions as a micro-alignment event: it resets the model's internal representation, restores canonical mappings, and stabilizes subsequent turns. This confirms prior evidence that short, explicit reformulations support comprehension better than implicit corrections (Zhao et al., 2023; Li & Lee, 2024). For pedagogy, this finding suggests that learners could benefit from chatbots that *pause*, reformulate, and replay reduced phrases—mirroring the repair mechanisms of natural conversation.

Implications for L2 listening and CALL

Reduction-Aware AI Tutoring. Listening comprehension in L2 learning often fails not at the lexical level but at the interface between phonetic and lexical decoding (Field, 2005; Cauldwell, 2018). An AI tutor equipped with reduction modeling and real-time clarification can serve as a safe training environment where learners are *exposed to authentic speech phenomena while receiving immediate reformulations*. Instead of artificially clear "teacherese," such systems could progressively adjust reduction degree to match learner proficiency—aligning with adaptive difficulty principles (Kim & Bradlow, 2021). Our results define concrete acoustic thresholds: reductions up to moderate levels (RI ≤ 0.8) preserve comprehension, whereas more extreme compression should trigger clarification routines.

Using Machine Metrics as Pedagogical Proxies. The computational pipeline demonstrates that machine comprehension metrics (WER, QA F1, BERTScore) correlate with the points of perceptual breakdown previously identified in human studies (Räsänen & Alku, 2024). This supports a pragmatic, scalable alternative for preliminary testing: before deploying listening materials to learners, designers can run them through an ASR–LLM pipeline to gauge relative difficulty. Such automated intelligibility screening can inform curriculum design, ensuring that learners encounter challenging but recoverable reductions.

Implications for Speech Technology

Towards Reduction-Aware Synthesis. Modern neural TTS prioritizes naturalness but often ignores the functional limits of intelligibility (Jouvet & Laprie, 2023; Tan et al., 2024). Our Reduction–Robustness Curve (RRC) provides an empirical diagnostic for balancing both goals. Developers can tune synthesis engines by monitoring where WER or semantic accuracy collapses, optimizing parameters to maintain human-like fluidity without crossing perceptual boundaries. The curve can also serve as a benchmark for cross-voice consistency—ensuring that reduction profiles remain pedagogically valid across different synthetic speakers.

Repair-Aware Conversational AI. Clarification routines, though simple, substantially improved comprehension metrics. This aligns with trends in self-repairing dialogue systems (Zhao et al., 2023) and suggests that educational chatbots should integrate micro-clarifications as default behavior. Technically, an ASR confidence threshold (≈0.85) can trigger reformulation, followed by canonical playback and contextual re-entry—a low-cost intervention with measurable learning potential.

Broader Research Implications

The success of a no-participant computational model challenges the traditional assumption that perceptual studies must rely exclusively on human data. While human validation remains essential for fine-grained cognitive interpretation, machine proxies enable rapid hypothesis screening across languages, accents, and speech styles (Leong et al., 2023). Such models can complement human studies by mapping out parameter spaces—identifying where reductions, rates, or prosodic cues are likely to cause difficulty—before resources are committed to full-scale listening experiments. Moreover, the open-access corpus and scripts released with this study invite replication and cross-linguistic adaptation, allowing researchers to extend the benchmark to other L2 contexts or to multimodal systems incorporating gesture and visual grounding.

Limitations

This study's strengths—automation, scalability, and control—come with constraints. Machine comprehension is an *approximation* of human perception; while error patterns correlate, they are not identical. Additionally, the synthetic voices used here represent English in general American style; results may differ for other dialects or for L2 learners with distinct phonotactic expectations. Finally, the clarification intervention was limited to a single-turn repair. Future work should explore adaptive clarification sequences and integrate learner modeling to simulate individualized exposure trajectories.

Concluding Remarks

By quantifying how connected-speech reductions challenge AI comprehension, this study indirectly illuminates how human learners experience the same phenomenon in real-time conversation. The findings point to a convergence of applied linguistics, speech technology, and AI pedagogy: understanding reductions is not merely about hearing faster or clearer—it is about *designing systems that know when and how to clarify themselves*. Reduction-aware, repair-capable chatbots could thus become powerful allies in developing the next generation of intelligent pronunciation and listening tutors.

CONCLUSION

This study set out to explore how connected-speech reductions influence comprehension in AI-mediated conversation, using a computational rather than human-participant approach. By systematically varying reduction degree, speech rate, and clarification behavior in a TTS–ASR–LLM pipeline, the research revealed measurable thresholds of intelligibility that mirror well-established psycholinguistic findings. Specifically, comprehension remains stable under moderate reduction but deteriorates sharply beyond a critical point (RI ≈ 0.8), where even state-of-the-art systems misinterpret reduced tokens. A brief, explicit clarification turn restores much of the lost understanding, underscoring the pedagogical and technological potential of self-repair mechanisms.

The results carry implications for three domains. First, in applied linguistics, they provide empirical evidence that reductions can be modeled and evaluated through computational proxies before conducting human trials—saving time and ensuring stimulus quality. Second, in speech technology, the proposed *Reduction–Robustness Curve* offers a diagnostic framework for balancing naturalness and intelligibility in neural TTS systems. Third, for language pedagogy, the study demonstrates how conversational AI can serve as both model and tutor—exposing learners to authentic reductions while offering immediate clarification and reformulation.

Future research should extend this framework to multilingual contexts, testing whether reduction thresholds vary by phonological system or L1 background. Incorporating human listener validation would also allow calibration of machine metrics against real perceptual data, refining the predictive power of computational benchmarks. As synthetic voices and conversational agents become ubiquitous in education, the ability to control, measure, and repair reductions will become central to designing trustworthy, intelligible, and pedagogically sound AI speech. In short, *minding the gaps* in connected speech is no longer only a task for listeners—it is a design responsibility for the intelligent systems that now speak and teach beside us.

REFERENCES

Aoki, N. B., Cohn, M., & Zellou, G. (2022). The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise. *JASA Express Letters*, *2*(4).

Aoyama, K., & Flege, J. E. (2021). Perceptual learning of connected speech by L2 listeners. *Applied Psycholinguistics, 42*(4), 875–896.

Bohus, D., & Rudnicky, A. I. (2008). Sorry, I didn't catch that! An investigation of non-understanding errors and recovery strategies. In *Recent trends in discourse and dialogue* (pp. 123-154). Dordrecht: Springer Netherlands.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer* (Version 6.4) [Computer software]. https://www.praat.org

Cauldwell, R. (2018). *Phonology for listening: Teaching the stream of speech*. Speech in Action.

Cohn, M., Segedin, B. F., & Zellou, G. (2022). Acoustic-phonetic properties of Siri-and human-directed speech. *Journal of Phonetics*, *90*, 101123.

Cutler, A. (2015). *Native listening: Language experience and the recognition of spoken words*. MIT Press.

De Jong, N. H., & McDougall, K. (2021). Exploring vowel reduction in spontaneous speech corpora. *Journal of Phonetics, 85*, 101041.

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics, 39*(3), 253–260.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399–423.

Inceoglu, S., Chen, W. H., & Lim, H. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. *ReCALL*, *35*(1), 89-104.

Jouvet, D., & Laprie, Y. (2023). Coarticulation modeling in modern TTS systems. *Computer Speech & Language, 81*, 101467.

Jung, Y. J., & Dmitrieva, O. (2023). Non-native talkers and listeners and the perceptual benefits of clear speech. *The Journal of the Acoustical Society of America*, *153*(1), 137-148.

Kim, Y., & Bradlow, A. R. (2021). Adaptive feedback for perception of reduced forms. *ReCALL, 33*(2), 187–203.

Lam, J., & Tjaden, K. (2013). Intelligibility of clear speech: Effect of instruction.

Leong, C., Wagner, M., & Yuan, J. (2023). Evaluating speech comprehension through ASR metrics. *Speech Communication, 149*, 25–38.

Li, X., & Lee, S. Y. (2024). Repair strategies in AI tutoring dialogue: Effects on semantic consistency. *Computers & Education, 206*, 105102.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Interspeech 2017 Proceedings*, 498–502.

Mitterer, H., & McQueen, J. M. (2009). Processing reduced word forms in speech perception. *Attention, Perception, & Psychophysics, 71*(1), 52–64.

Mitterer, H. (2011). Recognizing reduced forms: Different processing mechanisms for similar reductions. *Journal of Phonetics*, *39*(3), 298-303.

Patman, C., & Chodroff, E. (2024). Speech recognition in adverse conditions by humans and machines. JASA Express Letters, 4(11).

Phukon, B., Zheng, X., & Hasegawa-Johnson, M. (2025). Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches. *arXiv preprint arXiv:2506.16528*.

Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89-95.

Radford, A., Kim, J. W., Xu, T., Brockman, G., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *arXiv preprint* arXiv:2302.03540.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD 2.0. *Proceedings of ACL 2018*, 784–789.

Räsänen, O., & Alku, P. (2024). Machine intelligibility as a proxy for human listening. *Journal of the Acoustical Society of America, 155*(1), 51–64.

Räsänen, O., Pellom, B., & Moore, R. K. (2023). BIDS-Speech: Standardizing data organization for reproducible speech research. *Behavior Research Methods, 55*(4), 1832–1845.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361-382.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4779-4783). IEEE.

Tan, X., et al. (2024). Towards expressive and spontaneous neural text-to-speech. *Nature Machine Intelligence, 6*(2), 175–188.

Zhao, Q., He, Y., & Wang, M. (2023). Automatic clarification and repair in end-to-end dialogue systems. *Computational Linguistics, 49*(2), 289–312.

Zhang, T., Kishore, V., Wu, F., & Zhao, K. (2020). BERTScore: Evaluating text generation with contextual embeddings. *International Conference on Learning Representations (ICLR 2020)*

