SpeechSpectrum: A Linguistic Fidelity Spectrum for Accountable Speech-to-Text

ANONYMOUS AUTHOR(S)

Speech-to-text (STT) systems are increasingly embedded in everyday technologies, yet they largely continue to treat transcription as a technical problem of accuracy, assuming a single "correct" representation of speech. This overlooks that speech can be transcribed in multiple legitimate ways, and that different contexts demand different balances of fidelity, conciseness, and emphasis. We contribute *SpeechSpectrum*, a framework reconceptualizing STT as cross-modal translation along a continuum of representational fidelity that makes these representational decisions explicit and user-controllable. Through theoretical analysis and empirical investigation, we show that existing STT systems already impose spectrum-based choices without user input, indicating the normative significance of who controls transcription outcomes. Our user study (N=52) demonstrates that granting users explicit control over transcript representation improves task support, while a comparative study shows that large language models fail to capture the diversity and context-sensitivity of human preferences. We derive implications and recommendations for building STT systems that prioritize user agency in representational decisions, and release open-source code – including the `speechspectrum` Python package – and a prototype to support future research. Our work positions control over transcription fidelity as a core site of user agency in speech technologies, and shows that system-imposed defaults constitute an accountability gap.

CCS Concepts: • **Human-centered computing** → **Interaction paradigms**; **Accessibility**; • **Computing methodologies** → **Speech recognition**.

Additional Key Words and Phrases: speech-to-text, linguistic fidelity, automatic speech recognition, accessible technology

ACM Reference Format:

Anonymous Author(s). 2026. SpeechSpectrum: A Linguistic Fidelity Spectrum for Accountable Speech-to-Text. In *Proceedings of 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26).* ACM, New York, NY, USA, 44 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

When a user dictates a voice message, participates in a virtual meeting, or speaks to a voice assistant via a voice user interface, they engage with Speech-to-Text (STT) systems.[1] These STT systems transform the user's spoken words into written text. Yet this modality of translation – from the rich, temporal, and contextually embedded nature of speech to the standardized, persistent format of text – involves countless implicit decisions about what information to preserve, modify, or discard entirely [31, 78]. For example, should disfluencies like "um" and "uh" be removed, or kept because they can provide important information about a speaker's confidence level? And, how should stylistic differences be resolved? For example, *"w- what he was sayin"'* and *"what, what he was saying"* are both correct transcriptions, varying only in style [129, 133].

These stylistic differences reflect deeper questions about user preferences and contextual needs. Consider the diverse scenarios in which people use speech-to-text technology and their various requirements: A court stenographer

[1]Speech-to-Text (STT) refers both to the *systems* that perform speech-to-text translation, and the *task* of speech-to-text translation. STT encompasses Automatic Speech Recognition (ASR) plus downstream processing; while ASR denotes the core acoustic-to-text conversion, STT captures the full pipeline to user-facing output.

Manuscript submitted to ACM

documenting legal proceedings requires verbatim preservation of every utterance, including hesitations and false starts that may carry legal significance; A user dictating casual messages wants natural-sounding text that doesn't burden the recipient with the artifacts of spontaneous speech production; A d/Deaf person requires rich annotations (e.g., in Netflix captions [4, 14]) to fully understand conversational nuances. Each of these use cases demands a different balance between fidelity to the original speech signal and adaptation to the user's informational needs – yet current STT interfaces typically provide users with little control over this fundamental representational choice. This disconnect is not merely a technical limitation but a civic one: when systems impose representational choices without user input, they deny users the autonomy to access and comprehend spoken information in ways that meet their needs – an issue of particular consequence for populations relying on STT for basic communication access.

This disconnect between diverse user needs and how STT systems operationalize available capabilities results in a systematic denial of user agency, creating an accountability gap. This accountability gap is not merely a usability issue but a matter of algorithmic fairness: when systems impose representational choices without user input, they encode particular linguistic norms and communicative values while denying users control over how speech is rendered. The question of who controls transcription outcomes – system designers or users – has direct implications for whose speech patterns are accommodated and whose are normalized away. While STT systems have improved in technical accuracy, they remain oblivious to contextual factors determining whether transcripts serve user goals and individualized needs [167, 202]. The field has focused on reducing transcription errors while neglecting the important question of transcription purpose. We argue this misunderstands STT conversion as mechanical transcription task rather than cross-modal translation involving choices about representation and information structure. Throughout this paper, we distinguish "STT conversion" (technical signal transformation) from "STT translation" (the interpretive choices about fidelity and style).

Drawing from theoretical frameworks in linguistics of modality differences, we propose reconceptualizing STT output not as a single "correct" transcription, but as one point along a continuous spectrum of possible representations. This **linguistic fidelity spectrum** – where fidelity, borrowed from translation studies' distinction between source-oriented versus target-oriented translation [47, 139], refers to the degree of faithfulness to source material characteristics – ranges from highly compressed summaries that extract key semantic content to verbatim transcriptions that preserve much acoustic detail, with numerous intermediate points representing different balances between spoken language and written language conventions. Each point on this spectrum serves different user needs and contexts, and the optimal choice depends not on context-independent STT accuracy metrics like Word Error Rate (WER),[2] which assumes one correct transcription, but on the specific informational requirements of the user's task.

We introduce **SpeechSpectrum**, a framework that operationalizes this linguistic fidelity spectrum for STT system design. Rather than pursuing a one-size-fits-all approach to transcription, SpeechSpectrum envisions interfaces that give users explicit control over where their STT output shoul fall on this fidelity spectrum. Such systems treat representation level as a designable parameter, allowing users to navigate between different information densities and linguistic conventions as their needs require. This approach not only better serves individual users but also addresses broader questions of algorithmic fairness by making visible the representational choices that are currently hidden within system architectures.

We contribute: **SpeechSpectrum (§3)**, a continuum-based framework for understanding STT conversion, validated through theoretical analysis and empirical investigation; **case studies (Appendix B)** demonstrating how

[2]WER measures edit distance between predicted and reference transcripts: WER = $\frac{S+D+I}{N}$ (substitutions, deletions, insertions over total reference words).

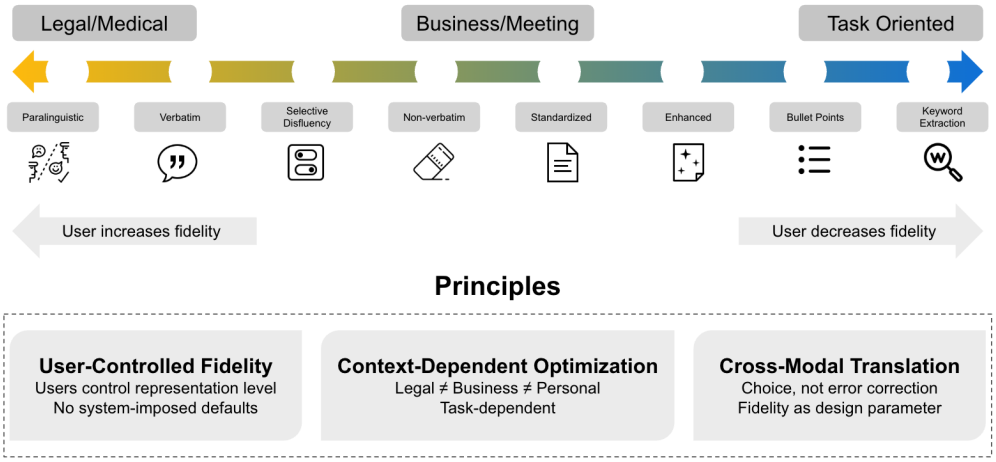


Fig. 1. **The SpeechSpectrum framework conceptualizes STT conversion as a continuous spectrum of verbatimicity levels rather than a single transcription target.** The gradient bar represents the continuum from high-fidelity paralinguistic preservation (left) to low-fidelity keyword extraction (right), with eight reference points showing commonly used representations. Domain indicators above show typical usage: Legal/Medical contexts require high fidelity, Business/Meeting contexts use moderate fidelity, and Task-Oriented applications prefer low fidelity. Bidirectional arrows emphasize user control over fidelity based on context and needs. Three core principles form SpeechSpectrum: (1) User-Controlled Fidelity – users specify their preferred representation level rather than accepting system defaults; (2) Context-Dependent Optimization – optimal fidelity varies by domain, task, and user goals; and (3) Cross-Modal Translation – STT is deliberate representational choice rather than minimizing transcription errors against a single correct target, treating fidelity as a designable parameter.

real-world STT systems implicitly implement spectrum-based choices through their design decisions; **user study evidence (§4)** that explicit control over transcript representation improves user experience; **LLM study findings (§4.2)** revealing that LLMs – increasingly proposed as proxies for human judgment in evaluation tasks – fail to accurately model diverse user preferences across fidelity levels; **design recommendations, R1–R9 (§4.3, Appendix G)** addressing conceptual challenges and providing concrete guidance on hybrid architectures, alignment-based metrics, data collection, and multimodal extensions; and **open-source resources (Appendix H)** including the experimental code (https://anonymous.4open.science/r/SpeechSpectrum-A3D4), the demo website for the user studies (https://speechspectrum.org/), and the `speechspectrum` Python package to translate transcripts along the fidelity spectrum.

In §2, we describe background on speech fidelity across diverse communicative contexts and user populations, examining how technical systems encode particular values about whose speech patterns are accommodated, highlighting gaps that SpeechSpectrum addresses. In §3, we introduce the SpeechSpectrum framework, and provide an overview of key components, from *paralingustic* transcription to *bullet point* translation. In §4, we present user studies demonstrating context-dependent fidelity preferences and LLM limitations in modeling human preferences. In §5, we conclude with implications for future speech technology research and design. We provide case studies of existing STT applications and technical implementation guidance in the appendix.

2 Background

A fundamental disconnect persists in how STT systems are studied, evaluated, and governed. Speech technology research has largely framed transcription as a technical problem of accuracy, optimizing for quantitative metrics while treating representational choices as neutral or incidental. In parallel, human-centered computing research has

often examined STT impacts without interrogating the technical assumptions and institutional defaults embedded in these systems, leaving affected users without visibility or the ability to contest representational decisions. **This divide obscures the normative significance of transcription decisions: choices about what counts as a 'correct' transcript encodes values, redistributes epistemic authority, and shapes downstream judgments in domains such as law, medicine, and accessibility.** We draw on AI ethics, HCI, speech technology, and NLP literatures to surface these hidden value judgments and to reframe STT not as a purely technical pipeline, but as a site of accountability where representational power is exercised.

Communication as a Spectrum. Linguistics recognizes that human communication exists along a spectrum rather than discrete spoken versus written categories, with numerous hybrid forms occupying intermediate positions [142]. This continuum justifies treating STT output as existing at different points of oral-literate mediation rather than pursuing a single 'optimal' representation. Translation studies [5] conceptualize STT conversion as cross-modal translation involving choices about fidelity and adaptation. Nida's [139] distinction between formal equivalence (preserving source language structures) and dynamic equivalence (preserving communicative effect) maps onto STT design: verbatim transcription prioritizes formal equivalence to speech, while cleaned output prioritizes dynamic equivalence for written consumption. This reveals that *quality* in STT cannot be defined without reference to intended function. Treating fidelity as a designable parameter highlights that these decisions are normative: selecting one representation over another determines which aspects of speech are preserved, erased, or institutionalized. Among many fields, computational linguistics demonstrates that disfluencies (speech production features like 'um,' 'uh,' false starts, and repetitions, which we detail in §Appendix A) are not errors to be corrected but meaningful features serving communicative functions [17, 44, 54, 182, 183]. Hesitations signal processing difficulty; false starts and repairs reveal real-time negotiation; filled pauses serve discourse management. As a result, disfluency removal systematically suppresses cues related to uncertainty, confidence, and agency – features that can be consequential in evaluative and institutional settings. **Our contribution in this domain is to propose *SpeechSpectrum*, a framework reconceptualizing STT as cross-modal translation along a continuum of representational fidelity.**

Speech Interface Design. Voice interface research [96, 135] has focused on naturalness and intent recognition, implicitly treating STT conversion as black box pre-processing. This works for command-based interactions, but breaks down when users need to review, edit, or reference the textual output of their spoken interactions. Current interfaces provide users with minimal visibility into speech interpretation and no control over representational choices. Despite recognition of user diversity, systems provide users with minimal control over the linguistic representation of their speech. Most commercial services include profanity filtering [68] and punctuation insertion [69] with limited disfluency removal [156]. RevAI offers verbatim and non-verbatim transcription [24], but few services provide finer-grained fidelity control, representing a significant gap in user agency.

Research on personalization in speech interfaces remain limited [157, 171], despite evidence that speech recognition systems perform worse for speakers from marginalized communities [99, 130, 213]. Work on user-specific STT models [26, 206] typically focuses on improving contextual accuracy [53] rather than allowing users to specify representational preferences. Post-processing work on enhancing STT output readability [115] similarly doesn't address stylistic transcription. The accessibility community has made the most progress recognizing diversity in transcript preferences. Live captioning research [18, 41, 103] typically examines surface-level preferences (font size, timing) rather than fundamental representational questions. Meeting transcription tools have revealed preferences for different detail depending on context. Users require varying summary lengths [59], indicating no single approach serves all use cases even within domain. **Our contribution is a set of agency-forward design recommendations (R1-R9) for STT systems.**

Speech Transcription Choices. WER's dominance in STT evaluation reflects a conception of transcription as mechanical reproduction rather than representational choice [3, 25, 58, 74, 130, 134, 169, 184]. This metric family (detailed in §subsection G.2) traditionally assumes a single "correct" transcription without justifying why one choice should be privileged or how such privileging can be considered legitimate across contexts – a result of engineering limitations rather than principle. This assumption becomes problematic when evaluating contextually appropriate representations. Consider "I, I think we should go" versus "I think we should go." Against a reference of "I think we should go," the first yields a higher WER due to repetition, yet for a legal professional assessing speaker confidence, the disfluent version may be more valuable. Conversely, for a business meeting summary, the cleaned version better serves user needs. WER's singular ground truth cannot accommodate this contextual variation. Recent work acknowledges that multiple valid transcriptions exist for the same speech [57, 95, 98, 129, 151, 161, 177], but offers limited systematic alternatives accounting for user context.

The STT community has implicitly recognized our argument through domain-specific systems. Medical STT [39, 40, 181] optimizes for different features – e.g., verbatim features are needed to diagnose a fluency disorder – than conversational STT systems [52, 123, 150]. Legal STT [102, 117, 158] preserves disfluencies general-purpose systems remove, while meeting tools [36, 165, 174] incorporate summarization inappropriate for forensic applications [121]. However, these approaches are framed as separate technical problems rather than instances of a broader fidelity design space. The ubiquity of post-processing also provides strong evidence. Text normalization, punctuation restoration, and disfluency removal – standard but inconsistently implemented across services without norms [130] – all represent spectrum movements, yet are treated as separate technical problems, obscuring the insight that they are collectively implementing a continuum of representations.

Recent work by Teleki et al. [186] and Mei et al. [130] provides compelling evidence. Their STT comparisons reveal that platforms by design preserve or remove different disfluencies, placing outputs at different spoken-written continuum points. **Our contribution is to conduct a user study to assess the usefulness of our proposed SpeechSpectrum framework. We demonstrate that optimal fidelity choices vary by user expertise, content type, and downstream task, directly challenging the assumption that a single representation can be justified as universally correct.**

3 The SpeechSpectrum Framework

Current STT paradigms suffer from three fundamental flaws that limit their ability to serve diverse user needs. First, the notion of "accuracy" remains acontextual, assuming that technical fidelity to some a single predetermined ground truth constitutes meaningful performance regardless of user goals or application context. Second, STT research exhibits pervasive linguistic naivety, treating STT conversion as mechanical reproduction rather than the complex process of cross-modal translation that it actually represents. Third, there exists a profound evaluation disconnect between the technical metrics that dominate STT research and the actual value that users derive from these systems in practice.

These limitations stem from a particular conceptual orientation: treating STT as transcription rather than translation. We propose instead understanding STT conversion as modality translation along a continuous spectrum of representational fidelity that compress, transform, and restructure information from the original speech signal according to different communicative purposes and user needs. SpeechSpectrum is a framework that prioritizes user agency in order to effectively meet widely varied user information needs with respect to spoken content. Critically, all representations along SpeechSpectrum are derived from the original speech signal: all representations must be derivable from what was actually said, without adding extraneous information, inferring unstated meanings, or incorporating external context. The framework navigates how to represent spoken content, not whether to

augment it. Rather than requiring users to choose between separate specialized tools (e.g., transcription services for verbatim output, summarization tools for condensed content), SpeechSpectrum enables users to navigate fidelity levels within a unified interface, making representational trade-offs explicit and controllable. A visual representation of SpeechSpectrum is shown in Figure 1, and an example transcript and its representations along SpeechSpectrum are shown in Figure 5.

3.1 A Continuum for Representing Speech-to-Text

We introduce the concept of **verbatimicity** (our operationalization of fidelity for STT conversion) – the degree to which textual output preserves the structural, lexical, and paralinguistic characteristics of the original speech signal. Unlike binary notions of accuracy (detailed more in §Appendix G), verbatimicity operates along a continuous spectrum that encompasses multiple dimensions of fidelity simultaneously, from prosodic preservation to information compression. Below we introduce the three foundational principles that guide how users navigate the verbatimicity spectrum as indicated in Figure 1.

User-Controlled Fidelity. Central to SpeechSpectrum is the principle that users should control where their STT output falls along the verbatimicity spectrum. This user agency recognizes that optimal representation depends on context, purpose, and individual communicative needs that cannot be predetermined by system designers. A legal professional documenting testimony requires different verbatimicity than a business executive reviewing meeting highlights, yet current STT systems provide no mechanism for users to specify their representational preferences.

Context-Dependent Optimization. Different domains, tasks, and user goals demand fundamentally different approaches to transcript representation. Legal contexts may require high fidelity to preserve hesitations that may indicate witness uncertainty [22], while medical triage documentation may benefit from concise bullet points that highlight critical symptoms [175]. Meeting transcription serves different purposes for real-time note-taking versus post-hoc review, and accessibility applications must balance speed with information richness. Rather than optimizing for universal accuracy metrics, SpeechSpectrum systems should adapt their representational choices to the specific communicative context and user objectives.

Cross-Modal Translation. Movement along the verbatimicity spectrum involves systematic decisions about information preservation and transformation during cross-modal conversion from speech to text. At high verbatimicity levels, systems preserve paralinguistic information such as hesitations and prosodic markers that may indicate speaker certainty, emotional state indicating emphasis (e.g., numerical annotations for pitch contours) and emotional affect (e.g., [laughs], [sighs]) or processing difficulty. At lower verbatimicity levels, summarization prioritizes factual information extraction over subtle emotional indicators communicated by disfluencies (i.e. *um*, *uh*), such as a perceived lack of confidence. Each compression step involves implicit judgments about what constitutes "relevant" information, making these choices inherently political and contextually dependent [33].

3.2 Components

Along the SpeechSpectrum, lexical and nonlexical (e.g., *um/uh*) tokens can be used to produce different representations. Each level serves particular communicative functions and user needs that cannot be fully replaced by other positions on the spectrum. SpeechSpectrum conceptualizes representation as a continuous spectrum with multiple possible pathways rather than strictly linear sequence. While we present common fidelity levels in approximate order, the spectrum accommodates branching paths (e.g., selective disfluency preservation) and context-dependent navigation rather than requiring stepwise progression through all intermediate forms. The levels we describe here represent commonly used reference points rather than discrete categories; our prototype implements only a subset of widely recognized representations for practical user study purposes.

Paralinguistic. The highest verbatimicity level extends beyond textual transcription to include paralinguistic signals: emotional expressions (laughter, crying), physiological sounds (yawns, coughs), and prosodic patterns (pitch variation, tone, pacing, volume changes) [20]. Effortful speech indicates communication difficulty [83]; even *emojilization* incorporates paralinguistic information [81]. This level serves specialized contexts requiring maximal communicative context: discourse analysis, therapeutic interaction documentation, forensic applications. Research on accessible speech interfaces for d/Deaf and hard-of-hearing (d/DHH) populations [93] offers fine-grained environmental, emotional, and spatio-temporal information via paralinguistic signal [41, 94, 128]. LLMs can process these paralinguistic aspects [88, 114], with recent research integrating them into speech language model architectures [92, 110, 198].

Verbatim. The verbatim transcript is the most faithful textual version, comprehensively including disfluencies: fillers or filled pauses (*"um"*, *"uh"*), repetitions (*"I, I think"*), false starts (*"We should go- let's leave"*), and repairs (*"turn left, I mean right"*). These differ from informal contractions (*"gonna"*, *"wanna"*) or dialect variations (*"y'all"*), as disfluencies represent real-time speech production processes rather than stable linguistic choices. Technically, STT models under-transcribe disfluencies by design or due to limited training data [186]. A user-centric challenge with obtaining verbatim transcripts is that STT models must handle the phenomenon of *good-enough word selection* [100], wherein speakers choose words based on cognitive accessibility rather than semantic precision, potentially leading to transcripts that accurately capture imprecise speech rather than intended meaning. These transcripts offer valuable contextual information for DHH individuals, with fine-grained emotion conveyed via the provided disfluency signal [94, 128].

Selective Disfluency Preservation. Between verbatim and enhanced transcription lies a customizable middle ground where users can toggle preservation of specific disfluency types. Users might choose to preserve meaningful hesitations while removing filled pauses, or maintain false starts while eliminating repetitions. This granular control acknowledges that different paralinguistic features serve different communicative functions and may be relevant for different user purposes. These may represent branching paths from the main verbatimicity spectrum rather than simple linear progression.

Non-Verbatim. Disfluency removal creates more readable text while preserving lexical content and basic syntactic structure. This level serves users who need access to semantic content without the cognitive overhead of processing production artifacts. However, the cleaning process necessarily involves interpretation decisions about which features constitute "errors" versus meaningful linguistic choices, potentially reflecting bias against non-standard linguistic practices. For instance, the removal of double negatives may seem like grammatical correction, but in legal contexts, whether one "not" is preserved or dropped can fundamentally alter sentence meaning and ensuing legal interpretations.

Standardized. Moving further along the spectrum, standardized transcription converts informal speech patterns to conventional written forms, transforming reductions (e.g. *gonna* to *going to*), informal expressions, and colloquialisms into their standard equivalents. This level bridges the gap between conversational and formal registers while attempting to maintain the speaker's essential content and structure.

Enhanced. Enhanced transcripts involve deliberate post-processing that improves word choice, sentence structure, and coherence while preserving intended meaning. Parliamentary proceedings exemplify this, where transcribers add omitted protocol elements like 'Madame Speaker' [196]. This addresses use cases where speakers want polished versions of their spontaneous speech – such as lawyers preparing statements or professionals creating documentation from informal discussion.

Legal Domain Imagine you are a case judge reading through a deposition transcript
Q1: Did the defendant seem confident about the details of the crash?
Q2: What were the events leading up to the crash?
Medical Domain Imagine you are a doctor looking over a triage dictation provided by a nurse
Q3: What are the main symptoms the patient is exhibiting?
Q4: Has the chest pain been going on for exactly three days, or could it have been longer/shorter?
Business Domain Imagine you are a team leader reading a meeting transcript
Q5: Does the team seem like they will meet the December deadline?
Q6: What are the action items from the meeting?

Fig. 2. **User study questions across three domains (legal, medical, business)**; full text in §Appendix C.

Bullet Points. Bullet points are ultra-condensed summaries. High-compression approaches prioritize functional over formal equivalence, extracting key information while abandoning surface linguistic features. These representations serve task-oriented contexts where users need actionable information rather than detailed linguistic content. However, the summarization process inevitably reflects assumptions about what information is "important," potentially marginalizing perspectives or concerns that don't fit dominant organizational narratives.

Keyword Extraction. At the extreme low-verbatimicity end, keyword extraction reduces speech to essential terms and concepts, serving contexts where users need rapid content identification or indexing capabilities rather than readable text.

Each level thus involves trade-offs between information preservation and usability, fidelity and accessibility, linguistic expression and standardized norms. The optimal choice depends entirely on user context, purpose, and values – decisions that only users themselves can make appropriately.

Having established the theoretical foundation of SpeechSpectrum and its components, we now turn to empirical validation. While existing STT applications implicitly operate at different points along this spectrum (detailed in §Appendix B), the question remains whether users would benefit from explicit control over these fidelity choices. The following section presents two complementary studies examining this question.

4 Empirical Studies

4.1 Study 1: User Study of Contextual Fidelity Preferences

4.1.1 Designing the SpeechSpectrum Prototype. To make the SpeechSpectrum framework concrete for empirical evaluation, we designed a simplified prototype that served as a research probe rather than a full end-user system. The prototype instantiated four representative transcript fidelities – Verbatim, Non-Verbatim, Enhanced, and Bullet Points – chosen to balance manageability for participants with coverage of the framework's conceptual space, allowing us to preserve the core principles of SpeechSpectrum while keeping the study tasks tractable.

We implemented the prototype as a lightweight web application to ensure accessibility and consistency across participants, shown in Figure 5. A web-based delivery lowered barriers to participation and guaranteed platform-independence: users could access the system through a standard browser without installing software. This decision was especially important for engaging less computationally-engaged professionals such as medical or legal experts, whose perspectives were central to evaluating domain-specific transcription needs.

The interface was organized around domain-specific scenarios – Legal, Medical, and Business – reflecting professional contexts where speech-to-text technologies are commonly applied. Participants could navigate across fidelity levels using a clickable "spectrum" interface and switch domains through a menu bar. Standard web libraries (CSS, Bootstrap, jQuery) were used to maintain visual consistency and responsiveness, but our primary design

goal was to surface representational trade-offs, not to demonstrate technical novelty. In this way, the prototype operationalized SpeechSpectrum as an interactive artifact, enabling us to empirically investigate whether different tasks and domains demand different points along the verbatimicity spectrum.

4.1.2 Study Design. We conduct a user study with N=52 participants through convenience sampling from academic and professional networks. We collect demographic data on participants' professional backgrounds (see Table 4)–beyond domain expertise, we collected information on participants' STEM and STT backgrounds to understand how technical familiarity with speech technologies might influence fidelity preferences. This allows us to examine whether domain experts, technical experts, and general users exhibit different navigation patterns across the spectrum. Participants engaged with the interactive SpeechSpectrum interface and completed scenario-based tasks. For each domain, we presented two distinct task scenarios requiring different information extraction approaches. Participants were asked to select which transcript version (Verbatim , Non-Verbatim , Enhanced , or Bullet Points) best supported answering each of the six independent questions shown in Figure 2; participants selected which transcript version best supported each task.

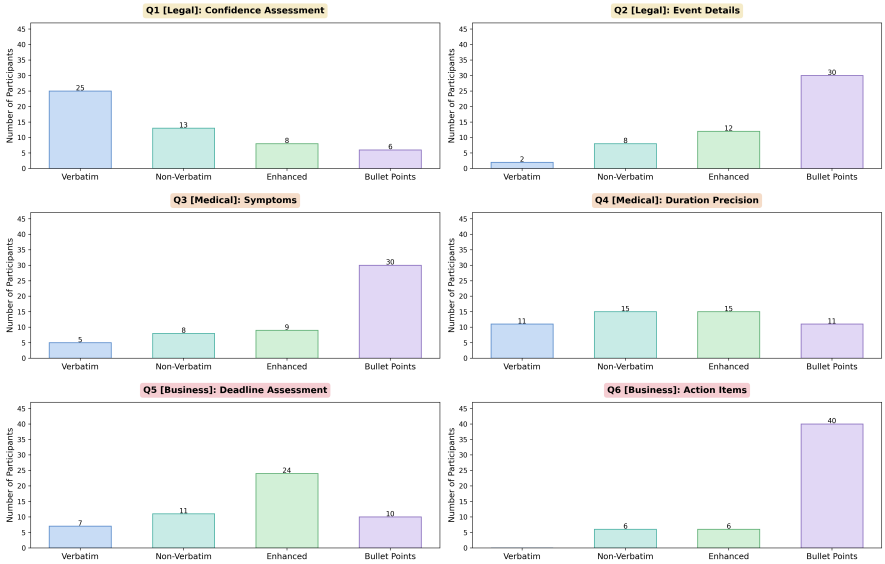
4.1.3 Are user preferences context-dependent? Our results, shown in Figure 3a, demonstrate clear evidence for context-dependent fidelity preferences, supporting the core argument for user-controllable representation. Participants revealed distinct preference patterns based on task requirements rather than universal preferences for higher or lower fidelity levels. Legal confidence assessment (Q1) favored Verbatim transcripts, while legal event details (Q2) preferred Bullet Points versions. Medical symptom identification (Q3) overwhelmingly chose Bullet Points , contrasting with medical duration precision (Q4) which showed more evenly-distributed preferences. Business deadline assessment (Q5) favored Enhanced transcripts, while action items (Q6) strongly preferred Bullet Points .

To evaluate whether human preferences differ from a uniform distribution across transcript representations, we applied a χ^2 goodness-of-fit test for each question Q_i, shown in Table 1. The χ^2 goodness-of-fit test evaluates departures from uniform preference distributions, while Cramér's V, derived from χ^2, provides a normalized measure of overall preference concentration. The dominance gap $\Delta(Q_i) = p_1 - p_2$ complements V by capturing the margin between the most- and second-most (p_1, p_2) selected representations. Together, these metrics differentiate broadly distributed preferences from cases with stronger local concentration.

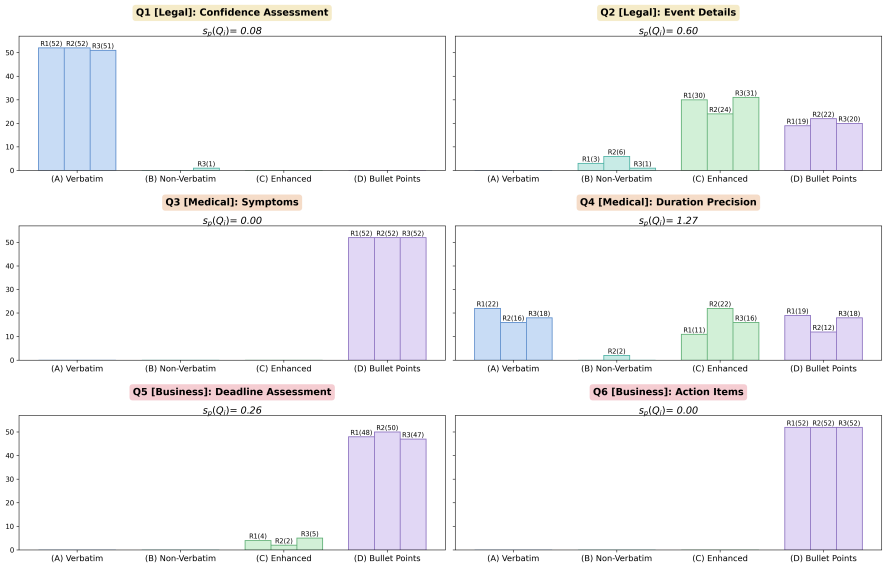
Human preference distributions are generally weakly concentrated, with small to moderate effect sizes ($V \leq 0.46$ in 5/6 questions) and modest dominance gaps ($\Delta(Q_i) \leq 0.40$ in 4.6 questions), indicating near-ties among competing representations rather than decisive single-choice dominance. However, in the Business domain, Q6 (*Action Items*) exhibits a strong concentration ($V = 0.70$, $\Delta = 0.65$), reflecting clear convergence on Bullet Points . Importantly, this contrast demonstrates that **while human preferences are generally distributed across multiple representations, sharper concentration can emerge for tasks with highly specific structural demands.**

The study results support our three framework principles. Participants demonstrated sophisticated reasoning about appropriate fidelity levels for different tasks, contradicting assumptions that users cannot make meaningful representational choices [23, 125]. Clear preference patterns emerged based on task requirements rather than participant demographics, demonstrating that optimal representation depends on use context rather than abstract accuracy metrics.

4.1.4 Are user preferences expertise-dependent? We examined whether domain expertise shapes fidelity preferences (detailed breakdown in Appendix Table 5). While sample sizes for specific expertise combinations are too small for robust statistical inference, we observed no strong systematic patterns, with substantial variation within each expertise group. Participant's open-ended responses illustrate this diversity; a medical expert noted:



(a) **Across six task scenarios, human participants selected different transcript versions depending on task needs, demonstrating that no single fidelity level universally serves all contexts.** For example, verbatim transcripts were preferred for confidence assessment in legal tasks (Q1), while bullet points dominated for identifying action items in business meetings (Q6). These results provide empirical support for SpeechSpectrum's core claim: transcript design should be treated as a context-dependent and user-controllable choice rather than as an optimization for a single accuracy metric.



(b) **When asked to complete the same six scenarios as human participants (Figure 3a), LLMs exhibited more extreme preference patterns, often converging on a single transcript type (e.g., consistently selecting bullet points for business tasks).** R1, R2, R3 indicate three experimental rounds with different random seeds to account for variability. This contrast highlights both the potential and the limitations of using LLMs as proxies for user evaluation in STT research. (See §4.2.1 – 4.2.2 for full statistical analysis details, and Appendix F.2 for an ablation study of the LLM temperature, τ.)

Fig. 3. **Comparison of human and LLM preference distributions across six task scenarios (Q1-Q6).**

Human vs. Uniform Preference Distributions						
Q_i [Domain]	χ^2	df	p	V	Most Freq. Representation (p_1)	$\Delta(Q_i)$ [$\Delta(Q_i) CI_{95\%}$]
Q_1 [Legal]	16.77	3	0.0008***	0.33	Verbatim	0.23 [0.02, 0.44]
Q_2 [Legal]	33.54	3	0.0000***	0.46	Bullet Points	0.35 [0.12, 0.56]
Q_3 [Medical]	30.31	3	0.0000***	0.44	Bullet Points	0.40 [0.17, 0.58]
Q_4 [Medical]	1.23	3	0.7456	0.09	Non-Verbatim / Enhanced	-
Q_5 [Business]	13.08	3	0.0045**	0.29	Enhanced	0.25 [0.04, 0.42]
Q_6 [Business]	76.62	3	0.0000***	0.70	Bullet Points	0.65 [0.44, 0.81]
LLM_{R1} vs. Uniform Preference Distributions						
Q_i [Domain]	χ^2	df	p	V	Most Freq. Representation (p_1)	$\Delta(Q_i)$ [$\Delta(Q_i) CI_{95\%}$]
Q_1 [Legal]	460.05	3	0.0000***	0.99	Verbatim	1.00 [1.00, 1.00]
Q_2 [Legal]	127.23	3	0.0000***	0.52	Enhanced	0.21 [0.02, 0.48]
Q_3 [Medical]	468.00	3	0.0000***	1.00	Bullet Points	1.00 [1.00, 1.00]
Q_4 [Medical]	47.64	3	0.0000***	0.32	Verbatim / Enhanced	0.06 [0.00, 0.29]
Q_5 [Business]	386.21	3	0.0000***	0.91	Bullet Points	0.85 [0.69, 0.96]
Q_6 [Business]	468.00	3	0.0000***	1.00	Bullet Points	1.00 [1.00, 1.00]

Table 1. **Human preferences are generally diffuse, often exhibiting near-ties among representations (mostly low-to-moderate Cramér's V, small $\Delta(Q_i)$), whereas LLM preferences are strongly concentrated with single-choice dominance.** χ^2 goodness-of-fit tests assess whether preference distributions differ from uniform (across verbatim, non-verbatim, enhanced, and bullet points), while Cramér's V – computed from χ^2 – provides a normalized, sample-size–robust effect size capturing the degree of overall preference concentration. The dominance gap $\Delta(Q_i) = p_1 - p_2$ complements V by quantifying the local margin between the most-preferred and runner-up representations. Together, V and $\Delta(Q_i)$ distinguish globally concentrated distributions from cases of decisive single-choice dominance.

> *It is nice to have the different output options. There are some situations [where] I would not want to dig through a verbatim dialogue in order to get some quick information.* [P20]

This highlights the value of concise representations for time-sensitive medical work.

A STT expert emphasized the practical benefits of mid-level fidelity:

> *Generally when reading a transcript, there a[re] rare cases in which the information I need is to see exact wording and thoughts, more often than not I need to know the general information, and the enhanced version fits well for that understanding in most cases.* [P11]

STEM professionals' responses also pointed to important trade-offs between detailed and summarized outputs. One participant explained:

> *...for some questions you need some of the more "soft" language aspects to help ascertain someones certainty, intent, etc. Like if someone is repeating themselves, stuttering, saying "I think", etc. Those are present in Verbatim (and somewhat Non-verbatim), but are largely missing in Enhanced/Bullet Points.* [P12]

A participant who did not prefer SpeechSpectrum outputs noted:

> *I prefer a full explanation in a person's own words in response to questions* [P13]

This shows how different fidelity levels highlight or suppress cues that matter for particular reasoning tasks.

Human vs. LLM_{R1} Preference Distributions					
Q_i [Domain]	χ^2	df	p	V	Top Choice Alignment
Q_1 [Legal]	36.47	3	0.0000***	0.59	✓
Q_2 [Legal]	14.46	3	0.0023**	0.37	✗
Q_3 [Medical]	27.90	3	0.0000***	0.52	✓
Q_4 [Medical]	21.42	3	0.0001***	0.45	✗
Q_5 [Business]	57.18	3	0.0000***	0.74	✗
Q_6 [Business]	13.57	2	0.0011**	0.36	✓

Table 2. Results of χ^2 goodness-of-fit analyses comparing preference distributions over four transcript representations, directly contrasting human and LLM distributions. Larger Cramér's V indicates greater distributional divergence between humans and the LLM. Results are shown for temperature, $\tau = 1.0$.

4.2 Study 2: Exploring LLMs as Proxies for Human Preferences

We conducted a follow-up study to examine whether LLMs can accurately model user preferences across fidelity levels. LLMs are increasingly proposed as scalable proxies for human evaluation in social science and NLP tasks [7, 64, 210], offering potential advantages for personalization: if LLMs could reliably predict which fidelity level suits different users and contexts, they could inform default settings or provide recommendations without requiring extensive human annotation. However, our results reveal important limitations to this approach, suggesting that while LLMs may be useful for generating candidate transcripts across the spectrum, preference modeling itself requires human judgment.

We created N=52 personas aligned with the respondents from our study, according to their self-identified professional expertise (see Appendix for preliminary study questions **P1-P3**), and add a *format instruction* to control the output format [162]:

> `Respond as a person who` **`[(P1) does/does not]`** `work in automatic speech recognition technology,` **`[(P2) does/does not]`** `work in STEM (science, technology engineering, mathematics), and` **`[(P3) has legal expertise/has medical expertise/does not have legal or medical expertise]`**`. Respond only with the letter for the answer choice.`

The format instruction corresponds to our mapping of the four categorical options onto ordinal values following a uniform distribution, which assumes an equidistant spacing of verbatimicity:

A=1 (Verbatim), B=2 (Non-Verbatim), C=3 (Enhanced), D=4 (Bullet Points)

For each persona, we asked **Q1-Q6** from our user study (detailed in Appendix 4.1.2). For each question, we concatenate the text data to the prompt as an alternative to UI-based interaction as in our user study (detailed in Appendix F). We prompt with the same persona-question pair three times to account for seed-based variability, shown as *Ri* on our figure. We used `gpt-5.1-2025-11-13` with default temperature, $\tau = 1.0$, and we used the *developer* role to control the LLM persona and response format, and the *user* role for the study questions.[3] We conduct an ablation study of $\tau = \{0.5, 1.0, 1.5\}$, shown in Appendix F.2, finding similar results to those in Figure 3b.

[3]A limitation of our LLM study is potential response bias, as research has shown LLMs exhibit 'yes bias' in grammatical judgments [42]. Future work should explore prompt variations and more nuanced preference elicitation methods to better understand the relationship between LLM and human preferences in transcript evaluation.

4.2.1 How much do LLM responses vary per persona? To summarize consistency across personas p for each question, we report the per-question pooled standard deviation, $s_p(Q_i)$, where $R = 3$ for the 3 rounds, $n_r = 52$ for 52 personas responding each round, and $s_r^2(Q_i)$ is the sample variance of persona responses for question Q_i in round r:[4]

$$s_p(Q_i) = \sqrt{\frac{\sum_{r=1}^{R}(n_r - 1)\, s_r^2(Q_i)}{\sum_{r=1}^{R}(n_r - 1)}}$$

We report results in Figure 3b. Lower values of $s_p(Q_i)$ indicate that personas produce consistent responses across rounds – e.g., if a persona responds with $\{A, A, A\}$ for a question, then $s_p(Q_i) = 0$ – while higher values reflect inconsistency in responses – e.g., $\{A, C, A\}$. Overall, **persona responses are largely stable ($s_p(Q_i) \leq 0.60$)** for 5/6 questions. Nonzero variability occurs for 4/6 of the questions ($s_p(Q_i) > 0.0$ for Q1, Q2, Q3, Q5), likely reflecting sensitivity to temperature and other sampling hyperparameters[5] – highlighting the value of running multiple inference rounds across hyperparameters to assess response stability in LLM-as-a-proxy experimental designs.[6]

4.2.2 Do users and LLMs have different preference distributions? To examine whether LLM and human responses follow similar patterns, we compare the distributions shown in Figures 3a and 3b using a χ^2 test of independence for contingency tables,[7] report the results in Table 2, with an extended discussion in Appendix D. Our results indicate that **LLMs and humans agreed on the most preferred representation in only 3 of 6 questions (Q1, Q3, Q6), revealing substantial divergence even at the level of top choices. Moreover, even when top choices aligned, LLMs produced more extreme or homogenized distributions than humans**, converging strongly on a single representation type (e.g., consistently selecting Bullet Points for Business tasks), whereas human preferences are more diverse and context-sensitive. Chi-squared analyses confirm that these differences are statistically significant across all six questions (all $p < 0.01$). High Cramér's V values across all six questions also show that LLMs converge prematurely on single fidelity levels where humans remain contextually pluralistic.

4.3 Agency-Aware Design Recommendations

Discussion. The results indicate that optimal transcript representation is highly context-sensitive, varying not only across domains but across tasks within the same domain. Even within Business scenarios, deadline assessment (Q5) favored Enhanced transcripts, whereas action item identification (Q6) strongly preferred Bullet Points, demonstrating that fidelity requirements are primarily task-driven rather than domain-determined.

These findings challenge STT evaluation practices relying on universal accuracy metrics like WER. Our results show that optimal representations depend on context: verbatim transcripts may be indispensable for assessing witness confidence in legal depositions, while bullet-point summaries better support physicians scanning triage notes. This reveals a fundamental limitation of WER – they assume a single "correct" representation exists when transcript value depends on intended use.

Importantly, the LLM study highlight limitations in using LLMs as proxies for human preference. Although LLMs matched the most-preferred human representation in some cases (3/6 questions), their preference distributions were consistently more concentrated, often exhibiting near-exclusive selection of a single representation. LLMs amplify

[4]Note that because the sample size is stable across personas ($n_r = 52$ always for the $R = 3$ rounds in our study), the pooled standard deviation reduces to taking the square root of the arithmetic mean of the per-round variances. We include the full formula in our main results for generalizability.
[5]See Appendix F.2 for an ablation study of temperature, $\tau = \{0.5, 1.0, 1.5\}$. As shown in this study, $s_p(Q_i)$ remains low across τ values.
[6]Another approach for future work is developing specialized LLM-based annotation frameworks that explicitly model uncertainty in preference judgments [64].
[7]We set H_0: The distribution of responses is the same for users and LLMs, and H_A: The distributions differ. Significance levels are reported at the $p < 0.5$ (*), $p < 0.01$ (**), $p < 0.001$ (***) levels.

professional roles to extremes rather than capturing the flexible, context-sensitive human expertise, suggesting preference modeling must remain grounded in actual user data rather than algorithmic proxies.

Design Recommendations. We synthesize these insights into four key recommendations:[8]

▷ **R1: Support multi-fidelity interaction.** As shown in Figure 3a, no single representation dominates across all questions: while some tasks elicit higher levels of convergence (e.g., Q6), others distributed more evenly across multiple verbamicities (e.g., Q4). SpeechSpectrum interfaces should therefore allow users to flexibly choose among verbamicities, rather than enforcing a single-output paradigm.

▷ **R2: Incorporate task-aware defaults.** In cases where Figure 3a shows plurality agreement (e.g., Q2, Q3, and Q6 strongly show Bullet Points as the preferred representation), interfaces could streamline user effort by providing task-aware defaults that match common preferences. As shown in Figure 3b, LLMs can approximate these preferences in only some cases; at the same time, defaults must remain adjustable to preserve user agency.

▷ **R3: Prioritize task-level defaults over domain heuristics.** Across both human and LLM results shown in Table 1, preference patterns vary more reliably by task than by domain. Even within the same domain, different tasks elicit distinct preference structures (e.g., Q1 vs. Q2 in Legal), ranging from weakly concentrated to strongly dominant distributions. SpeechSpectrum interfaces should therefore condition transcript defaults and affordances on task intent (e.g., assessment, extraction, verification) rather than relying on coarse domain-level assumptions.

▷ **R4: Provide educational scaffolding.** Tasks with more diffuse distributions (e.g., Q4 in Figure 3a) suggest that users may be uncertain about which representation best fits the task. Interfaces could incorporate educational scaffolding – such as interactive examples or lightweight guidance – to help users develop intuition about when to select different fidelities.

5 Conclusion

Speech-to-text systems now pervade everyday technologies, yet they continue to impose rigid transcription choices that often fail to reflect the variability of users' needs. Our work positions this not as a technical limitation but as an accountability gap with disparate impacts. STT systems used in legal and medical contexts systematically fail speakers from marginalized communities – including speakers of non-standard dialects [217] and patients with clinical speech impairments [130] – yet these users lack agency over how their speech is represented. SpeechSpectrum addresses this by enabling user control over transcript fidelity, providing a mechanism for users to navigate system limitations and contest representational decisions. While this cannot eliminate underlying performance disparities, it redistributes control from system designers to affected users, which is particularly consequential for populations who bear the greates harms from STT system failures. With SpeechSpectrum, we introduced a continuum-based framework that repositions transcription as a spectrum rather than a single outcome.

Looking forward, our findings suggest concrete directions for more user-centered speech systems: ones that flexibly present multiple transcription fidelities, expose choice to the user, and adapt to context rather than enforcing a single "correct" output. More fundamentally, SpeechSpectrum repositions questions of transcription fidelity as matters of user agency and algorithmic accountability. When systems impose representational choices without user input, they make normative judgments about communicative legitimacy – determining which speech features are preserved as meaningful and which are discarded as noise. By granting users explicit control over these choices, we distribute agency over consequential representational decisions rather than concentrating it in the hands of system designers.

[8] Additional technical recommendations (R5-R9) addressing system architecture, evaluation methodology, and data collection are provided in Appendix G.

Generative AI Usage Statement

This study responsibly employed AI technologies to enhance writing clarity such as refining sentence structure and assist with technical tasks such as LaTex table formatting and equation typesetting. All substantive intellectual contributions such as philosophical formalization and experimental design were produced by the authors.

References

[1] 1997. CALLHOME American English Speech (LDC97S42). Web Download. https://catalog.ldc.upenn.edu/LDC97S42

[2] 3Play Media. 2025. *3Play Media.* https://www.3playmedia.com/

[3] Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How might we create better benchmarks for speech recognition?. In *Proceedings of the 1st workshop on benchmarking: Past, present and future.* 22–34.

[4] Amani AlBoul, Ahmad S Haider, and Hadeel Saed. 2025. Enhancing Accessibility for Deaf and Hard-of-Hearing Viewers in the Arab World through Subtitling: Insights from Netflix's Original Saudi Movies. *Forum for Linguistic Studies* 7, 3 (2025), 906–926.

[5] Tanel Alumäe and Allison Koenecke. 2025. Striving for open-source and equitable speech-to-speech translation.

[6] Amazon. 2025. Alexa. https://www.amazon.com/dp/B0DCCNHWV5.

[7] Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234* (2025).

[8] Apple. 2025. Apple Watch. https://www.apple.com/watch/.

[9] Apple Inc. 2025. *Make a recording in Voice Memos on iPhone.* https://support.apple.com/guide/iphone/make-a-recording-iph4d2a39a3b/ios iPhone User Guide, Apple Support.

[10] Apple Inc. 2025. *Send and receive audio messages in Messages on iPhone.* https://support.apple.com/en-my/guide/iphone/iph2e42d3117/ios

[11] Apple Inc. 2025. *Set up your voicemail on iPhone.* https://support.apple.com/en-my/guide/iphone/iph3c99490e/ios

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

[13] Siddhant Arora, Jinchuan Tian, Hayato Futami, Jiatong Shi, Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe. 2025. Chain-of-Thought Reasoning in Streaming Full-Duplex End-to-End Spoken Dialogue Systems. *arXiv preprint arXiv:2510.02066* (2025).

[14] Mariana Arroyo Chavez, Molly Feanny, Matthew Seita, Bernard Thompson, Keith Delk, Skyler Officer, Abraham Glasser, Raja Kushalnagar, and Christian Vogler. 2024. How users experience closed captions on live television: quality metrics remain a challenge. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* 1–16.

[15] Ava. 2025. *Ava.* https://www.ava.me/

[16] Nguyen Bach and Fei Huang. 2019. Noisy BiLSTM-Based Models for Disfluency Detection. In *Proceedings of Interspeech 2019.* 4230–4234.

[17] Muhammad Yeza Baihaqi, Angel García Contreras, Seiya Kawano, and Koichiro Yoshino. 2025. Rapport-Building Dialogue Strategies for Deeper Connection: Integrating Proactive Behavior, Personalization, and Aizuchi Backchannels. In *Proceedings of Interspeech 2025.* 1083–1087.

[18] Keith Bain, Sara Basson, Alexander Faisman, and Dimitri Kanevsky. 2005. Accessibility, transcription, and access everywhere. *IBM systems journal* 44, 3 (2005), 589–603.

[19] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Proceedings of Interspeech 2023.* 4489–4493.

[20] Jennifer Balogh. 2001. Strategies for concatenating recordings in a voice user interface: what we can learn from prosody. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems.* 249–250.

[21] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.* 65–72.

[22] Kristen Bell, Jenny Hong, Catalin Voss, Graham Todd, and AJ Alvero. 2025. Using Machine Learning to Scrutinize Parole Release Hearings. *Berkeley Tech. LJ* 40 (2025), 233.

[23] Daniel Bennett, Oussama Metatla, Anne Roudaut, and Elisa D Mekler. 2023. How does HCI Understand Human Autonomy and Agency?. In *CHI'23: CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery.

[24] Nishchal Bhandari, Danny Chen, Miguel Ángel del Río Fernández, Natalie Delworth, Jennifer Drexler Fox, Migüel Jetté, Quinten McNamara, Corey Miller, Ondřej Novotný, Ján Profant, Nan Qin, Martin Ratajczak, and Jean-Philippe Robichaud. 2025. Reverb: Open-Source ASR and Diarization from Rev. arXiv:2410.03930 [cs.CL] https://arxiv.org/abs/2410.03930

[25] 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 (2001), 123–147.

[26] Theresa Breiner, Swaroop Ramaswamy, Ehsan Variani, Shefali Garg, Rajiv Mathews, Khe Chai Sim, Kilol Gupta, Mingqing Chen, and Lara McConnaughey. 2022. Userlibri: A dataset for asr personalization using only text. *arXiv preprint arXiv:2207.00706* (2022).

[27] CaseFleet. [n. d.]. *CaseFleet.*

[28] Minsuk Chang, Mina Huh, and Juho Kim. 2021. Rubyslippers: Supporting content-based voice navigation for how-to videos. In *Proceedings of the 2021 CHI conference on human factors in computing systems.* 1–14.

[29] Rohan Chaudhury, Maria Teleki, Xiangjue Dong, and James Caverlee. 2024. DACL: Disfluency augmented curriculum learning for fluent text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*

(LREC-COLING 2024). 4311–4321.

[30] Anshul Chavda, M Jagadeesh, Chintalapalli Raja Kullayappa, B Jayaprakash, Medchalimi Sruthi, and Pushpak Bhattacharyya. 2025. DRIVE: Disfluency-Rich Synthetic Dialog Data Generation Framework for Intelligent Vehicle Environments. *arXiv preprint arXiv:2507.19867* (2025).

[31] Tuochao Chen, Qirui Wang, Runlin He, and Shyamnath Gollakota. 2025. Spatial Speech Translation: Translating Across Space With Binaural Hearables. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[32] William Chen, Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe. 2024. Train Long and Test Long:Leveraging Full Document Contexts in Speech Processing. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 13066–13070. doi:10.1109/ICASSP48485.2024.10446727

[33] Anna Seo Gyeong Choi and Hoon Choi. 2025. Fairness of Automatic Speech Recognition: Looking Through a Philosophical Lens. *arXiv preprint arXiv:2508.07143* (2025).

[34] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. " Nobody speaks that fast!" An empirical study of speech rate in conversational agents for people with vision impairments. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[35] Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-Supervised Speech Representations are More Phonetic than Semantic. In *Proceedings of Interspeech 2024*. 4578–4582.

[36] Priyanjana Chowdhury, Nabanika Sarkar, Sanghamitra Nath, and Utpal Sharma. 2024. Analyzing the Effects of Transcription Errors on Summary Generation of Bengali Spoken Documents. *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, 9 (2024), 1–28.

[37] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, et al. 2019. The state of speech in HCI: Trends, themes and challenges. *Interacting with computers* 31, 4 (2019), 349–371.

[38] Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. 2025. Recent Advances in Speech Language Models: A Survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 13943–13970. doi:10.18653/v1/2025.acl-long.682

[39] Mykhailo Danilevskyi, Fernando Perez-Tellez, and Jelena Vasic. 2025. Towards an Accurate Domain-Specific ASR: Transcription for Pathology. In *International Conference on Text, Speech, and Dialogue*. Springer, 309–318.

[40] Gary C David, Angela Cora Garcia, Anne Warfield Rawls, and Donald Chand. 2009. Listening to what is said–transcribing what is heard: the impact of speech recognition technology (SRT) on the practice of medical transcription (MT). *Sociology of Health & Illness* 31, 6 (2009), 924–938.

[41] Caluã de Lacerda Pataca, Matthew Watkins, Roshan Peiris, Sooyeon Lee, and Matt Huenerfauth. 2023. Visualization of speech prosody and emotion in captions: Accessibility for deaf and hard-of-hearing users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[42] Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences* 120, 51 (2023), e2309583120.

[43] Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties Matter: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 12914–12929.

[44] Evgeniia Diachek and Sarah Brown-Schmidt. 2024. Linguistic features of spontaneous speech predict conversational recall. *Psychonomic Bulletin & Review* 31, 4 (2024), 1638–1649.

[45] Ivana Didirková. 2024. Disfluency in speech and language disorders. *Clinical linguistics & phonetics* 38, 4 (2024), 285–286.

[46] Jane A Edwards. 2005. The transcription of discourse. *The handbook of discourse analysis* (2005), 321–348.

[47] Peter G Emery. 2004. Translation, equivalence and fidelity: A pragmatic approach. *Babel* 50, 2 (2004), 143–167.

[48] ENCO. 2025. *enCaption*. https://www.enco.com/products/encaption

[49] Eva Duran Eppler and Eva Codó. 2016. Challenges for language and identity researchers in the collection and transcription of spoken interaction. In *The Routledge handbook of language and identity*. Routledge, 304–319.

[50] Evernote Corporation. 2025. *Evernote*. https://evernote.com/

[51] Daniele Falavigna, Matteo Gerosa, Diego Giuliani, and Roberto Gretter. 2010. An automatic transcription system of hearings in Italian courtrooms. In *Proceedings of the 2nd ACM workshop on Multimedia in forensics, security and intelligence*. 99–104.

[52] Manaal Faruqui and Dilek Hakkani-Tür. 2022. Revisiting the Boundary between ASR and NLU in the Age of Conversational Dialog Systems. *Computational Linguistics* 48, 1 (2022), 221–232.

[53] Jennifer Drexler Fox and Natalie Delworth. 2022. Improving contextual recognition of rare words with an alternate spelling prediction model. *arXiv preprint arXiv:2209.01250* (2022).

[54] Jean E Fox Tree. 2001. Listeners' uses of um and uh in speech comprehension. *Memory & cognition* 29, 2 (2001), 320–326.

[55] Freed Inc. 2025. *Freed AI SOAP Note Generator*. https://www.getfreed.ai/lp/soap-note-ai AI-powered tool that builds SOAP-format clinical notes from audio recordings; HIPAA-compliant, multi-platform.

[56] Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. How to evaluate reward models for rlhf. *arXiv preprint arXiv:2410.14872* (2024).

[57] Rita Frieske and Bertram E Shi. 2024. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572* (2024).

[58] Victoria A Fromkin. 1971. The non-anomalous nature of anomalous utterances. *Language* (1971), 27–52.

[59] Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Tn. 2024. Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization?. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*. 387–394.

[60] Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024. Speech Translation with Speech Foundation Models and Large Language Models: What is There and What is Missing?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 14760–14778. doi:10.18653/v1/2024.acl-long.789

[61] Marco Gaido, Sara Papi, Matteo Negri, Luisa Bentivogli, et al. 2024. Speech Translation with Speech Foundation Models and Large Language Models: What is There and What is Missing?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14760–14778.

[62] Garmin. 2025. Setting Up Voice Assistant on Your Garmin Watch. https://support.garmin.com/en-US/?faq=B0n9YwrwMg4j7yEgevIWgA.

[63] Ginger Labs, Inc. [n. d.]. *Notability*. https://notability.com/

[64] Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candes, and Dan Jurafsky. 2025. Can Unconfident LLM Annotations Be Used for Confident Conclusions?. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 3514–3533.

[65] John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, Vol. 1. IEEE Computer Society, 517–520.

[66] Google. 2025. Google Assistant. https://assistant.google.com/.

[67] Google. 2025. Google Home. https://home.google.com/welcome/.

[68] Google Cloud. 2025. Enable the profanity filter | Cloud Speech-to-Text V2 documentation. https://cloud.google.com/speech-to-text/v2/docs/profanity-filter. Last updated 2025-09-09 UTC.

[69] Google Cloud. 2025. Get automatic punctuation | Cloud Speech-to-Text V2 documentation. https://cloud.google.com/speech-to-text/v2/docs/automatic-punctuation. Last updated 2025-09-09 UTC.

[70] Google Cloud. 2025. Google Cloud Speech-to-Text. https://cloud.google.com/speech-to-text.

[71] Google DeepMind. [n. d.]. Project Astra. https://deepmind.google/models/project-astra/.

[72] Google LLC. 2025. *Google Keep*. https://keep.google.com/ Note-taking service with support for text, lists, images, audio recording and transcription.

[73] Google LLC. 2025. *Send a voice message in Google Chat*. https://support.google.com/chat/answer/14763931?hl=en\let\begingroup\escapechar\m@ne\let\def{\@@par} Google Chat Help Center (Android version).

[74] Ludmila Gordeeva, Vasily Ershov, Oleg Gulyaev, and Igor Kuralenok. 2021. Meaning Error Rate: ASR domain-specific metric framework. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 458–466.

[75] Alex Gorodetski, Ilan Dinstein, and Yaniv Zigel. 2019. Speaker diarization during noisy clinical diagnoses of autism. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2593–2596.

[76] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594* (2024).

[77] Happy Broadcast, Inc. 2025. *Lid: AI-Powered Voice Journaling*. https://www.getlid.co/

[78] David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025. Lost in moderation: How commercial content moderation apis over-and under-moderate group-targeted hate speech and linguistic variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.

[79] Judith Holler. 2025. Facial clues to conversational intentions. *Trends in Cognitive Sciences* (2025).

[80] Paul Hömke, Stephen C Levinson, Alexandra K Emmendorfer, and Judith Holler. 2025. Eyebrow movements as signals of communicative problems in human face-to-face interaction. *Royal Society Open Science* 12, 3 (2025), 241632.

[81] Jiaxiong Hu, Qianyao Xu, Limin Paul Fu, and Yingqing Xu. 2019. Emojilization: An automated method for speech to emoji-labeled text. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[82] Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A Benchmark Dataset for Meeting Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 16409–16423.

[83] Lena-Marie Huttner, Jeppe H Christensen, Gitte Keidser, Tobias May, Torsten Dau, and Sergi Rotger-Griful. 2025. Does effortful speech production indicate communication difficulty caused by noise and hearing aid support?. In *Proceedings of Interspeech 2025*. 1088–1092.

[84] Koji Inoue, Yukoh Wakabayashi, Hiromasa Yoshimoto, and Tatsuya Kawahara. 2014. Speaker diarization using eye-gaze information in multi-party conversations. In *Proceedings of INTERSPEECH 2014*. 562–566.

[85] Instagram (Meta Platforms, Inc.). 2025. *Send a voice message in chats on Instagram*. https://help.instagram.com/805014243174268/?helpref=related_articles Instagram Help Center — FAQ article.

[86] Hyunhoon Jung, Hee Jae Kim, Seongeun So, Jinjoong Kim, and Changhoon Oh. 2019. TurtleTalk: An educational programming game for children with voice user interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[87] Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, et al. 2024. Frozen large language models can perceive paralinguistic aspects of speech. *arXiv preprint arXiv:2410.01162* (2024).

[88] Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, and Ozlem Kalinli. 2025. Frozen Large Language Models Can Perceive Paralinguistic Aspects of Speech. In *Proceedings of INTERSPEECH 2025*. 4323–4327.

[89] Wonjune Kang and Deb Roy. 2024. Prompting Large Language Models with Audio for General-Purpose Speech Summarization. In *Proceedings of Interspeech 2024 (interspeech$_2$024)*. *ISCA*, 1955˜1959.

[90] Fahad Khan, Yufeng Wu, Julia Dray, Bronwyn Hemsley, and A Baki Kocaballi. 2025. Conversational Agents to Support People with Communication Disability: A Co-design Study with Speech Pathologists. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.

[91] Subhendu Khatuya, Koushiki Sinha, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2024. Instruction-guided bullet point summarization of long financial earnings call transcripts. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2477–2481.

[92] Heeseung Kim, Soonshin Seo, Kyeongseok Jeong, Ohsung Kwon, Soyoon Kim, Jungwhan Kim, Jaehong Lee, Eunwoo Song, Myungwoo Oh, Jung-Woo Ha, et al. 2024. Paralinguistics-aware speech-empowered large language models for natural conversation. *Advances in Neural Information Processing Systems* 37 (2024), 131072–131103.

[93] JooYeong Kim, SooYeon Ahn, and Jin-Hyuk Hong. 2023. Visible nuances: A caption system to visualize paralinguistic speech cues for deaf and hard-of-hearing individuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[94] JooYeong Kim and Jin-Hyuk Hong. 2025. OnomaCap: Making Non-speech Sound Captions Accessible and Enjoyable through Onomatopoeic Sound Representation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[95] Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L Seltzer. 2021. Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. In *Proceedings of Interspeech 2021*. 1977–1981.

[96] Yelim Kim, Mohi Reza, Joanna McGrenere, and Dongwook Yoon. 2021. Designers Characterize Naturalness in Voice User Interfaces: Their Goals, Practices, and Challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 242, 13 pages. doi:10.1145/3411764.3445579

[97] Frederic Kirstein, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2024. What's under the hood: Investigating Automatic Metrics on Meeting Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 6709–6723.

[98] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1672–1681.

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

[100] Mark J Koranda, Martin Zettersten, and Maryellen C MacDonald. 2022. Good-enough production: Selecting easier words instead of more accurate ones. *Psychological Science* 33, 9 (2022), 1440–1451.

[101] Korbinian Kuhn, Verena Kersken, and Gottfried Zimmermann. 2025. Communication Access Real-Time Translation Through Collaborative Correction of Automatic Speech Recognition. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[102] Param Kulkarni, Yingchi Liu, Hao-Ming Fu, Shaohua Yang, Isuru Gunasekara, Matt Peloquin, Noah Spitzer-Williams, Xiaotian Zhou, Xiaozhong Liu, Zhengping Ji, et al. 2025. Auto-Drafting Police Reports from Noisy ASR Outputs: A Trust-Centered LLM Approach. In *Companion Proceedings of the ACM on Web Conference 2025*. 2859–2862.

[103] Raja S Kushalnagar, Walter S Lasecki, and Jeffrey P Bigham. 2013. Captions versus transcripts for online video content. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. 1–4.

[104] Adrien Lardilleux and Yves Lepage. 2017. Charcut: Human-targeted character-based mt evaluation with loose differences. In *Proceedings of IWSLT 2017*.

[105] Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan Tn. 2023. Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 343–352.

[106] Khai Le-Duc, Khai-Nguyen Nguyen, Long Vo-Dang, and Truong-Son Hy. 2024. Real-time Speech Summarization for Medical Conversations. In *Proceedigns of Interspeech 2024*. 1960–1964.

[107] Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023. From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–16.

[108] Dawon Lee, Jongwoo Choi, and Junyong Noh. 2025. OptiSub: Optimizing Video Subtitle Presentation for Varied Display and Font Sizes via Speech Pause-Driven Chunking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–12.

[109] Qisheng Li and Shaomei Wu. 2024. " I Want to Publicize My Stutter": Community-led Collection and Curation of Chinese Stuttered Speech Data. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–27.

[110] Huan Liao, Qinke Ni, Yuancheng Wang, Yiheng Lu, Haoyue Zhan, Pengyuan Xie, Qiang Zhang, and Zhizheng Wu. 2025. NVSpeech: An Integrated and Scalable Pipeline for Human-Like Speech Modeling with Paralinguistic Vocalizations. *arXiv preprint arXiv:2508.04195* (2025).

[111] Robin Lickley. 2017. Disfluency in typical and stuttered speech. *Book series Studi AISV* 3 (2017), 373–387.

[112] Belle Lin. 2025. AI Voice Agents Are Ready to Take Your Call. *The Wall Street Journal* (2025). https://www.wsj.com/articles/ai-voice-agents-are-ready-to-take-your-call-a62cf03b Accessed: 2025-09-11.
[113] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
[114] Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-yi Lee, and Ivan Bulyko. 2024. Paralinguistics-enhanced large language modeling of spoken dialogue. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 10316–10320.
[115] Susan Lin, Jeremy Warner, J.D. Zamfirescu-Pereira, Matthew G Lee, Sauhard Jain, Shanqing Cai, Piyawat Lertvittayakumjorn, Michael Xuelin Huang, Shumin Zhai, Bjoern Hartmann, and Can Liu. 2024. Rambler: Supporting Writing With Speech via LLM-Assisted Gist Manipulation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1043, 19 pages. doi:10.1145/3613904.3642217
[116] Yun Liu, Lu Wang, William R. Kearns, Linda Wagner, John Raiti, Yuntao Wang, and Weichao Yuwen. 2021. Integrating a voice user interface into a virtual therapy platform. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.
[117] Debbie Loakes. 2022. Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication* 7 (2022), 803452.
[118] Paria Jamshid Lou and Mark Johnson. 2017. Disfluency Detection using a Noisy Channel Model and a Deep Neural Language Model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 547–553.
[119] Paria Jamshid Lou and Mark Johnson. 2020. End-to-End Speech Recognition and Disfluency Removal. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2051–2061.
[120] Paria Jamshid Lou and Mark Johnson. 2020. Improving disfluency detection by self-training a self-attentive model. *arXiv preprint arXiv:2004.05323* (2020).
[121] Robbie Love and David Wright. 2021. Specifying challenges in transcribing covert recordings: Implications for forensic transcription. *Frontiers in Communication* 6 (2021), 797448.
[122] Brian MacWhinney. 2007. The talkbank project. In *Creating and digitizing language corpora: Volume 1: Synchronic databases*. Springer, 163–180.
[123] Gaurav Maheshwari, Dmitry Ivanov, Théo Johannet, and Kevin El Haddad. 2025. Asr benchmarking: Need for a more representative conversational dataset. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
[124] Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3 (LDC99T42). Web Download.
[125] Adrienne L Massanari. 2010. Designing for imaginary friends: information architecture, personas and the politics of user-centered design. *new media & society* 12, 3 (2010), 401–416.
[126] Roselyn Mathew. 2024. Disfluencies vs. Dysfluencies: Types, Causes, and Differences Between Them. https://www.torontospeechtherapy.com/blog/2024/disfluencies-vs-dysfluencies-types-causes-and-differences-between-them Accessed: 2025-09-11.
[127] Maven AGI. 2025. *Maven AGI*. https://www.mavenagi.com/
[128] Lloyd May, Keita Ohshiro, Khang Dang, Sripathi Sridhar, Jhanvi Pai, Magdalena Fuentes, Sooyeon Lee, and Mark Cartwright. 2024. Unspoken sound: identifying trends in non-speech audio captioning on YouTube. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
[129] Quinten McNamara, Miguel Ángel del Río Fernández, Nishchal Bhandari, Martin Ratajczak, Danny Chen, Corey Miller, and Migüel Jetté. 2024. Style-agnostic evaluation of ASR using multiple reference transcripts. arXiv:2412.07937 [cs.CL] https://arxiv.org/abs/2412.07937
[130] Katelyn Xiaoying Mei, Anna Seo Gyeong Choi, Hilke Schellmann, Mona Sloane, and Allison Koenecke. 2025. Addressing Pitfalls in Auditing Practices of Automatic Speech Recognition Technologies: A Case Study of People with Aphasia. *arXiv preprint arXiv:2506.08846* (2025).
[131] Meta. [n. d.]. Meta AI Glasses. https://www.meta.com/ai-glasses/.
[132] Meta Research. 2015. The Not-So-Universal Language of Laughter. https://research.facebook.com/blog/2015/8/the-not-so-universal-language-of-laughter/. Meta Research Blog.
[133] Corey Miller, Danielle Silverman, Vanesa Jurica, Elizabeth Richerson, Rodney Morris, and Elisabeth Mallard. 2018. Embedding Register-Aware MT into the CAT Workflow. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, Janice Campbell, Alex Yanishevsky, Jennifer Doyon, and Doug Jones (Eds.). Association for Machine Translation in the Americas, Boston, MA, 275–282. https://aclanthology.org/W18-1920/
[134] Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition.. In *Proceedings of Interspeech 2004*. 2765–2768.
[135] Christine Murad, Cosmin Munteanu, Benjamin R Cowan, and Leigh Clark. 2019. Revolution or evolution? Speech interaction and HCI design guidelines. *IEEE Pervasive Computing* 18, 2 (2019), 33–45.
[136] Varun Nathan, Ayush Kumar, and Jithendra Vepa. 2023. Investigating the Role and Impact of Disfluency on Summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 541–551.
[137] Max Nelson, Shannon Wotherspoon, Francis Keith, William Hartmann, and Matthew Snover. 2024. Cross-Lingual Conversational Speech Summarization with Large Language Models. *arXiv preprint arXiv:2408.06484* (2024).
[138] Nextpoint, Inc. 2025. *Nextpoint*. https://www.nextpoint.com/
[139] Eugene Albert Nida and Charles Russell Taber. 1974. *The theory and practice of translation*. Vol. 8. Brill Archive.
[140] NVIDIA. 2025. Parakeet ASR. https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2.
[141] Jeesun Oh, Wooseok Kim, Sungbae Kim, Hyeonjeong Im, and Sangsu Lee. 2024. Better to Ask Than Assume: Proactive Voice Assistants' Communication Strategies That Respect User Agency in a Smart Home Environment. In *Proceedings of the 2024 CHI Conference on Human*

Factors in Computing Systems (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 846, 17 pages. doi:10.1145/3613904.3642193

[142] Walter J Ong and John Hartley. 2013. *Orality and literacy*. Routledge.

[143] OpenAI. 2025. GPT-5.1 Model. https://platform.openai.com/docs/models/gpt-5.1.

[144] OpenAI. 2025. Speech to text. https://platform.openai.com/docs/guides/speech-to-text.

[145] Otter.ai, Inc. 2025. *Otter.ai*. https://otter.ai/

[146] Sharon Oviatt. 1994. Predicting and managing spoken disfluencies during human-computer interaction. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

[147] 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*. 311–318.

[148] Taejin Park, Huck Yang, Kyu Han, and Shinji Watanabe. 2025. Beyond End-to-End ASR: Integrating Long-Context Acoustic and Linguistic Insights. In *Interspeech 2025 Tutorial* (Rotterdam, The Netherlands). ISCA. Tutorial presented at Interspeech 2025.

[149] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language* 72 (2022), 101317.

[150] Hannaneh B Pasandi and Haniyeh B Pasandi. 2022. Evaluation of asr systems for conversational speech: A linguistic perspective. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 962–965.

[151] Bornali Phukon, Xiuwen Zheng, and Mark Hasegawa-Johnson. 2025. Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches. *Proceedings of Interspeech 2025* (2025), 5708–5712.

[152] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[153] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.

[154] Fabian Retkowski, Maike Züfle, Andreas Sudmann, Dinah Pfau, Shinji Watanabe, Jan Niehues, and Alexander Waibel. 2025. Summarizing Speech: A Comprehensive Survey. arXiv:2504.08024 [cs.CL] https://arxiv.org/abs/2504.08024

[155] Rev. 2025. *Rev*. https://www.rev.com/

[156] Rev.ai. 2025. Features | Rev.ai API Documentation. https://docs.rev.ai/api/features/. Accessed: 2025-09-10.

[157] Samantha Robertson and Mark Díaz. 2022. Understanding and being understood: User strategies for identifying and recovering from mistranslations in machine translation-mediated chat. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 2223–2238.

[158] Hadeel Saadany, Catherine Breslin, Constantin Orăsan, and Sophie Walker. 2022. Better transcription of uk supreme court hearings. *arXiv preprint arXiv:2211.17094* (2022).

[159] Elizabeth Salesky, Susanne Burger, Jan Niehues, and Alex Waibel. 2018. Towards fluent translations from disfluent speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 921–926.

[160] Elizabeth Salesky, Matthias Sperber, and Alex Waibel. 2019. Fluent Translations from Disfluent Speech in End-to-End Speech Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2786–2792.

[161] Zitha Sasindran, Harsha Yelchuri, TV Prabhakar, and Supreeth Rao. 2023. H eval: A new hybrid evaluation metric for automatic speech recognition tasks. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 1–7.

[162] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608* (2024).

[163] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[164] Aviv Shamsian, Aviv Navon, Neta Glazer, Gill Hetz, and Joseph Keshet. 2024. Keyword-Guided Adaptation of Automatic Speech Recognition. In *Proceedings of Interspeech 2024*. 732–736.

[165] Roshan Sharma, Suwon Shon, Mark Lindsey, Hira Dhamyal, and Bhiksha Raj. 2024. Speech vs. Transcript: Does It Matter for Human Annotators in Speech Summarization?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14779–14797.

[166] Peng Shen, Xugang Lu, and Hisashi Kawai. 2025. Retrieval-Augmented Speech Recognition Approach for Domain Challenges. arXiv:2502.15264 [cs.CL] https://arxiv.org/abs/2502.15264

[167] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.

[168] Suwon Shon, Kwangyoun Kim, Yi-Te Hsu, Prashant Sridhar, Shinji Watanabe, and Karen Livescu. 2024. DiscreteSLU: A large language model with self-supervised discrete speech units for spoken language understanding. *arXiv preprint arXiv:2406.09345* (2024).

[169] Elizabeth Shriberg. 1996. Disfluencies in switchboard. In *Proceedings of international conference on spoken language processing*, Vol. 96. IEEE Philadelphia, PA, 11–14.

[170] Elizabeth Ellen Shriberg. 1994. Preliminaries to a theory of speech disfluencies. *Doctoral dissertation, University of California at Berkeley* (1994).

[171] Miroslav Sili, Markus Garschall, Martin Morandell, Sten Hanke, and Christopher Mayer. 2016. Personalization in the user interaction design: Isn't personalization just the adjustment according to defined user preferences?. In *International Conference on Human-Computer Interaction*. Springer, 198–207.

[172] Soap AI. 2025. *Soap AI.* https://www.soapnote.ai/ AI-powered medical scribing tool for HIPAA-compliant, EMR-ready SOAP notes (multilingual, differential diagnosis, mobile web).

[173] Soliloquy Apps Limited. 2025. *AudioDiary.* https://audiodiary.ai/ AI-powered multi-platform voice journaling app (transcription, analysis, goal-setting, privacy-focused).

[174] Yuanfeng Song, Di Jiang, Xuefang Zhao, Xiaoling Huang, Qian Xu, Raymond Chi-Wing Wong, and Qiang Yang. 2021. Smartmeeting: Automatic meeting transcription and summarization for in-person conversations. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2777–2779.

[175] Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*. 717–729.

[176] SpeakWrite. 2025. *SpeakWrite.* https://www.speakwrite.com/

[177] Charan Sridhar and Shaomei Wu. 2025. Jjj-just Stutter: Benchmarking Whisper's Performance Disparities on Different Stuttering Patterns. In *Proceedings of Interspeech 2025*. 3753–3757.

[178] Radina Stoykova, Kyle Porter, and Thomas Beka. 2024. The AI Act in a law enforcement context: The case of automatic speech recognition for transcribing investigative interviews. *Forensic Science International: Synergy* 9 (2024), 100563.

[179] Bo-Hao Su, Hui-Ying Shih, Jinchuan Tian, Jiatong Shi, Chi-Chun Lee, Carlos Busso, and Shinji Watanabe. 2025. Reasoning Beyond Majority Vote: An Explainable SpeechLM Framework for Speech Emotion Recognition. *arXiv preprint arXiv:2509.24187* (2025).

[180] Jiwon Suh, Injae Na, and Woohwan Jung. 2024. Improving Domain-Specific ASR with LLM-Generated Contextual Descriptions. In *Proceedings of Interspeech 2024*. 1255–1259.

[181] Hanna Suominen, Liyuan Zhou, Leif Hanlen, and Gabriela Ferraro. 2015. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. *JMIR medical informatics* 3, 2 (2015), e4321.

[182] Benjamin Swets, Susanne Fuchs, Jelena Krivokapić, and Caterina Petrone. 2021. A cross-linguistic study of individual differences in speech planning. *Frontiers in psychology* 12 (2021), 655516.

[183] Benjamin Swets, Matthew E Jacovina, and Richard J Gerrig. 2013. Effects of conversational pressures on speech planning. *Discourse Processes* 50, 1 (2013), 23–51.

[184] Piotr Szymański, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. 2023. Why aren't we NER yet? Artifacts of ASR errors in named entity recognition in spontaneous speech transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1746–1761.

[185] Maria Teleki, Xiangjue Dong, and James Caverlee. 2024. Quantifying the Impact of Disfluency on Spoken Content Summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 13419–13428. https://aclanthology.org/2024.lrec-main.1175

[186] Maria Teleki, Xiangjue Dong, Soohwan Kim, and James Caverlee. 2024. Comparing ASR systems in the context of speech disfluencies. *Proceedings of Interspeech 2024* (2024), 4548–4552.

[187] Maria Teleki, Sai Janjur, Haoran Liu, Oliver Grabner, Ketan Verma, Thomas Docog, Xiangjue Dong, Lingfeng Shi, Cong Wang, Stephanie Birkelbach, Jason Kim, Yin Zhang, and James Caverlee. 2025. DRES: Benchmarking LLMs for Disfluency Removal. arXiv:2509.20321 [cs.CL] https://arxiv.org/abs/2509.20321

[188] Maria Teleki, Sai Janjur, Haoran Liu, Oliver Grabner, Ketan Verma, Thomas Docog, Xiangjue Dong, Lingfeng Shi, Cong Wang, Stephanie Birkelbach, Jason Kim, Yin Zhang, and James Caverlee. 2025. DRES: Benchmarking LLMs for Disfluency Removal. In *arXiv*.

[189] Maria Teleki, Sai Janjur, Haoran Liu, Oliver Grabner, Ketan Verma, Thomas Docog, Xiangjue Dong, Lingfeng Shi, Cong Wang, Stephanie Birkelbach, Jason Kim, Yin Zhang, and James Caverlee. 2025. Z-Scores: A Metric for Linguistically Assessing Disfluency Removal. arXiv:2509.20319 [cs.CL] https://arxiv.org/abs/2509.20319

[190] Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving Statistical Significance in Human Evaluation of Automatic Metrics via Soft Pairwise Accuracy. In *Proceedings of the Ninth Conference on Machine Translation*. 1222–1234.

[191] TranscribeGlass. 2025. TranscribeGlass. https://www.transcribeglass.com/.

[192] James P Trujillo and Judith Holler. 2024. Conversational facial signals combine into compositional meanings that change the interpretation of speaker intentions. *Scientific Reports* 14, 1 (2024), 2286.

[193] Rahmat Ullah, Ikram Asghar, Gareth Evans, Rab Nawaz, Saeed Akbar, and Dorothy Anne Roberts. 2024. Enhancing Speaker Diarization in Forensic Audio: A Comparative Analysis of Machine Learning Algorithms for Gender Classification. In *International Conference on Smart Systems and Emerging Technologies*. Springer, 150–161.

[194] Verbit. 2025. *Verbit.* https://verbit.ai

[195] Dominik Wagner, Sebastian P Bayerl, Ilja Baumann, Korbinian Riedhammer, Elmar Nöth, and Tobias Bocklet. 2024. Large language models for dysfluency detection in stuttered speech. *arXiv preprint arXiv:2406.11025* (2024).

[196] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 993–1003.

[197] Feng Wang, Wei Chen, Zhen Yang, Qianqian Dong, Shuang Xu, and Bo Xu. 2018. Semi-supervised disfluency detection. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3529–3538.

[198] Qiongqiong Wang, Hardik B Sailor, Jeremy HM Wong, Tianchi Liu, Shuo Sun, Wenyu Zhang, Muhammad Huzaifah, Nancy Chen, and Ai Ti Aw. 2025. Incorporating Contextual Paralinguistic Understanding in Large Speech-Language Models. *arXiv preprint arXiv:2508.07273* (2025).

[199] Shaolei Wang, Zhongyuan Wang, Wanxiang Che, Sendong Zhao, and Ting Liu. 2021. Combining self-supervised learning and active learning for disfluency detection. *Transactions on Asian and Low-Resource Language Information Processing* 21, 3 (2021), 1–25.

[200] Shaoyue Wen, Songming Ping, Jialin Wang, Hai-Ning Liang, Xuhai Xu, and Yukang Yan. 2024. AdaptiveVoice: Cognitively adaptive voice interface for driving assistance. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[201] WhatsApp (Meta Platforms, Inc.). 2025. *How to send voice messages.* https://faq.whatsapp.com/657157755756612/?cms_platform=web WhatsApp Help Center — FAQ article.

[202] Shaomei Wu, Kimi Wenzel, Jingjin Li, Qisheng Li, Alisha Pradhan, Raja Kushalnagar, Colin Lea, Allison Koenecke, Christian Vogler, Mark Hasegawa-Johnson, et al. 2025. Speech AI for All: Promoting Accessibility, Fairness, Inclusivity, and Equity. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[203] Xander. 2025. Xander Captioning Glasses. https://www.xanderglasses.com/xanderglasses.

[204] Chih-Kai Yang, Neo S Ho, and Hung-yi Lee. 2025. Towards holistic evaluation of large audio-language models: A comprehensive survey. *arXiv preprint arXiv:2505.15957* (2025).

[205] Jingfeng Yang, Diyi Yang, and Zhaoran Ma. 2020. Planning and generating natural and diverse disfluent texts as augmentation for disfluency detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1450–1460.

[206] Karren Yang, Ting-Yao Hu, Jen-Hao Rick Chang, Hema Swetha Koppula, and Oncel Tuzel. 2023. Text is all you need: Personalizing asr models using controllable speech synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[207] Menglin Yang, Jialin Chen, Yifei Zhang, Jiahong Liu, Jiasheng Zhang, Qiyao Ma, Harshit Verma, Qianru Zhang, Min Zhou, Irwin King, et al. 2024. Low-rank adaptation for foundation models: A comprehensive review. *arXiv preprint arXiv:2501.00365* (2024).

[208] Kamer Ali Yuksel, Thiago Ferreira, Ahmet Gunduz, Mohamed Al-Badrashiny, and Golara Javadi. 2023. A reference-less quality metric for automatic speech recognition via contrastive-learning of a multi-language model with self-supervision. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 1–5.

[209] Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency detection using a bidirectional LSTM. *arXiv preprint arXiv:1604.03209* (2016).

[210] Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. 2025. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157* (2025).

[211] Xinlei Zhang, Takashi Miyaki, and Jun Rekimoto. 2020. WithYou: automated adaptive speech tutoring with context-dependent speech recognition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[212] Xinlei Zhang, Takashi Miyaki, and Jun Rekimoto. 2021. JustSpeak: Automated, user-configurable, interactive agents for speech tutoring. *Proceedings of the ACM on Human-Computer Interaction* 5, EICS (2021), 1–24.

[213] Robin Zhao, Anna SG Choi, Allison Koenecke, and Anaïs Rameau. 2025. Quantification of automatic speech recognition system performance on d/deaf and hard of hearing speech. *The Laryngoscope* 135, 1 (2025), 191–197.

[214] Yijun Zhao, Jiangyu Pan, Jiacheng Cao, Jiarong Zhang, Yan Dong, Yicheng Wang, Preben Hansen, and Guanyun Wang. 2025. Unlocking the Power of Speech: Game-Based Accent and Oral Communication Training for Immigrant English Language Learners via Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–24.

[215] Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2025. Factual Dialogue Summarization via Learning from Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*. 4474–4492.

[216] Zoho. 2025. *Zoho.* https://www.zoho.com/

[217] Maryam Zolnoori, Sasha Vergez, Zidu Xu, Elyas Esmaeili, Ali Zolnour, Krystal Anne Briggs, Jihye Kim Scroggins, Seyed Farid Hosseini Ebrahimabad, James M Noble, Maxim Topaz, et al. 2024. Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. *JAMIA open* 7, 4 (2024), ooae130.

[218] Maike Züfle, Sara Papi, Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, and Jan Niehues. 2025. Nutshell: A dataset for abstract generation from scientific talks. *arXiv preprint arXiv:2502.16942* (2025).

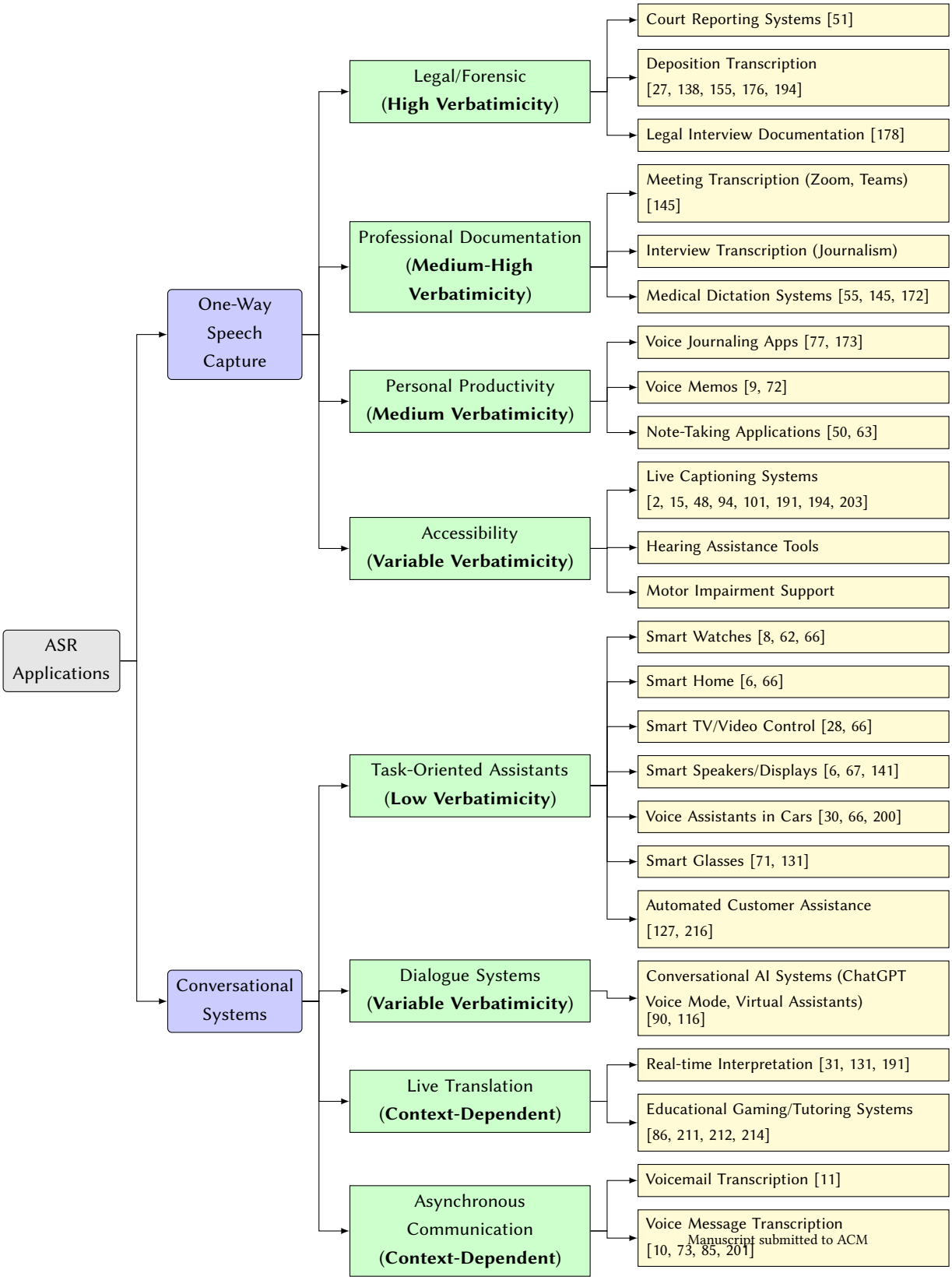


Fig. 4. **Taxonomy of Current Automatic Speech Recognition (ASR) Applications Showing Implicit Fidelity Choices.** Example applications are categorized by interaction modality (One-Way Speech Capture vs. Conversational) and typical verbatimicity requirements. Current systems make these fidelity choices at design time without user control, motivating the need for the SpeechSpectrum framework.

A Technical Clarifications on Verbatimicity and Spectrum Components

This appendix provides technical details on key concepts introduced in §3, including definitional clarifications, scope limitations, and orthogonal features that can be incorporated across multiple points of the SpeechSpectrum.

The concept of **verbatimicity** – defined in the main text as the degree to which textual output preserves the structural, lexical, and paralinguistic characteristics of the original speech signal – relates to but differs from existing concepts in forensic linguistics and translation studies. Verbatimicity differs from *(de)naturalized transcription* in forensic contexts [121]. While both concepts acknowledge that transcription involves representational choices, (de)naturalized transcription focuses primarily on legal admissibility and evidentiary standards in courtroom settings. In contrast, verbatimicity emphasizes user-controlled representation across diverse contexts beyond legal applications, treating fidelity as a designable parameter rather than a fixed procedural requirement. Verbatimicity relates to but extends beyond the concept of *fidelity* in translation studies. While fidelity in translation studies typically refers to faithfulness to source meaning or form along a single dimension – such as semantic equivalence versus structural preservation – verbatimicity specifically captures the degree of preservation across *multiple simultaneous dimensions*: lexical choice, syntactic structure, and paralinguistic features. This multidimensional approach reflects the complexity of cross-modal translation from speech to text, where decisions about one dimension (e.g., removing disfluencies) may interact with others (e.g., loss of prosodic information signaled through hesitations).

As noted in §3.2, the paralinguistic level represents the highest verbatimicity point on the SpeechSpectrum. However, standard text-based representations have inherent limitations in capturing the full richness of paralinguistic signals. While standard text can approximate some paralinguistic features through punctuation, capitalization, or emoticons [132], full preservation often requires additional annotation systems beyond conventional orthography. Higher-fidelity representations exist beyond our scope – such as International Phonetic Alphabet (IPA) phonetic transcription or detailed prosodic annotation systems – but these serve narrower research purposes and typically exhibit low inter-transcriber reliability due to their complexity and specialized nature. Our framework focuses on representations most relevant to everyday STT applications, balancing expressiveness with practical usability. For specialized research contexts requiring maximal acoustic detail, domain-specific annotation schemes (e.g., ToBI for prosody, CA transcription conventions for conversation analysis) remain more appropriate than general-purpose STT systems.

We use the term "disfluencies" for typical speech production phenomena in normal conversation, following speech technology literature. This differs from "dysfluencies," which refers to speech disruptions characteristic of speech disorders such as stuttering or cluttering [126]. This terminological distinction is important: disfluencies are universal features of spontaneous speech production that occur across all speakers; dysfluencies are clinical manifestations that may indicate underlying communicative impairments requiring therapeutic intervention. In the context of SpeechSpectrum, we focus on disfluencies as natural features of everyday speech rather than pathological markers, though we acknowledge that systems designed for clinical populations (e.g., speech-language pathology applications) may require different treatment of these phenomena.

Consensus on the precise definition of disfluency remains elusive in both linguistics and speech technology [45, 111, 170]. What counts as a hesitation, filler, or repair varies by speaker, context, and annotator perspective. This definitional ambiguity has practical implications for SpeechSpectrum systems. Discourse markers like *"like"* or *"you know"* may be considered disfluent noise in one context (e.g., formal presentations) but pragmatically meaningful in another (e.g., casual conversation where they serve discourse-structuring functions). Similarly, repetitions may represent planning disfluencies or emphatic stress depending on prosodic realization. This variability contributes to transcription style variation across human annotators and ASR systems. Rather than enforcing a single definition,

SpeechSpectrum embraces this multiplicity: different fidelity levels can accommodate different interpretations of what constitutes meaningful versus expendable speech features, with users controlling which interpretation best serves their needs.

Speaker diarization – the partitioning of audio by speaker identity [75, 149, 193] – is an important feature for multi-party conversations that is *technically orthogonal to verbatimicity*. While not a core component of the verbatimicity spectrum itself, diarization can be incorporated at any fidelity level, making it a valuable user-controllable feature for SpeechSpectrum systems. SpeechSpectrum systems should treat diarization as an independent, user-controllable feature that can be toggled on or off at any fidelity level depending on task requirements. This independence reflects a broader design principle: some representational features (like diarization, timestamps, or confidence scores) operate orthogonally to verbatimicity and should be configurable separately rather than bundled into specific fidelity levels.

B Case Studies

This appendix section provides concrete examples of how current STT applications implicitly implement different verbatimicity levels, validating the need for the SpeechSpectrum framework presented in the main text.

Current STT applications already operate at different points along the verbatimicity spectrum, but these choices are made implicitly at design-time without user control. Figure 4 presents our taxonomy of STT applications, revealing how different domains and interaction modalities naturally gravitate toward different fidelity levels. This analysis demonstrates both the validity of our framework and the limitations of current one-size-fits-all approaches.

Legal and Forensic Applications. Legal contexts demand maximum fidelity to protect the integrity of records. Court reporting systems [51] and deposition transcription services [27, 138, 155, 194] preserve disfluencies, hesitations, and even paralinguistic features because these elements carry legal significance. A witness's *"um"* or false start might indicate uncertainty relevant to credibility assessment. Recent work on legal interview documentation [178] further demonstrates that verbatim transcription is not merely technically achievable but professionally mandatory in certain contexts.

Professional Documentation. Professional settings like medical dictation and meeting transcription occupy a middle ground. Medical applications demonstrate this complexity clearly: clinical dictation systems [55, 172] prioritize semantic accuracy and readability, actively cleaning disfluencies to produce professional documentation, while speech-language pathology applications require comprehensive disfluency preservation for therapeutic analysis. Projects like TalkBank [122] and CALLHOME [1] demand higher fidelity, preserving precise timing, overlaps, and paralinguistic features for research purposes. Meeting transcription platforms [145] similarly navigate this balance, often providing both real-time "rough" captioning and post-processed "clean" versions, acknowledging that immediate access and polished records serve different needs. Journalistic interview transcription prioritizes readability and semantic content while maintaining speaker authenticity necessary for accurate quote attribution.

Research and Academic Applications. Academic research contexts can demonstrate highly granular verbatimicity requirements. Discourse analysis demands fine-grained pause notation, overlap marking, and detailed prosodic annotation to study conversational dynamics [46]. Sociolinguistic research requires phonetic detail expressed in orthography, accent preservation, and paralinguistic markers for language documentation and dialectal studies [49]. Ethnographic fieldwork may need environmental sound notation and multilingual code-switching preservation that standard ASR systems cannot provide.

Personal Productivity Tools. Voice journaling apps [77, 173] and note-taking applications [50, 63] prioritize usability over strict fidelity. Voice memo systems [9, 72] actively clean speech to produce readable text, assuming users want polished output rather than verbatim records. However, this assumption may not hold for all users or contexts – a researcher documenting field observations might need different fidelity than someone creating a shopping list.

Accessibility Systems. Accessibility applications reveal complex fidelity requirements. Live captioning systems [2, 15, 191, 203] must balance multiple competing needs: speed, accuracy, readability, and information richness. Recent work on onomatopoeia transcription [94] extends beyond traditional text to convey paralinguistic information through creative representations (e.g., *"bu u u wa ang"* for engine sounds). Captioning tool OptiSub [108] recognizes that even presentation format affects accessibility, offering customizable display options with pause-based chunking for naturalistic caption breaking. Semi-automated approaches [101] have been proposed to mitigate high word error rate in real-time captioning. These innovations implicitly acknowledge that accessibility is not monolithic – different users need different representations.

Task-Oriented Assistants. Smart speakers [6, 67, 141], voice assistants [66], and smart glasses [71, 131] operate at the low-verbatimicity end of the spectrum. These systems aggressively compress speech to extracted intents and entities, discarding most linguistic detail. Voice assistants in cars [30, 200] face additional constraints of safety and attention management. Smart watches [8, 62] and smart TV controls [28] further demonstrate how constrained interaction models fundamentally differ from natural conversation – users must learn specific command structures the system understands. This is reflected in user interactions, as user interactions with computer systems are noticeably more fluent than human-human interactions [146].

Dialogue Systems. Conversational AI platforms and agents supporting users with disabilities [90, 116] demonstrate more sophisticated fidelity management. Automated customer assistance systems [127, 216] must balance maintaining conversation flow with accurate understanding, implicitly adjusting their processing based on context. Voice user interfaces in automated phone systems and call centers represent another application domain operating at low verbatimicity, where systems must extract caller intent while managing conversation flow efficiently [112].

Real-Time Communication. Systems providing real-time cross-language communication demonstrate complex fidelity decisions, navigating between source fidelity and target language naturalness. Real-time interpretation services [31, 131, 191] must balance accuracy with temporal constraints while preserving communicative intent across linguistic boundaries. Educational gaming and tutoring systems [86, 211, 214] represent a specialized case, requiring enough detail to assess pronunciation and fluency, particularly for language learners requiring accent understanding and accurate transcription of fast speech.

Asynchronous Communication. Applications for delayed message review have distinct fidelity requirements from real-time interaction. Voicemail transcription [11] typically provides clean, readable text since users review messages asynchronously and prioritize comprehension over production artifacts. Voice chat transcription in messaging apps [10, 73, 85, 201] faces different constraints – balancing processed speech with accuracy while preserving enough speaker personality to maintain social context in casual communication.

This examination of current STT applications reveals a fundamental paradox: while the industry has already evolved to provide different verbatimicity levels across different application domains, individual users remain locked into whatever fidelity level designers predetermined for their specific use case. A lawyer receives verbatim transcripts in court reporting software but cleaned text in meeting transcription tools, regardless of whether those defaults match their needs in that moment. The implicit recognition that different contexts require different fidelity levels – evident in the diversity of approaches across our taxonomy – makes the absence of user control even more striking. To understand whether users would benefit from explicit control over these fidelity choices, we conducted empirical studies examining user preferences and task performance across different verbatimicity levels.

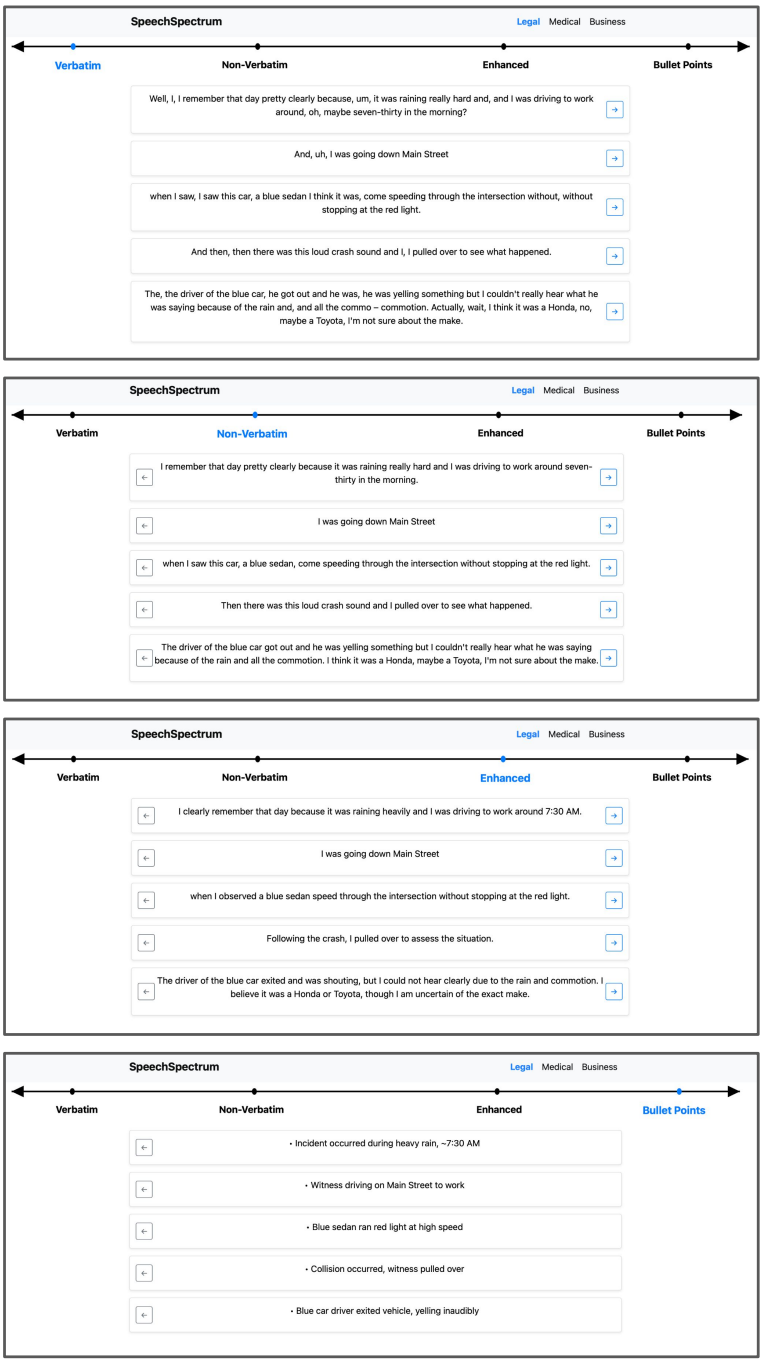


Fig. 5. **SpeechSpectrum instantiation for the user study with an illustrative transcript in the legal domain.** The SpeechSpectrum interface positions transcript versions along the fidelity spectrum (Verbatim, Non-Verbatim, Enhanced, Bullet Points), navigable via clickable labels at the top of the screen. Users can also switch between three example domains – Legal, Medical, and Business – via the top-right menu. Each fidelity level shows five distinct transcript examples to demonstrate the range of representational choices at that verbatimicity level. By surfacing multiple representations within the same interactive space, the prototype demonstrates how SpeechSpectrum operationalizes user-controlled fidelity, allowing participants to explore how different transcript forms better support different tasks and contexts.

C SpeechSpectrum Examples

All transcript variants used in the study are shown in Table 3: The set of transcript representations used in the user study across four fidelity levels (Verbatim , Non-Verbatim , Enhanced , Bullet Points) in the Legal , Medical , and Business domains. The study included both handcrafted transcripts ($N = 23$) and SpeechSpectrum-generated transcripts ($N = 29$; generation details in Appendix H), for a total of $N = 52$ participants.

Table 3. All transcript variants provided in the user study.

Domain	Granularity	Handcrafted Text	`speechspectrum`-Generated Text
Legal	(A) Verbatim	Well, I, I remember that day pretty clearly because, um, it was raining really hard and, and I was driving to work around, oh, maybe seven-thirty in the morning? And, uh, I was going down Main Street when I saw, I saw this car, a blue sedan I think it was, come speeding through the intersection without, without stopping at the red light. And then, then there was this loud crash sound and I, I pulled over to see what happened. The, the driver of the blue car, he got out and he was, he was yelling something but I couldn't really hear what he was saying because of the rain and, and all the commo – commotion. Actually, wait, I think it was a Honda, no, maybe a Toyota, I'm not sure about the make.	Well, I, I remember that day pretty clearly because, um, it was raining really hard and, and I was driving to work around, oh, maybe seven-thirty in the morning? And, uh, I was going down Main Street when I saw, I saw this car, a blue sedan I think it was, come speeding through the intersection without, without stopping at the red light. And then, then there was this loud crash sound and I, I pulled over to see what happened. The, the driver of the blue car, he got out and he was, he was yelling something but I couldn't really hear what he was saying because of the rain and, and all the commo – commotion. Actually, wait, I think it was a Honda, no, maybe a Toyota, I'm not sure about the make.
Legal	(B) Non-Verbatim	I remember that day pretty clearly because it was raining really hard and I was driving to work around seven-thirty in the morning. I was going down Main Street when I saw this car, a blue sedan, come speeding through the intersection without stopping at the red light. Then there was this loud crash sound and I pulled over to see what happened. The driver of the blue car got out and he was yelling something but I couldn't really hear what he was saying because of the rain and all the commotion. I think it was a Honda, maybe a Toyota, I'm not sure about the make.	I remember that day pretty clearly because it was raining really hard and I was driving to work around maybe seven-thirty in the morning. I was going down Main Street when I saw this car, a blue sedan I think it was, come speeding through the intersection without stopping at the red light. Then there was this loud crash sound and I pulled over to see what happened. The driver of the blue car got out and he was yelling something but I couldn't really hear what he was saying because of the rain and all the commotion. Actually, I think it was a Honda, no, maybe a Toyota, I'm not sure about the make.
Legal	(C) Enhanced	I clearly remember that day because it was raining heavily and I was driving to work around 7:30 AM. I was going down Main Street when I observed a blue sedan speed through the intersection without stopping at the red light. Following the crash, I pulled over to assess the situation. The driver of the blue car exited and was shouting, but I could not hear clearly due to the rain and commotion. I believe it was a Honda or Toyota, though I am uncertain of the exact make.	I remember that day clearly because it was raining heavily and I was driving to work at around seven-thirty in the morning. I was heading down Main Street when I saw a blue sedan—possibly a Honda or maybe a Toyota—speed through the intersection without stopping at the red light. I heard a loud crash, so I pulled over to see what had happened. The driver of the blue car got out and started yelling, but I couldn't make out what he was saying because of the rain and the general commotion.
Legal	(D) Bullet Points	* Incident occurred during heavy rain, ~7:30 AM * Witness driving on Main Street to work * Blue sedan ran red light at high speed * Collision occurred, witness pulled over * Blue car driver exited vehicle, yelling inaudibly	- It was raining heavily. - The narrator was driving to work at around 7:30 a.m. - They were heading down Main Street. - They saw a blue sedan, possibly a Honda or Toyota, speed through an intersection without stopping at a red light. - They heard a loud crash and pulled over to see what had happened. - The driver of the blue car got out and started yelling. - The narrator could not understand what the driver was saying because of the rain and general commotion.

Continued on next page

		Handcrafted	speechspectrum-Generated
Medical	(A) Verbatim	Okay, so, um, the patient came in today and, uh, she was complaining of, you know, chest pain that's been going on for about, let's see, um, three days now. And, uh, she said it gets worse when she, when she breathes deeply or, or coughs. So I, I examined her and found some, uh, some tenderness in the, the inter – intercostal muscles on the right side. Her vital signs were, um, let me think, blood pressure was one-twenty over eighty, heart rate was, was seventy-two, and temperature was normal at, uh, ninety-eight point six. I'm thinking this is probably, you know, costochon – costochondritis rather than anything, anything more serious like a cardiac event.	Okay, so, um, the patient came in today and, uh, she was complaining of, you know, chest pain that's been going on for about, let's see, um, three days now. And, uh, she said it gets worse when she, when she breathes deeply or, or coughs. So I, I examined her and found some, uh, some tenderness in the, the inter – intercostal muscles on the right side. Her vital signs were, um, let me think, blood pressure was one-twenty over eighty, heart rate was, was seventy-two, and temperature was normal at, uh, ninety-eight point six. I'm thinking this is probably, you know, costochon – costochondritis rather than anything, anything more serious like a cardiac event.
Medical	(B) Non-Verbatim	The patient came in today and she was complaining of chest pain that's been going on for about three days now. She said it gets worse when she breathes deeply or coughs. I examined her and found some tenderness in the intercostal muscles on the right side. Her vital signs were blood pressure was one-twenty over eighty, heart rate was seventy-two, and temperature was normal at ninety-eight point six. I'm thinking this is probably costochondritis rather than anything more serious like a cardiac event.	The patient came in today and she was complaining of chest pain that's been going on for about three days now. She said it gets worse when she breathes deeply or coughs. I examined her and found some tenderness in the intercostal muscles on the right side. Her vital signs were blood pressure one-twenty over eighty, heart rate seventy-two, and temperature was normal at ninety-eight point six. I'm thinking this is probably costochondritis rather than anything more serious like a cardiac event.
Medical	(C) Enhanced	The patient came in today complaining of chest pain that has been ongoing for three days. She reports it worsens with deep breathing and coughing. On examination, I found tenderness in the intercostal muscles on the right side. Vital signs: blood pressure 120/80, heart rate 72, temperature 98.6°F. Assessment is likely costochondritis rather than a cardiac event.	The patient presented today with a three-day history of chest pain that worsens with deep inspiration and coughing. On examination, there was tenderness of the intercostal muscles on the right side. Vital signs were as follows: blood pressure 120/80 mmHg, heart rate 72 beats per minute, and temperature 98.6°F. The clinical picture is most consistent with costochondritis rather than a more serious condition such as a cardiac event.
Medical	(D) Bullet Points	* Patient: 3-day chest pain history * Triggers: deep breathing, coughing * Physical findings: right intercostal tenderness * Vitals: BP 120/80, HR 72, temp 98.6°F * Diagnosis: likely costochondritis vs cardiac	- Patient has a three-day history of chest pain. - Chest pain worsens with deep inspiration and coughing. - Physical examination shows tenderness of the intercostal muscles on the right side. - Vital signs: blood pressure 120/80 mmHg, heart rate 72 bpm, temperature 98.6°F. - Clinical picture is most consistent with costochondritis. - Costochondritis is considered more likely than a serious condition such as a cardiac event.
Business	(A) Verbatim	So, um, we need to talk about the quarterly numbers and, and where we stand with the project timeline. I know, I know we've been behind schedule but, but Sarah's team has been working really hard on the, the user interface redesign. And, uh, Mike, didn't you say something about the database issues being, being mostly resolved now? Or was that the ser – server issues? No, database, right. I think, I think we're in a much better position than we were last month but we still need to, to figure out the budget allocation for the next quarter and, and maybe hire one or two more developers if, if we want to meet the December deadline. Actually, was it December or January? I think December.	So, um, we need to talk about the quarterly numbers and, and where we stand with the project timeline. I know, I know we've been behind schedule but, but Sarah's team has been working really hard on the, the user interface redesign. And, uh, Mike, didn't you say something about the database issues being, being mostly resolved now? Or was that the ser – server issues? No, database, right. I think, I think we're in a much better position than we were last month but we still need to, to figure out the budget allocation for the next quarter and, and maybe hire one or two more developers if, if we want to meet the December deadline. Actually, was it December or January? I think December.

Continued on next page

		Handcrafted	speechspectrum-Generated
Business	(B) Non-Verbatim	We need to talk about the quarterly numbers and where we stand with the project timeline. I know we've been behind schedule but Sarah's team has been working really hard on the user interface redesign. Mike, didn't you say something about the database issues being mostly resolved now? I think we're in a much better position than we were last month but we still need to figure out the budget allocation for the next quarter and maybe hire one or two more developers if we want to meet the December deadline.	We need to talk about the quarterly numbers and where we stand with the project timeline. I know we've been behind schedule, but Sarah's team has been working really hard on the user interface redesign. Mike, didn't you say something about the database issues being mostly resolved now? Or was that the server issues? No, database, right. I think we're in a much better position than we were last month, but we still need to figure out the budget allocation for the next quarter and maybe hire one or two more developers if we want to meet the December deadline. Actually, was it December or January? I think December.
Business	(C) Enhanced	We need to discuss the quarterly numbers and our current position on the project timeline. We have been behind schedule, but Sarah's team has been working hard on the user interface redesign. Mike mentioned that the database issues are mostly resolved. We are in a better position than last month, but we still need to determine budget allocation for next quarter and consider hiring additional developers to meet the December deadline.	We need to review the quarterly numbers and assess our position against the project timeline. Although we have been behind schedule, Sarah's team has made significant progress on the user interface redesign. Mike has reported that the database issues are now mostly resolved, which puts us in a much better position than we were in last month. However, we still need to determine the budget allocation for next quarter and consider hiring one or two additional developers if we want to meet the planned December deadline.
Business	(D) Bullet Points	* Review quarterly numbers & project timeline * Behind schedule but UI redesign progressing well * Database issues mostly resolved * Improved position vs. last month * Action items: budget allocation, potential hiring for December deadline	- Review the quarterly numbers and assess position relative to the project timeline - Project has been behind schedule - Sarah's team has made significant progress on the user interface redesign - Mike reports the database issues are now mostly resolved - Current situation is much better than last month - Budget allocation for next quarter still needs to be determined - Hiring one or two additional developers is being considered to meet the planned December deadline

D Additional Human vs. LLM$_{R1}$ Preference Distribution Details

Table 2 compares human and LLM$_{R1}$ preference distributions across six task-domain scenarios and reveals systematic divergence in how transcript fidelity is valued. For all questions, χ^2 tests indicate statistically significant differences between human and LLM distributions (all $p \leq 0.0023^{**}$), demonstrating that LLM preferences do not mirror human judgment patterns. Top-choice alignment (✓) occurs in only three of six cases (Q1, Q3, Q6), and even in these aligned scenarios, moderate to high Cramér's V and nontrivial Jensen–Shannon divergence indicate substantial distributional mismatch. In the remaining cases (Q2, Q4, Q5), the LLM top choice contradicts the human top choice (✗), with the largest divergence observed for business deadline assessment (Q5), where both effect size ($V = 0.74$) and JSD are highest. Positive entropy differences across all tasks ($\Delta H > 0$) show that human preferences are consistently more diffuse and heterogeneous, whereas LLM responses are more concentrated and peaked. Taken together, these results show that even when LLMs occasionally select the same top option as humans, they fail to reproduce the overall shape, spread, and uncertainty of human preference distributions, highlighting the limits of LLMs as proxies for user judgment in fidelity-sensitive speech-to-text design.

E Additional User Study Details

We present the results in Table 4. We also show the introduction text for the user study in Figure 6.

Category	Response	Count	Percent
ASR Technology	Yes	7	13.5%
	No	45	86.5%
STEM	Yes	44	84.6%
	No	8	15.4%
Domain Expertise	Legal	7	13.5%
	Medical	6	11.5%

Table 4. Breakdown of participant demographics ($N = 52$). The majority of participants did not work in STT technology and most reported STEM backgrounds. Roughly one-third total reported domain-specific expertise in legal or medical contexts.

> First, open your browser, and navigate to https://SpeechSpectrum.org.
> There are 4 points along the SpeechSpectrum – Verbatim, Non-Verbatim, Enhanced, Bullet Points – which contain different versions of the same transcript. You can navigate the transcript versions by clicking on the labels at the top of the screen, or by using the arrows within the individual boxes.
> We provide 3 example transcripts: Legal, Medical, and Business. You can navigate to each of these examples using the top right menu bar.
> We will now ask you to perform a few tasks [enclosed in square brackets], and answer questions about your experience and opinions related to SpeechSpectrum.

Fig. 6. **Introduction text for our user study.** This text is first displayed to users as part of the user study form.

The study involved voluntary surveys about non-sensitive topics in computer science. The research posed minimal risk and collected no personally identifying information.

ASR	STEM	Legal	Medical	N	Q_1 [Legal]	Q_2 [Legal]	Q_3 [Medical]	Q_4 [Medical]	Q_5 [Business]	Q_6 [Business]
✗	✗	✗	✗	1	Non-Verbatim	Non-Verbatim	Enhanced	Enhanced	Verbatim	Bullet Points
✗	✗	✗	✓	1	Verbatim	Bullet Points	Bullet Points	Non-Verbatim	Non-Verbatim	Bullet Points
✗	✗	✓	✗	6	Enhanced	Bullet Points	Bullet Points	Non-Verbatim	Bullet Points	Bullet Points
✗	✓	✗	✗	32	Verbatim	Bullet Points	Bullet Points	Non-Verbatim	Enhanced	Bullet Points
✗	✓	✗	✓	5	Verbatim	Enhanced	Bullet Points	Non-Verbatim	Enhanced	Bullet Points
✓	✓	✗	✗	6	Verbatim	Bullet Points	Bullet Points	Enhanced	Enhanced	Bullet Points
✓	✓	✗	✓	1	Verbatim	Bullet Points	Non-Verbatim	Verbatim	Non-Verbatim	Bullet Points

Table 5. **Professional demographic characteristics sometimes yield differing top-choice profiles**. Unique participant demographics are indicated by profiles, e.g., *[ASR=✗, STEM=✗, Legal=✓, Medical=✗]* indicates that there were $N = 6$ of the 52 participants who did not have ASR, STEM, or Medical expertise, and their dominant preference for Q_1 [Legal] was Enhanced. Note that these participant demographics are more specific than those in Table 4. The sample sizes are small (for three profiles, $N = 1$), hence future work should further investigate the impact of specific professional profiles.

F Additional LLM Study Details

In this section, we provide additional details for the LLM study.

F.1 Text Structure

Figure 7 illustrates the prompt structure used in the LLM study, which differs from the interface-based presentation used in the user study. Rather than interacting with transcript variants through a graphical interface, the LLM was provided with all relevant information as a single textual prompt.

Each prompt consisted of three components. First, we included a persona-conditioning instruction that specified whether the model should respond as someone with experience in STT, experience in STEM fields, and/or legal or medical expertise. This instruction was used to align the model's responses with the same professional dimensions collected from human participants.

Second, we provided a domain-specific task description (legal, medical, or business), framing the downstream question the model was asked to answer. This task context mirrors the scenarios used in the user study but is presented textually rather than through interactive navigation.

Third, we appended the four candidate transcript representations – Verbatim, Non-Verbatim, Enhanced, and Bullet Points – each corresponding to a distinct position along the SpeechSpectrum. The model was instructed to select the single transcript representation that would be most helpful for answering the given task and to respond only with the letter corresponding to its choice (a format instruction), ensuring a controlled and comparable output format.

This textual concatenation replaces the interactive election used in the user study and allows the LLM to evaluate transcript representations solely through prompt-based reasoning. By holding transcript content constant and varying only persona conditioning and task context, this structure enables a direct comparison between human preference distributions and LLM-generated selections across fidelity levels.

```
Developer Instruction:
Respond as a person who [does/does not] work in automatic speech recognition technology,
[does/does not] work in STEM (science, technology engineering, mathematics), and [has
legal expertise/has medical expertise/does not have legal or medical expertise]. Respond
only with the letter for the answer choice.

User Input:
Imagine you are a doctor looking over a triage dictation provided by a nurse. Which
version of the transcript (i.e. point) is the most helpful for you to answer the following
question: What are the main symptoms the patient is exhibiting?

(A) VERBATIM: Okay, so, um, the patient came in today and, uh, she was complaining of,
you know...
(B) NON-VERBATIM: The patient came in today and she was complaining of chest pain that's...
(C) ENHANCED: The patient came in today complaining of chest pain that has been ongoing...
(D) BULLET POINTS:
* Patient: 3-day chest pain history
* Triggers: deep breathing...
```

Fig. 7. **Prompt used in the LLM study, combining persona conditioning, task context, and four alternative transcript representations from which the model was instructed to select a single option.** Presented with Q3, a medical example.

F.2 Temperature Ablation Study

In this section, we conduct an additional ablation study on $\tau = \{0.5, 1.0, 1.5\}$, to determine the impact of temperature on the LLM preferences. Results for $\tau = 1.0$ are available in the main paper, and results for $\tau = 0.5, 1.5$ are shown here.

As shown in Figures 8 and 9, extreme LLM preference distributions remain extreme with varying τ values.

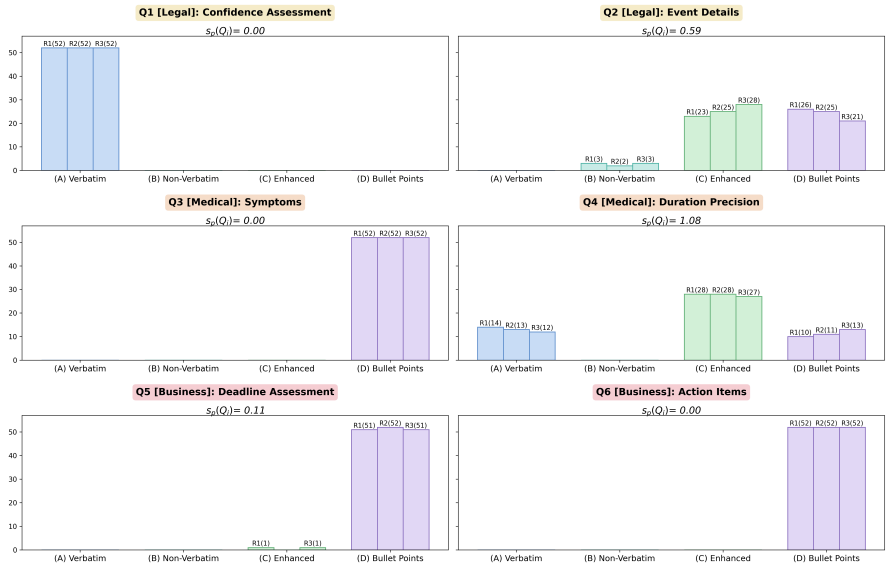


Fig. 8. **Results for $\tau = 0.5$.**

G Components for Designing SpeechSpectrum Systems

This appendix provides technical guidance for implementing SpeechSpectrum systems. While the main body establishes the framework and demonstrates its value through user studies, here we examine the architectural components, evaluation methodologies, and data collection strategies that enable practical realization of multi-fidelity STT interfaces. We present four additional design recommendations (R5-R9) that address technical implementation concerns.

While SpeechSpectrum provides a conceptual framework for understanding STT as a continuum of representational choices, realizing this vision requires practical tools and architectures that can generate, transform, and align transcripts across fidelity levels. We treat ASR, DRM, LLM, and SLM systems – detailed next – not simply as technical models, but as design components that are important for enabling users to navigate and control their place on the fidelity spectrum. Importantly, our findings from §4.2 show that while automated systems such as LLMs can suggest fidelity preferences, ultimate control must remain with users, whose contextual understanding and personal needs cannot be fully captured by algorithmic approaches.

This raises the practical challenge of how to technically implement systems that can fluidly generate multiple representations along the verbatimicity spectrum. In this section, we examine how existing tools can be composed into modular pipelines or end-to-end architectures, highlighting their trade-offs in flexibility, interpretability, and user alignment. As shown in Figure 10, multiple pathways exist for producing different points on the SpeechSpectrum;

(i) LLM_{R1} vs. Uniform Preference Distributions for $\tau = 0.5$						
Q_i [Domain]	χ^2	df	p	V	Most Frequent Representation (p_1)	Δ[95% CI]
Q_1 [Legal]	468.00	3	0.0000***	1.00	(A) Verbatim	1.00 [1.00, 1.00]
Q_2 [Legal]	126.67	3	0.0000***	0.52	(D) Bullet Points	0.06 [0.00, 0.33]
Q_3 [Medical]	468.00	3	0.0000***	1.00	(D) Bullet Points	1.00 [1.00, 1.00]
Q_4 [Medical]	89.28	3	0.0000***	0.44	(C) Enhanced	0.27 [0.04, 0.48]
Q_5 [Business]	452.21	3	0.0000***	0.98	(D) Bullet Points	0.96 [0.88, 1.00]
Q_6 [Business]	468.00	3	0.0000***	1.00	(D) Bullet Points	1.00 [1.00, 1.00]
(ii) Human vs. LLM_{R1} Preference Distributions for $\tau = 0.5$						
Q_i [Domain]	χ^2	df	p	V	Human Δ[95% CI]	LLM Δ[95% CI]
Q_1 [Legal]	36.47	3	0.0000***	0.59	0.23 [0.02, 0.44]	1.00 [1.00, 1.00]
Q_2 [Legal]	8.02	3	0.0457*	0.28	0.35 [0.12, 0.56]	0.06 [0.00, 0.33]
Q_3 [Medical]	27.90	3	0.0000***	0.52	0.40 [0.17, 0.58]	1.00 [1.00, 1.00]
Q_4 [Medical]	19.34	3	0.0002***	0.43	0.00 [0.00, 0.19]	0.27 [0.04, 0.48]
Q_5 [Business]	66.72	3	0.0000***	0.80	0.25 [0.04, 0.42]	0.96 [0.88, 1.00]
Q_6 [Business]	13.57	2	0.0011**	0.36	0.65 [0.44, 0.81]	1.00 [1.00, 1.00]

Table 6. **For $\tau = 0.5$**, results of χ^2 goodness-of-fit tests and associated effect sizes evaluating (i) deviations from uniform preference distributions and (ii) divergence between human and LLM preference distributions across four transcript types. For panel (i), Cramér's V quantifies global dispersion versus concentration of preferences across representations, with larger V indicating greater concentration; in panel (ii), larger V indicates greater divergence between human and LLM distributions. $\Delta(Q_i) = p_1 - p_2$ denotes the local dominance gap between the most and second-most frequent representations, with larger values indicating stronger local concentration. Δ [95% CI] are estimated via nonparametric bootstrap resampling (10,000 iterations).

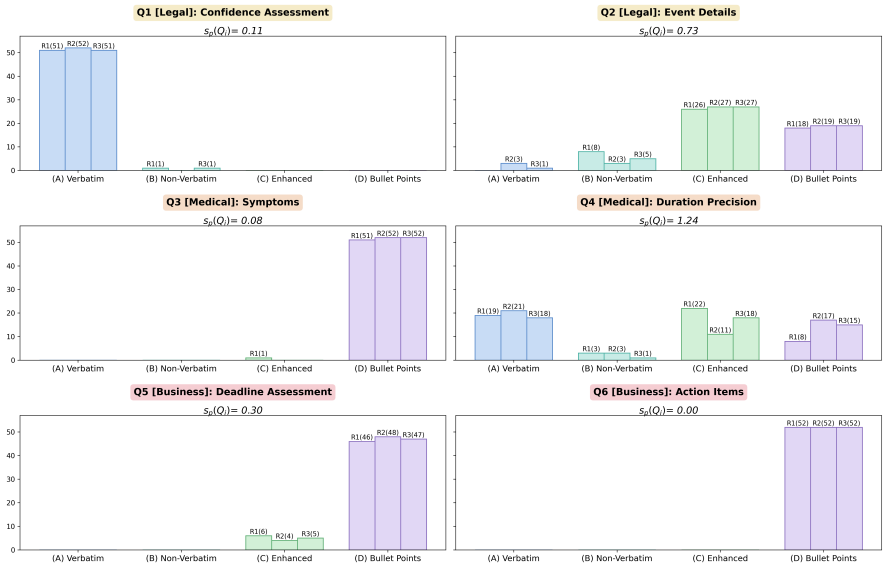


Fig. 9. **Results for $\tau = 1.5$.**

Table 8 provides exemplars of these approaches. Our goal in this section is to provide a technical foundation for understanding how SpeechSpectrum interfaces can be implemented. Rather than prescribing a single architecture,

(i) LLM_{R1} vs. Uniform Preference Distributions for $\tau = 0.5$						
Q_i [Domain]	χ^2	df	p	V	Most Frequent Representation (p_1)	Δ[95% CI]
Q_1 [Legal]	452.21	3	0.0000***	0.98	(A) Verbatim	0.96 [0.88, 1.00]
Q_2 [Legal]	95.49	3	0.0000***	0.45	(C) Enhanced	0.15 [0.00, 0.38]
Q_3 [Medical]	460.05	3	0.0000***	0.99	(D) Bullet Points	0.96 [0.88, 1.00]
Q_4 [Medical]	39.23	3	0.0000***	0.29	(C) Enhanced	0.06 [0.00, 0.29]
Q_5 [Business]	359.54	3	0.0000***	0.88	(D) Bullet Points	0.77 [0.58, 0.92]
Q_6 [Business]	468.00	3	0.0000***	1.00	(D) Bullet Points	1.00 [1.00, 1.00]
(ii) Human vs. LLM_{R1} Preference Distributions for $\tau = 0.5$						
Q_i [Domain]	χ^2	df	p	V	Human Δ[95% CI]	LLM Δ[95% CI]
Q_1 [Legal]	33.18	3	0.0000***	0.56	0.23 [0.02, 0.44]	0.96 [0.88, 1.00]
Q_2 [Legal]	10.16	3	0.0173*	0.31	0.35 [0.12, 0.56]	0.15 [0.00, 0.38]
Q_3 [Medical]	24.84	3	0.0000***	0.49	0.40 [0.17, 0.58]	0.96 [0.88, 1.00]
Q_4 [Medical]	11.93	3	0.0076**	0.34	0.00 [0.00, 0.19]	0.06 [0.00, 0.29]
Q_5 [Business]	51.94	3	0.0000***	0.71	0.25 [0.04, 0.42]	0.77 [0.58, 0.92]
Q_6 [Business]	13.57	2	0.0011**	0.36	0.65 [0.44, 0.81]	1.00 [1.00, 1.00]

Table 7. **For $\tau = 1.5$**, results of χ^2 goodness-of-fit tests and associated effect sizes evaluating (i) deviations from uniform preference distributions and (ii) divergence between human and LLM preference distributions across four transcript types. For panel (i), Cramér's V quantifies global dispersion versus concentration of preferences across representations, with larger V indicating greater concentration; in panel (ii), larger V indicates greater divergence between human and LLM distributions. $\Delta(Q_i) = p_1 - p_2$ denotes the local dominance gap between the most and second-most frequent representations, with larger values indicating stronger local concentration. Δ [95% CI] are estimated via nonparametric bootstrap resampling (10,000 iterations).

we survey the landscape of available components – ASRs, DRMs, LLMs, and SLMs – and examine their trade-offs. This foundation is essential for designers to make informed architectural decisions based on their specific context, whether prioritizing interpretability, performance, or user control. In operationalizing SpeechSpectrum, we recognize the imperative to understand the underlying technology. In this section, we summarize the underlying technological components, with the aim of bridging conceptual design with implementable systems.

Automatic Speech Recognition System (ASR). ASR models are used for translating the raw speech-audio waveform to text transcriptions. A key challenge for ASR systems is the correct transcription of *domain-specific keywords* [164, 166, 180]; decoding methods are often used to guarantee correctness of domain-specific keyword transcription, but these methods are rigid and often rely on retrieved documents. ASR systems struggle to handle noisy, accented, overlapping, stuttered, or fast[9] speech [107, 109], particularly in real-world environments. ASR systems, however, are also efficient and scalable, enabling low-latency transcription across large volumes of speech. In our framework, ASR represents the core transcription component within broader STT systems. While ASR specifically handles speech-to-text conversion, STT encompasses the full pipeline from audio input to final user-facing output, which may include post-processing, formatting, and transformation stages.

Disfluency Removal Model (DRM). Disfluency removal models – implemented as either lightweight classification models [119, 120, 205] or large language models (LLM-as-DRMs) [187] – convert verbatim transcriptions into non-verbatim, fluent text via disfluency removal according to the Shriberg definition [170]. Once disfluencies are removed, the resulting text approximates the conventions of edited written language – characterized by complete sentences, standard punctuation, and absent production artifacts – making it more suitable for text-based NLP tools

[9]It has been shown that people with vision impairments – who are used to interpreting fast speech via screen readers – speak quickly when interacting with conversational agents, and this fast speech is a cause of system error [34].

trained primarily on written corpora [159, 209], enabling more effective downstream processing. Traditional DRM evaluation relies on word-level precision, recall, and F1 scores, which highlight failure modes such as over-deletion and under-deletion [189]. However, newer metrics such as $\mathcal{Z}$-Scores[10] [189] offer a more linguistically grounded assessment by revealing which disfluent node types – such as interjections, parentheticals, or edited nodes – have been removed. Most models are trained and benchmarked on the Switchboard corpus [65], but this dataset's age and demographic limitations hinder generalization. A central challenge for DRMs lies in generalizing beyond their training domains while preserving linguistically meaningful phenomena rather than mistakenly removing them. Recently, in the LLM-as-a-DRM approach, reasoning has been shown to cause an *over-removal* failure mode [187]. At the same time, DRMs' principal strength is the ability to generate fluent, concise text that enhances the effectiveness of downstream tasks such as summarization and information extraction.

Large Language Model (LLM). LLMs are conditioned on prior text tokens $x_1, x_2, \ldots, x_t$, such that $P(x_{t+1} \mid x_{1:t})$ effectively performs next-word prediction for language generation tasks. LLMs are primarily used in the prompt and adaptation (via low rank adapters [207]) setups. A key challenge for LLMs is susceptibility to hallucination and lack of grounding in the input audio. A key strength for LLMs is their flexibility and capacity for semantic reasoning, enabling them to reframe transcripts for diverse user needs.

Speech Language Model (SLM). In contrast to LLMs which are only conditioned on text tokens, SLMs[11] are conditioned jointly on prior text tokens $x_1, x_2, \ldots, x_t$ and speech tokens $s_1, s_2, \ldots, s_t$, such that $P(x_{t+1} \mid x_{1:t}, s_{1:m})$ performs next-word prediction using fusion-based architectures, such as cross-attention mechanisms that integrate speech and text embeddings, or joint encoder-decoder models that process both modalities simultaneously. Unlike ASR systems which treat speech as input to be converted, SLMs maintain speech as a persistent representational modality throughout processing, allowing them to leverage prosodic, intonational, and other acoustic cues for semantic understanding. SLMs are generally used for end-to-end approaches, and can incorporate prosodic and other information available in the audio modality (in contrast to LLMs). While SLMs can theoretically produce verbatim transcripts, they are typically optimized for semantic understanding and contextual processing rather than pure transcription fidelity, making direct speech-to-verbatim conversion less aligned with their architectural strengths, and differentiating SLM from ASR. Cui et al. [38], Gaido et al. [60], and Arora et al. [12] survey recent SLM architectural approaches in detail, while Retkowski et al. [154] survey speech summarization approaches. A key challenge for SLMs is their computational cost, which make them difficult to train and deploy at scale. A key strength for SLMs is their ability to leverage prosody, intonation, and other speech cues to generate more contextually accurate transcriptions.

[10] Disambiguation: These are *not* z-scores in the sense of standard or normal scores in statistics; $\mathcal{Z}$-Scores are a specialized disfluency removal metric detailed in [189].
[11] While the more general Multimodal LLMs (MLLMs) model text, audio, speech, and image, in contrast, SLMs model only text and audio.

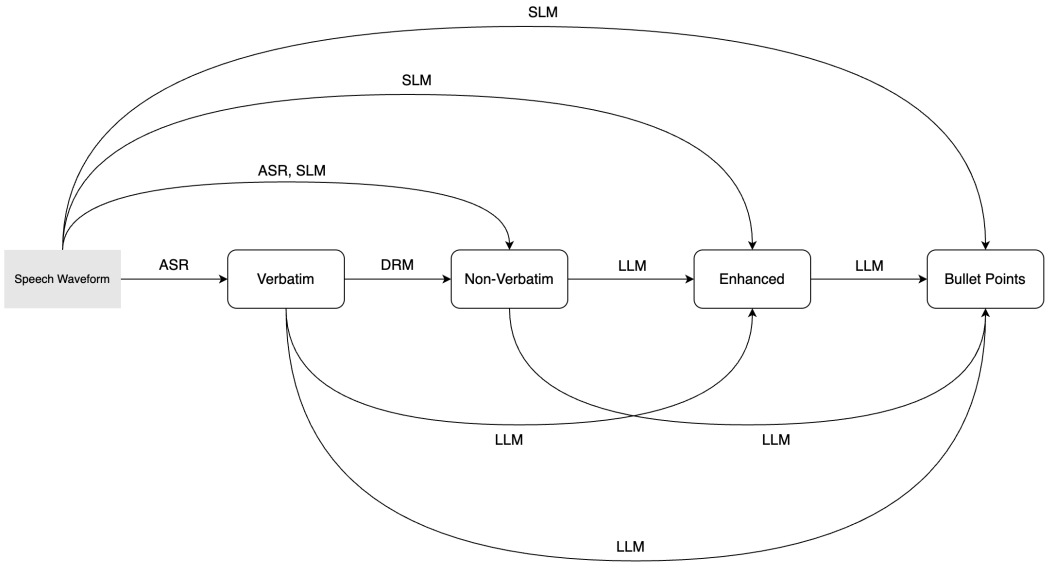


Fig. 10. **Design Pathways for Producing SpeechSpectrum Components.** This diagram illustrates how different tools – ASR, DRM, LLM, and SLM – can be composed to generate transcript representations across fidelity levels. Arrows indicate transformation flows between components (e.g., Verbatim to Non-Verbatim to Enhanced), highlighting how modular pipelines and end-to-end approaches support different routes along the SpeechSpectrum. Rather than a single optimal pathway, the figure emphasizes flexibility in technical design to enable user-controlled navigation of transcript fidelity. Arrows are drawn unidirectionally to indicate that it is only possible to faithfully translate to a lower fidelity level from the original audio.

Component	Tools	Exemplars
	Modular	
Speech Waveform → Verbatim	ASR	Whisper(X) [19, 152], GoogleASR [70], Parakeet-v2 [140]
Verbatim → Non-Verbatim	DRM	LLM-as-a-DRM [187], Synthetic Curriculum Learning [29], BERT-Based Parser [120], Planner-Generator [205], Student-Teacher [199], Bi-LSTM [16], Semi-Supervised [197], Noisy Channel [118]
Non-Verbatim → Enhanced	LLM	Fuse [136], Repair [185], Survey [154]
Enhanced → Bullet Points	LLM	Survey [154]
	End-to-End	
Speech Waveform → Non-Verbatim	ASR, SLM	GoogleASR [70], Acoustic-Lexical [195], LSTM/NiN [160], E2E [119]
Speech Waveform → Enhanced	SLM	Medical RTSS [106], NUTSHELL [218], LongHuBERT [32]
Speech Waveform → Bullet Points	SLM	No specialized systems, can use SLM prompt-based approach.
Verbatim → Enhanced	LLM	Contrastive Student-Teacher [215], Prompt-Based [137], Prompt-Based [97], Chapterization [105]
Verbatim → Bullet Points	LLM	FLAN-FinBPS [91], Aligned [89], MeetingBank [82]
Non-Verbatim → Bullet Points	LLM	No specialized systems, can use LLM prompt-based approach.

Table 8. **Examples of Tools Supporting SpeechSpectrum Components.** The table provides exemplars of how modular and end-to-end approaches can generate different transcript forms along the fidelity spectrum. Modular pipelines (top) separate responsibilities across components, while end-to-end systems (bottom) map directly from speech to higher-level representations such as Non-Verbatim, Enhanced, or Bullet Points. Rather than an exhaustive catalog, the table highlights representative methods that can be mobilized as components to support SpeechSpectrum's design principles: user-controlled fidelity, context-dependent optimization, and cross-modal translation.

G.1 Designing Across Modular and End-to-End Systems

A central design question in SpeechSpectrum is whether to adopt a *modular pipeline* or an *end-to-end* architecture for generating linguistic representations. In modular systems, components such as ASRs, DRMs, LLMs, and SLMs operate sequentially in a cascaded pipeline. In contrast, end-to-end systems map directly from speech input to task output with a single model. While both paradigms are viable, they embody different trade-offs in terms of flexibility, interpretability, and user alignment. There is increasing recognition of multimodal speech-language models that jointly process speech audio and text as a distinct approach from traditional sequential pipelines, evidenced by the rise of Spoken Language Models (SLM) [12, 204]. This work acknowledges that speech and text have fundamentally different linguistic properties that cannot be collapsed into a single representational approach – that is, a single model or pipeline cannot optimally serve all points along the verbatimicity spectrum, as the information preserved in verbatim transcription differs qualitatively from that in enhanced or summarized forms. Additionally, research has demonstrated that summarization of speech transcripts differs fundamentally from summarization of written text transcripts [154], primarily due to the gap in LLM knowledge: LLMs are trained on written data, which is distributionally different from speech data. This research supports our framework's emphasis on treating STT as cross-modal translation rather than mechanical reproduction.

End-to-end models, including recent SLMs, have demonstrated strong task performance. However, emerging evidence suggests that their representations remain more phonetic than semantic. For example, Choi et al. [35] show that near-homophones such as *dog* and *dig* are closely clustered, while synonyms such as *dog* and *puppy* remain more distant. This mismatch can be problematic for tasks requiring semantic fidelity. Modular pipelines address this by allowing specialization of components, such that each component can maintain responsibility for various aspects of the representation. For example, dedicated ASR modules can be fine-tuned for domain-specific vocabulary [164, 166, 180], a task where large, general-purpose models still struggle [148]. Additionally, a modular architecture allows for robust *debugging* practices [99], an advantage for long-term software maintenance.

Beyond specialization, modularity offers advantages in transparency and accountability. Intermediate outputs make it possible to perform fine-grained error analysis, which is difficult in monolithic end-to-end models. Similarly, modular components support auditing – an increasingly important consideration for systems like SpeechSpectrum, where fairness, bias detection, and accountability are central. Modular, cascaded pipelines remain the most widely adopted approach in practice [154], in part because they afford this kind of inspection and adaptation.

Consequently, we suggest to ▷ **R5: Pursue hybrid architectures that combine the interpretability of modular pipelines with the performance advantages of end-to-end models (e.g. [13, 87, 168, 179]).** For contexts requiring interpretability, domain adaptation or auditing, modular pipelines may be preferable. In contrast, in settings where efficiency and simplicity are the priority, end-to-end systems may offer advantages. By pursuing hybrid architectures, SpeechSpectrum systems can move beyond the dichotomy of modular versus end-to-end, toward adaptive systems that reflect the situated needs of their users.

G.2 Evaluating Fidelity Beyond Accuracy for STT Systems

Evaluation methodology plays a central role in shaping how users experience STT systems. Yet existing metrics constrain how performance is understood, often privileging a singular ground truth reference over the multiplicity of outputs users may find acceptable. ASR systems widely treat the speech-to-text transformation as a technical problem of achieving *accuracy*, optimizing for metrics like WER which assumes a single, universal notion of what constitutes the "correct" textual representation of speech. Semantic-style ASR metrics like BLEU, METEOR, and CHARCUT (detailed below) have been proposed to mitigate the weaknesses of the exact-matching paradigm of WER. While these methods can resolve the issue of *legitimate semantic preservation* in transcription, they do not resolve the issue of *legitimate stylistic differences* in transcription – e.g., as previously raised, *w- what he was sayin'* and *what, what he was saying* are both correct transcriptions which vary only in *style* [129]. A new STT metric, MULTIREFERENCE [129], allows for these differences, but is expensive to obtain, requiring multiple ground-truth human annotation references. Hence, there is a gap in *stylistic evaluation methodology for automatic speech recognition systems* [37].

Table 9 provides an overview of commonly adopted ASR and Machine Translation (MT) metrics, illustrating how they differ by domain, unit of analysis, and evaluation principle. These metrics – ranging from word-level edit distance (WER) to character-level overlap (CER) and n-gram precision/recall measures (BLEU, ROUGE) – were originally designed for either ASR or MT and later adapted across contexts. While each provides a useful baseline, they share a common limitation: they assume strict evaluation against a single reference as the definitive measure of success.

This singular reference-centric assumption becomes problematic when multiple transcriptions may be equally valid and differ only stylistically. Synonymity-based measures like METEOR offer improvements by rewarding semantic similarity, but they remain focused on surface-level textual similarity – measuring lexical overlap and n-gram matches – rather than deeper dimensions such as fluency, style, or contextual appropriateness. As Gaido et al. [61] note, these constraints limit the interpretive value of evaluation for speech-based systems.

Metric	Domain	Unit of Analysis	Evaluation Principle	Distinctive Features
WER	ASR	word-level	edit distance	All error types are *penalized equally.*
CER	ASR	character-level	edit distance	Adapted version of WER, all error types are *penalized equally.*
BLEU [147]	MT	word-level	n-gram precision-based	Utilizes a weighted geometric mean based on n-gram precision with a *brevity penalty.*
ROUGE-N [113]	MT	word-level	n-gram recall-based	Strictly allows *exact* word matching.
METEOR [21]	MT	word-level (primarily)	unigram F-based	Includes semantic matching for *synonyms*, and correlates well with human evaluations.
CHARCUT [104]	ASR, MT	character-level	n-gram F-like via a longest common subsequence operation	Used for *segment visualization* in interactive ASR user interfaces, where character-level alignment enables users to see precisely which portions of the transcript differ from reference text, supporting error analysis and correction workflows.

Table 9. **Overview of common STT evaluation metrics, organized by domain, unit of analysis, and evaluation principle.** The table highlights how different metrics – ranging from edit-distance measures (WER, CER) to n-gram and semantic similarity approaches (BLEU, ROUGE, METEOR, CHARCUT) – emphasize particular types of errors. As shown by our results, this reliance on single-reference correctness overlooks the stylistic and contextual variation that users value in transcripts, revealing the need for evaluation approaches aligned with SpeechSpectrum's principles.

Recent work in large language model (LLM) optimization highlights an alternative paradigm: *preference-based evaluation.* Alignment methods such as direct preference optimization (DPO) [153] and proximal policy optimization (PPO) [163] – as well as many others – illustrate how preference signals can be used to navigate large solution spaces. Translating this into evaluation, preference-based methods assess alignment with human or LLM judgments rather than a singular ground truth. This shift is particularly relevant to ASR systems, where outputs occupy a broad solution space and stylistic variation is not error but an important part of user experience.

Hence, an appropriate metric for the STT solution space is Pairwise Ranking Accuracy (PRA) [56]. Previously proposed for automatic speech recognition [208], PRA is a meta-metric that measures how often an automated metric agrees with human (or LLM) preferences when comparing two outputs. PRA reframes evaluation around *preference alignment* rather than singular ground-truth matching. PRA is defined as:

$$\text{PRA} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[r(x_i^a, x_i^b) = h(x_i^a, x_i^b)\right] \tag{1}$$

where N is the total number of pairwise comparisons, x_i^a, x_i^b are candidate outputs, $r(\cdot)$ is the metric ranking, and $h(\cdot)$ is the human or LLM ranking (including ties), and $\mathbf{1}[\cdot]$ is the indicator function. In essence, PRA captures the average agreement between $r(\cdot)$ and $h(\cdot)$ across all pairs of outputs, measuring the alignment of the metric with human or LLM preferences. By capturing preferences rather than correctness, the learned signal forces no notion of binary correctness, reframing evaluation around preference alignment.

More sophisticated alternatives extend this framework: Soft Pairwise Accuracy (SPA) incorporates statistical significance [190], while Deutsch et al. [43] explicitly model ties. These pairwise methods can be operationalized through human ratings or via LLM-as-a-Judge frameworks [76]. While promising, each route has trade-offs: human preference ratings require annotator effort and cost, whereas LLM-based preferences may diverge from human judgments (as seen in the differences between Figure 3a and Figure 3b), potentially exaggerating or homogenizing rankings. *Importantly, however, preference-based evaluation reframes the human role: rather than constructing "gold-standard" transcripts under rigid annotation rules, humans can instead rank candidate outputs of variable verbatimicity – a cognitively lighter task.*

Hence, we recommend to ▷ **R6: Include preference-based evaluation methods like Pairwise Ranking Accuracy (PRA) in STT evaluation to move beyond the assumption of a singular ground truth.** By aligning evaluation with human judgments, STT systems can better reflect the wide space of valid outputs encountered in practice.

Related to preference-based evaluation, our empirical findings in §subsection 4.2 reveal important limitations when using LLMs to model these preferences. ▷ **R7: Exercise caution when using LLMs to model user preferences for transcript fidelity.** While LLMs can approximate aggregate patterns, they tend toward extreme or homogenized preferences that don't capture the diversity and nuance of human judgment. LLMs may be useful for generating candidate transcripts across fidelity levels, but ultimate preference modeling and evaluation should involve human users. This recommendation reinforces that evaluation frameworks must remain grounded in actual user needs rather than algorithmic proxies.

G.3 Reframing Disfluency Corpora as Design Resources

Disfluencies – i.e. filled pauses (*uh, um*), false starts, repetitions, repairs, etc. – are common in everyday speech and often reflect natural interactional processes like planning, hesitation, or emphasis [170]. From a user perspective, these features may not merely be "errors," but could be resources that shape how conversation unfolds.

Rather than treating annotator disagreement as error, future datasets could model such variation explicitly – capturing multiple annotator perspectives, cultures, contextual dependencies, and stylistic preferences. This reframing shifts the goal from enforcing a singular ground truth toward supporting flexibility, positioning DRMs as adaptive tools that reflect the diversity of real-world communication.

Existing disfluency removal datasets [65, 124] have primarily relied on linguistic annotators to mark the disfluencies. While this paradigm provides consistency, it overlooks the situated expertise of domain professionals in areas such as law or medicine, where expectations for "fluent" speech differ substantially. In these domains, what counts as an error is not only linguistic but also contextual and task-dependent.

Systematically capturing inter-rater reliability offers a valuable design signal for incorporating disagreement. Cohen's κ and Krippendorff's α are established inter-rater reliability metrics that can be used here. Utterances with high inter-rater reliability values may support confident automatic processing, while those with low inter-rater reliability values could be used to trigger human-in-the-loop review or display multiple renderings. In this way, disagreement becomes a resource for supporting user awareness of ambiguity.

Therefore, we recommend to ▷ **R8: Expand disfluency removal datasets to both incorporate annotator disagreement (i.e., multiple interpretations of the same utterance) in the form of inter-rater reliability, and to include domain expertise, in addition to linguistic annotation.** This broader approach would enable the development of DRM systems that are not only technically accurate, but also contextually sensitive and responsive to the diverse communicative practices found across domains.

G.4 Extending The SpeechSpectrum Beyond Speech-to-Text

While SpeechSpectrum focuses on speech-to-text conversion, we acknowledge that spoken communication is inherently multimodal, incorporating visual signals such as gaze, gestures, facial expressions, and body posture that carry meaning not fully recoverable from audio alone [79, 80, 84, 192]. Future systems should extend the notion of representational fidelity to these modalities, enabling users to control not only how speech is rendered into text, but also how non-verbal cues are preserved, summarized, or omitted. As with speech, representational choices over visual signals involve normative judgments about relevance, salience, and interpretability. Providing user-controllable fidelity over multimodal cues can improve accessibility (e.g., for d/Deaf or neurodivergent users), enhance interpretive accuracy in high-stakes contexts such as legal or medical settings, and reduce the risk of systems imposing hidden assumptions about which communicative signals "matter." Treating multimodal representation as a spectrum rather than a fixed extraction pipeline generalizes SpeechSpectrum's core principle: accountability requires making representational decisions explicit and contestable rather than implicit and system-defined. Therefore, we recommend to ▷ **R9: Extend fidelity control beyond speech-to-text to multimodal communication signals.**

H The `speechspectrum` Python Package (Available via PyPI)

We provide an open-source Python package, `speechspectrum` (v1.0.1), which implements the transcript transformation pipeline. The package operationalizes the SpeechSpectrum framework by enabling generation of multiple speech-to-text representations along a linguistic fidelity continuum, from verbatim transcripts to compressed summaries.

H.1 Installation and Usage

The package is distributed via the Python Package Index and can be installed using:

```
pip install speechspectrum
```

Source code for the package is available at https://anonymous.4open.science/r/SpeechSpectrum-A3D4. Users must provide valid OpenAI API credentials for the underlying language and speech models at runtime. Example usage demonstrating end-to-end transcript generation is also provided in the accompanying Jupyter notebooks included in the repository.

The package is released under the MIT License and is intended to support reproducibility, further experimentation, and future research on user-controllable speech-to-text representations.

H.2 Implementation Details

The package is implemented in Python (Python $\geq$ 3.8) and relies on OpenAI large language models for downstream text transformations. Audio-to-text transcription is performed using `gpt-4o-mini-transcribe` [144], while subsequent transformations are carried out using instruction-controlled `gpt-5.1-2025-11-13` [143]. The transformation stages are implemented as independent functions, allowing users to invoke individual steps or compose custom pipelines.

H.3 Prompt Formulation Details

Corresponding to the prompts shown in Figure 11, we provide details about how the

Verbatim → Non-Verbatim. This stage uses a specialized prompt and a configuration similar to that shown to perform well in the Disfluency Removal Evaluation Suite (DRES) Teleki et al. [188], but implemented with

`gpt-5.1-2025-11-13` (a newer model). The disfluency definitions and structural categories used in this prompt follow Shriberg's [170] framework:

- Reparandum: the segment to be deleted
- Interruption point: where the speaker cuts off the reparandum
- Interregnum: fillers or repair cues (e.g., *uh*, *um*, restarts)
- Repair: intendended/fluent speech to be kept

The reference examples used in this prompt can be found on the following pages of Shriberg [170]:

- Example 1: Page 9
- Example 2: Page 14
- Example 3: Page 27
- Example 4: Page 66
- Example 5: Page 68

Non-Verbatim → Enhanced. This step meets the needs of downstream users who expect high-quality output (e.g., customer requests).

Enhanced → Bullet Points. Convert an enhanced transcript into concise bullet points using a structured extraction prompt. This stage reflects customer demand for rapid distillation of spoken content (e.g., industry use cases), similarly to medical-scribe workflows such as generating SOAP-note style summaries [172].

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Developer Instruction
You are an expert in linguistics.

Verbatim → Non-Verbatim
Using a transcript of spontaneous speech below, clean it by removing disfluencies in line with Shriberg's structure: identify the reparandum (the portion to be deleted), interruption point, and interregnum (filled pauses, self-repair cues) so that the remaining repair constitutes the speaker's intended fluent sentence. Disfluencies must be deleted to arrive at the speaker's intended sequence.

Specifically:
- Remove filler words and sounds (e.g., um, uh, you know) when they occur as interregnum material.
- Remove repeated/self-repaired segments (reparandum) up to the interruption point; keep only the repair portion.
- Do not remove material that constitutes the repair (the intended utterance) or change meaning.
- Preserve meaning, tone, and speaker intent, and maintain grammatical correctness and readability.
- Do not add any new content or reinterpret the speaker's words.
- Output only the cleaned transcript, with no commentary or annotations.

Example 1:
Input: Show me flights from boston on um monday
Output: Show me flights from boston on monday

Example 2:
Input: Show me the – which early flights go to boston
Output: Which early flights go to boston

Example 3:
Input: which flights leave after eleven – leave after noon
Output: which flights leave after noon

Example 4:
Input: um i guess we're going to talk describe uh job benefits
Output: we're going to describe job benefits

Example 5:
Input: he – she – she went
Output: she went

Here is the transcript: [TEXT]

Non-Verbatim → Enhanced
Rewrite the following transcription it so it is clear, readable, and well-structured, retaining single paragraph formatting. Enhance grammar, flow, and clarity.

Here is the text: [TEXT]

Enhanced → Bullet Points
Extract the key points from the following text. Deliver them as clear, concise bullet points. Not necessarily atomic facts, but condensed bullet points. Do not add anything that isn't explicitly stated.

Here is the text: [TEXT]

Fig. 11. **Prompts used for `speechspectrum` tool.**