SpeechSpectrum: A Framework for Speech-to-Text Representation Along the Linguistic Fidelity Spectrum

ANONYMOUS AUTHOR(S)

Speech-to-text (STT) systems are increasingly embedded in everyday technologies, yet they 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. Through theoretical analysis and empirical investigation, we show how existing STT applications already implicitly make spectrum-based choices, validating the framework. Our user study demonstrates that explicit control over transcript representation improves user experience; we also find that LLMs fail to simulate the wide variance in user preferences. We derive design implications and recommendations for STT systems, and provide open-sourced code and prototypes, to support future research. Our work positions fidelity as a designable parameter, advancing both conceptualization and practical design of speech technologies.

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

Additional Key Words and Phrases: voice user interface, speech-to-text, linguistic fidelity, automatic speech recognition

ACM Reference Format:
Anonymous Author(s). 2026. SpeechSpectrum: A Framework for Speech-to-Text Representation Along the Linguistic Fidelity Spectrum. In *Proceedings of ACM CHI 2026 (CHI '26).* ACM, New York, NY, USA, 26 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 [27, 73]. 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* [124, 128].

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 documenting legal proceedings requires verbatim preservation of every utterance, including hesitations and false starts that may carry legal significance; A professional attending a virtual meeting may prefer cleaned transcripts that remove disfluencies while preserving the essential semantic content for later reference; A user dictating casual messages wants natural-sounding text that doesn't burden the recipient with the artifacts of spontaneous speech production; A sociolinguist requires paralinguistic signals alongside output speech, to study artifacts of identity and conversational dynamics; A d/Deaf person requires rich annotations (e.g., in Netflix captions [4, 13]) 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 no control over this fundamental representational choice.

Central to these representational choices is the treatment of *disfluencies*[2], which encompass various speech phenomena including 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 other non-standard speech features like contractions (*"gonna", "wanna"*) or dialect

[1] Speech-To-Text (STT) refers both to the *systems* that perform speech-to-text translation, and the *task* of speech-to-text translation.

[2] We use the term "disfluencies" to refer to typical speech production phenomena that occur in normal conversation, following standard usage in speech technology literature. This differs from "dysfluencies," which specifically refers to speech disruptions associated with stuttering or other speech disorders [121]. While both involve breaks in speech flow, disfluencies are universal features of spontaneous speech production, whereas dysfluencies are clinical manifestations that may require therapeutic intervention.

variations (*"y'all"*), as disfluencies typically represent real-time speech production processes rather than stable linguistic choices. While there are existing disfluency definitions [41, 107, 161], consensus remains elusive. What counts as a hesitation, filler, or repair often varies by speaker, context, and annotator, limiting the generalizability of current datasets. For example, tokens such as *"like"* or *"you know"* may be treated as noise in one context but as pragmatically meaningful cues in another. These definitional ambiguities contribute to the stylistic variation in transcription approaches.

This disconnect between diverse user information needs, technological capability, and user agency represents a critical gap in human-computer interaction design. While STT systems have achieved remarkable improvements in technical accuracy, they remain largely oblivious to the contextual factors that determine whether a transcript or transcript artifact actually serves the user's goals and individualized needs [189]. The field has focused primarily on reducing transcription errors while neglecting the equally important question of transcription purpose. We argue that this approach fundamentally misunderstands the nature of STT conversion,[3] treating it as a mechanical transcription task rather than as a form of cross-modal translation that necessarily involves choices about representation, emphasis, and information structure.

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, refers to the degree of faithfulness to source material characteristics [43] – ranges from highly compressed summaries that extract key semantic content to verbatim transcriptions that preserve every 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)[4], 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 falls 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.

Increasingly, state-of-the-art STT systems extend beyond verbatim transcription, integrating downstream transformations such as summarization, meeting-minutes generation, caption enrichment, or task-specific extraction. End-to-end STT systems map directly from speech input to final text output using a single model, while modular STT systems use separate components (ASR → disfluency removal → text processing) in a pipeline architecture. With either approach, these STT systems are not merely converting speech into verbatim transcripts but are engaging in layered representational choices that collapse, reframe, or augment the spoken modality according to design goals. Accordingly, our framework treats "'speech-to-text" as a larger category encompassing both traditional modular systems and end-to-end systems, as detailed in §6. This broader scope highlights the limits of current evaluation practices [33] – accuracy measures like WER, or even other more semantically-meaningful metrics, only capture a fraction of the design space. This reinforces the need to conceptualize STT as a spectrum of representational fidelity that applies across transcription, translation, and summarization tasks alike.

We contribute:

- **SpeechSpectrum, a continuum-based framework** for understanding STT conversion, the value of which we demonstrate through both theoretical analysis and empirical investigation.
- **Analysis of case studies** showing that real-world STT systems implicitly implement spectrum-based choices, validating the need for the SpeechSpectrum framework.
- **Evidence from a user study** showing that providing explicit control over transcript representation improves user experience with STT output.
- **Findings from an LLM study** revealing that LLMs simulating human preferences fail to capture the empirically observed human desire for granular representation.

[3] Throughout this paper, we distinguish between "STT conversion" (the technical process of transforming speech signals into text) and "STT translation" (the interpretive process involving representational choices about fidelity, style, and information structure). This distinction emphasizes that STT systems should perform cross-modal translation with meaningful design choices rather than mechanical signal conversion.

[4] WER is the gold standard metric used to evaluate performance of different STT systems, calculated as the edit distance between the predicted transcript and a reference "ground truth" transcript, normalized by the number of words in the reference: WER = $\frac{S+D+I}{N}$ where S, D, I are substitutions, deletions, and insertions respectively, and N is the total number of words in the reference.

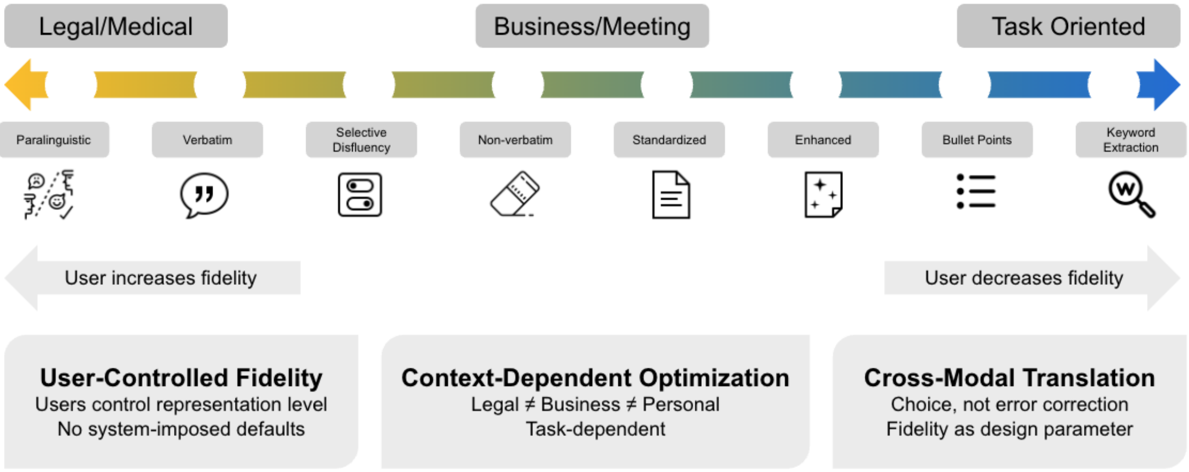


Fig. 1. **The SpeechSpectrum framework conceptualizes STT conversion as a continuous spectrum of verbatimicity level rather than a single transcription target.** The gradient bar represents the continuum from high-fidelity paralinguistic preservation (left) to low-fidelity keyword extraction (right), with eight reference points showing commonly used representation types. Domain context indicators above the spectrum show typical usage patterns: Legal/Medical contexts requiring high fidelity, Business/Meeting contexts using moderate fidelity, and Task-Oriented applications preferring low fidelity. Bidirectional arrows emphasize user control over fidelity levels based on context and communicative needs. The three core principles establish the framework's foundation: (1) User-Controlled Fidelity enables individuals to specify their preferred representation level rather than accepting system defaults; (2) Context-Dependent Optimization recognizes that optimal fidelity varies by domain, task, and user goals; and (3) Cross-Modal Translation reframes STT as deliberate representational choice rather than mechanical error correction, treating fidelity as a designable parameter.

- **Design recommendations (R1–R7)** for STT systems, including both the conceptual challenges of realizing spectrum-based STT systems and concrete guidance on hybrid architectures, alignment-based metrics, and contestable disfluency removal data collection to support researchers and practitioners in designing, implementing, evaluating, and advancing SpeechSpectrum interfaces.
- **Open-source resources**, including our code (https://anonymous.4open.science/r/SpeechSpectrum/) and prototype (https://[redacted-for-anonymity].github.io/SpeechSpectrum).

In §2, we review related work in HCI, speech technologies, and natural language processing, 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 analyze existing STT applications across domains. In §5, we present user studies demonstrating context-dependent fidelity preferences and LLM limitations in modeling human preferences. In §6, we examine technical approaches for implementing SpeechSpectrum systems, including modular and end-to-end architecture, evaluation methodologies, and data collection strategies. In §7, we conclude with implications for future speech technology research and design.

2 Literature Review

The HCI community has highlighted system designer concerns with STT interpretations of ASR errors [130], yet a fundamental disconnect persists between HCI research and speech technology development. Speech researchers often overlook the importance of user-centered design in speech systems, while HCI researchers frequently treat STT as a black box without engaging with its technical constraints and possibilities. This divide is particularly pronounced in the speech domain, where technical metrics dominate evaluation practices while user needs and preferences remain underexplored. To address this issue, we consider a wide range of perspectives, integrating concepts and findings from HCI, speech technologies, and natural language processing studies.

Communication as a Spectrum. Foundational work in linguistics has long recognized that human communication exists along a spectrum rather than discrete spoken versus written categories. Cultures and individuals navigate between purely oral and highly literate communicative practices, with numerous hybrid forms occupying intermediate positions [137]. This continuum provides theoretical justification for treating ASR output as existing at different points of oral-literate mediation rather than pursuing a single "optimal" representation.

Drawing from translation studies [5], we conceptualize STT conversion as a form of cross-modal translation that necessarily involves choices about fidelity and adaptation. Eugene Nida's [134] distinction between formal equivalence (preserving source language structures) and dynamic equivalence (preserving communicative effect) maps directly onto ASR design choices: verbatim transcription prioritizes formal equivalence to the speech signal, while cleaned or summarized output prioritizes dynamic equivalence for written consumption. This translation framework reveals that "accuracy" in STT systems cannot be defined without reference to the intended function of the output. Just as a literary translation serves different purposes than a technical manual translation, different STT use cases require different approaches to cross-modal adaptation. The key insight is that fidelity – understood as faithfulness to source modality characteristics – represents a designable parameter rather than a fixed system goal.

Computational linguistics research, pragmatics and discourse analysis specifically, has demonstrated that disfluencies are not simply errors to be corrected but meaningful linguistic features that serve specific communicative functions [16, 40, 50, 172, 173]. Hesitations mark processing difficulty and signal upcoming complex information; false starts and repairs reveal speakers' real-time meaning negotiation; filled pauses serve discourse management functions. This work shows that the decision to preserve or remove disfluencies fundamentally changes the informational content of the transcript.

Our contribution is to propose *SpeechSpectrum*, a framework which reconceptualizes STT as cross-modal translation along a continuum of representational fidelity.

Speech Interface Design. Research on voice user interfaces [92, 130] has primarily focused on naturalness, or intent recognition and task completion, implicitly treating the STT conversion as a black box pre-processing step. This approach works well for command-based interactions, but breaks down when users need to review, edit, or reference the textual output of their spoken interactions. Current voice interfaces provide users with minimal visibility into how their speech has been interpreted and no control over the representational choices embedded in that interpretation.

Despite growing recognition of user diversity in speech interface design, current systems provide users with minimal control over the linguistic representation of their speech. Most commercial services include profanity filtering [64] and including basic punctuations [65], and some offer limited disfluency removal [151], but few services allow for finer-grained user selection of fidelity. It is not standard practice for services to offer a spectrum of parameters ranging from verbatim transcription, cleaned text, enhanced formatting, or summary-level representation. This represents a significant gap in user agency over a fundamental aspect of the interface experience.

While some HCI research on personalization has explored how interfaces can adapt to individual user needs and preferences [152, 162], this work remains limited in the speech domain. Some work has investigated user-specific ASR models that adapt to individual speaking patterns and vocabulary [22, 193]. However, this personalization typically focuses on improving recognition accuracy [49] rather than allowing users to specify their preferred representational style or fidelity level. Post-processing of journaling speech has also been explored [111], but also does not focus on stylistic transcription.

The accessibility research community has made the most progress in recognizing user diversity in transcript preferences. Work on live captioning and hearing accessibility tools has explored customizable transcript features, though typically examining surface-level preferences (font size, display timing) rather than fundamental questions about linguistic representation [17, 37, 99]. There has also been significant progress in accessibility by expanding input modalities through silent speech interfaces[5] [74, 141, 142, 169, 186, 199], but this thrust of research still does not allow users to consider different types of textual outputs.

The rise of meeting transcription tools has revealed user preferences for different levels of transcript detail depending on context and purpose. Users require varying lengths of meeting summaries [55], indicating that a single representational approach cannot serve all use cases even within the same domain. **Our contribution is to propose a set of design recommendations (R1-R7) for STT systems.**

Speech Transcription Choices. Meanwhile, the dominance of WER in automatic speech recognition system evaluation reflects a conception of transcription as mechanical reproduction rather than representational choice [3, 21, 54, 70, 125, 129, 160, 174]. This family of metrics (detailed in §6.2) largely assumes the existence of a single "correct" transcription against which all systems should be measured, yet provides no theoretical justification for why one representational choice should be privileged over others. One can glean that this setup is a result of practical engineering limitations more than anything else. Recent

[5] Silent speech interfaces allow for imperceptible or near-imperceptible whispered speech input to preserve users' privacy in public spaces, while still naturally interacting with an STT system; these interfaces allow users with speech disorders to interface with speech systems. Silent speech interfaces can also be more *socially acceptable* alternatives to speech interfaces [142].

work has begun to acknowledge these circumstances [53, 91, 94, 146, 155, 167], but offers limited systematic alternatives that account for user context and task requirements.

Interestingly, the STT community has implicitly recognized our core argument through the development of domain-specific systems. Medical STT systems [35, 36, 171] optimize for different representational features – e.g. verbatim features are needed to diagnose a fluency disorder – than conversational STT systems [48, 119, 145], effectively implementing different points on our proposed spectrum. Legal STT systems [98, 113, 153] preserve disfluencies that general-purpose systems remove, while meeting transcription tools [32, 158, 165] often incorporate summarization features that compress information in ways that would be inappropriate for forensic applications [117]. However, these domain-specific approaches are typically framed as separate technical problems rather than instances of a broader design space around fidelity choices.

Perhaps the strongest evidence towards the value of our framework comes from the ubiquity of post-processing in STT pipelines. Text normalization, punctuation restoration, and disfluency removal systems – standard but inconsistently implemented across services without any norms [125] – all represent movements along our proposed spectrum, yet they are typically treated as separate technical problems rather than coordinated representational choices. This fragmentation obscures the fundamental insight that these systems are collectively implementing a continuum of possible representations.

Recent work by Teleki et al. [176] and Mei et al. [125] provides compelling evidence for our framework. Their comparison of automatic speech recognition systems reveals that different platforms by design preserve or remove different types of disfluencies, effectively placing their outputs at different points on the spoken-written continuum. Crucially, they demonstrate that optimal fidelity choices vary by content type and downstream task, directly supporting our position that representation should be context-dependent rather than universally optimized.

Our contribution is to conduct a user study to assess the usefulness of our proposed SpeechSpectrum framework for domain experts; we find that the SpeechSpectrum framework serves the needs of domain experts and non-domain experts alike.

3 The SpeechSpectrum Framework

Current ASR 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 imagined ground truth constitutes meaningful performance regardless of user goals or application context. Second, ASR 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 ASR research and the actual value that users derive from these systems in practice.

These limitations stem from a particular conceptual orientation: treating ASR as transcription rather than translation. We propose instead understanding STT conversion as modality translation along what we term a "spectrum of compression" – a continuous range of representational choices 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. A visual representation of SpeechSpectrum is shown in Figure 1, and an example transcript and its representations along SpeechSpectrum are shown in Figure 3.

3.1 A Continuum for Representing Speech-to-Text

We introduce the concept of **verbatimicity** – the degree to which textual output preserves the structural, lexical, and paralinguistic characteristics of the original speech signal.[6] Unlike binary notions of accuracy (detailed more in §6), 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

[6] This concept relates to but differs from (de)naturalized transcription in forensic contexts [117], which focuses on legal admissibility rather than user-controlled representation.

testimony requires different verbatimicity than a business executive reviewing meeting highlights, yet current ASR 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 require high fidelity to preserve hesitations that may indicate witness uncertainty, while medical triage documentation benefits from concise bullet points that highlight critical symptoms. 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, 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 [29].

3.2 Components

Along the SpeechSpectrum, different 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. It is important to emphasize that SpeechSpectrum conceptualizes representation as a true continuum with infinite possible gradations between any two points. The levels we describe here represent commonly used reference points rather than discrete categories, and our prototype interface implements only a subset of the most widely recognized representations for practical user study purposes.

Paralinguistic. The highest verbatimicity level extends beyond textual transcription[7] to include imputed descriptions of paralinguistic signals, such as emotional expressions (laughter, crying, snorting, grunting), physiological sounds (yawns, coughs), and prosodic patterns (lexical contrast features like pitch variation, tone, pacing, pausing, volume changes) [19]. For example, effortful speech (in terms of higher fundamental frequency, volume, etc.) indicates communication difficulty [80]. Even *emojilization* of text has been proposed to incorporate paralinguistic information [78]. This level also includes environmental information, such as overlapping speech, and acoustic context. This level serves specialized contexts such as discourse analysis, therapeutic interaction documentation, or forensic applications where maximal communicative context must be preserved. A key HCI thrust has centered on information presentation for d/Deaf and hard-of-hearing (d/DHH) [89], offering fine-grained environmental, emotional, and spacio-temporal information via the paralinguistic signal [37, 90, 123]. It has been shown that LLMs are capable of processing and understanding these paralinguistic speech aspects [84, 110], and research has focused recently on integrating these aspects into LLM and speech language model architectures [88, 106, 184].

Verbatim. The verbatim transcript is the most faithful textual version to the speech audio, and should comprehensively include disfluencies. A current challenge with obtaining verbatim transcripts is that ASR models under-transcribe disfluencies [176] – sometimes by design. A challenge with obtaining verbatim transcripts is that ASR models must handle the phenomenon of *good-enough word selection* [96], 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 d/Deaf and hard of hearing (DHH) individuals, with fine-grained emotion conveyed via the disfluency signal [90, 123].

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.

[7] While standard text can approximate some paralinguistic features through punctuation, capitalization or emoticons [127], full paralinguistic preservation often requires additional annotation systems.

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 decisions.

Standardized. Moving further along the spectrum, standardized transcription converts informal speech patterns to conventional written forms, transforming contractions (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 maintaining the speaker's essential content and structure.

Enhanced. Enhanced transcripts represent a form of deliberate post-processing or editing that, in most cases, improves word choice, sentence structure, and overall coherence while preserving the speaker's intended meaning. Parliamentary proceedings exemplify this approach, where transcribers add omitted protocol elements like 'Madame Speaker' to maintain institutional conventions [182]. This level addresses use cases where speakers want to present a more polished version of their spontaneous speech – such as lawyers preparing statements or professionals creating documentation from informal discussion.

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, between fidelity and accessibility, between linguistic expression and standardized communication norms. The optimal choice depends entirely on user context, purpose, and values – decisions that only users themselves can make appropriately.

While our framework primarily addresses STT conversion, we acknowledge that language processing is inherently multimodal [76], with interpretation often grounded in *visual signals* such as gaze, facial expressions, and gestures. Extending our framework to incorporate the vision modality opens promising directions for future work in the paralinguistic direction. Prior work shows that visual cues can enrich the interpretation of speaker intent [75, 179] and help identify communicative breakdowns [77]. Moreover, *gaze tracking* has proven effective for supporting ASR diarization and turn-taking analysis [81]. By integrating vision as an even richer signal than the audio signal alone, future speech systems may achieve more robust disambiguation, greater sensitivity to social context, and ultimately more natural interactions.

4 Case Studies

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 2 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 [47] and deposition transcription services [23, 133, 150, 180] 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 [168] 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 [51, 163] prioritize semantic accuracy and readability, actively cleaning disfluencies to produce professional documentation, while speech-language pathology

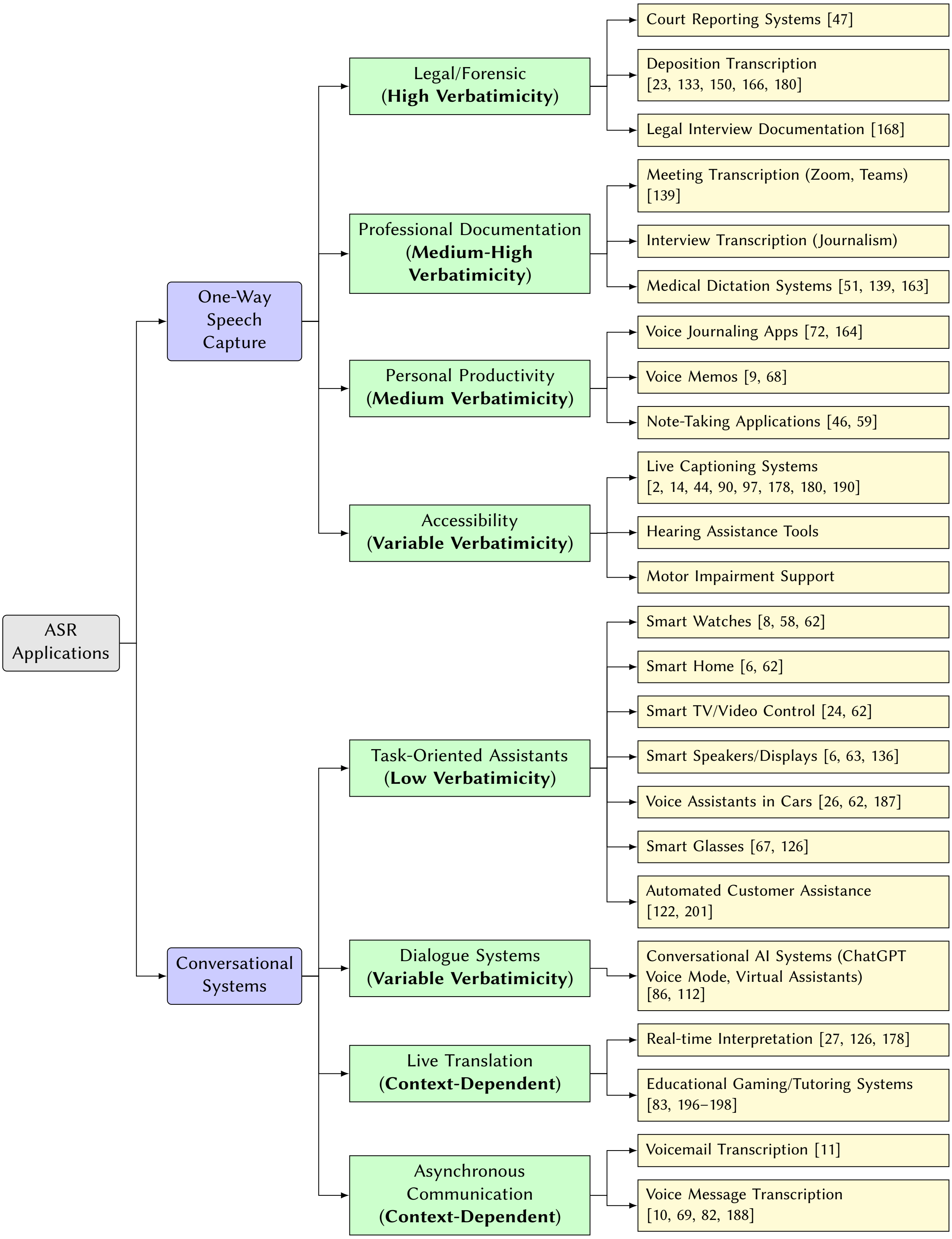

ASR Applications
One-Way Speech Capture
Conversational Systems
Legal/Forensic (High Verbatimicity)
Professional Documentation (Medium-High Verbatimicity)
Personal Productivity (Medium Verbatimicity)
Accessibility (Variable Verbatimicity)
Task-Oriented Assistants (Low Verbatimicity)
Dialogue Systems (Variable Verbatimicity)
Live Translation (Context-Dependent)
Asynchronous Communication (Context-Dependent)
Court Reporting Systems [47]
Deposition Transcription [23, 133, 150, 166, 180]
Legal Interview Documentation [168]
Meeting Transcription (Zoom, Teams) [139]
Interview Transcription (Journalism)
Medical Dictation Systems [51, 139, 163]
Voice Journaling Apps [72, 164]
Voice Memos [9, 68]
Note-Taking Applications [46, 59]
Live Captioning Systems [2, 14, 44, 90, 97, 178, 180, 190]
Hearing Assistance Tools
Motor Impairment Support
Smart Watches [8, 58, 62]
Smart Home [6, 62]
Smart TV/Video Control [24, 62]
Smart Speakers/Displays [6, 63, 136]
Voice Assistants in Cars [26, 62, 187]
Smart Glasses [67, 126]
Automated Customer Assistance [122, 201]
Conversational AI Systems (ChatGPT Voice Mode, Virtual Assistants) [86, 112]
Real-time Interpretation [27, 126, 178]
Educational Gaming/Tutoring Systems [83, 196–198]
Voicemail Transcription [11]
Voice Message Transcription [10, 69, 82, 188]

Fig. 2. **Taxonomy of Current 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.

applications require comprehensive disfluency preservation for therapeutic analysis. Projects like TalkBank [118] and CALLHOME [1] demand higher fidelity, preserving precise timing, overlaps, and paralinguistic features for research purposes. Meeting transcription platforms [139] 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 millisecond-level pause notation, overlap marking, and detailed prosodic annotation to study conversational dynamics [42]. Sociolinguistic research requires phonetic detail, accent preservation, and paralinguistic markers for language documentation and dialectal studies [45]. Ethnographic fieldwork may need environmental sound notation and multilingual code-switching preservation that standard ASR systems cannot provide.

Personal Productivity Tools. Voice journaling apps [72, 164] and note-taking applications [46, 59] prioritize usability over strict fidelity. Voice memo systems [9, 68] 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, 14, 178, 190] must balance multiple competing needs: speed, accuracy, readability, and information richness. Recent work on onomatopoeia transcription [90] extends beyond traditional text to convey paralinguistic information through creative representations (e.g., *"bu u u wa ang"* for engine sounds). OptiSub [104] recognizes that even presentation format affects accessibility, offering customizable display options with pause-based chunking for naturalistic caption breaking. Semi-automated approaches [97] have been proposed to mitigate high automatic speech recognition 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, 63, 136], voice assistants [62], and smart glasses [67, 126] 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 [26, 187] face additional constraints of safety and attention management. Smart watches [8, 58] and smart TV controls [24] 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 [140].

Dialogue Systems. Conversational AI platforms and agents supporting users with disabilities [86, 112] demonstrate more sophisticated fidelity management. Automated customer assistance systems [122, 201] 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 [108].

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 [27, 126, 178] must balance accuracy with temporal constraints while preserving communicative intent across linguistic boundaries. Educational gaming and tutoring systems [83, 196, 198] 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, 69, 82, 188] 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.

5 Empirical Studies

5.1 Study 1: User Study of Contextual Fidelity Preferences

5.1.1 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. This reduction allowed us to preserve the core principles of SpeechSpectrum while keeping the study tasks focused and tractable.

We implemented the prototype as a lightweight web application to ensure accessibility and consistency across participants. A web-based delivery lowered barriers to participation and guaranteed platform-independence: users could access the system through a standard browser without installing software or requiring specialized hardware. 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.

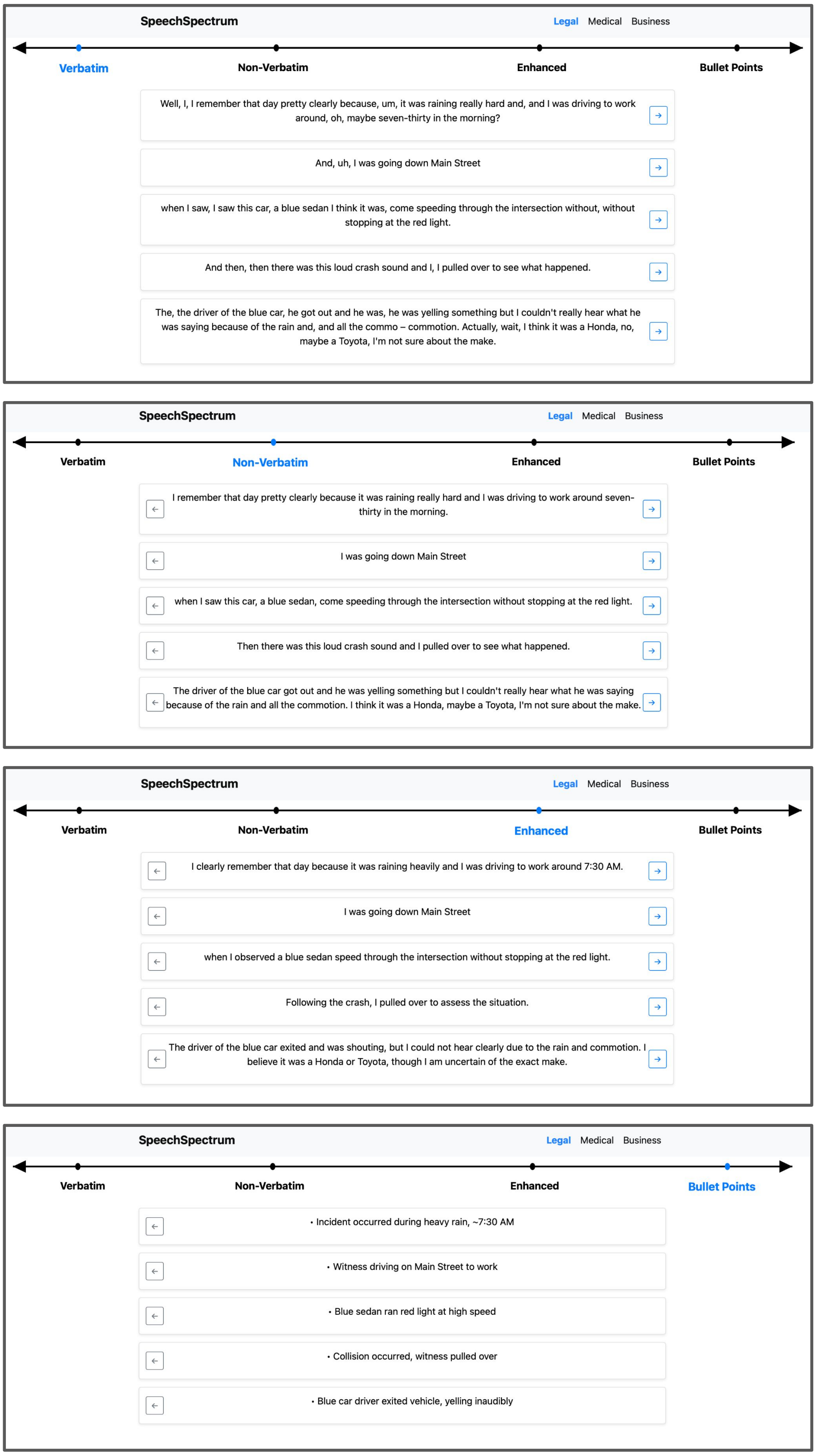
The interface was organized around domain-specific scenarios – Legal, Medical, and Business – reflecting professional contexts where speech-to-text technologies are commonly applied. Participants could navigate across fidelity levels using a clickable "spectrum" interface and switch domains through a 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.

5.1.2 Study Design. We conduct a user study with 23 participants through convenience sampling from academic and professional networks. We collect demographic data on participants' professional backgrounds and no other PII were collected.

Participants engaged with the interactive SpeechSpectrum interface for approximately 3 minutes on average, with most making fidelity preferences quickly upon reviewing transcript examples, and completed scenario-based tasks. For each domain, we presented two distinct task scenarios requiring different information extraction approaches. We ask each user the following six **Questions (Q)** which refer to one of the three domain-specific examples **[Domain Example]**, annotated with color and formatting for ease of comparison:

> **Q1 [Legal Example].** Imagine you are a case judge reading through a deposition transcript. Which version of the transcript (i.e. point) is the most helpful for you to answer the following question: **Did the defendant seem confident about the details of the crash?**
>
> **Q2 [Legal Example].** Imagine you are a case judge reading through a deposition transcript. Which version of the transcript (i.e. point) is the most helpful for you to answer the following question: **What were the events leading up to the crash?**
>
> **Q3 [Medical Example].** 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?**
>
> **Q4 [Medical Example].** 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: **Has the chest pain been going on for exactly three days, or could it have been longer/shorter?**
>
> **Q5 [Business Example].** Imagine you are a team leader reading a meeting transcript. Which version of the transcript (i.e. point) is the most helpful for you to answer the following question: **Does the team seem like they will meet the December deadline?**
>
> **Q6 [Business Example].** Imagine you are a team leader reading a meeting transcript. Which version of the transcript (i.e. point) is the most helpful for you to answer the following question: **What are the action items from the meeting?**



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.

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 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.

- 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

Fig. 3. **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.

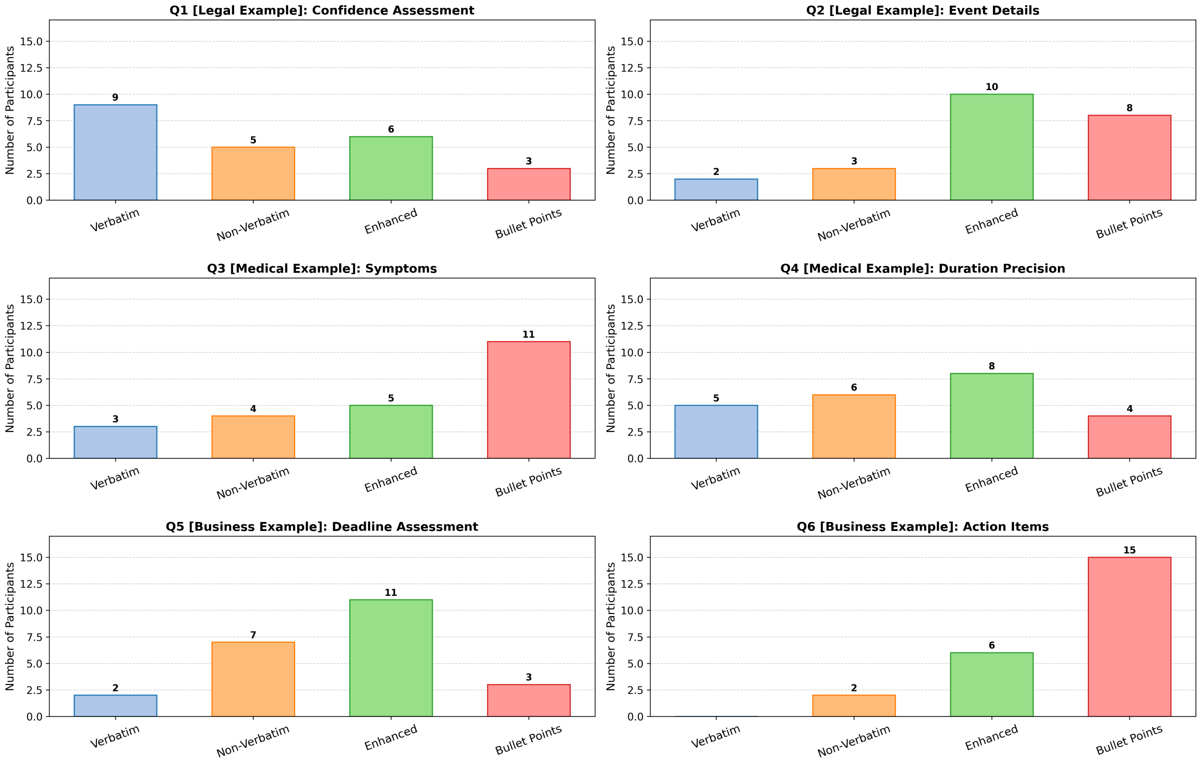


Fig. 4. **User Study Participants' Context-Dependent Fidelity Preferences Across Domains.** Across six task scenarios, participants selected different transcript versions depending on task needs, demonstrating that no single fidelity level universally serves all contexts. For example, verbatim transcripts 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.

Participants selected which transcript version best supported each task. Participants also answered an overall preference question comparing SpeechSpectrum's multi-fidelity approach to traditional single-output ASR systems.

5.1.3 Results. Our results demonstrate clear evidence for context-dependent fidelity preferences, supporting the core argument for user-controllable representation. As shown in Figure 4, participants showed distinct preference patterns based on task requirements rather than universal preferences for higher or lower fidelity levels. Legal confidence assessment (Q1) strongly favored Verbatim transcripts, while legal event details (Q2) preferred Enhanced versions. Medical symptom identification (Q3) overwhelmingly chose Bullet Points, contrasting with medical duration precision (Q4) which showed distributed preferences. Business deadline assessment (Q5) favored Enhanced transcripts, while action items (Q6) strongly preferred Bullet Points.

The study results provide empirical support for our three framework principles. Participants demonstrated sophisticated reasoning about appropriate fidelity levels for different tasks, contradicting assumptions that users cannot make meaningful representational choices. The variation in preferences across scenarios indicates that users possess contextual knowledge about their information need that systems cannot predetermine. Clear preference patterns emerged based on task requirements rather than participant demographics or universal fidelity preferences. This supports our argument that optimal representation depends on use context rather than abstract accuracy metrics.

Participant reasoning revealed understanding that different transcript versions serve different communicative functions, supporting our reconceptualization of STT as translation rather than transcription.

Across professional subgroups, most participants expressed a preference for SpeechSpectrum outputs compared to generic ASR (**Q7.** `Do you prefer SpeechSpectrum outputs to generic ASR outputs?`). All legal experts (N=4) preferred SpeechSpectrum, reflecting the importance of access to verbatim details in legal contexts. Medical experts (N=4) also showed strong support (75%), with one noting:

> *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.

Among ASR experts (N=4), 75% preferred SpeechSpectrum, with one emphasizing 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 (N=18) showed similar patterns, with approximately 72% preferring SpeechSpectrum. Their 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]

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

5.1.4 Discussion. **Our study provides the first empirical evidence that users can meaningfully navigate transcript fidelity choices and that these choices shape both task performance and satisfaction.** Rather than treating transcript representation as a static system output, participants actively selected fidelity levels that aligned with their situational goals. This supports our central argument that speech-to-text systems should treat representation as a user-controllable design parameter rather than a backend optimization target.

These findings challenge dominant practices in STT – specifically ASR – evaluation, which typically rely on universal accuracy metrics such as WER. While such metrics capture transcription fidelity in the narrow sense of word matching, they overlook the broader question of whether a transcript actually serves its intended purpose. Consider two equally valid transcriptions of the same utterance: "I, I think we should go" versus "I think we should go." Against a reference ground truth of "I think we should go," the first would yield a higher WER due to the repetition, yet for a legal professional assessing speaker confidence, the disfluent version may be more valuable than the "cleaner" alternative. Conversely, for a business meeting summary, the cleaned version better serves the user's needs despite being penalized by traditional metrics when evaluated against a verbatim reference. Our results demonstrate that optimal representations are not universal, but context-dependent: a verbatim transcript may be indispensable for assessing a witness's confidence in a legal deposition, while a bullet-point summary may better support a physician scanning a triage note. This reveals a fundamental limitation of WER and similar metrics—they assume a single "correct" representation exists, when in reality the value of a transcript depends entirely on the user's contextual needs and intended use. This shift reframes system quality as a function of how well platforms support diverse user needs across the verbatimicity spectrum.

Importantly, the study reveals that users possess sophisticated intuitions about the relationship between transcript fidelity and task demands. Participants reasoned about when soft linguistic cues (e.g., hesitations, repetitions) mattered, and when concise summaries were preferable. This challenges assumptions that users cannot or should not make representational choices. Instead, our evidence points to the value of increasing user agency in STT interfaces, moving away from architectures that embed fixed representational decisions without user input.

Taken together, these findings position SpeechSpectrum as both a conceptual and practical contribution: it demonstrates that fidelity should be understood as a designable, controllable parameter and shows how giving users agency over this parameter can reshape how STT technologies are evaluated and experienced.

5.2 Study 2: Exploring LLMs as Proxies for Human Preferences

We also conducted a small LLM study, to determine if the preferences of LLMs align with the preferences of users. Recently, LLMs have been proposed as a proxy for humans in social science studies [7, 60, 195]. We aim to discover whether LLMs can act as proxies for human participants in selecting representations with appropriate fidelity levels across SpeechSpectrum.

We created N=23 personas aligned with the respondents from our study, according to their self-identified professional expertise (see Appendix for preliminary study questions **P1-P3**):

```
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.
```

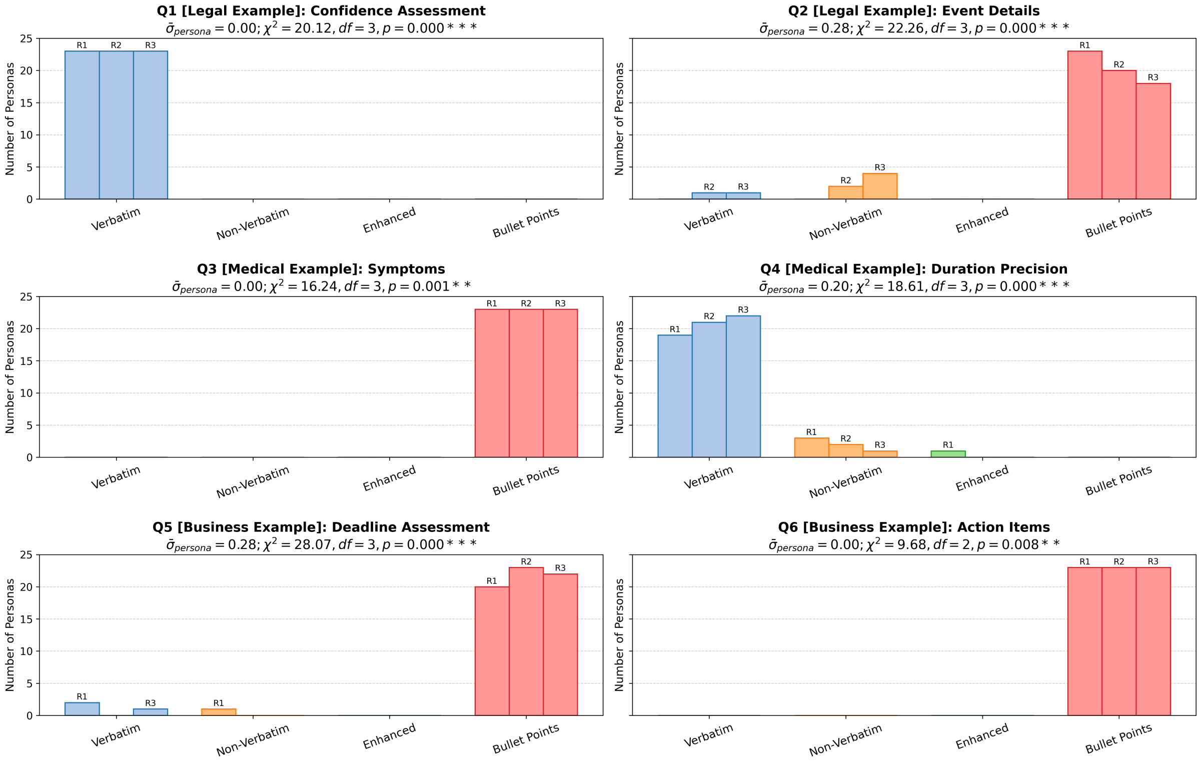


Fig. 5. **LLM Study Shows Extreme Persona Preference Patterns.** When asked to complete the same six scenarios as human participants (Figure 4), LLMs exhibited more extreme preference patterns, often converging on a single transcript type (e.g., consistently selecting bullet points for business tasks). R1, R2, R3 indicate three experimental rounds with different random seeds to account for variability. Compared to humans, who displayed greater variation and nuanced reasoning across contexts, LLMs produced polarized responses. This contrast highlights both the potential and the limitations of using LLMs as proxies for user evaluation in STT research.

For each persona, we asked **Q1-Q6** from our user study detailed in §5.1.2 three times to account for seed-based variability, shown as *Ri* on our figure. We used *gpt-5-mini-2025-08-07*, and we use the *developer* role to control the LLM persona and response format, and the *user* role for the study questions [138].

To measure how much persona responses vary across seed rounds, we mapped the four categorical options (A–D) onto ordinal values (A=1, B=2, C=3, D=4). For each persona and question, we then calculated the standard deviation of responses across seed rounds. For example, if a persona answers $\{A, A, A\}$ for a question, then $\sigma_{persona}(Q_i) = 0$.

To summarize consistency across personas p for each question, we report the average standard deviation:

$$\bar{\sigma}_{persona}(Q_i) \;=\; \frac{1}{N}\sum_{p=1}^{N}\sigma(\text{responses}_{p,Q_i})$$

Lower values of $\bar{\sigma}_{persona}(Q_i)$ indicate that personas converge on a stable response type, while higher values reflect inconsistency in responses – e.g. $\{A, C, A\}$ – across rounds.

Clear patterns emerged by question type:

- **Q1, Q3, Q6:** $\bar{\sigma}_{persona}(Q_i) = 0$, showing perfectly stable responses.
- **Q2, Q4, Q5:** $\bar{\sigma}_{persona}(Q_i) \geq 0.20$, showing small but noticeable variability.

Overall, persona responses are largely stable, though variability may arise for certain questions – highlighting the value of per-question consistency checks in LLM-as-a-proxy setups.

5.2.1 Discussion. As shown in Figure 5, our findings suggest that while LLMs can simulate preference judgments, they tend to produce exaggerated or homogenized choices. Chi-squared analyses confirmed that human and LLM distributions differed significantly across all six questions (all $p < .01$). In our study, LLMs often converged on a single "best" option, showing low variation across personas. In contrast, human participants – who do show consensus in preferences in the aggregate – displayed greater heterogeneity, in line with the findings of Anthis et al. [7]. For example, some lawyers expressed a strong preference for verbatim transcripts, while others preferred concise bullet points. This highlights a key distinction: LLMs, when asked to adopt a role, often amplify the role to an extreme, whereas humans embody roles more flexibly, with personal variation and contextual nuance. This contrast shows both the promise and the limitation of persona-based prompting: while LLMs mirror human patterns in aggregate, they amplify roles to extremes. Without empirical grounding in real user data, systems that rely on LLM outputs risk encoding narrow or distorted assumptions about preferences.

A limitation of our LLM study is potential response bias, as research has shown LLMs exhibit 'yes bias' in grammatical judgments [38]. 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.

5.3 Design Recommendations

Taken together, the patterns observed in Figure 4 and Figure 5 suggest concrete implications for the design of multi-fidelity SpeechSpectrum interfaces. We synthesize these insights into four key recommendations:

- **(R1) Support multi-fidelity interaction.** As shown in Figure 4, no single representation dominates across all questions: while some tasks elicited higher levels of convergence (e.g., Q1, Q3, Q5, Q6), others distributed more evenly across multiple formats (e.g., Q2, Q4). SpeechSpectrum interfaces should therefore allow users to flexibly choose among representations, rather than enforcing a single-output paradigm.
- **(R2) Incorporate task-aware defaults.** In cases where Figure 4 shows clear convergence (e.g., Q1 shows majority Verbatim, Q6 shows overwhelmingly Bullet Points), interfaces could streamline user effort by providing task-aware defaults that match common preferences. As shown in Figure 5, LLMs can approximate these preferences; at the same time, defaults must remain adjustable to preserve user agency.
- **(R3) Enable domain-specific customization.** Both figures reveal systematic domain differences: legal and medical questions show more distributed outputs (e.g., Q2 and Q4 split across several formats), while business questions are sharper and more consistent (e.g., Q6 dominated by Bullet Points). Interfaces should therefore allow tailoring of output style and fidelity to domain-specific needs—for example, structured precision in medical documentation versus concise summaries in business contexts.
- **(R4) Provide educational scaffolding.** Tasks with more diffuse distributions (e.g., Q2 and Q4 in Figure 4) 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.

6 Components for Designing SpeechSpectrum Systems

While SpeechSpectrum provides a conceptual framework for understanding speech-to-text 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 §5.2 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 6, multiple pathways exist for producing different points on the SpeechSpectrum; Table 1 provides exemplars of these approaches. Our goal is to show how such technical components can be mobilized in service of SpeechSpectrum's principles – user-controlled fidelity, context-dependent optimization, and cross-modal translation – thereby 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* [157, 159, 170]; 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[8]

[8] 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 [30].

speech [103, 105], particularly in real-world environments. ASR systems, however, are also efficient and scalable, enabling low-latency transcription across large volumes of speech.

Disfluency Removal Model (DRM). Specialized (often small) disfluency removal models are used for translating verbatim transcriptions to non-verbatim text, where disfluency is generally defined according to the Shriberg definition [161]. Post-disfluency removal, the "fluent" text can be treated as *approximately written text*, and can therefore be processed by an LLM. LLMs are generally not used out-of-the-box as DRMs, because the word-based precision, recall, and F1-score evaluation requires alignment of the ground-truth sequence with the candidate sequence. Instead, sequence classification over the input sequence provides a more natural alignment for evaluation, and these methods dominate the DRM approach space. The DRM models are often fine-tuned and evaluated on the Switchboard dataset [61]. Despite its impact, Switchboard is an outdated and demographically narrow resource. A key challenge for DRMs is generalizing beyond training domains and avoiding removal of meaningful speech phenomena. A key strength for DRMs is their ability to produce fluent, concise text that better supports downstream tasks such as summarization or information extraction.

Large Language Model (LLM). LLMs are conditioned on prior text tokens $x_1, x_2, \ldots, x_t$, such that $P(x_{t+1} \mid x_{1:t})$ to effectively perform next-word prediction for language generation tasks. LLMs are primarily used in the prompt and adaptation (via low rank adapters) 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[9] 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 $P(x_{t+1} \mid x_{1:t}, s_{1:m})$ using fusion-based architectures. 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. Arora et al. [12], Cui et al. [34], Gaido et al. [56] survey recent SLM architectural approaches in detail, while Retkowski et al. [149] survey speech summarization approaches. A key challenge for SLMs is their computational cost and strict audio-speech data requirements, 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.

[9] While the more general Multimodal LLMs (MLLMs) model text, audio, speech, and image, in contrast, SLMs model only text and audio.

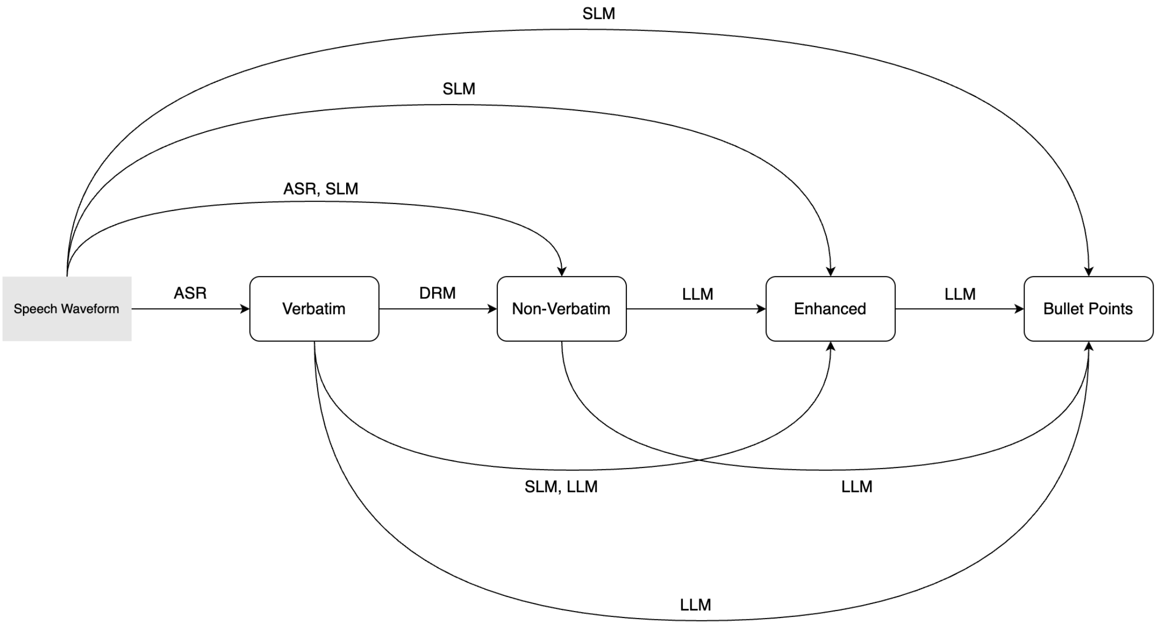


Fig. 6. **Design Pathways for Producing SpeechSpectrum Components.** This diagram illustrates how different tools – ASR, DRM, LLM, and SLM – can be composed to generate transcript representations across fidelity levels. Arrows indicate transformation flows between components (e.g., Verbatim to Non-Verbatim 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) [18, 147], GoogleASR [66], Parakeet-v2 [135]
Verbatim ⟶ Non-Verbatim	DRM	Synthetic Curriculum Learning [25], BERT-Based Parser [116], Planner-Generator [192], Student-Teacher [185], Bi-LSTM [15], Semi-Supervised [183], Noisy Channel [114]
Non-Verbatim ⟶ Enhanced	LLM	Fuse [131], Repair [175], Survey [149]
Enhanced ⟶ Bullet Points	LLM	Survey [149]
	End-to-End	
Speech Waveform ⟶ Non-Verbatim	ASR, SLM	GoogleASR [66], Acoustic-Lexical [181], LSTM/NiN [154], E2E [115]
Speech Waveform ⟶ Enhanced	SLM	Medical RTSS [102], NUTSHELL [202], LongHuBERT [28]
Speech Waveform ⟶ Bullet Points	SLM	No specialized systems, can use SLM prompt-based approach.
Verbatim ⟶ Enhanced	SLM, LLM	Contrastive Student-Teacher [200], Prompt-Based [132], Prompt-Based [93], Chapterization [101]
Verbatim ⟶ Bullet Points	LLM	FLAN-FinBPS [87], Aligned [85], MeetingBank [79]
Non-Verbatim ⟶ Bullet Points	LLM	No specialized systems, can use LLM prompt-based approach.

Table 1. **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.

6.1 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, 191]. This work acknowledges that speech and text have fundamentally different linguistic properties that cannot be collapsed into a single representational approach. Additionally, research has demonstrated that summarization of speech transcripts differs fundamentally from summarization of written text transcripts [149], 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 performance. However, emerging evidence suggests that their representations remain more phonetic than semantic. For example, Choi et al. [31] 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 [157, 159, 170], a task where large, general-purpose models still struggle [144]. Additionally, a modular architecture allows for robust *debugging* practices [95], 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 [149], 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.** 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.

6.2 Evaluating Fidelity Beyond Accuracy for ASR 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* [124]. A new ASR metric, MULTIREFERENCE [124], 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* [33].

Table 2 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 surface-oriented, focusing on lexical overlap rather than deeper dimensions such as fluency, style, or contextual appropriateness. As Gaido et al. [57] 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 [143]	MT	word-level	n-gram precision-based	Utilizes the weighted geometric mean with a *brevity penalty.*
ROUGE-N [109]	MT	word-level	n-gram recall-based	Strictly allows *exact* word matching.
METEOR [20]	MT	word-level (primarily)	unigram F-based	Includes semantic matching for *synonyms*, and correlates well with human evaluations.
CHARCUT [100]	ASR, MT	character-level	n-gram F-like via a longest common subsequence operation	Used for *segment visualization* in interactive ASR user interfaces.

Table 2. **Overview of common ASR 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* [148] and *proximal policy optimization* [156] – 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 ASR solution space is Pairwise Ranking Accuracy (PRA) [52]. Previously proposed for ASR [194], 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 [177], while Deutsch et al. [39] explicitly model ties. These pairwise methods can be operationalized through human ratings or via LLM-as-a-Judge frameworks [71]. 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 4 and Figure 5), 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 ASR evaluation to move beyond the assumption of a singular ground truth.** By aligning evaluation with human judgments, ASR systems can better reflect the wide space of valid outputs encountered in practice.

6.3 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 [161]. From a user perspective, these features are not merely "errors," but 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 [61, 120] have primarily relied on linguistic annotators. 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 **(R7) 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.

7 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 has highlighted this gap as not just a technical shortcoming, but a fundamental HCI challenge of agency, transparency, and of designing for user information needs. With SpeechSpectrum, we introduced a continuum-based framework that repositions transcription as a spectrum rather than a single outcome. Through theoretical framing and an empirical user study, we demonstrated how this perspective can better align system behavior with user expectations and task demands.

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. In the future, STT systems can layer in visualizations of model confidence and offer users finer-grained control over fidelity, so that these systems can better reflect the nuance of spoken communication – even including multimodal signals such as gaze [81] and other visual cues [75, 179]. By reframing transcription as a design space, we aim to inform the next generation of speech technologies—making them not only more accurate, but also more accountable, empowering, and responsive to human needs.

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] 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.

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

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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).

[23] CaseFleet. [n. d.]. *CaseFleet.*

[24] 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.

[25] 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.

[26] 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).

[27] 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.

[28] 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

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

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

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

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

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

[45] 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.

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

[47] 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.

[48] 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.

[49] 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).

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

[51] 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.

[52] 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).

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

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

[55] 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.

[56] 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

[57] 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.

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

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

[60] 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.

[61] 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.

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

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

[64] 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.

[65] 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.

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

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

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

[69] 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).

[70] 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.

[71] 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).

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

[73] 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.

[74] Hirotaka Hiraki and Jun Rekimoto. 2025. SilentWhisper: inaudible faint whisper speech input for silent speech interaction. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems.* 1–6.

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

[76] Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in cognitive sciences* 23, 8 (2019), 639–652.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

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

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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).

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

[108] 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.

[109] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[110] 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.

[111] 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

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] Paria Jamshid Lou and Mark Johnson. 2020. Improving disfluency detection by self-training a self-attentive model. *arXiv preprint arXiv:2004.05323* (2020).

[117] Robbie Love and David Wright. 2021. Specifying challenges in transcribing covert recordings: Implications for forensic transcription. *Frontiers in Communication* 6 (2021), 797448.

[118] Brian MacWhinney. 2007. The talkbank project. In *Creating and digitizing language corpora: Volume 1: Synchronic databases*. Springer, 163–180.

[119] 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.

[120] Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3 (LDC99T42). Web Download.

[121] 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.

[122] Maven AGI. 2025. *Maven AGI.* https://www.mavenagi.com/

[123] 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.

[124] 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

[125] 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).

[126] Meta. [n. d.]. Meta AI Glasses. https://www.meta.com/ai-glasses/.

[127] 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.

[128] 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/

[129] 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.

[130] 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.

[131] 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.

[132] 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).

[133] Nextpoint, Inc. 2025. *Nextpoint.* https://www.nextpoint.com/

[134] Eugene Albert Nida and Charles Russell Taber. 1974. *The theory and practice of translation*. Vol. 8. Brill Archive.

[135] NVIDIA. 2025. Parakeet ASR. https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2.

[136] 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

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

[138] OpenAI. 2025. *GPT-5 System Card*. Technical Report. OpenAI. https://cdn.openai.com/gpt-5-system-card.pdf.

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

[140] 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.*

[141] Laxmi Pandey and Ahmed Sabbir Arif. 2024. MELDER: The Design and Evaluation of a Real-time Silent Speech Recognizer for Mobile Devices. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–23.

[142] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of speech and silent speech input methods in private and public. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[143] 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.

[144] 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.

[145] 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.

[146] 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.

[147] 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.

[148] 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.

[149] 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

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

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

[152] 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.

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

[154] 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.

[155] 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.

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

[157] 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.

[158] 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.

[159] 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

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

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

[162] 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.

[163] 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).

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

[165] 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.

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

[167] 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.

[168] 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.

[169] Zixiong Su, Shitao Fang, and Jun Rekimoto. 2023. Liplearner: Customizable silent speech interactions on mobile devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

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

[171] 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.

[172] 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.

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

[174] 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.

[175] 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

[176] 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.

[177] 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.

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

[179] 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.

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

[181] 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).

[182] 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.

[183] 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.

[184] 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).

[185] 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.

[186] Xue Wang, Zixiong Su, Jun Rekimoto, and Yang Zhang. 2024. Watch your mouth: Silent speech recognition with depth sensing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–15.

[187] 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.

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

[189] 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.

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

[191] 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).

[192] 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.

[193] 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.

[194] 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.

[195] 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).

[196] 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.

[197] 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.

[198] 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.

[199] Juntao Zhou, Dian Ding, Yijie Li, Yu Lu, Yida Wang, Yongzhao Zhang, Yi-Chao Chen, and Guangtao Xue. 2025. M2SILENT: Enabling Multi-user Silent Speech Interactions via Multi-directional Speakers in Shared Spaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[200] 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.

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

[202] 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).

A Additional User Study Details

We aim to understand the professional background of participants with the questions: (i) `Do you work in ASR technology?`, (ii) `Do you work in STEM (science, technology, engineering, and math)?`, (iii) `Do you work in expertise related to legal or medical?`.

```
First, open your browser, and navigate to https://[redacted-for-anonymity].github.io/SpeechSpectrum.
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. 7. **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.

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