A Cross-Community Agenda for Speech AI

Maria Teleki[1,*], Kimi V. Wenzel[2,*], Anna Seo Gyeong Choi[3,*], Tobias Weinberg[4,*], Shree Harsha Bokkahalli Satish[5,*], Stephanny Sanchez[6,*], Belu Ticona[7,*], Ariadna Sanchez[8,*], Yash Sonkar[9,*], Aarti Mathur[10,*], Christoph Minixhofer[11,*], Abraham Glasser[12,*], Raja Kushalnagar[12,*], James Caverlee[1,*], Minha Lee[13,*], Shaomei Wu[14,*], Alyssa Hillary Zisk[15,*], Éva Székely[5,*], Dylan Gaines[16,*], Angelika Seeschaaf Veres[17,*], Seray Ibrahim[18,*], Nicholas Cummins[18,*], Allison Koenecke[3,4,*]

[1]Texas A&M University, [2]Carnegie Mellon University, [3]Cornell University, [4]Cornell Tech, [5]KTH Royal Institute of Technology, [6]University of Alberta, [7]George Mason University [8]University of Edinburgh, [9]IIIT Hyderabad, [10]Chegg Inc., [11]Deepgram, [12]Gallaudet University, [13]Eindhoven University of Technology, [14]AImpower.org, [15]AssistiveWare, [16]Kennesaw State University, [17]OCAD University, [18]King's College London

*** All authors contributed equally. Correspondence:** mariateleki@tamu.edu, kwenzel@andrew.cmu.edu

Abstract

Speech AI, any AI system that recognizes, transforms, or generates speech, is built and evaluated across two communities with only a small overlap: technical natural language processing (NLP) venues (e.g., *ACL, ICASSP, Interspeech), and sociotechnical HCI venues (e.g., ASSETS, CHI, FAccT). In this position paper, we work toward a cross-community synthesis, organizing our critique around three problems: speech AI operates with an incomplete model of communication; it operates with an incomplete model of identity; and its standard metrics measure the wrong constructs. We draw on augmentative and alternative communication (AAC) as a setting where these failures are most visible and their stakes highest, alongside other underserved speakers — people who stutter, multilingual speakers, and non-binary and transgender users. For each problem we offer solution sketches oriented toward designing for human variability, nearly all of which require quantitative and qualitative methods in combination. We close on the venue structures that hold these methods apart, and on what program committees and individual authors can do to bring them together.

Introduction

Speech AI[1] is built and evaluated in two places at once. Technical venues like *ACL, ICASSP, and Interspeech produce the models, the corpora, and the metrics by which progress is declared. Sociotechnical venues like ASSETS, CHI, and FAccT produce the accounts of how people use these systems and corresponding desiderata. The two communities examine the same artifacts, and they overlap only thinly in citation practice, method, and reviewer pool. This position paper argues that the most consequential open problems in speech AI sit in the gap between them, and that closing those problems requires work that neither community can carry out alone.

Speech makes this divide costly in a way that other modalities do not, because speech is uniquely indexical. Unlike text, spoken language *unavoidably* encodes social identity — accent, intonation, speech style, and prosody signal group affiliation, personality, health status, and stance in ways that written language does not (Eckert, 2019; Székely

[1]In this work, we use *speech AI* to refer to the range of systems that recognize, transform, or generate speech, at any level from the individual component to the deployed application. This includes single components operating on their own — ASR transcribing speech to text, TTS synthesizing text into speech, and voice conversion transforming one voice into another. It also includes *chained* pipelines, where an ASR module transcribes speech to text, an LLM operates on that text, and a TTS module synthesizes the response back into speech with each stage handled by a separate model. Further examples of speech AI are newer *full-duplex* spoken dialogue models, which process and generate speech directly and continuously in both directions, without discrete turn-taking between separate modules. Applications built on these components include AAC devices, voice assistants, captioning services, and clinical scribes. See reviews on clinical speech AI (Ng et al., 2026) and SpeechLLMs (Arora et al., 2025; Cui et al., 2025; Yang et al., 2025).

et al., 2025b; Foulkes and Docherty, 2006; Eckert, 2008; Teleki et al., 2025; Weinberg et al., 2025b). A system that recognizes or generates speech is therefore always doing two things simultaneously: performing a measurable technical operation, and making a social judgment about a speaker.

This divide is apparent in how both communities negotiate diversity and fairness. Both communities invoke 'diversity' consistently, yet the term indexes fundamentally different commitments within each research community. In technical venues, diversity is operationalized as demographic breadth in training corpora, (e.g. multiple accents, languages, or speaker age groups), and its adequacy is judged by whether a system generalizes across a predefined taxonomy of speaker categories, evaluated through aggregate metrics disaggregated by group (Dheram et al., 2022; Liu et al., 2022; Papakyriakopoulos et al., 2023). Diversity, in this framing, is a property of *data*. Where atypical speech enters at all, it does so predominantly within a pathological frame, as something to be detected, corrected, or accommodated through better data augmentation (Tobin et al., 2025; Tomanek et al., 2023). The social meaning of a voice is never fully captured by the demographic category it is assigned to, and designing for diversity without accounting for this indexical richness reproduces the very exclusions it claims to remedy.

HCI communities treat diversity instead as a property of *experience*: who is heard, on whose terms, and with what consequences for identity, self-expression, and participation (Erete et al., 2018; Ibrahim et al., 2018; Wu et al., 2025; Michel et al., 2025; Weinberg et al., 2026a). Being represented in a corpus and being served by a system are not equivalent, and conflating them has attracted sustained critique under the label of 'diversity-washing': the risk that superficial inclusion efforts obscure structural inequities or bypass community consent (Cunningham et al., 2025).

The NLP community has begun to grapple with this. Recent work argues that diversity in language technology is an ambiguous and underspecified concept admitting potentially conflicting operationalizations, and calls for the field to explicitly negotiate what diversity *means* before debating how to measure it (Fazelpour and De-Arteaga, 2022; Nguyen and Ploeger, 2025; Bowman and Dahl, 2021). Speech AI has yet to mount a comparable reckoning in defining diversity.

In this position paper, we set an agenda for **a cross-community synthesis**, across technical and sociotechnical venues. We present **a critique of both speech AI communities organized around three problems:**

Problem 1: Speech AI operates with an incomplete model of communication.

Problem 2: Speech AI operates with an incomplete model of identity.

Problem 3: Speech AI metrics measure the wrong constructs.

For each of these three failures, we present **a set of solution sketches**[2] in Speech AI that center human variability, offering a path forward for both scientific communities. We ground each problem and solution sketch in a case study of augmentative and alternative communication (AAC),[3] as this is a speech AI use case where issues are particularly salient and consequential to users. To help researchers see where they can contribute, we tag each solution sketch by the kind of work it invites — from empirical study (USER STUDIES , FIELD STUDIES) and measurement (METRIC CREATION , BENCHMARKING) to system building (SYSTEM BUILDING) and community practice (PARTICIPATORY DESIGN , FIELD-BUILDING); detailed descriptions of each tag are provided in Appendix A.

Problem 1: Speech AI operates with an incomplete model of communication.

We begin with the model of communication that conversational speech AI implicitly assumes, since the commitments made there propagate into how these systems represent identity and how the field measures its own progress. A core idea in contemporary studies of social interaction is that speech is just one ingredient of communication (Levinson, 2025), deeply nestled within a wider ecology of multimodal streams, cooperative acts, and co-constructed practices of two or more parties (Clark, 1996; Goodwin and Heritage, 1990). Wider streams that impact speech and communication can

[2] Similarly to Bowman and Dahl (2021).

[3] Definitions of AAC vary, but generally cover the use of communication methods other than speech for disability-related reasons. AAC supports include communication boards, books, devices, and more; speech AI is relevant to AAC systems with speech output. For AAC users, their AAC is a large part of their identity and presentation (Tuttleturtle (S. Fuller), 2020), making speech AI an important issue for many AAC users.

include suprasegmental features (e.g., stress, intonation, rhythm), non-spoken modes that nonetheless convey linguistic information in signed languages (e.g., facial expression, gaze, gesture) (Hill, 2020), and importantly, collaborative patterns between interactants to establish mutual understanding (Goodwin, 2000). When deciding how we might investigate or design potential speech AI supports, it is important to acknowledge these wider processes.

These dynamics are constitutive of communication; active-listening signals, turn-taking cues, humor, and embodied feedback are part of how speakers assert presence, negotiate meaning, and remain recognizable as themselves across an interaction (Weinberg et al., 2025a,b; Jurafsky et al., 1997). Current speech AI systems, even full-duplex models, largely treat these as noise or omit them entirely, optimizing instead for the verbal channel alone. This is not merely a coverage gap; it reflects a fundamentally impoverished model of what communication is.

Popular metrics reinforce this lack of attention to context and communication. Ultimately, because these models are evaluated using rigid tests that ignore context, high accuracy scores on curated datasets hide a severe measurement gap: their real-world inability to grasp the deeply context-dependent nature of human interaction (Zhang et al., 2025). Transitioning towards scenario-specific evaluation metrics offers a promising path to resolving these blind spots. By developing benchmarks that measure performance within defined social and environmental contexts, future models can successfully bridge this gap, toward achieving contextual intelligence.

Considering speech AI for AAC, numerous prior studies have shown that speech generation via AAC devices can produce understandability issues between interactants (Bloch and Wilkinson, 2004; Higginbotham and Caves, 2002; Ibrahim et al., 2018). These wider streams and conversational practices impact the perception of AAC-generated speech during social interactions and how far AAC users can agentively express themselves during conversations (Valencia et al., 2020; Weinberg et al., 2025b). For AAC users in particular, the inability to produce timely non-verbal cues — whether a backchannel, an expressive reaction, or a well-timed humorous remark — is not only a conversational inconvenience but a constraint on identity expression itself (Weinberg et al., 2025a,b). However, a major focus of speech AI for AAC has been on enhancing speech efficiency with less focus on interactional issues such as reciprocity, conversational management, or identity expression (Ibrahim et al., 2026). We highlight issues of identity expression directly in Problem 2.

Solution Sketch 1

SYSTEM BUILDING USER STUDIES PARTICIPATORY DESIGN

To address these limitations, we suggest prioritizing more realistic benchmarks that measure how well a model responds to intent and social environment, moving beyond isolated transcription scores. This shift in evaluation can, in turn, drive architectural improvements like unified acoustic-semantic embeddings, long-term conversational memory buffers, and the explicit injection of situational metadata.

Future speech AI models should also include other dimensions of communication; for example, emotional expression. Emotion models fit to third-party labels transfer poorly to self-reported emotion (El-Tawil et al., 2026), and closing that gap requires models personalized to the individual (Tavernor and Provost, 2026); we take this as the design problem for expressive AAC output. A personalized expressive model, trained on data the user contributes and governs, would give an AAC user a vocabulary of prosodic states they recognize as their own and can deploy deliberately in conversation. Facial expression and physiological data from a wearable can serve as inputs for users who opt into them, on the terms those users set. Evaluation follows the same logic; the user reports what they intended to convey, communication partners report what they received, and the gap between these is examined under success criteria the user defines.

Caveat: *Sensing of this kind carries documented bias (Holliday and Reed, 2025). Personalization without user control over the model and its data collapses into surveillance, and an expressive system that optimizes a physiological proxy reproduces the failure mode described in Problem 3.*

Problem 2: Speech AI operates with an incomplete model of identity.

We attribute speech AI's incomplete identity model to reductive user models (Problem 2A) and inaccurate representation for speech synthesis users (Problem 2B). Both of these issues feed back into Problem 1, insofar as speech AI's model of communication lacks appropriate recognition and operationalization of user identity.

Problem 2A: Speech AI user models are reductive.

Technologies have co-produced language ideologies in which users are expected to fit certain taxonomies, and variation is treated as pathology instead of data. Models that treat 'atypical' users as pathological, ignorable, or treatable with sufficient data augmentation simplify away the variation that is the data, e.g., the true norm of multilingualism (Schneider, 2022) in favor of national monolingual ideologies. Variations in vocal quality associated with identity markers (Wenzel et al., 2023; Ahmed et al., 2022) or disability (Preece et al., 2024; Li et al., 2025a) are averaged out without anyone asking the users being averaged what they think about it (Radford et al., 2022; Li et al., 2024).

Creating accurate user models is inherently challenged by human identities being neither discrete nor constant. Current taxonomies often rely on reductive frameworks that fail to capture the fluid spectrum of identities individuals embody or the intersectionality of their experiences, such as the compounding acoustic variations of gender, age, and regional dialect. An individual's voice, vocabulary, and speaking style evolve gradually with age (Rojas et al., 2020), dialect drift (Harrington, 2006), gender fluidity (Netzorg et al., 2024), or progressive shifts in communication ability such as with dysarthria (Rosen et al., 2012). They can also adapt immediately to context such as through code-switching (Poplack, 1980) or audience adaptation. Speech AI devices themselves can also impact intra-speaker variation, for example, a device enabling faster input may lead a speaker to be more verbose, while a slower one may encourage brevity.

Resisting any of this through personalization introduces its own tension. Training a model on one's own communication data can better reflect individual tone, vocabulary, and cultural identity, but raises unresolved questions of authorship, data governance, and contextual appropriateness (Rincón et al., 2021; Li et al., 2025b; Weinberg et al., 2026a). Technical obstacles compound the normative ones: post-training methods need data that represents individuals across their multiple identity dimensions, personalized evaluation frameworks barely exist, and many popular LLMs do not publish the model weights that post-training would need to modify.

Solution Sketch 2A

PARTICIPATORY DESIGN SYSTEM BUILDING

Popular AI models are designed and owned by few companies with their own interests and agendas, for whom authentic individual user identities might not represent an interest. Grassroots communities and cross-institutional collaborations that center systematically neglected identities could foster spaces to develop new alternative ways of technology design and model development. Examples include the Masakhane community that develops AI for and by Africans (Nekoto et al., 2020), the Te Hiku Media Project that develops ASR for and by Māori people (Leoni et al., 2024), among others. The Mozilla Data Collective recently released a *community compensation* feature, allowing communities to get paid for their data (Collective, 2026). Community ownership addresses who builds the model; Sketch 3C takes up the complementary question of how such a model would be evaluated across the fluid and intersectional identities described above.

Problem 2B: Speech synthesis users need a voice that accurately represents them.

Where Problem 2A concerned the reductive user models which form the basis of speech AI systems, Problem 2B points towards issues of system output and user experience. When examining speech synthesis output, we encounter issues of both avowed identity (how an individual identifies themself, reflexively) and ascribed identity (how others attribute identities to an individual) (Bucholtz and Hall, 2005; Martin and Nakayama, 2010).

Speech AI systems are trained on vast quantities of data, and the large foundational models at their core generate text and synthesize speech according to the patterns most common across that data. The result is an implicit *averaging effect* wherein individual voices and identity expressions are flattened

(echoing similar concerns in LLM systems Agarwal et al. (2025)). Speech AI tools can reduce and limit users' self-expression (Xu et al., 2025) and the uniqueness of their avowed identities, reducing authorial voice and pushing users toward average or otherwise reductive speech representations, a form of erasure (Wenzel and Kaufman, 2024). This is especially prominent in cases where people feel insecure about their skills or pressured to perform a particular identity (e.g. in AI-led job interviews), as they may surrender their cognitive load to the system (Shaw and Nave, 2026) and conform to a self-representation different from their avowed identity.

Another issue for avowed identity regards how speech AI tools privilege Western cisgender voices. Current English speech synthesis systems are dominated by American and British English accents, and multilingual systems generally are dominated by high-resource languages (Xinyuan et al., 2025; Zhong et al., 2025; Ogun et al., 2024). This means users from other linguistic backgrounds may get a voice that sounds nothing like them or even need to adapt a voice for another language (such as in Callahan and Hanson (2023), where a Czech voice was used for the Native American Lakota), reinforcing linguistic privilege. Systems purporting to generate synthetic AI voices, if not carefully designed and evaluated, risk amplifying existing linguistic hierarchies and accent-based discrimination (Michel et al., 2025). This points to a representational ceiling that technical improvements to synthesis alone cannot solve: the barrier for the most underserved users lies not in generation quality but in the absence of voices that reflect their cultural register, linguistic identity, and sense of self in the first place.

Even with a diverse and representative speech dataset, creating an accurate measure of identity alignment remains difficult due to the nuances of voice identity utilization. For example, AAC users may consider how long they've been using, and grown to identify with, a given voice, how many (possibly more famous) others use the same voice, and the situational use of different voices (Preece et al., 2024) alongside the more typical language, accent, age, gender, and expressiveness. Voice identity in AAC is not fixed at the moment of selection; it accumulates over time through use, accruing social history, relational meaning, and familiarity that can make a long-used voice feel like one's own even when it was never a good fit — and make a better-fitting voice feel wrong precisely because it is unfamiliar (Weinberg et al., 2026b). At the same time, AAC users do not inhabit a single stable identity; self-presentation is often context-dependent, with users navigating different registers across relationships, settings, and life roles (Weinberg et al., 2026b,a). These fluid identity needs surface in concrete voice choices that go beyond the typical dimensions of language, accent, age, and gender. AAC is part of gender expression, so it is part of transgender expression (Tuttleturtle (S. Fuller), 2020; Zisk, 2021; Danielescu et al., 2023; Weinberg et al., 2026b); likewise, a disabled person may want an audibly disabled voice, and may prefer a synthetic one that most listeners find easier to understand than their embodied speech (Preece et al., 2024; Weinberg et al., 2026b). Creating a metric that accurately encompasses the nuances of voice identity alignment remains an unaddressed challenge in speech AI.

When AAC users cannot access voices that match their preferences on all axes, they are forced to pick from a selection of incorrect voices to determine a (possibly situational) least wrong option (Preece et al., 2024; Tuttleturtle (S. Fuller), 2020; Zisk, 2021; Weinberg et al., 2026b). This may lead to unexpected choices in prioritization, such as attempting to sidestep synthetic voices (Weinberg et al., 2026a; Sellwood et al., 2024), a multilingual AAC user selecting a voice on the basis of prosody matching most closely with their identity, even if sacrificing accurate pronunciation (Callahan and Hanson, 2023), or using voices of different genders between different languages (Zisk, Forthcoming). Linguistic minorities may need to adapt voices from other languages and dialects, and some non-binary AAC users have needed to adapt voices from other genders (Tuttleturtle (S. Fuller), 2020; Zisk, 2021) as non-binary AAC voices have often not been available (Danielescu et al., 2023). Across all of these cases, the choice of voice is never purely aesthetic — it is a negotiation between what the system can offer, what the user's social context demands, and what identity the user is trying to project or protect.

When users cannot aptly choose their avowed identity, they impact their ascribed identity and how others may perceive them by their synthetic voice. Listeners readily assign race, gender, and regional identity to synthetic and voice-assistant speech, and those assignments carry the same social consequences — credibility judgments, warmth, and

competence — that accompany perceived identity in human speech (Holliday, 2021, 2023). Importantly, how listeners perceive and socially categorize a given synthetic voice is independent of whether the speaker behind it recognizes it as theirs. This means a voice can be a poor match on one axis and a reasonable match on another; a voice a user identifies with may still be miscategorized by listeners, and a voice built to be broadly intelligible or 'neutral' to listeners may fail to feel like the user's own. The gap between these perspectives is measurable in adjacent settings. Models trained on third-party emotion labels transfer poorly to self-reported emotion (El-Tawil et al., 2026), and closing that gap requires models personalized to the individual, since systems fit to consensus annotations cannot be assumed to recover what a speaker reports about themselves (Tavernor and Provost, 2026). A user may thus select a voice that feels like their own and still be misread by everyone who hears it, which means voice selection is a negotiation across two audiences whose criteria need not agree.

Solution Sketch 2B

SYSTEM BUILDING METRIC CREATION USER STUDIES

Designing better speech AI tools requires giving users accessible ways to articulate, explore, and refine identity-linked vocal preferences over time (Ahmed et al., 2022; Netzorg et al., 2024; Weinberg et al., 2026b). This includes creating an accurate metric for voice identity alignment, and adjusting such metrics to account for both avowed and ascribed voice identity.

Problem 3: Speech AI metrics measure the wrong constructs.

The sub-problems below share a shape: something easy to count is treated as the thing we actually care about. We count how many words a system got wrong, and call that a measure of whether the transcript served the speaker (Problem 3A). We measure how close a synthesized voice sounds to a recording, and call that a measure of whether the voice fits its user (Problem 3B). We average scores across unrelated benchmarks, and call that a measure of how good a system is (Problem 3C). Each number may accurately quantify its chosen proxy, but none fully captures the construct for which it is commonly taken to stand.

Goodhart's law is an old adage, commonly stated as '*When a measure becomes a target, it ceases to be a good measure*.' Once a benchmark is widely adopted, optimization pressure concentrates on the narrow slice of behavior it happens to score, and the metric drifts from the construct it was meant to measure to the proxy itself (Goodhart, 1984; Manheim and Garrabrant, 2018; El-Mhamdi and Hoang, 2024). We hold this is not a marginal concern in speech AI, but the dominant failure mode of how the field currently evaluates its own progress.

Problem 3A: Speech recognition metrics obscure real-world harm.

Speech recognition systems have traditionally been evaluated using quantitative metrics such as word error rate (WER), which quantifies normalized edit distance between model predictions and reference transcriptions, or match error rate (MER), which measures the errors between aligned words in the reference and model output (Morris et al., 2004). While these metrics provide a standardized way to compare model performance, they are known to be inadequate in capturing the downstream user impact of transcription errors. For example, WER treats all substitutions equally, failing to distinguish between harmful substitutions (e.g., profanity, identity-related slurs, or threatening language) and semantically benign errors such as homophones (Wu et al., 2019). To address some of these limitations, more quantitative metrics have been introduced to quantify the semantic distance between the prediction and the reference, including BERTScore (Zhang et al., 2020) and SemScore (Aynetdinov and Akbik, 2024).

Going beyond redesigning error metrics and reference transcripts, evaluation must also directly measure the consequences of errors in situated interaction. For example, can users reliably tell the difference between a speech-to-text systems with average WERs of 5% versus 10%? How do similar recognition errors affect users differently across contexts such as medical documentation, workplace communication, or casual texting? How do different groups interpret and react to the transcription failures differently? These questions suggest that the limitation of current benchmarking metrics are not solely technical, but also contextual and social. Existing metrics have allowed speech AI developers to abstract away the lived experience of interacting with speech AI systems, obscuring

how system failures may reinforce marginalization, exclusion, or psychological harm.

Previous HCI research has empirically explored some of these questions. For example, Wenzel et al. (2023) showed that voice assistant errors produce disparate psychological impacts for Black and white users. Relatedly, Kadoma et al. (2026) found that erroneous transcription of speech reduces both speaker and content evaluations, potentially disproportionately impacting minoritized speakers. Other work has documented how fluent transcription of stuttered speech can reinforce fluency dominance and erase the speaker's identity as a person who stutters (Li et al., 2024), while in contrast, aphasia patients reported preferring their transcriptions to be written in the 'cleanest' form, erasing stutters (Mei et al., 2026). We argue that speech AI evaluation must move beyond aggregate recognition accuracy and toward user-centered evaluation frameworks that explicitly measure contextualized harms and lived user experience.

Developing such evaluation frameworks requires community-centered research that identifies which forms of recognition failure matter most to affected users. Studies involving people with aphasia and people who stutter have shown that generative ASR systems hallucinate more frequently for these populations, sometimes generating degrading, threatening, or violent content that causes tangible harms (Koenecke et al., 2024; Sridhar and Wu, 2025). These findings call for new evaluation measures such as hallucination rate and harmful hallucination rate. Recent work with people who stutter further challenges normative assumptions embedded in conventional ASR evaluation (Li et al., 2025a; Sridhar and Wu, 2025). Existing transcription practices often prioritize 'intended speech' over verbatim speech production, implicitly framing disfluencies as errors to be removed rather than meaningful aspects of communication (Teleki et al., 2024; Radford et al., 2022; Choi et al., 2026b,a), impacting SpeechLLMs (Teleki et al., 2026). The stuttering community expressed a desire for greater control over whether and how speech disfluencies are preserved in transcripts, motivating proposals such as verbatim WER, which evaluates recognition performance against verbatim rather than normalized transcripts (Li et al., 2025a; Sridhar and Wu, 2025).

Developing community-centered evaluation metrics also requires new datasets and annotation methodologies. Existing ASR training and benchmarking datasets routinely exclude so-called 'non-verbal' audio segments and prioritize transcription of intended speech rather than faithful representation of speech production (Radford et al., 2022; Romana et al., 2024). As a result, speech blocks, repetitions, prolongations, and other forms of disfluency that frequently occur in certain groups, such as people who stutter, are often absent from existing datasets. These omissions not only limit model performance for marginalized speech communities, but also encode normative assumptions about what constitutes valid or desirable speech.

Finally, we also need to complement the benchmarking metrics, which are often calculated over clean, static benchmarking datasets, with direct measurement on user experience and sentiment in situated interactions. Prior work by Wenzel et al. (2023), for instance, measured users' affective responses to voice assistants with varying error rates, including self assessments of self-consciousness and self-esteem. Such approaches directly examine the psychological and social impact of speech AI systems, closing the gap between static metrics and lived experience and user well-being.

Solution Sketch 3A

FIELD STUDIES SYSTEM BUILDING PARTICIPATORY DESIGN

Future work should focus on contextualizing performance benchmarks within users' lived experience. This necessitates conducting community-embedded research, and creating new datasets and annotation methods.

Problem 3B: Personalized voice evaluations don't consider AAC user needs.

The same distance between measured performance and lived experience opens on the synthesis side, where the available instruments are fewer and the AAC context is served by none of them. There may be limited quantitative metrics for speech synthesis in general (Wagner et al., 2019), and most speech synthesis is not designed for the AAC context in any case (Pullin et al., 2025). Whether designed for speech synthesis or adapted to speech synthesis from speech recognition, this limits the metrics used in AAC contexts, and we are not aware of any *designed for* the AAC context. Metrics adapted from other areas of speech AI for use in the AAC space include character error rate, word error rate, the structural similarity index, and Speaker

Encoder Cosine Similarity (Székely et al., 2025a; Sanchez and King, 2025). Character error rate, word error rate and structural similarity were used to evaluate the effectiveness of prompting an AAC system with dysarthric speech, possibly analogous to the automation of revoicing (Preece et al., 2024; Pullin et al., 2025; Sellwood et al., 2024). Speaker Encoder Cosine Similarity was used to evaluate similarity to recordings for an AAC user who previously spoke and may also be appropriate to evaluate the similarity of synthesized voices before and after speech engine updates. However, there may not be an appropriate comparator when evaluating and selecting new voices for AAC users who never spoke orally (Preece et al., 2024) or do not want an AAC voice that sounds like their embodied voice (Tuttleturtle (S. Fuller), 2020; Zisk, 2021).

A broader family of objective speech evaluation metrics is also available, including intrusive quality measures such as PESQ (ITU-T P.862) and its successors (Rix et al., 2001), which score a synthesised utterance against a clean recording of the same content in the target voice. This is a requirement often inappropriate in an AAC context, since for a user who never spoke orally (Preece et al., 2024) or who does not want their embodied voice (Tuttleturtle (S. Fuller), 2020; Zisk, 2021), no such reference exists. Reference-free predictors such as UTMOS (Saeki et al., 2022) attempt to predict human preference. While they are non-intrusive, they inherit typical-speech and unfamiliar-listener assumptions instead.

The main limitation with current speech synthesis evaluation for AAC is that it follows standard speech synthesis evaluation methods, which rely on listeners' perceptions of voices that they are not familiar with, when familiarity has been shown to significantly change the perception of cloned voices (Rosi et al., 2025). Even though recent work has proposed a framework for evaluating personalized speech synthesis that aims for a more situationally framed, preference-based approach (Sanchez et al., 2026), it is still a non-user centric evaluation. Although these evaluations may still provide insight into the technical performance of personalized speech synthesis, they do not capture what AAC users themselves want from their voices. As future work, an evaluation for personalised speech synthesis developed alongside AAC users is long overdue.

Solution Sketch 3B

BENCHMARKING USER STUDIES

We propose evaluating personalized voices with three types of listeners: (1) the user of the AAC system, (2) listeners familiar with the user of the voice, and (3) unfamiliar listeners to the user of the voice. For listeners (2) and (3), evaluations should be contextually-framed (Edlund et al., 2024) and mimic real-life use cases wherein the AAC user will communicate with listeners. Evaluating these voices from three distinct listeners can provide more insights, not only based on preference and fidelity to the AAC user, but also in terms of usability in different contexts (e.g., talking to your friend, or ordering a coffee).

Problem 3C: Leaderboards prioritize single-number evaluation.

The limitations in 3A and 3B share a structure: a single number stands in for a construct it partially captures, and the field then optimizes the number.

Widely used community leaderboards such as the HuggingFace Open ASR Leaderboard[4] reinforce this dynamic; a single average of Word Error Rates (WER%) ranks systems across dozens of unrelated benchmarks and conditions. The resulting ranking becomes a reference point for model development, despite potentially reflecting no real deployment context in particular. It can also privilege organizations with the resources to optimize across many benchmarks simultaneously. Examples of other such proxy metrics and associated problems, apart from WER (Phukon et al., 2025) include Mean Opinion Scores (MOS) (Le Maguer et al., 2024; O'Mahony et al., 2021), speaker-similarity cosine distances (Carbonneau et al., 2025), and Multiple Choice Question Answering (MCQA) measures for bias (Lin et al., 2024), as in NLP (Parrish et al., 2022; Nadeem et al., 2021; Jin et al., 2025).

These metrics underplay what users care about, *e.g.*, intelligibility-in-context, voice naturalness and feeling heard, having one's voice and identity reflected back faithfully, whether AI responses are fair and useful to them — or more domain-specific metrics, such as hospital preferences for AI medical scribes to transcribe patient notes using correct medical terminology and integrating seamlessly

[4] Hugging Face Open ASR Leaderboard

with electronic health record systems (Koenecke et al., 2026).

We do not propose abandoning quantitative evaluation; we propose pluralizing it, adding complementing evaluations to it and being explicit about *which metrics matter to whom, and for what decisions*, following Blodgett et al. (2020).

Solution Sketch 3C

BENCHMARKING FIELD-BUILDING PARTICIPATORY DESIGN METRIC CREATION

We propose three commitments: (1) Treat every benchmark as a partial, situated metric and report robustness across voices, prompts, decoding settings, and task formats by default, not as an ablation, so that single-number leaderboards stop being a credible currency (Bokkahalli Satish et al., 2026a, 2025b; Frisch et al., 2026; Ethayarajh and Jurafsky, 2020; Wang et al., 2024).(2) Pair every aggregate metric with at least one evaluation in which affected communities and end-users define and evaluate what forms of evidence matter, how success should be evaluated and how the resulting outcomes are interpreted. For instance, this could mean qualitative studies of a small set of deliberately diverse $N = 1^*$ cases. Recognize and design evaluations for realistic scenarios, human variability and design measurements keeping in mind the people who currently fail or are failed by the system (Bokkahalli Satish et al., 2026b; Bondi et al., 2021; Zee et al., 2024; Justin et al., 2026; Mei et al., 2026), and accept that we are intervening in a complex adaptive system rather than tuning a static objective. (3) Build evaluation *interfaces*, not just scores, that let stakeholders interrogate model behavior directly under controlled perturbations of voice (Bokkahalli Satish et al., 2025a; Li et al., 2025c; Wei et al., 2026). Such interfaces resist value capture because they refuse to compress the evaluation down to a single optimizable number.

[*] $N = 1$ in the sense of highly-intersectional, complex users; consider, for example, a user with a unique set of health conditions, who can realistically say *"there is no one like me."*

A Call for Structural Change

The three problems follow from a division of labor: one community holds the instruments that make speech systems measurable, the other the vocabulary that makes their social consequences describable, and neither suffices alone for a modality that encodes identity in every utterance. Whose voice is heard, on whose terms, and with what consequences for identity are questions that broader corpora and disaggregated metrics can inform but cannot settle by themselves — which is why nearly every sketch we offer carries tags from more than one methodological family, and why measuring what these systems do to people requires benchmarks and field studies, metrics and co-design, in combination.

AAC anchors the argument because it is where the stakes are least deniable: a user whose voice is the system's output, whose identity is negotiated through a menu of imperfect options, and who is judged by listeners on the result. What AAC makes vivid is present wherever speech AI is deployed. Much remains unsettled — what diversity should mean here, whether instruments imported from text and vision can be repaired or need replacing, who holds authority over a voice once it has entered a model — and each of these requires both communities in the room.

Getting them into the room is a matter of incentives, and those incentives are set by reviewers and program committees. We therefore address them directly. The routes available to the kind of work we describe are uneven: *ACL venues admit position work into the main track and route speech submissions through a dedicated speech area, while ICASSP and Interspeech call for technical papers in a single format with no contribution type for a critical argument or reflection, so a claim about what the field should measure reaches those reviewers once it has been instantiated as results. FAccT, CHI, and ASSETS accept critique and construction alike, at some remove from the practitioners who maintain the benchmarks in question. Organizers hold the levers that close this distance: seating area chairs with cross-community expertise, recruiting reviewer pools that can assess both kinds of contribution, creating evaluation tracks and special themes where this work has a natural home, and stating in calls for papers that methodological pluralism is expected. Where those levers go unpulled, the sketches we offer will be written by people

whose venues will not reward them for it, and the technical and sociotechnical communities will continue to publish past each other.

We close with an invitation to build the shared vocabulary this work depends on. At the scale of a community, that means joint workshops and boundary-spanning papers of every kind: position papers such as this one, and equally the mixed-methods studies written to be read on both sides of the divide. At the scale of a single researcher it is smaller, and available without anyone's permission: write the next paper with the other half of the room in mind. Cite across the divide. Recruit a co-author trained differently. Change accumulates one paper at a time.

Acknowledgments

AI tools were used to assist with the language of the paper. We thank everyone who attended and participated in the SpeechAI4All Workshop at CHI '26 (Wenzel et al., 2026a). NC is part-funded by the National Institute for Health and Care Research (NIHR) Biomedical Research Centre (BRC): Maudsley. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. SHBS was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Limitations

This is a position paper, and our contributions are solution sketches: early ideas meant to open cross-community work, not finished methods or evaluated systems. We are excited about the directions they point toward, and we offer them in the expectation that the shape each one finally takes will be worked out by the researchers who take them up — refined, recombined, and put in contact with real users, real data, and real constraints. We see the sketches as starting points, and we look forward to what emerges as different communities build on them.

Our critique draws on AAC as its anchoring case. Given AAC users' reliance on speech AI, we believe this case study may be an instantiation of the curb-cut effect, wherein optimizing for users who rely on these technologies most can yield benefits for the majority of users (Blackwell, 2017). Nonetheless, AAC users are not a monolith, and lessons drawn from this setting will not transfer cleanly to every underserved speaker. Future research may explore the needs of the other communities we invoke more deeply, such as people who stutter, non-binary and transgender voice users, multilingual speakers, speakers of low-resource languages, older adults, children, and others. Our account of 'both communities' likewise flattens real internal heterogeneity within the technical and socio-technical venues we name, and the boundary between them is more porous than our framing sometimes suggests; we use this lens as we believe cross-community collaboration is central to progress. This framing helps to highlight each community's strength and weakness, and opportunities to assist one another.

Ethical Considerations

Several of our sketches carry risks that must be addressed as that work unfolds. Problem 1, in particular, admits physiological and facial sensing as opt-in inputs to an expressive model; such sensing is invasive by construction and raises acute concerns around consent, surveillance, and data governance. The personalization directions in Problem 2 raise parallel questions about who owns a cloned voice and who controls the model trained on someone's speech. Our treatment of Problem 3 also assumes an actor with the resources to evaluate however it chooses. In practice, much of the existing evaluation apparatus is already difficult for smaller speech AI companies to commit resources to partaking in, and our critique of what those instruments measure says nothing about who is in a position to run them. A fuller account would treat the metrics we discuss as constrained by cost as well as by construct. We flag these tensions throughout, but we do not resolve them, and we caution against reading any sketch as a mandate to collect intimate data absent community-defined safeguards.

References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pages 1–21.

Alex A Ahmed, Levin Kim, and Anna L Hoffmann. 2022. 'this app can help you change your voice': Authenticity and authority in mobile applications for transgender voice training. *Convergence*, 28(5):1283–1302.

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*.

Ansar Aynetdinov and Alan Akbik. 2024. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity. *Preprint*, arXiv:2401.17072.

Cynthia L. Bennett, Erin Brady, and Stacy M. Branham. 2018. Interdependence as a Frame for Assistive Technology Research and Design. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '18, pages 161–173, New York, NY, USA. Association for Computing Machinery.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8.

Angela Glover Blackwell. 2017. The curb-cut effect. *Stanford Social Innovation Review*, 15(1):28–33.

Steven Bloch and Ray Wilkinson. 2004. The understandability of AAC: A conversation analysis study of acquired dysarthria. *Augmentative and Alternative Communication*, 20(4):272–282.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5454–5476.

Shree Harsha Bokkahalli Satish, Gustav Eje Henter, and Éva Székely. 2025a. Hear Me Out: Interactive evaluation and bias discovery platform for speech-to-speech conversational AI. In *Proc. Interspeech*, pages 2151–2152. International Speech Communication Association.

Shree Harsha Bokkahalli Satish, Gustav Eje Henter, and Éva Székely. 2025b. When Voice Matters: Evidence of Gender Disparity in Positional Bias of Speech-LLMs. In *International Conference on Speech and Computer*, pages 25–38.

Shree Harsha Bokkahalli Satish, Gustav Eje Henter, and Éva Székely. 2026a. Do Bias Benchmarks Generalise? Evidence from Voice-Based Evaluation of Gender Bias in SpeechLLMs. In *Proc. ICASSP*, pages 4566–4570. IEEE.

Shree Harsha Bokkahalli Satish, Christoph Minixhofer, Maria Teleki, James Caverlee, Ondrej Klejch, Peter Bell, Gustav Eje Henter, and Éva Székely. 2026b. The Voice Behind the Words: Quantifying Intersectional Bias in SpeechLLMs. *arXiv preprint arXiv:2603.16941*.

Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A Killian. 2021. Envisioning communities: a participatory approach towards ai for social good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 425–436.

Samuel Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

Janet Callahan and Elizabeth K Hanson. 2023. Reconnecting indigenous language for a child using augmentative and alternative communication. *Language, Speech, and Hearing Services in Schools*, 54(2):387–394.

Marc-André Carbonneau, Benjamin van Niekerk, Hugo Seuté, Jean-Philippe Letendre, Herman Kamper, and Julian Zaïdi. 2025. Analyzing and improving speaker similarity assessment for speech synthesis. *arXiv preprint arXiv:2507.02176*.

Anna Seo Gyeong Choi, Maria Teleki, James Caverlee, Miguel del Rio, Corey Miller, and Hoon Choi. 2026a. Beyond single ground truth: Reference monism as epistemic injustice in ASR evaluation. In *Preprint*.

Anna Seo Gyeong Choi, Maria Teleki, Miguel del Rio, James Caverlee, Corey Miller, and Allison Koenecke. 2026b. Speechspectrum: A linguistic fidelity spectrum for accountable speech-to-text. In *Preprint*.

Herbert H. Clark. 1996. *Using Language*. "Using" Linguistic Books. Cambridge University Press.

Mozilla Data Collective. 2026. Get a sneak preview of Mozilla Data Collective's community compensation feature! LinkedIn Post. Accessed: 2026-07-09.

Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. MIT press.

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)*, pages 13943–13970.

Jay L Cunningham, Kevin Zhongyang Shao, Rock Yuren Pang, and Nathanael Elias Mengist. 2025. Advancing NLP data equity: Practitioner responsibility and accountability in NLP data practices.

In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 679–691.

Andreea Danielescu, Sharone A Horowit-Hendler, Alexandria Pabst, Kenneth Michael Stewart, Eric M Gallo, and Matthew Peter Aylett. 2023. Creating inclusive voices for the 21st century: A non-binary text-to-speech for conversational assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, pages 1–23, New York, NY, USA. Association for Computing Machinery.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. In *Proc. Interspeech*, pages 1268–1272.

Penelope Eckert. 2008. Variation and the indexical field 1. *Journal of sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2019. The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language*, 95(4):751–776.

Jens Edlund, Christina Tånnander, Sébastien Le Maguer, and Petra Wagner. 2024. Assessing the impact of contextual framing on subjective TTS quality. In *Proc. Interspeech*, pages 1205–1209.

El-Mahdi El-Mhamdi and Lê-Nguyên Hoang. 2024. On goodhart's law, with an application to value alignment. *arXiv preprint arXiv:2410.09638*.

Yara El-Tawil, Aneesha Sampath, and Emily Mower Provost. 2026. What you feel is not what they see: On predicting self-reported emotion from third-party observer labels. *arXiv preprint arXiv:2601.21130*.

Sheena Erete, Aarti Israni, and Tawanna Dillahunt. 2018. An intersectional approach to designing in the margins. *interactions*, 25(3):66–69.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machine learning systems. *Big Data & Society*, 9(1):20539517221082027.

Paul Foulkes and Gerard Docherty. 2006. The social life of phonetics and phonology. *Journal of phonetics*, 34(4):409–438.

Ursula Martius Franklin. 2014. *Ursula Franklin speaks: Thoughts and afterthoughts*. McGill-Queen's Press-MQUP.

Blade Frisch, Will Wade, Dylan Gaines, Michelle Kinsella, Betts Peters, Tamara Broderick, and Keith Vertanen. 2026. It's complicated: On the design and evaluation of ai-powered aac interfaces. In *Speech AI for All: The What, How, and Who of Measurement Workshop at the CHI Conference on Human Factors in Computing Systems*.

Patricia Garcia, Tonia Sutherland, Marika Cifor, Anita Say Chan, Lauren Klein, Catherine D'Ignazio, and Niloufar Salehi. 2020. No: Critical refusal as feminist data practice. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, pages 199–202.

Charles A. E. Goodhart. 1984. Problems of Monetary Management: The U.K. Experience. *Papers in Monetary Economics*, 1:91–121.

Charles Goodwin. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10):1489–1522.

Charles Goodwin and John Heritage. 1990. Conversation Analysis. *Annual Review of Anthropology*, 19(1):283–307. _eprint: https://doi.org/10.1146/annurev.an.19.100190.001435.

Jonathan Harrington. 2006. An acoustic analysis of 'happy-tensing'in the queen's christmas broadcasts. *Journal of Phonetics*, 34(4):439–457.

D Jeffery Higginbotham and Kevin Caves. 2002. AAC performance and usability issues: The effect of AAC technology on the communicative process. *Assistive Technology*, 14(1):45–57.

Joseph Hill. 2020. Do deaf communities actually want sign language gloves? *Nature Electronics*, 3(9):512–513.

Nicole Holliday. 2023. Siri, you've changed! acoustic properties and racialized judgments of voice assistants. *Frontiers in Communication*, 8:1116955.

Nicole R Holliday. 2021. Perception in black and white: Effects of intonational variables and filtering conditions on sociolinguistic judgments with implications for asr. *Frontiers in Artificial Intelligence*, 4:642783.

Nicole R Holliday and Paul E Reed. 2025. Gender and racial bias issues in a commercial "tone of voice" analysis system. *PloS one*, 20(2):e0314470.

Seray B Ibrahim, Tom Griffiths, Michael Clarke, Simon Judge, Petr Slovak, Graham Pullin, and Jeff Higginbotham. 2026. Before the Technological Fix: Scoping AI and AAC for Social Futures. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, pages 1–10, New York, NY, USA. Association for Computing Machinery.

Seray B. Ibrahim, Asimina Vasalou, and Michael Clarke. 2018. Design Opportunities for AAC and Children with Severe Speech and Physical Impairments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 227:1–227:13, New York, NY, USA. ACM.

Jiho Jin, Woosung Kang, Junho Myung, and Alice Oh. 2025. Social bias benchmark for generation: A comparison of generation and qa-based evaluations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11215–11228.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 88–95. IEEE.

Bakoubolo Essowe Justin, Catherine Nana Nyaah Essuman, Messan Agbobli, Ahoefa Kansiwer, Eli Jean Doumeyan, Julie Pato, Notou Your Timibe, Emile KOGBEDJI Agossou, and Guedela Bakouya. 2026. Eyaa-tom 26, yodi-mantissa and lom bench: A community benchmark for tts in local languages. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 264–270.

Kowe Kadoma, Priyal Shrivastava, and Mor Naaman. 2026. Lost in transcription: Subtitle errors in automatic speech recognition reduce speaker and content evaluations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

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*, FAccT '24, page 1672–1681, New York, NY, USA. Association for Computing Machinery.

Allison Koenecke, John-Jose Nunez, Anaïs Rameau, and Irene Y. Chen. 2026. Perspective: Listening to users when auditing medical ai scribes. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pages 1619–1631. PMLR.

Sébastien Le Maguer, Simon King, and Naomi Harte. 2024. The limits of the mean opinion score for speech synthesis evaluation. *Computer Speech & Language*, 84:101577.

Gianna Leoni, Lee Steven, Tūreiti Keith, Keoni Mahelona, Peter-Lucas Jones, and Suzanne Duncan. 2024. Solving failure modes in the creation of trustworthy language technologies. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages @ LREC-COLING 2024*, pages 325–330, Torino, Italia. ELRA and ICCL.

Stephen C. Levinson. 2025. *The Interaction Engine: Language in Social Life and Human Evolution*, 1 edition. Cambridge University Press.

Jingjin Li, Qisheng Li, Rong Gong, Lezhi Wang, and Shaomei Wu. 2025a. Our collective voices: The social and technical values of a grassroots chinese stuttered speech dataset. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 2768–2783, New York, NY, USA. Association for Computing Machinery.

Jingjin Li, Peiyao Liu, Rebecca Lietz, Ningjing Tang, Norman Makoto Su, and Shaomei Wu. 2025b. Govern with, not for: Understanding the stuttering community's preferences and goals for speech ai data governance in the us and china. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1548–1560.

Jingjin Li, Shaomei Wu, and Gilly Leshed. 2024. Re-envisioning remote meetings: Co-designing inclusive and empowering videoconferencing with people who stutter. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, page 1926–1941, New York, NY, USA. Association for Computing Machinery.

Minzhi Li, William Barr Held, Michael J Ryan, Kunat Pipatanakul, Potsawee Manakul, Hao Zhu, and Diyi Yang. 2025c. Mind the gap: Static and interactive evaluations of large audio models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8749–8766.

Yi-Cheng Lin, Wei-Chih Chen, and Hung-yi Lee. 2024. Spoken stereoset: on evaluating social bias toward speaker in speech large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 871–878. IEEE.

Zhe Liu, Irina-Elena Veliche, and Fuchun Peng. 2022. Model-based approach for measuring the fairness in ASR. In *Proc. ICASSP*, pages 6532–6536. IEEE.

David Manheim and Scott Garrabrant. 2018. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*.

Judith N Martin and Thomas K Nakayama. 2010. *Intercultural communication in contexts*. United States: The McGraw-Hill Companies.

Katelyn X Mei, Anna Seo Gyeong Choi, Hilke Schellmann, Mona Sloane, and Allison Koenecke. 2026. Addressing auditing pitfalls in automatic speech recognition technologies: A case study of people with aphasia. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 3422–3465.

Shira Michel, Sufi Kaur, Sarah Elizabeth Gillespie, Jeffrey Gleason, Christo Wilson, and Avijit Ghosh. 2025. "it's not a representation of me": Examining accent

bias and digital exclusion in synthetic ai voice services. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–245.

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 *Interspeech*, 4-8, page 2004.

Christine Murad, Humaira Tasnim, and Cosmin Munteanu. 2022. "voice-first interfaces in a gui-first design world": Barriers and opportunities to supporting vui designers on-the-job. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–10.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. 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)*, pages 5356–5371.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

Robin Netzorg, Alyssa Cote, Sumi Koshin, Klo Vivienne Garoute, and Gopala Krishna Anumanchipalli. 2024. Speech after gender: a trans-feminine perspective on next steps for speech science and technology. *arXiv preprint arXiv:2407.07235*.

Si-Ioi Ng, Lingfeng Xu, Ingo Siegert, Nicholas Cummins, Nina R Benway, Julie Liss, and Visar Berisha. 2026. An end-to-end overview of clinical speech ai. *IEEE Transactions on Audio, Speech and Language Processing*.

Dong Nguyen and Esther Ploeger. 2025. We need to measure data diversity in NLP—better and broader. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8823–8832.

Sewade Ogun, Abraham T Owodunni, Tobi Olatunji, Eniola Alese, Babatunde Oladimeji, Tejumade Afonja, Kayode Olaleye, Naome A Etori, and Tosin Adewumi. 2024. 1000 african voices: Advancing inclusive multi-speaker multi-accent speech synthesis. *arXiv preprint arXiv:2406.11727*.

Johannah O'Mahony, Pilar Oplustil-Gallegos, Catherine Lai, and Simon King. 2021. Factors affecting the evaluation of synthetic speech in context. In *11th ISCA Speech Synthesis Workshop*, pages 148–153.

Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. 2023. Augmented datasheets for speech datasets and ethical decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 881–904.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

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. *arXiv preprint arXiv:2506.16528*.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching. *Linguistics*, 18:581–618.

Jamie Preece, Emma Sullivan, Fin Tams-Gray, and Graham Pullin. 2024. Making my voice and owning its future. *Medical Humanities*, 50(4):624–634.

Graham Pullin, Kevin Williams, Darryl Sellwood, Jamie Preece, Emma Sullivan, Meredith Allan, Todd Hutchinson, Katie Brown, Fin Tams-Gray, and Jeff Higginbotham. 2025. Discussing future AAC: technology, interactions and ownership. *Augmentative and Alternative Communication*, pages 1–7.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Cami Rincón, Os Keyes, and Corinne Cath. 2021. Speaking from experience: Trans/non-binary requirements for voice-activated ai. *Proceedings of the ACM on human-computer interaction*, 5(CSCW1):1–27.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE international conference on acoustics, speech, and signal processing*, volume 2, pages 749–752.

Sandra Rojas, Elaina Kefalianos, and Adam Vogel. 2020. How does our voice change as we age? a systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age. *Journal of Speech, Language, and Hearing Research*, 63(2):533–551.

Amrit Romana, Minxue Niu, Matthew Perez, and Emily Mower Provost. 2024. Fluencybank timestamped: An updated data set for disfluency detection and automatic intended speech recognition. *Journal of Speech, Language, and Hearing Research*, 67(11):4203–4215.

Kristin M Rosen, Joanne E Folker, Adam P Vogel, Louise A Corben, Bruce E Murdoch, and Martin B Delatycki. 2012. Longitudinal change in dysarthria associated with friedreich ataxia: a potential clinical endpoint. *Journal of neurology*, 259(11):2471–2477.

Victor Rosi, Emma Soopramanien, and Carolyn McGettigan. 2025. Perception and social evaluation of cloned and recorded voices: Effects of familiarity and self-relevance. *Computers in Human Behavior: Artificial Humans*, 4.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Ariadna Sanchez and Simon King. 2025. Can We Reconstruct a Dysarthric Voice with the Large Speech Model Parler TTS? In *Proc. Interspeech*, pages 4138–4142.

Ariadna Sanchez, Christoph Minixhofer, Korin Richmond, Ondrej Klejch, Peter Bell, and Simon King. 2026. An Evaluation Framework for Text-to-Speech Voice Reconstruction. In *To appear at Interspeech 2026*.

Britta Schneider. 2022. Multilingualism and ai: The regimentation of language in the age of digital capitalism. *Signs and Society*, 10(3):362–387.

Darryl Sellwood, Lateef McLeod, Kevin Williams, Katie Brown, and Graham Pullin. 2024. Imagining alternative futures with augmentative and alternative communication: a manifesto. *Medical Humanities*, 50(4):620–623.

Steven D Shaw and Gideon Nave. 2026. Thinking-fast, slow, and artificial: How ai is reshaping human reasoning and the rise of cognitive surrender. *Available at SSRN 6097646*.

Jaisie Sin, Heloisa Candello, Leigh Clark, Benjamin R Cowan, Minha Lee, Cosmin Munteanu, Martin Porcheron, Sarah Theres Völkel, Stacy Branham, Robin N Brewer, et al. 2023. Cui@ chi: Inclusive design of cuis across modalities and mobilities. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–5.

Jaisie Sin, Sho Conte, Cosmin Munteanu, Jules Maitland, Novia Nurain, Sayan Sarcar, and Wei Zhao. 2025. Beyond deficit-based design: Re-imagining co-creation approaches with equity-denied groups. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–5.

Jaisie Sin, Rachel L. Franz, Cosmin Munteanu, and Barbara Barbosa Neves. 2021. Digital design marginalization: New perspectives on designing inclusive interfaces. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Charan Sridhar and Shaomei Wu. 2025. J-j-j-just stutter: Benchmarking whisper's performance disparities on different stuttering patterns. In *Proc. Interspeech*, pages 3753–3757.

Éva Székely, Péter Mihajlik, Máté Soma Kádár, and László Tóth. 2025a. Voice reconstruction through large-scale tts models: Comparing zero-shot and fine-tuning approaches to personalise tts in assistive communication. In *Proc. Interspeech*, pages 2735–2739.

Éva Székely, Jura Miniota, and Míša Michaela Hejná. 2025b. Will AI shape the way we speak? the emerging sociolinguistic influence of synthetic voices. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 335–340.

James Tavernor and Emily Mower Provost. 2026. It is personal: The importance of personalization for recognizing self-reported emotion. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16202–16206. IEEE.

Maria Teleki, Xiangjue Dong, Soohwan Kim, and James Caverlee. 2024. Comparing asr systems in the context of speech disfluencies. In *Interspeech 2024*, pages 4548–4552.

Maria Teleki, Xiangjue Dong, Haoran Liu, and James Caverlee. 2025. Masculine defaults via gendered discourse in podcasts and large language models. In *ICWSM 2025*.

Maria Teleki, Sai Janjur, Haoran Liu, Oliver Grabner, Ketan Verma, Thomas Docog, Xiangjue Dong, Lingfeng Shi, Cong Wang, Stephanie Birkelbach, Jason Kim, Yin Zhang, Éva Székely, and James Caverlee. 2026. Conversational speech reveals structural robustness failures in speechllm backbones.

Jimmy Tobin, Katrin Tomanek, and Subhashini Venugopalan. 2025. Towards a single ASR model that generalizes to disordered speech. In *Proc. ICASSP*, pages 1–5. IEEE.

Katrin Tomanek, Katie Seaver, Pan-Pan Jiang, Richard Cave, Lauren Harrell, and Jordan R Green. 2023. An analysis of degenerating speech due to progressive dysarthria on ASR performance. In *Proc. ICASSP*, pages 1–5. IEEE.

Jutta Treviranus. 2018. The three dimensions of inclusive design: Part one. *Medium*.

Jutta Treviranus. 2025. We need new research metrics if we are to address disparity. LinkedIn Post. Accessed: 2026-05-02.

Tuttleturtle (S. Fuller). 2020. My AAC is part of my gender presentation. In *AAC in the Cloud*.

Stephanie Valencia, Amy Pavel, Jared Santa Maria, Seunga (Gloria) Yu, Jeffrey P. Bigham, and Henny Admoni. 2020. Conversational Agency in Augmentative

and Alternative Communication. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, New York, NY, USA. Association for Computing Machinery.

Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, et al. 2019. Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*.

Angelina Wang, Aaron Hertzmann, and Olga Russakovsky. 2024. Benchmark suites instead of leaderboards for evaluating ai fairness. *Patterns*, 5(11).

Sheng-Lun Wei, Yu-Ling Liao, Yen-Hua Chang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2026. Bias in the ear of the listener: Assessing sensitivity in audio language models across linguistic, demographic, and positional variations. *arXiv preprint arXiv:2602.01030*.

Tobias M Weinberg, Ricardo E Gonzalez Penuela, Stephanie Valencia, and Thijs Roumen. 2026a. I, robot? exploring ultra-personalized ai-powered AAC; an autoethnographic account. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Tobias M Weinberg, Kowe Kadoma, Ricardo E Gonzalez Penuela, Stephanie Valencia, and Thijs Roumen. 2025a. Why so serious? exploring timely humorous comments in aac through ai-powered interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Tobias M Weinberg, Aaleyah Lewis, Ricardo E Gonzalez Penuela, Weicong Hong, Jennifer Mankoff, and Thijs Roumen. 2026b. Me, myself, and my voice: Exploring cultural and linguistic identity in aac ai-generated voices. *arXiv preprint arXiv:2605.24337*.

Tobias M Weinberg, Claire O'Connor, Ricardo E. Gonzalez Penuela, Stephanie Valencia, and Thijs Roumen. 2025b. One Does Not Simply 'Mmhmm': Exploring Backchanneling in the AAC Micro-Culture. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '25, pages 1–14, New York, NY, USA. Association for Computing Machinery.

Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. 2023. Can voice assistants be microaggressors? cross-race psychological responses to failures of automatic speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Kimi Wenzel and Geoff Kaufman. 2024. Designing for harm reduction: Communication repair for multicultural users' voice interactions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Kimi Wenzel, Alisha Pradhan, Maria Teleki, Tobias M Weinberg, Robin Netzorg, Alyssa Hillary Zisk, Anna Seo Gyeong Choi, Jingjin Li, Raja Kushalnagar, Colin Lea, Abraham Glasser, Christian Vogler, Ly Xundefinednzhèn M. Zhundefinedngsūn Brown, Nan Bernstein Ratner, Allison Koenecke, Karen Nakamura, and Shaomei Wu. 2026a. Speech ai for all: The what, how, and who of measurement. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, New York, NY, USA. Association for Computing Machinery.

Kimi Wenzel, Alisha Pradhan, Maria Teleki, Tobias M Weinberg, Robin Netzorg, Alyssa Hillary Zisk, Anna Seo Gyeong Choi, Jingjin Li, Raja Kushalnagar, Colin Lea, et al. 2026b. Speech ai for all: The what, how, and who of measurement. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Shaomei Wu, Lindsay Reynolds, Xian Li, and Francisco Guzmán. 2019. Design and evaluation of a social media writing support tool for people with dyslexia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

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*, pages 1–6.

Henry Li Xinyuan, Zexin Cai, Ashi Garg, Kevin Duh, Leibny Paola García-Perera, Sanjeev Khudanpur, Nicholas Andrews, and Matthew Wiesner. 2025. Scalable controllable accented tts. *arXiv preprint arXiv:2508.07426*.

Yiwen Xu, Monideep Chakraborti, Tianyi Zhang, Katelyn Eng, Aanchan Mohan, and Mirjana Prpa. 2025. Your voice is your voice: Supporting self-expression through speech generation and llms in augmented and alternative communication. *arXiv preprint arXiv:2503.17479*.

Zhengdong Yang, Shuichiro Shimizu, Yahan Yu, and Chenhui Chu. 2025. When large language models meet speech: A survey on integration approaches. *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20298–20315.

Anna Zee, Marc Zee, and Anders Søgaard. 2024. Group fairness in multilingual speech recognition models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2213–2226.

Linhao Zhang, Jian Zhang, Bokai Lei, Chuhan Wu, Aiwei Liu, Wei Jia, and Xiao Zhou. 2025. Wildspeech-bench: Benchmarking end-to-end speechllms in the wild. *arXiv preprint arXiv:2506.21875*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Jinzuomu Zhong, Korin Richmond, Zhiba Su, and Siqi Sun. 2025. Accentbox: Towards high-fidelity zero-shot accent generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alyssa Hillary Zisk. 2021. "why are you using a male voice?". *CommunicationFIRST Perspectives*.

Alyssa Hillary Zisk. Forthcoming. Trying to say 怎么fucking了: code-mixing with AAC (augmentative and alternative communication). In Ruchi Kapila and Mai Ling Chan, editors, *Community Engagement for Speech-Language Pathology: From Lived Experience to Clinical Expertise*. Cognella Publishing.

A Contributor Tags

Each solution sketch in the main text carries one or more *contributor tags* that signal the kind of work the sketch invites, so that readers can scan for the sketches where their own methods apply. The seven tags fall into four color families: *empirical study* (USER STUDIES , FIELD STUDIES), *measurement* (METRIC CREATION , BENCHMARKING), *system building* (SYSTEM BUILDING), and *community practice* (PARTICIPATORY DESIGN , FIELD-BUILDING). Most sketches span more than one: a single piece of work can, for instance, build a system and evaluate it with a user study. Here, we define each tag:

USER STUDIES User Studies

Controlled, participant-facing empirical work: perception and listening tests, lab experiments, surveys, and other structured evaluations of how people perceive or respond to a system.

FIELD STUDIES Field Studies

Situated, in-context empirical work in real settings: interviews, ethnography, lived-experience accounts, and longitudinal or community-embedded observation.

METRIC CREATION Metric Creation

Designing and validating a new measure or instrument: formalizing a construct and showing it captures what it claims to.

BENCHMARKING Benchmarking

Building, running, and reporting evaluations against a measure: datasets, protocols, robustness reporting, and leaderboard practice.

SYSTEM BUILDING System Building

Designing and implementing models, architectures, interfaces, datasets, or pipelines.

PARTICIPATORY DESIGN Participatory Design

Work done *with* affected communities: co-design, data governance, ownership, and community-defined success criteria.

FIELD-BUILDING Field-Building

Work that changes how the field itself evaluates and rewards progress: cross-community infrastructure such as shared evaluation tracks, joint workshops, boundary-spanning position papers, and shifts in what the field treats as credible evidence.

B Inclusive Design Practices

If we are serious about reducing inequity and supporting meaningful innovation, we need to rethink what we measure. Alternative metrics should focus on the extent to which efforts reach and make a substantive difference for people facing the greatest barriers to inclusion, and the degree to which projects engage with complex, high-effort challenges that may take longer but have the potential for deeper, systemic impact. This reframing aligns with critiques by Jutta Treviranus (Treviranus, 2025), who argues that prevailing evaluation systems undermine inclusive innovation by rewarding ease and scalability over necessity and equity.

Several overlapping traditions make a related argument from different starting points: design justice centers the leadership and needs of communities most harmed by design processes (Costanza-Chock, 2020), data feminism names how supposedly neutral data practices encode existing power hierarchies (Garcia et al., 2020), and participatory AI calls for affected communities to shape system design and evaluation directly rather than being consulted after the fact (Birhane et al., 2022). We draw here specifically on Treviranus's (Treviranus, 2018) inclusive design framework, which advances an approach grounded in recognizing and designing for human variability, rather than conformity to an assumed norm. It emphasizes open, transparent, and participatory processes, including co-design with individuals who are often excluded or underserved by conventional design practices and technologies as defined by Ursula Franklin, "*Technology involves organizations, procedures, symbols, new words, equations, and most of all, Mindsets*." (Franklin, 2014). Situated within complex adaptive systems, Inclusive Design and

wider situated design frameworks (e.g., Bennett et al., 2018; Costanza-Chock, 2020; Delgado et al., 2023; Birhane et al., 2022) understand design outcomes as emergent, contextualized, and interdependent, challenging the assumption that they can be optimized through standardization or efficiency-driven models. These frameworks call for alternative evaluation criteria: examining the extent to which technologies meaningfully engage those facing the greatest barriers to inclusion, and the degree to which projects address complex, systemic challenges with the potential for transformative impact. In this way, inclusive design not only reorients design practice but also redefines what counts as success and what mindset we apply to technology.

C Exclusionary Practices

Often, deficit-based design sees disability as something to *fix* — with technology's help (Sin et al., 2025). The problem is how pervasive uninclusive designs, i.e., designs that perpetuate exclusion of certain communities, introduce marginalization in which excluded communities experience deeper marginalization due to the increased lack of access; this can be, for instance, introduced inability to access healthcare apps or social media, which can be detrimental when considering enlarging social and psychological exclusion over time, a process termed as digital design marginalization (Sin et al., 2021). While the problem is noted and inclusive practices are understood to be important (Sin et al., 2023; Wu et al., 2025; Wenzel et al., 2026b), there are additional hurdles, such as voice-first design principles being difficult to identify and implement, as VUI designers often rely on their prior GUI design experience for voice-based interaction design (Murad et al., 2022).