The Conversational Context Problem: Why Speech Datasets Violate Consent by Design

ANNA SEO GYEONG CHOI, Cornell University, USA

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

ACM Reference Format:

Anna Seo Gyeong Choi. 2025. The Conversational Context Problem: Why Speech Datasets Violate Consent by Design. In *Proceedings of CSCW 2025 (CSCW '25).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

As foundation models increasingly incorporate speech modalities, ethical frameworks for dataset curation have been adapted from text and image domains without accounting for the unique properties of spoken language. While existing work on responsible dataset curation has addressed consent, privacy, and representation [1, 5, 6], these frameworks fundamentally assume data can be attributed to individual subjects who can provide meaningful consent. This assumption breaks down entirely for speech data, which is inherently interactional, temporally dynamic, and socially embedded in ways that resist traditional notions of individual data ownership.

Speech datasets capture not individual utterances, but traces of complex social interactions involving multiple participants, temporal dependencies, and contextual information that traditional consent mechanisms cannot address. This creates what I term the "conversational context problem": the fundamental incompatibility between current ethical frameworks and the relational nature of speech data.

2 The Relational Nature of Speech and Its Consent Implications

Unlike text or images, speech exists within conversational contexts involving multiple participants and complex interdependencies. When someone speaks, their utterance is shaped by and responsive to social context, including previous speakers, implied audiences, and ongoing conversational dynamics [7]. This creates relationships extending far beyond the primary speaker whose voice is recorded.

Consider speech data from medical consultations. While traditional consent focuses on the patient, their utterances are fundamentally responsive to the healthcare provider's questions and prompts. The patient's speech contains traces of the provider's communication style, medical terminology, and diagnostic approach. The dataset captures not just the patient's voice, but embedded information about the healthcare provider who has not consented to their communicative patterns being used for model training.

Author's Contact Information: Anna Seo Gyeong Choi, Cornell University, Ithaca, NY, USA, sc2359@cornell.edu.

Manuscript submitted to ACM

This relational quality also has temporal dimensions. Speech unfolds over time, with speakers often building on or modifying previous statements [4]. A speaker might begin with general information and gradually reveal sensitive details, or shift their stance in response to other participants. Traditional consent mechanisms involving a single decision point cannot account for these dynamic revelations and evolving conversational contexts.

3 Two Critical Consent Challenges

The relational nature of speech creates two specific challenges that current ethical frameworks cannot address [3]. First, the "conversational debt" problem arises because individual utterances are semantically dependent on contributions from other speakers. When someone responds to a question, their utterance only makes sense within the context of the preceding question. Current consent frameworks focus exclusively on the primary speaker while ignoring these "ghost participants" whose contributions are embedded within collected utterances. This is particularly evident in speech data from parole hearings, where individuals respond to board members' questions about rehabilitation and future plans. The collected responses contain information about not only the speaker, but also the types of questions asked, framing used by board members, and institutional context shaping responses. When this data trains models, it potentially encodes perspectives and biases of questioning authorities who never consented to their approaches being captured.

Second, the "temporal consent" problem emerges because speech unfolds dynamically, with speakers often revealing information gradually or changing their stance during utterances. Someone might begin answering with public information but realize they're moving into private territory, or evolve their thinking as they speak. Traditional consent assumes speakers can provide informed consent for their entire contribution before speaking, ignoring the unpredictable nature of speech production. This becomes particularly problematic with clinical speech data, where patients might begin discussing general symptoms but gradually reveal sensitive information about mental health or personal struggles as they become comfortable. Static consent obtained initially cannot account for these evolving revelations, yet current practices treat entire recordings as uniformly consented content.

4 Toward Conversational Ethics Frameworks

Addressing these challenges requires moving beyond individual-focused consent models toward "conversational ethics" frameworks accounting for the relational, temporal, and multi-dimensional nature of speech data. Such frameworks would need mechanisms for collective consent acknowledging multiple participants embedded within speech datasets, even when only one voice is prominent. This might involve obtaining consent from all participants whose contributions are embedded, or developing technical approaches that identify and remove traces of unconsented participants while preserving linguistic value. The temporal consent challenge suggests need for "streaming consent" mechanisms allowing speakers to modify consent decisions as conversations unfold. Rather than treating consent as binary and static, such approaches would provide ongoing agency over contributions, potentially allowing retroactive exclusion of specific segments based on what speakers actually revealed rather than anticipated revealing.

As the field moves toward larger foundation models trained on diverse speech corpora, developing ethical frameworks addressing speech's unique challenges is urgent. Without such frameworks, we risk perpetuating power imbalances, privacy violations, and consent failures embedded in current speech dataset curation approaches [2, 8], ultimately undermining speech technologies' potential to serve human flourishing equitably.

References

[1] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. 2023. Ethical considerations for responsible data curation. *Advances in Neural Information Processing Systems* 36 (2023), 55320–55360.

[2] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1536–1546.

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

[4] Herbert H Clark. 1996. *Using language*. Cambridge university press.

[5] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[6] 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*. 881–904.

[7] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language* 50, 4 (1974), 696–735.

[8] Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

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