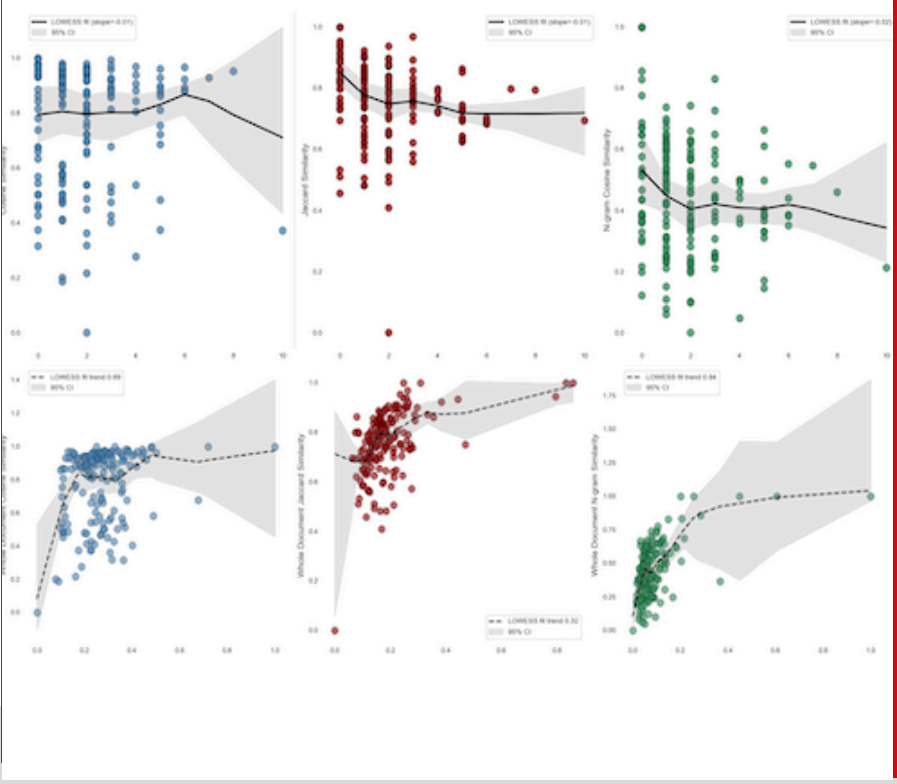
AI VS. HUMAN TRANSCRIPTION: EVALUATING ACCURACY AND MEANING IN POLICE BODY-WORN CAMERA FOOTAGE

Journal of Experimental Criminology (2026)
https://doi.org/10.1007/s11292-026-09774-0

This study was funded by the National Institute of Justice (grant # 15PNIJ-23-GG-05490-RESS)



Authors:

Savannah A. Reid
School of Criminology & Criminal Justice
Northeastern University

Stephen Abeyta
Marron Institute of Urban Management
New York University

Eric L. Piza
School of Criminology & Criminal Justice
Northeastern University

Nathan T. Connealy
Department of Criminology & Criminal Justice
University of Tampa

Victoria A. Sytsma
Department of Sociology
Queen's University

Key Takeaways:

- AI-generated transcripts of body-worn camera footage capture most of the same words as human-edited transcripts, but are much worse at identifying who is speaking
- AI transcripts identify only about two-thirds as many speakers as human-edited transcripts on average (3.3 vs. 4.8 per document), and never detect more than 7 speakers even when humans identify as many as 16
- AI tends to consolidate or drop less-prominent voices (bystanders, radio traffic, secondary civilians), collapsing back-and-forth exchanges into fewer, longer speaker turns
- Document-level content similarity (the overall words captured) stays high even when speaker attribution is poor; AI can "get the words right" while "getting people wrong"

AI vs. Human Transcription: Evaluating Accuracy and Meaning in Police Body-Worn Camera Footage

Research Summary:

Body-worn cameras (BWCs) are now standard in American policing, but the resulting volume of footage makes systematic review difficult without automation. AI-powered automatic speech recognition (ASR) promises to convert this footage into searchable, analyzable text at scale, and vendors such as Axon have already partnered with ASR providers (Rev.ai) to offer automated transcription to law enforcement clients. But the value of any downstream AI-driven analysis of BWC footage depends entirely on whether the underlying transcript accurately captures what was said and who said it.

This study uses a sample of 73 police incidents (176 BWC videos, roughly 23 hours of footage) from the Kansas City, Missouri Police Department. Each video's audio was transcribed twice through Rev.ai: once through fully automated AI transcription, and once through Rev's human-edited service, in which professional transcribers review and correct the AI draft.

The analysis proceeded in three steps: extracting each speaker and their quotations from both transcript types, comparing how closely individual speaker profiles aligned across the AI and human-edited pairs (using Jaccard and cosine text-similarity metrics, including bi-gram/tri-gram sequencing), and comparing whole-document similarity to assess overall lexical overlap.

The clearest finding is a systematic under-counting of speakers by the AI system. AI-generated transcripts identified only about two-thirds as many unique speakers as human-edited versions (an average of 3.3 versus 4.8 per document), never surpassing 7 speakers even in encounters where human transcribers identified as many as 16. AI transcripts tended to consolidate the conversation into fewer speakers with longer, merged quotations. Despite this, whole-document similarity scores between AI and human-edited transcripts were comparatively high.

Together, these patterns indicate that AI transcription accurately captures content but struggles with attribution, especially in complex, multi-party scenes with background noise, overlapping speech, or less-prominent voices such as bystanders and radio traffic. This distinction matters because different uses of BWC transcripts have very different tolerance for attribution error. Content-level applications can likely rely on AI transcription alone, since aggregate word overlap is high and individual misattributions are unlikely to skew large-sample results. But applications that hinge on knowing who spoke, how often, and in what sequence—such as internal investigations, use-of-force reviews, officer performance evaluations, and procedural-justice or de-escalation research—require accurate speaker attribution and are vulnerable to being compromised by AI-only transcripts.

These results extend a broader body of ASR research showing that speech-recognition accuracy varies substantially across audio environments and speaker characteristics. The study also speaks to rapidly developing criminal justice research using AI to score BWC footage for officer professionalism and communication quality, which has generally found that AI-derived behavioral codes can correlate with human ratings and even reproduce treatment effects from earlier trials—but with notable discrepancies on finer-grained distinctions and across software platforms. Where most of that literature evaluates AI performance at the level of behavioral codes or outcomes, this study shifts the focus to the transcription layer those codes ultimately depend on. In doing so, the study adds an important qualification to an otherwise optimistic literature on AI-assisted BWC review: alignment at the outcome level does not guarantee that the underlying text faithfully represents who spoke and how an interaction actually unfolded.