societies

MDPI

Concept Paper

Do We Need a Voice Methodology? Proposing a Voice-Centered Methodology: A Conceptual Framework in the Age of Surveillance Capitalism

Laura Caroleo

Faculty of Political, Economic and Social Sciences, University of Milan, 20122 Milan, Italy; laura.caroleo@unimi.it

Abstract

This paper explores the rise in voice-based social media as a pivotal transformation in digital communication, situated within the broader era of chatbots and voice AI. Platforms such as Clubhouse, X Spaces, Discord and similar ones foreground vocal interaction, reshaping norms of participation, identity construction, and platform governance. This shift from text-centered communication to hybrid digital orality presents new sociological and methodological challenges, calling for the development of voice-centered analytical approaches. In response, the paper introduces a multidimensional methodological framework for analyzing voice-based social media platforms in the context of surveillance capitalism and AI-driven conversational technologies. We propose a high-level reference architecture machine learning for social science pipeline that integrates digital methods techniques, automatic speech recognition (ASR) models, and natural language processing (NLP) models within a reflexive and ethically grounded framework. To illustrate its potential, we outline possible stages of a PoC (proof of concept) audio analysis machine learning pipeline, demonstrated through a conceptual use case involving the collection, ingestion, and analysis of X Spaces. While not a comprehensive empirical study, this pipeline proposal highlights technical and ethical challenges in voice analysis. By situating the voice as a central axis of online sociality and examining it in relation to AI-driven conversational technologies, within an era of post-orality, the study contributes to ongoing debates on surveillance capitalism, platform affordances, and the evolving dynamics of digital interaction. In this rapidly evolving landscape, we urgently need a robust vocal methodology to ensure that voice is not just processed but understood.

Keywords: voice methodology; fairness; political transcription; ethical AI; digital methods; speech communication; post orality; voice data; surveillance capitalism; conversational AI

check for updates

Academic Editors: Felice Addeo and Gabriella Punziano

Received: 30 June 2025
Revised: 24 August 2025
Accepted: 29 August 2025
Published: 30 August 2025

Citation: Caroleo, L. Do We Need a Voice Methodology? Proposing a Voice-Centered Methodology: A Conceptual Framework in the Age of Surveillance Capitalism. *Societies* **2025**, *15*, 241. https://doi.org/10.3390/soc15090241

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Digital communication has undergone a profound transformation, with voice-based interactions, encompassing social media platforms like Clubhouse, X Spaces (formerly Twitter), and Discord, as well as conversational AI chatbots—emerging as a central feature of the contemporary media landscape. In this paper, I propose a conceptual and methodological framework paper rather than a complete empirical study, presenting a multidimensional methodological framework to analyzing digital oral interactions within the context of surveillance capitalism, supported by a reference architecture of a machine

learning pipeline to illustrate feasibility [1]. The shift from text-centered communication to hybrid digital orality reshapes norms of participation, identity construction, and governance, posing unique sociological and methodological challenges. In this work, I introduce a reference machine learning pipeline integrating digital methods, automatic speech recognition (ASR) models, and natural language processing (NLP) models within an ethically reflexive framework. To illustrate its potential, I conducted a preliminary PoC (Proof of Concept) by collecting approximately 250,000 unique X Spaces, testing the data collection, processing and analysis steps. While not a comprehensive empirical study, this pipeline proposal provides initial insights into technical and ethical challenges of voice analysis. This article aims to contribute to debates on platform affordances, digital identity, and the ethical implications of voice data. By designing, developing, and evaluating a voice-centered methodology, I seek to address the complexities of digital oral interactions, prioritizing nuanced understanding over mere data processing.

While social media and AI continue to establish itself as some of the most powerful and pervasive tools in human history, contemporary critiques have adopted multiple theoretical perspectives to analyze its technological, social, philosophical, and economic implications. The very concept of the computational system, traditionally associated with computer science, algorithm design, software, hardware, and computing systems, has shifted in meaning. It is no longer merely an object of study but a way of thinking and living. This logical and algorithmic approach is now applied across all domains of society, including healthcare, banking, urban planning, and personnel time management. Our social interactions are mediated by digital platforms employing algorithms, and our online presence is shaped by systems and computations that place us within recommendation engines and preference analytics. Moreover, the way we access information and organize culture and knowledge is profoundly influenced by computational processes [2–5]. We live in a timeless time, where linear time has been replaced by fluid timetime, because digital allow near-simultaneous connection across time zones, creating a deferred simultaneity [6,7]. This perspective on digital temporality also asserts that the 24/7 regime of social media produces a temporal mode that erodes traditional distinctions between work and rest [8], thus introducing a peculiar temporality characterized by a tension between planned obsolescence and an eternal present [9].

This shift is further explored by Luciano Floridi, who examines how the «infosphere» is reconfiguring our understanding of reality and humanity itself, we are not merely living with information technologies; we are becoming «inforgs», informational organisms interconnected within these systems [10]. Floridi's concept of onlife, a neologism describing the hybrid reality in which the distinction between online and offline blurs, provides a fundamental interpretive framework. The onlife experience represents a new mode of existence where barriers between real and virtual, natural and artificial, dissolve into a continuum. This ontological transformation is further developed in the Onlife Manifesto, where Floridi and colleagues argue that information and communication technologies are no longer mere tools but environments in which we live [11]. José van Dijck offers an in-depth analysis of the socio-technical nature of digital platforms, focusing on how they are not simply technical tools, but complex systems that structure social, economic, and cultural interactions. Her perspective is based on the idea that digital platforms are socio-technical constructions that encode mechanisms of sociality in their architectures [12–14]. Platforms exert subtle yet pervasive forms of governance, shaping user behavior through algorithmic curation, content moderation, and interface design, thus reinforcing their role as active mediators rather than neutral intermediaries [15–17]. Digital platforms, through habits and updates, create a repetitiveness masked by novelty, and shape sociality through automated connectivity, perpetuating neoliberal power structures and surveillance [9].

From an intersectional perspective, feminist critics have pointed out that algorithms reproduce and amplify existing social inequalities, as these technological systems incorporate and perpetuate biases, often operating in ways that appear neutral. Search and recommendation algorithms are deeply embedded with racial, gender, and class biases, which systematically work to reinforce discrimination. Often, the datasets themselves lack diversity or are collected in ways that marginalize oppressed groups [18–21].

South Korean philosopher Byung-Chul Han, who lives in Germany and is known for his critical analyses of contemporary society with a focus on the impact of digital technologies, neoliberalism, and the culture of acceleration, explores how digital capitalism and the performance society are transforming human experience, social relations, and subjectivity. The constant pressure to perform pushes individuals to internalize the need to optimize themselves, resulting in fatigue, depression, and burnout. This ties in with the concept of habitual new media, in which repetitive digital practices (e.g., constant updates) exhaust users, as well as the analysis of the productivity imposed by platforms [9,14,22]. For Han the contemporary demand for transparency, driven by digital technologies and neoliberal ideology, has created a "smooth" society in which everything is exposed and commodified [23]. Furthermore, we have transitioned from biopolitics to psychopolitics, wherein the mind is controlled by exploiting desires and emotions to manipulate behavior, as a result, individuals lose their individuality and become part of a decentralized, fragmented, and chaotic digital swarm [24,25]. In this way, platforms generate new forms of psychological distress by promoting addiction and constant social comparison. They are psychosocial environments that shape individual subjectivity, not simple communication tools. This phenomenon is linked to "connected loneliness," characterized by constant digital connectivity alongside growing emotional distance in relationships, thus leads to digital nihilism and a sense of emptiness induced by platform design [26–29].

Evgeny Morozov warns against technological solutionism and digital utopianism. He highlights that this digital manipulation is intertwined with opaque data collection practices that fuel commercial profit, enable mass surveillance, and manipulate public opinion [30,31]. Although social media offer considerable transformative potential, enabling collective organization and social participation, this potential does not automatically guarantee positive outcomes. Their impact depends on how they are used and the objectives behind them, with significant risks of abuse or counterproductive effects [32]. From a critical Marxist perspective, social platforms represent a new form of digital capitalism that monetizes users' digital labor. Interactions, personal data, and user-generated content are exploited as economic resources, mainly through advertising and value extraction mechanisms, highlighting the dynamics of exploitation inherent in these platforms. Only a digital humanism that prioritizes human values, cooperation, and social justice over profit and surveillance can offer an alternative [33–35].

Digital transformations call for a critical reflection that goes beyond the opportunities offered by new technologies, examining their social, economic, and ethical implications. The digital revolution is not merely technological but also anthropological; it redefines the conditions of production, consumption, and power relations, while also acquiring a strong ideological dimension [36–38].

Within this context, surveillance capitalism emerges as a central concern in critiques of digital media: by monetizing personal data and generating structural vulnerabilities that deeply affect privacy, autonomy, and security, it unilaterally transforms human experience into free raw material to be translated into behavioral data. This culture of surveillance normalizes tracking as a daily practice embedded in social routines [39,40].

As "our life is lived in media, rather than with media" [41,42] in an environment where surveillance permeates social experience, power and visibility dynamics within online re-

lationships are redefined [43]. These dynamics, exacerbated by algorithmic opacity and dataism, amplify social inequalities and threaten individual autonomy, posing an existential challenge to democracy, as evidenced by digital manipulation and the exploitation of digital labor on platforms [12,13,30,31,39,44].

2. Surveillance Capitalism: The New Frontier of Voice Analysis

Surveillance capitalism has evolved from a system primarily based on textual data analysis to one increasingly incorporating natural language processing and voice analysis. Behavioral surplus now extends well beyond textual traces, capturing the tonal and emotional richness of the human voice [39]. This shift represents a qualitative leap in the ability to extract and analyze behavioral data. While standard machine learning techniques such as topic modeling, text classification, sentiment analysis, named entity recognition, and textual analysis focus on linguistic patterns, word frequencies, and semantic structures, voice monitoring enables the capture of emotional nuances, moods, and situational contexts with unprecedented granularity. The integration of textual and vocal data allows for deeper and more invasive user profiling, underscoring the urgent need for critical reflection on the ethical implications of these technologies [40]. The incorporation of voice analysis into digital surveillance systems has opened new frontiers for «affective computing» as the voice carries biometric and emotional markers that text cannot capture, offering surveillance systems an even more intimate window into human subjectivity [45–47].

The shift from text analysis to voice analysis represents not only a technological evolution but also a fundamental transformation in the regime of value extraction within surveillance capitalism. The appropriation of voice as data marks a new frontier in the colonization of human experience, transforming personal interactions into economic resources [48].

With the implementation of voice analysis on social media platforms, originally centered on text and images such as voice-based spaces on X, and the rise in voice-only platforms such as Clubhouse, the analysis of voice communication presents methodological, technical, and ethical challenges that clearly distinguish it from text analysis. While digital text analysis benefits from decades of methodological refinement, the systematic study of online voice communication remains in its early stages. These limitations are particularly evident in multilingual contexts, where speech recognition systems exhibit significant performance disparities, with word error rate (WER) increasing substantially for minority languages and non-standard dialects. Word error rate is a measure of accuracy in speech recognition systems, indicating the percentage of words transcribed incorrectly [49–51].

Analytical reliability is significantly influenced by factors such as overall audio quality, variable acoustic conditions in recording environments, background noise, people's cross-talk, and the quality of input devices, particularly on platforms that support real-time group communication.

These technical challenges are closely intertwined with ethical concerns regarding privacy and consent: voice contains unique biometric markers that can enable individual identification without explicit consent, requiring ethical and legal frameworks that extend beyond those developed for textual data.

Semantic speech analysis presents additional complexities, including contextual ambiguity, the inherent temporality of speech, and the multimodality of vocal communication, all of which pose significant obstacles to automation. These difficulties are mirrored in content moderation systems, which, being primarily designed for textual content [16], struggle to detect problematic speech, particularly on platforms hosting live voice conversations.

Accessibility issues also emerge as a central concern. The emphasis on voice communication can introduce new barriers for users with hearing impairments or speech disorders, raising important questions about inclusivity and fairness. At the same time, research

shows that voice-based systems may offer benefits for blind users and older adults with visual impairments [52]. Finally, the volatility of voice data underscores how the ephemeral nature of vocal communication complicates historical research and the analysis of long-term trends.

3. From Textual Hegemony to Digital Orality: The Transformation of Voice-Based Platforms and Their Social Implications

The social media ecosystem has undergone a significant evolution, shifting from the dominance of text-based platforms, such as early versions of Twitter and Facebook, to image- and audiovisual-centered platforms like Instagram and TikTok, and more recently, to voice-centered platforms such as Clubhouse and X Spaces[1]. The advent of digital technologies has introduced new communication practices and created unprecedented spaces for social interaction. Digital platforms now offer a variety of hybrid expressive forms in which voice and orality, once marginalized, regain a central role, redefining relationships among individuals and between individuals and institutions. This transition reflects a fundamental change in digital socialization, as voice-based platforms introduce immediacy and authenticity, profoundly altering the dynamics of mediated social interaction. The differences between text-based and voice-based social media go beyond technical aspects, deeply influencing social practices and the construction of digital identity.

Whereas text-based social media encourage reflexivity and curated self-presentation, audio-based platforms prioritize spontaneity and performative authenticity [52–55]. The concept of the "looking-glass self," introduced by sociologist Charles Horton Cooley in 1902, describes the process by which individuals form their identities based on how they believe others perceive and respond to them [56]. This concept has been applied to social media to explain how people construct their online identities through the feedback they receive from others. Social media mirror the mechanisms of the looking-glass self, offering various "mirrors" through which individuals present themselves, perceive others' judgments, such as likes and follows, and further develop their sense of self [55–57].

On voice platforms, the absence of physical self-presentation may offer users greater freedom, liberating them from the constraints of corporeality. In this context, individuals may feel more comfortable expressing themselves, focusing more on the content of their communication than on visual appearance or self-styling. For example, an individual may excel in a purely vocal context, such as a radio broadcast or other oral interaction environments, due to their strong communicative skills, while being disadvantaged in face-to-face or video-call interactions because of a physical appearance perceived as not conforming to dominant aesthetic standards or due to a physical impairment. Brewer et al. (2024) show how voice-based communities can provide alternative spaces of interaction for blind and low-vision older adults, highlighting diverse participation roles and the potential of voice-only platforms to mitigate certain accessibility barriers [52]. In such cases, the absence of the visual channel reduces potential biases, allowing the content and quality of the voice to come to the fore. However, for transgender and gender non-conforming individuals, voice itself can become a site of vulnerability, as vocal pitch, timbre, or speech patterns may be perceived as incongruent with their gender identity, potentially exposing them to misgendering, stereotyping, or discrimination in audio-only environments, a dynamic often linked to voice dysphoria and the gendered perception of vocal characteristics [58–60]. This distinction is also evident in content moderation practices and misinformation dynamics, as voice content raises fundamentally different technical and ethical challenges compared to textual content, requiring new approaches to platform governance. Moreover, the study of oral interaction environments must also take into account the role of algorithmic curation in shaping communicative dynamics. Platforms with distinct primary

purposes such as Discord, originally designed for gaming and teamwork, and Clubhouse or X Spaces, as well as TikTok's recent video-off livestreams (which tend to emulate the voice-only spaces of Clubhouse and X Spaces) oriented toward social exchange and deliberation, deploy recommendation and moderation algorithms that can create clustered narrative environments. These algorithmic architectures influence which voices are amplified or silenced, how conversations are surfaced to users, and the degree of ideological or thematic homogeneity within communities. In this sense, algorithms do not merely mediate access to content; they actively participate in defining political discourse and in shaping the interpretive context around speech and voice. On a broader horizon, the growing integration of AI-driven conversational agents and chatbots into these environments may alter the pragmatics and stylistic registers of online speech, introducing hybrid forms of interaction in which human and machine voices co-construct meaning. Such shifts call for methodological frameworks capable of tracing not only the content of vocal interactions, but also the algorithmic and AI-driven infrastructures that structure their visibility and circulation [18,61–63]. The coexistence of these communicative modalities, balancing the permanence of text with the ephemeral nature of voice, creates new forms of social and cultural interaction and memory. The ephemeral nature of oral communication fosters more fluid and dynamic social interactions, characterized by the transience of messages and an emphasis on the present moment.

These dynamics resonate with the contemporary digital media landscape, where communicative practices oscillate between the persistence and stability of text and the spontaneity and transience of voice. Voice, in particular, encourages the co-creation of shared and temporary meanings, contributing to the development of a fluid and dynamic social memory. Finally, affordance theory offers further insight into the role of voice in digital contexts: ephemerality and real-time synchronization shape interaction and social memory, creating spaces for the momentary co-construction of meaning and relationships [64,65].

The evolution of social media from predominantly textual platforms to environments that integrate and prioritize oral communication marks a significant transformation in the landscape of digital communication. Oral interaction environments allow the creation of a virtual presence that users often compare to the intimacy of a living room, fostering "familiar" micro-communities that promote closeness. In these spaces, users engage in intellectual discussions, share knowledge, and demonstrate high levels of interactivity [53,54,66].

However, like text-based platforms, voice-based platforms such as Clubhouse and X reproduce ideological polarization dynamics similar to those already observed in textual and visual environments [53,66].

Moreover, these platforms integrate with other digital tools, such as links and chats, recalling convergence culture theory, where old and new media interact in increasingly complex and fluid ways. There is no longer a single "black box" controlling the flow of information and communication; rather, we live in an era of constant interaction and exchange across multiple platforms [67,68]. Voice-based social platforms are characterized by their simplicity, core features include the use of virtual spaces, oral communication, real-time interaction, and the ability for listeners to become speakers, enhancing participation. Anyone can open a space and schedule events in advance and the user who opens the space automatically becomes the host or moderator and can appoint co-moderators, these moderators can decide who may speak, mute microphones, move users back to the audience, or remove them entirely. Multiple participants may be "on stage" as speakers (on X, the limit is 13 people, including the host), and the speaker list can be changed at any time. Additionally, participants can take on different roles: host, co-host, speakers (with microphones on or off), and listeners. Spaces can be recorded for later listening, and

links or parallel chats can be activated during events. Oral communication demonstrates its unique impact on social dynamics, particularly in fostering empathy and the perception of others' humanity. Kraus has shown that voice-only communication enhances empathic accuracy compared to multimodal forms, as it increases attention to paralinguistic cues such as tone, cadence, and intensity, which convey emotion. This focus on vocal detail allows participants to more accurately interpret others' emotional states, an essential aspect of conflict resolution and social connection [69]. Spoken language also humanizes the interlocutor more effectively than written text, making mental capacities such as rationality and emotionality more perceivable. In disagreement contexts, hearing the voice reduces dehumanization by emphasizing cognitive and emotional traits associated with human nature [69]. Finally, vocal production and perception, rooted in psychobiological processes and shaped by sociocultural factors, play a crucial role in the evolution of human communication, enabling the transmission of emotion even without visual cues [70]. These offline studies confirm the unique advantages of voice as a privileged communication channel for building understanding and trust in social interaction. They are echoed by media richness theory, which suggests that communication platforms offering more signals are more effective in maintaining relationships and managing complex tasks within organizations [71], with synchronous communication playing a key amplifying role [72,73].

The integration of voice into online communities represents one of the most significant transformations in digital communication. Video games, in particular, paved the way for this shift through the introduction of Voice over Internet Protocol (VoIP) technology, enabling the transmission of vocal signals online [74,75]. Massively multiplayer online games (MMOs) such as World of Warcraft, Final Fantasy, and EVE Online have provided fertile ground for the adoption of voice communication, facilitating new forms of collaboration and socialization [76,77]. These environments foster shared cultures in which voice communication becomes fundamental to building individual and collective identities and organizing group actions, also functioning as autonomous social spaces where players meet, converse, and coordinate beyond the gaming experience [78–87]. Moreover, VoIP use in gaming has influenced other fields such as education and training [88,89], showing how video games can serve not only as entertainment but also as platforms for innovation and personal growth [90].

Recently, social audio platforms have further expanded the possibilities for online voice interaction. Services like Discord, initially developed for gamers, have become gathering spaces for diverse communities, demonstrating how voice communication can foster meaningful social bonds in digital environments, supporting the emergence of hierarchical structures and defined social roles [76,90–93].

A recent study has explored the challenges faced by moderators on platforms like Discord, where the ephemeral nature of voice communications complicates the collection of evidence for problematic behavior, such as hate speech or voice raids. The lack of permanent recordings forces moderators to rely on personal impressions and indirect testimony, raising issues of transparency and fairness in decision-making. These dynamics require a reconsideration of moderation tools for voice platforms to ensure inclusive and safe environments [94].

This evolution can also be understood through the lens of domestication theory, which examines how technologies are adopted and integrated into everyday life [95]. According to this framework, digital platforms are not only technological innovations but also the result of cultural, social, and individual negotiations that shape their use and evolution. The domestication of voice platforms reflects their capacity to adapt to users' needs and practices, transforming from niche tools into central spaces in online communication dynamics.

On the one hand, some platforms were created exclusively as voice communication spaces. Clubhouse, launched in 2020, is a prominent example: initially conceived as a purely voice-based social network without textual content, it introduced a real-time interaction model prioritizing verbal dialog. However, as its user base grew and demand for greater versatility increased, Clubhouse later integrated features such as text chat and direct messaging, expanding its range of possibilities. Similarly, Yalla-a popular voice social networking platform in the Middle East and North Africa-offers voice chat rooms for social interactions and games. Platforms originally designed for textual or microblogging purposes have gradually incorporated oral dimensions to diversify interaction modes. A notable example is Twitter (now X), which introduced Spaces, real-time audio conversation rooms that extend its core text-based functions. This integration reflects a hybrid approach in which voice communication complements the platform's original textual nature, enhancing user engagement and enabling new forms of community-building. Telegram has introduced voice chats in groups for real-time conversation, while Discord-initially developed for gamers-has become a reference point for communities using both voice and text channels to interact.

In China, the vocal dimension has also emerged on platforms like Little Red Book (Xiaohongshu, in Chinese 小红书), known for textual and visual lifestyle content. Here, the introduction of voice features enables users to provide oral reviews or participate in voice discussions, enriching the user experience. Other platforms such as WhatsApp, Telegram, LinkedIn, and TikTok have adopted various strategies to integrate voice. WhatsApp, already widespread for asynchronous communication thanks to its voice messaging and audio call functions, has introduced real-time voice rooms for groups of over 33 people. LinkedIn, focused on professional interactions, has added voice messages to facilitate more personal and effective communication, along with audio events. TikTok-originally a short-form video platform-has experimented with live streaming functionalities that allow for voice interactions between creators and audiences, further expanding engagement opportunities.

Native voice platforms like TeamSpeak (used primarily in gaming) or Zello (which turns devices into push-to-talk walkie-talkies) offer targeted approaches to real-time voice communication. Emerging platforms such as Loud and the broader concept of the metaverse-including virtual spaces like VRChat, highlight how voice is becoming central to digital interaction.

Moreover, voice affordances have gained significance in dating platforms. Apps like Tinder, Bumble, and Hinge-originally designed to facilitate connections through profiles and images-have introduced voice messaging features to make the process of getting to know someone more authentic. These features allow users to convey emotions and intonation, surpassing the limitations of written communication and improving the overall interaction experience.

This dual trajectory between native voice platforms and the gradual implementation of voice affordances into pre-existing contexts, underscores the growing importance of voice as a tool for social connection and interaction. It reflects not only a technological evolution but also a transformation in how people relate to each other and build online communities.

4. The Rise in Voice-Based Social Media: From Sociological Implications to Methodological Challenges

This research explores the use of voice-based platforms and their role in transforming social and communicative dynamics, with a particular focus on a methodology that combines digital methods and a qualitative approach. Specifically, it addresses the fol-

lowing key research question: how can voice-centered methodologies be developed and applied to study the specificities, cultural impacts, and social dynamics of oral platforms? What insights can a voice-first methodological approach offer compared to traditional text-based analysis?

The proposed methodology aims to combine digital data obtained through digital methods [96–98], via public APIs provided by the platforms or via scraping implemented through tool such as 4CAT (a platform developed by the Open Intelligence Lab at the University of Amsterdam) or by custom developed module inside the pipeline with Python libraries such as Beautiful Soup, Playwright and others, in order to analyze networks and main topics, and to understand user movements through vast volumes of nested data. These tools not only map existing dynamics but also identify emerging trends and potential future directions in the voice-based platform landscape.

However, while text-based platforms allow for the extensive collection of structured data, voice-based social media introduce complexity in the analysis and interpretation of content, thereby challenging traditional surveillance practices. Platforms such as Clubhouse, X Spaces, and Discord exemplify this evolution. These voice-based digital environments require innovative research methodologies to address the complexities of audio analysis, including speaker diarization, accurate transcription, and real-time semantic understanding of content [99,100].

Such approaches leverage technological advances in deep learning models like wav2vec 2.0 [101], Whisper [102], WhisperX [103] and SeamlessM4T [104] and other models are employed for ASR and translation tasks, and are combined with models for voice emotion classification and representation, as well as with standard NLP techniques for the semantic analysis of complex audio interactions. These technological innovations enable real-time data processing, a key element for understanding the dynamics of voice interactions and their sociocultural implications.

According to Bourdieu (1991), the concept of linguistic capital becomes crucial in the context of voice-based social media, as voice conveys social and cultural indicators, such as accent, intonation, and manner of speech, that often remain invisible in textual communication [105]. These indicators contribute to the creation of new forms of social stratification and exclusion, with direct implications for power dynamics in digital spaces [105].

This phenomenon aligns with Giddens's (1984) structuration theory, which argues that technology not only enables social action but also constrains it [106]. In voice-based social media, for example, the affordances of the platform determine users' modes of interaction, influencing who can participate and under what conditions [107]. This underscores the idea that voice platforms are not neutral tools, but active agents in shaping social dynamics.

The analysis of voice communication requires a rethinking of traditional research methodologies. While natural language processing (NLP) with the advent of LLMs (Large Language Models), for textual content has reached advanced levels, the handling of audio streams presents significant challenges. These include contextual understanding, prosodic feature analysis, handling overlapping speech and the mapping of conversational flows [108,109]. These tools represent a fundamental evolution but remain limited in their capacity to grasp the complexity of vocal interactions.

The collection and analysis of voice data also raise important ethical concerns, particularly regarding surveillance and privacy. Drawing on Foucault's theories[2] (1977), the recording of voice conversations represents a new form of social control that demands careful regulation to prevent abuse [110]. The ephemeral nature of voice interactions, often unrecorded or unarchivable, further complicates their analysis, requiring methodological approaches that balance data collection with the protection of user privacy.

Moreover, technical difficulties related to speaker diarization and content attribution remain among the main obstacles for researchers. Accurately attributing individual contributions within voice conversations is essential to ensure the validity and reliability of analyses. Without scalable and precise tools to address these challenges, valuable insights into emerging social phenomena risk being lost [111–114].

The growing spread of voice-based social media demands a rethinking of digital research methodologies. A methodological pipeline for voice content analysis should include tools for automated transcription, advanced speech recognition algorithms, and technologies for semantic analysis of audio data. Furthermore, it is essential to develop ethical frameworks that ensure the responsible use of collected information, in line with data protection regulations (e.g., the GDPR) and critical surveillance theories. Only through interdisciplinary integration can we fully understand the implications of this phenomenon and develop adequate methodological responses to the challenges of the digital age.

5. A Multidimensional Pipeline for Vocal Content Analysis in Social Media: Methodology and Applications

This methodological proposal is structured along a multidimensional framework that combines the analysis of textual and network components within digital platforms with an innovative general approach to managing voice content in humanities and social science. While textual metadata analysis is a well-established aspect of research, the central element of this methodology lies in the development of a pipeline designed to address the specific challenges posed by processing and analyzing voice data. This pipeline, is organized into distinct nodes of a DAG (Direct Acyclic Graph), comprises: (1) continuous data collection (when communication occurs in real time) or standard data collection via API or batch scraping, (2) audio waveform processing using advanced speech recognition technologies, and (3) analysis through natural language processing (NLP) techniques.

Step 1: Continuous Data Collection

Data collection and ingestion represents the first critical step in ensuring accurate monitoring of voice interactions on social media. This phase can leverage the official APIs of social media platforms, such as X, Discord, Telegram and others, or perform scraping. Using the APIs simplifies the application of keyword-based filters to target specific communities or groups under study, and, when combined with Boolean operators, these filters further refine the search. In contrast, the scraping process requires greater effort to design and implement software modules capable of parsing undocumented APIs, and it is more complex due to potential interface changes and the need to comply with rate limiting. Both approaches can be scheduled at regular intervals of five minutes or less, depending on the performance of the system hosting the pipelines. This process not only captures the real-time audio stream and stores it locally for further analysis, but also continuously updates and stores associated metadata in JSON format, ensuring a comprehensive view of user activity throughout the duration of the voice interaction.

Key data collection sub-steps includes:

- User Profiling: Audio-based social media can reveal relevant information about users. Collecting and tagging user and demographic metadata helps contextualize interactions by enabling better correlation and analysis of this data.

Fields metadata such as creator, host usernames, and participants counter enable the reconstruction of interaction networks, mapping relationships between creators, co-hosts, and audiences over time. Attributes like lang, title, and creator bio support topic modeling and discourse analysis, allowing for the identification of dominant themes, linguistic patterns, and cross-linguistic trends. Temporal markers (created_at, scheduled start, updated

at) facilitate the study of event dynamics, including scheduling strategies and engagement lifecycles. Additional profile-level metrics (followers, following and listed counter) provide measures of creator influence and reach, which can be correlated with participation rates and conversational topics. By integrating these data points, it becomes possible to examine not only individual event characteristics, but also broader patterns in network structures, thematic convergence, and the diffusion of ideas across the platform.

- Audio streaming ending: When a conversation ends, the pipeline detects it and saves the audio file in persistent storage with the related metadata, preserving the original content and context for subsequent analysis phases. It is important to note that discrepancies may occur between the archived audio and the corresponding metadata, as a speaker may not be detected during data collection but still be present in the original audio file.

Step 2: Audio Processing with Automatic Speech Recognition (ASR) models

Once audio data is acquired and saved as an audio artifact, the pipeline proceeds to convert the audio file into a plain text format using automatic speech recognition (ASR) models. Due to the fast advancement in this field, I advocate that the pipeline must be as agnostic as possible from the ASR model. This approach guarantees a modular design with the possibility to replace the ASR model with an improved one without changes in the pipeline. With the transcribed plain text, with time marks, diarization and speaker recognition (speaker labeling) the pipeline can proceed further to vectorize text to produce text embeddings, this part of the DAG can be split into this core parts:

- Audio Transcoding: Converting audio formats into a standard one to ensure compatibility with the downstream module of the pipeline.
- Audio transcription: produce text transcription with the corresponding timestamp at word level (if supported by the model).
- Speaker Diarization: The process of partitioning an audio stream containing human speech into homogeneous segments based on speaker identity, thereby distinguishing between different participants in the conversation.
- Speaker Recognition and Voice Activity Detection (VAD): Uniquely identify speakers and detect the presence or absence of human speech, enabling granular analyses such as tracking individual sentiment, engagement, and participation patterns. The contextual integrity of voice data is preserved during processing, facilitating an accurate understanding of conversational dynamics.

Step 3: LLMs and Natural Language Processing (NLP) analysis

Textual data in plain text or in a word embedding format can be used with NLP techniques and models or as knowledge base for a RAG (Retrieved Augmented Generation) system to extract meaningful information from conversations. This task can be customized based on the scope of the project with additional models to perform:

- Entity Recognition: Identifying individuals, places, and key events mentioned in the discussion [108].
- Topic Modeling: Categorizing the main themes emerging from the conversation using topic modeling techniques, such as FastTopic [115].
- Sentiment and Content Analysis: Measuring emotional tone and identifying signals of polarization or conflict in discourse, crucial for understanding group dynamics and collective opinions [109].

Pipeline Applications: Longitudinal Studies and Cross-Platform Analysis.

This pipeline (Figure 1) not only addresses immediate challenges related to voice content but also provides tools for conducting longitudinal and cross-platform studies. For ex-

ample, tracking user interactions over time allows researchers to observe the evolution of group dynamics and community formation processes. Additionally, cross-platform analysis enables comparisons of interaction patterns across different digital spaces, identifying communication norms unique to each platform and studying how social networks develop across diverse digital environments.

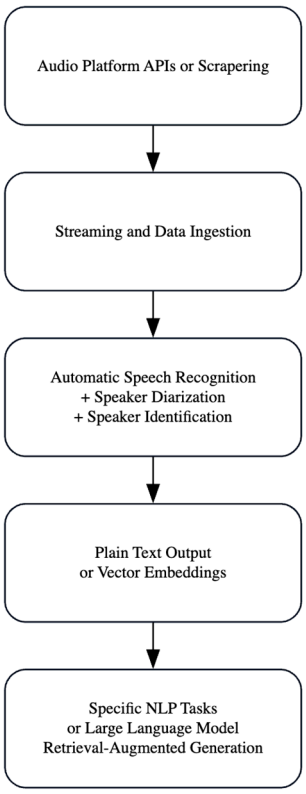


Figure 1. High-level architecture of the proposed three-phase pipeline for voice-based social media analysis, integrating real-time data ingestion, speech-to-text processing, and advanced NLP/LLM techniques to extract actionable insights from conversational audio.

The integration of such pipelines into humanities and social science research strengthens the analysis and study of voice interaction online, providing a robust methodology to address the technical and methodological complexities of voice data. By leveraging advanced deep learning techniques such as ASR, NLP, and LLMs, combined with real-time data collection, this approach opens new avenues for examining social interactions in digital contexts, thereby enhancing our understanding of contemporary communication phenomena.

5.1. Use Case on X Spaces

To demonstrate the feasibility of the proposed pipeline for analyzing digital oral interactions, including voice-based social media platforms, I develop a PoC (Proof of Concept) pipeline using Python 3.11.3 and DVC 2.54.0 (Data Vision Control) for acquiring approximately 249,584 unique X Spaces between 3 April and 13 June 2023, over a 70-day timeframe. Data collection occurred before X transitioned its APIs to a paid model, utilizing public APIs queried every five minutes with 100 keyword-based filters and Boolean operators. The pipeline collected and stored metadata for about 25.43 GB, capturing 3,168,371 observations.

The dataset (Table 1) contained complex metadata, including essential fields such as the unique identifier of each space (id), participant count (participant_count), event status (state: active, completed, or canceled), primary language (lang), event title (title),

and creator details (e.g., creator_id, creator_username, creator_verified, followers_count, tweet_count). Additional fields included co-host information (host_usernames, host_usernames_count) and timestamps (created_at, scheduled_start, updated_at). These metadata enabled quantitative analyses of social dynamics, creator influence, and engagement trends.

Table 1. Summary of Step 1 Data Collection and Dataset Characteristics (including Future Planned Metrics).

Parameter	Value/Description
Observation period	3 April–13 June 2023 (70 days)
Platform	X Spaces
Data collection method	Public APIs, queried every 5 min with keyword-based filters and Boolean operators
Number of unique Spaces collected	249,584
Total observations (entries)	3,168,371
File size	25.43 GB JSON
First record timestamp	3 April 2023, 14:08:16
Last record timestamp	13 June 2023, 11:16:34
Metadata fields	Space ID, participant count, event status (active/completed/canceled), primary language, title, creator ID/username/verification, followers count, tweet count, co-host usernames/count, timestamps (created_at, scheduled_start, updated_at)
Potential analyses	Social dynamics, creator influence, engagement trends
Current limitations	No audio completeness rate, WER, or diarization accuracy measured in the PoC
Future planned metrics	• Audio completeness rate (%) • Word Error Rate (WER) by language/dialect • Diarization accuracy (%) • Bias and fairness audits across linguistic varieties • Cross-platform comparability metrics

Following the metadata-based collection, additional tests were conducted to analyze the actual audio content of X Spaces events. Unlike the JSON metadata files described above, these audio recordings were retrieved from the Periscope infrastructure. The development immediately revealed significant technical challenges. First, an X Space can host up to 13 speakers simultaneously on stage, creating a highly complex audio environment for diarization. Even with advanced tools such as WhisperX [103], accurate speaker recognition proved problematic: multiple voices overlapped, rapid turn-taking occurred, and background noise often interfered with audio segmentation.

Moreover, during initial transcription trials, it became evident that automatic speech recognition (ASR) frequently produced inaccurate outputs. Some utterances were misattributed to the wrong speakers, and certain sentences were incompletely or incorrectly transcribed. Intonation and prosodic cues, central to interpreting the pragmatic meaning of spoken interactions, were often distorted or entirely lost in the transcription process.

A different scenario emerged when the stage was limited to only two active speakers. In these cases, both diarization and ASR accuracy improved substantially: speaker attribu-

tion was more reliable, error rates were lower, and the preservation of prosodic features was markedly better. This contrast underscores the importance of accounting for the number of simultaneous speakers in the design and evaluation of audio analysis pipelines for voice-based social media.

5.2. Ethics and Compliance

The collection and analysis of voice data in social media environments present significant ethical and legal challenges. Voice, in itself, does not constitute biometric personal data, as it does not meet the definition provided by the GDPR.

Article 4(1)(14) GDPR defines biometric data as: "personal data resulting from specific technical processing relating to the physical, physiological, or behavioral characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data."

However, as with images or video recordings, voice may be considered biometric data if certain specific criteria are met. These relate to the nature of the data, namely, that it refers to a physical, physiological, or behavioral characteristic of a person, and to the means and purposes of processing, which must involve technical processing for the purpose of unique identification. This requires the existence of a lawful basis under Article 6 GDPR and also one of the conditions set out in Article 9 (2) GDPR. This condition is necessary even in cases where voice falls under the scope of "special categories of personal data."

A voice recording may constitute sensitive data not only because of its content but also indirectly, as a result of its combination with other information (some of the data listed in Article 9 (1) GDPR). In such cases, voice must be processed as "special categories of personal data."

Each step of the proposed pipeline must therefore be mapped against applicable legal frameworks, platform terms of service, and institutional ethics review procedures. In line with Mantelero (2022) [116], lawful processing requires identifying a clear legal basis (Article 6 GDPR) and an explicit exemption under Art. 9 for biometric processing, such as explicit consent or research in the public interest with appropriate safeguards. The principle of data minimization (Article 5(1)(c) GDPR) mandates collecting only what is strictly necessary, while the principle of storage limitation (Article 5(1)(e)) requires defined retention periods and secure deletion procedures.

Platform-specific Terms of Service (e.g., X Spaces, Clubhouse, Discord) further restrict automated collection, impose respect for user privacy, and may prohibit scraping without prior authorization. A clear distinction must therefore be made between public events, where content is intentionally accessible to non-participants, and private or restricted-access rooms, where informed consent from all speakers is necessary. Incidental participants (those not directly addressed by the researcher) should be anonymized through voice de-identification techniques or lightweight real-time speaker diarization methods [111] or embedding extraction without raw audio storage [112].

To operationalize these safeguards, the pipeline should integrate an Ethics Decision Tree specifying which consent strategies apply and which anonymization protocols should be adopted. All recordings should undergo a risk assessment, including potential harms related to re-identification and discrimination. Metadata such as usernames, geolocation, and timestamps should also be treated as potentially identifiable and, where possible, pseudonymized or aggregated.

Finally, ethical reflexivity must extend to transcription practices. As Bucholtz [117] reminds us, transcription is a political and interpretive act; researchers must ensure that rendering spoken interactions into text does not introduce bias or stigmatization, particularly for minority language varieties [50,118]. A combined machine–human audit process,

with documented discrepancies and participatory validation when possible, can enhance both accuracy and fairness in representation.

5.3. Results Achieved

The proof-of-concept (PoC) implementation on X Spaces provides insights that go beyond a technical demonstration, highlighting the sociological significance of voice-centered methodologies. First, the scale of the dataset collected, approximately 249,000 unique Spaces within a 70-day period, demonstrates the intensity and pervasiveness of oral interactions in contemporary digital environments. This confirms that voice is not a marginal phenomenon but a central modality of online sociality, requiring tailored methodological approaches.

Second, the PoC revealed that platform affordances directly affect methodological reliability. While diarization and automatic speech recognition (ASR) were more effective in interactions with only two active speakers, accuracy sharply declined when multiple participants spoke simultaneously, as is common on X Spaces where up to 13 users can intervene. This finding illustrates how technological design both enables and constrains interaction, aligning with structuration theory and emphasizing the importance of situating methodological tools within platform architectures [106].

Third, the PoC highlights structural inequalities embedded in technical infrastructures. Transcription accuracy varied significantly across languages and accents, with higher word error rates for minority linguistic varieties. This is not only a technical limitation but also a sociological result: it reveals how algorithmic infrastructures may reproduce and reinforce linguistic hierarchies, amplifying dominant voices while marginalizing others. Such disparities underscore the importance of fairness audits and equity-aware annotation in future methodological development. These challenges also extend beyond compliance. In Foucault's terms, the recording and transcription of voice interactions constitute new forms of disciplinary surveillance, while Zuboff's surveillance capitalism demonstrates how prosodic and emotional traces of voice can be commodified as behavioral surplus. Recognizing these dynamics clarifies why ethical reflexivity cannot be separated from methodological design.

Finally, the contrast between metadata-based analysis and audio-based analysis points to fundamental epistemic trade-offs. Metadata enables large-scale mapping of participation networks, event dynamics, and engagement patterns, but it cannot capture the nuances of tone, prosody, and affect that are central to vocal interaction. Conversely, audio analysis foregrounds the interpretive richness of spoken communication but exposes significant challenges of accuracy, scalability, and bias. This tension illustrates the need for hybrid methodological strategies that integrate computational scalability with qualitative depth.

Together, these findings show that the proposed pipeline not only functions technically but also generates valuable sociological insights into participation, power, and inequality on voice-based platforms. The results confirm the urgency of developing a voice-centered methodology that is both technically robust and theoretically informed. Together, these findings show that the proposed pipeline not only functions technically but also generates valuable sociological insights into participation, power, and inequality on voice-based platforms. The results confirm the urgency of developing a voice-centered methodology that is both technically robust and theoretically informed.

6. Discussion

The rise in voice-based social media platforms has opened new possibilities for the analysis of digital communication, particularly in terms of sociological and methodolog-

ical frameworks. Platforms such as Clubhouse, X Spaces, Discord and others illustrate how voice as a medium of communication transforms interaction norms, fosters unique community structures, and reveals the complexities of identity and power within digital spaces [53,54]. However, at the point in this methodological pipeline where we decide to implement the transition from audio to text, we must recall-as Mary Bucholtz (2000) emphasizes in The Politics of Transcription-that transcription is not a neutral act [117]. It is deeply interpretive and shaped by the theoretical and methodological choices of the transcriber, often carrying biases that can distort the representation of oral interactions. This critique is particularly relevant in the context of voice-based platforms, where the challenge of accurately capturing the richness of spoken communication is heightened [117].

Bucholtz's work highlights that transcription practices do not merely document what is said; they actively influence how discourse is interpreted. Decisions regarding the inclusion or exclusion of pauses, hesitations, and non-standard speech forms affect the perceived identity of speakers, often reflecting broader social biases. For instance, transcribing non-standard grammar or accents without context risks reinforcing stereotypes and marginalizing certain voices. These challenges are further exacerbated in the analysis of voice-based social media, where the ephemeral and dynamic nature of conversations presents significant technical and ethical obstacles. Handling overlapping speech, speaker diarization, and transcription accuracy remain critical hurdles that underscore the limitations of current technologies.

While Bucholtz [117] highlights the political and interpretive nature of transcription, translating this awareness into methodological practice requires structured decision-making tools. I propose a Transcription Decision Matrix to guide researchers in selecting appropriate transcription styles according to research objectives, ethical considerations, and participant profiles. This matrix differentiates between orthographic transcription, suitable for large-scale NLP analysis with minimal prosodic detail, and conversation-analytic transcription, which encodes pauses, overlaps, intonation, and prosodic features essential for fine-grained interaction analysis [119]. In multilingual and multi-accent contexts, an equity annotation layer should capture accent, variety, or non-standard speech features without stigmatizing the speaker, following the bias-aware annotation practices recommended by Blodgett et al. and Hovy & Spruit [50,51]. This layer ensures that linguistic diversity is documented as a valuable communicative resource rather than a deviation from an arbitrary standard. To enhance accuracy and fairness, though only for small datasets, I recommend a dual-track transcription process: (1) initial machine transcription using ASR, followed by (2) human auditing for error correction, annotation, and contextual enrichment. All discrepancies between ASR output and human revisions should be logged, enabling reflexive evaluation of the transcription process and providing training data for improved models [113,118]. Where feasible, participatory validation-inviting speakers to review and comment on their transcribed speech-can further ensure accuracy, empower participants, and align with participatory research ethics [21]. Integrating these procedural safeguards into the pipeline operationalizes Bucholtz's theoretical critique, turning reflexive awareness into concrete practices that promote both analytical rigor and social equity.

In large-scale use cases on voice-based social networks, however, it is not realistic to adopt an exclusively orthographic approach; instead, reliance on automated models becomes necessary. Yet, this requirement belongs to a complex research domain: accurately identifying and marking pauses, syllable duration, and pitch variations would require a highly accurate and specialized ASR model. At present, Whisper does not natively support prosodic information extraction, offering it only through post-processing with linguistic tools such as Praat [102,120]. While commercial models such as Microsoft Azure Speech Service [121], Amazon Transcribe combined with Polly or Comprehend [122], and IBM

Watson Speech to Text [123] do provide prosodic markers as part of their output, they are not open-source, lie outside the scope of this paper, and their implementation would require additional technical development and dedicated resources, more suitable for a future standalone project. Similarly, while the concept of an equity annotation layer is theoretically compelling, it is not operationally achievable within the scope and timeline of the present work. Developing a system capable of registering accents and linguistic varieties without stigmatization would require extensive multilingual and multidialectal data experimentation, as well as collaboration between linguists, ethnographers, computer scientists, sociologists and fairness-in-ML experts amounting to an independent research objective.

Regarding procedural reflexivity through dual transcription (ASR plus human revision), discrepancy logging, and participatory validation, in the context of long-term, high-volume datasets (hundreds of hours of recordings), such an approach would not be scalable or sustainable. The magnitude of the challenge demands automated solutions at scale, while extensive manual validation would render processing the entire corpus impractical. Therefore, a mixed strategy is proposed, applying targeted manual checks to representative samples, balancing quality control with operational feasibility. While the systematic measurement of ASR performance across minority languages and varieties is undeniably important, it is not possible to implement a concrete measurement plan in this work due to the absence of a balanced multilingual dataset, requiring manually annotated speech data in multiple languages and dialects, and the time and resource constraints that make such an undertaking an independent research project in its own right.

Moreover, reliance on technologies such as Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) introduces additional complications. Although these tools can process large-scale data and identify patterns, they lack the human capacity for contextual and diagonal reading. Machines operate through linear information retrieval and content generation, rather than through genuine understanding. Comprehension, as Bucholtz's analysis suggests, involves reasoning and interpretation that go beyond the literal content of what is read or heard. Consequently, the lack of contextual awareness in machine-generated transcription risks producing superficial analyses that fail to capture the deeper socio-cultural dynamics of vocal interactions.

While technical pipelines integrating ASR, NLP, and LLMs offer unprecedented opportunities for large-scale analysis, not all researchers in the humanities and social sciences possess the technical expertise or resources required to implement them. In such cases, digital ethnography remains a valuable and accessible methodological approach, enabling in-depth, contextual understanding of voice-based interactions without relying on complex computational infrastructures. This approach can complement, rather than replace, technically intensive methods, ensuring that the study of vocal platforms remains open to diverse disciplinary perspectives and does not become the exclusive domain of computationally skilled researchers.

Ethical considerations also remain paramount in the analysis of voice data. Since voice is classified as biometric data under the GDPR, ensuring privacy and anonymization becomes essential. Managing incidental participants and guaranteeing informed consent are particularly challenging in environments where voice data is continuously exchanged. Once again, Bucholtz's critique is pertinent, highlighting the researcher's responsibility to represent participants' voices faithfully and ethically, without distortion or exploitation. The use of transcription in such contexts must be carefully designed to respect the integrity and privacy of those involved.

Despite these challenges, the proposed multidimensional pipeline offers a promising approach to analyzing voice-based platforms. By integrating quantitative methods, such as network analysis and large-scale interaction mapping, with qualitative approaches like

conversation analysis and ethnographic research, this pipeline balances breadth and depth. It supports longitudinal studies that track participant interactions and shifts in group dynamics over time, while also enabling cross-platform analyses to compare communication norms across different digital spaces. This hybrid methodology addresses the limitations of isolated quantitative and qualitative approaches, fostering a more comprehensive understanding of digital vocal interactions.

This hybrid methodology addresses the limitations of isolated quantitative and qualitative approaches, fostering a more comprehensive understanding of digital vocal interactions. The results achieved through the PoC confirm that voice-based social media platforms cannot be studied using the same methodological assumptions applied to text-centered environments. Voice introduces specific epistemological, sociological, and ethical challenges that reshape how digital communication should be analyzed.

At the methodological level, the comparison between metadata and audio analysis illustrates a central epistemic trade-off. Metadata offers scalability but reduces interaction to quantifiable traces; audio captures nuance and affect but introduces fragility and bias. Text-based methods remain valuable, but they obscure elements such as intonation, empathy, and vocal presence that are crucial to understanding digital oral interaction. A voice-centered methodology therefore extends digital methods by foregrounding dimensions of sociality that text cannot capture.

Finally, the PoC demonstrates that no single methodological approach is sufficient. While automated pipelines can scale to millions of interactions, qualitative and ethnographic perspectives remain indispensable to interpret the meanings of voice, particularly in multilingual and intersectional contexts. This hybrid strategy resists the temptation of solutionism, as warned by Morozov, by balancing computational analysis with reflexive and contextual approaches.

In this way, the discussion confirms that a voice-centered methodology is not merely a technical innovation but a theoretical and epistemological imperative. It bridges critical perspectives on surveillance, bias, and platform governance with practical tools for analyzing digital orality, showing how theory and method must co-evolve to address the challenges of contemporary communication.

Indeed, the theoretical perspectives outlined earlier provide not only a critical backdrop but also a direct rationale for the proposed methodological framework. The findings reinforce Floridi's notion of the onlife condition. Oral spaces, in fact, free users from the constraints that keep them constantly tied to the phone screen, allowing them to carry out everyday activities-such as running, shopping, cooking, walking, or moving from one place to another-without having to use their eyes or hands. In this sense, there is a return to the ear as the primary instrument of connection, within a hybrid condition that is distinctly onlife. Moreover, the sonic nature of these spaces enables users to be, symbolically, in many places at once: if a participant is cooking, one hears the clatter of dishes; if another is walking, the sounds of the city are captured; if someone joins from a café, the clinking of cups and background chatter become audible. In this way, the user's experience takes shape in a novel condition, within online environments that structure visibility and participation in real time. van Dijck's analysis of platforms as socio-technical infrastructures highlights the need to situate technical pipelines within broader environments of power, mediation, and governance. This means that our pipeline must not be seen as a neutral analytical device but as a tool embedded in, and responsive to, platform architectures that actively structure participation and visibility. Similarly, the inequalities revealed by ASR performance across languages and accents substantiate feminist critiques of algorithmic bias. The consistently higher error rates in transcribing minority linguistic varieties illustrate how voice technologies reproduce existing hierarchies of power and visibility. These

are not incidental flaws but symptoms of broader dynamics of digital inequality, aligning with critiques of algorithmic discrimination advanced by Noble, Benjamin, and D'Ignazio & Klein. Such insights provide a further rationale for embedding fairness audits into the pipeline. Voice is a deeply intersectional marker: accent, pitch, and rhythm can reinforce social inequalities. Methodologically, this requires monitoring word error rates across linguistic varieties, implementing equity-focused annotation layers, and explicitly documenting the social implications of model misrecognition. These critiques are directly reflected in our proof-of-concept. In particular, the significantly higher error rates observed for minority linguistic varieties demonstrate how algorithmic infrastructures amplify existing hierarchies of power and visibility. What feminist theorists describe as embedded bias becomes tangible in the unequal treatment of different voices within ASR systems, underscoring the urgency of fairness and equity layers as indispensable components of a voice-centered methodology. The technical difficulties encountered in transcription and diarization are not only methodological obstacles, but reflections of the socio-technical architectures that constrain interaction, in line with Giddens's structuration theory.

Similarly, Byung-Chul Han's critique draws attention to how online oral communication is shaped by pressures of performance and discipline, at times even more intensely than on textual and visual platforms. These dynamics can be further intensified by the simultaneity and real-time nature of oral conversations. On voice-based platforms, speakers are not merely exchanging information but are often compelled to present themselves as audible, engaging, and emotionally expressive, thereby conforming to platform expectations. In our methodological design, this entails recognizing that users contribute to the reproduction of regimes of control, determining which voices are intelligible and recordable, as well as regimes of self-optimization, as they adapt their vocal behavior to be more easily followed, appreciated, and valued within algorithmic systems.

Finally, Morozov's critique reminds us that our pipeline cannot be framed as a purely technical fix. Instead, it must be complemented by reflexive, qualitative, and ethnographic approaches that can capture the social meanings and lived experiences of vocal interactions. This explains why I propose a hybrid framework that integrates computational methods with qualitative ones, rather than privileging automation alone.

The results also confirm Zuboff's diagnosis of surveillance capitalism's expansion into the affective and biometric domain. By showing how tonal and prosodic cues are technically difficult to capture, the PoC reveals both the potential and the risks of affective computing. Platforms' capacity to appropriate emotional traces of voice further exemplifies how personal experience is commodified as behavioral surplus.

Limitations and Future Research

While the proposed voice-centered methodology demonstrates promising avenues for the analysis of oral interactions on digital platforms, several limitations must be acknowledged.

Technical constraints. Automatic Speech Recognition (ASR) and speaker diarization remain imperfect, particularly in multi-speaker environments where overlapping speech and background noise compromise accuracy. Current models also show significant disparities across languages, dialects, and accents, resulting in higher error rates for non-dominant varieties. These shortcomings risk reinforcing existing linguistic hierarchies rather than mitigating them.

Epistemological trade-offs. The pipeline highlights an unavoidable tension between scalability and depth. Metadata enables large-scale mapping of participation, while audio analysis captures affective nuance but introduces fragility and interpretive uncertainty. This duality illustrates that no single methodological approach is sufficient, and results

should always be interpreted with caution. These trade-offs resonate with Morozov's critique of solutionism: no purely technical pipeline can fully resolve the epistemic challenges posed by digital orality. The coexistence of scalable but reductive metadata analysis and nuanced but fragile audio analysis illustrates the limits of automation, confirming that hybrid approaches are not optional but necessary.

Ethical and legal challenges. Voice recordings may constitute biometric data under GDPR and other regulatory frameworks. Consent, anonymization, and proportionality remain critical concerns, particularly in environments where incidental participants are common. Moreover, transcription itself is an interpretive act, raising questions of fairness and representation that extend beyond compliance. These challenges also extend beyond compliance. In Foucault's terms, the recording and transcription of voice interactions constitute new forms of disciplinary surveillance, while Zuboff's surveillance capitalism demonstrates how prosodic and emotional traces of voice can be commodified as behavioral surplus. Recognizing these dynamics clarifies why ethical reflexivity cannot be separated from methodological design.

Resource limitations. The implementation of fairness and equity annotation, and participatory validation requires multilingual corpora, human expertise, and interdisciplinary collaboration that exceed the scope of the present study. These remain important goals for future development.

Looking ahead, future research should prioritize three directions. First, the systematic benchmarking of ASR and diarization models across multiple languages and accents to expose and reduce algorithmic bias. Second, the integration of computational pipelines with qualitative approaches such as digital ethnography, ensuring that large-scale data analysis remains grounded in lived social experience. Third, closer engagement with policy and governance debates, particularly around the regulation of biometric data, platform accountability, and the design of fairness standards for voice technologies.

By addressing these limitations, subsequent work can advance toward a more inclusive and reflexive voice-centered methodology that not only captures the richness of digital oral interaction but also challenges the inequalities and risks embedded in current infrastructures.

7. Conclusions

This study provides a conceptual and methodological foundation for the development of a new line of research centered on voice, which I might define as "oral sociology" or "vocal sociology". However, further work is needed to consolidate and refine this emerging branch of research and to advance what I define as a "voice methodology" within the proposed framework.

The empirical results of the proof-of-concept further reinforce this theoretical contribution. First, the sheer scale of oral interactions on X Spaces confirms that voice is not a marginal phenomenon but a central modality of online sociality. As observed, Spaces played a significant role in the 2024 U.S. electoral campaign, with a significant number of events hosted by candidates. For instance, Ron DeSantis launched his presidential race in a Space in May 2023, Vivek Ramaswamy even urinated during a live conversation with Elon Musk, and Robert F. Kennedy Jr. inaugurated the "Campaign Kitchen" format in November 2023 [124]. It is important to note that, in addition to the platforms already mentioned, dating applications have also progressively integrated a vocal component. Tinder, for example, enables the use of Voice Prompts and video chat, offering the possibility to exchange voice messages; Hinge allows users to enrich their profiles with voice clips; Bumble integrates in-app voice calls; while Azar, Badoo, and Happn include both voice and video functions. This development highlights, in continuity with the insights of early

scholars of orality such as McLuhan, Ong, and de Kerckhove, the growing centrality of the oral dimension [125–129]. There is also the entire sphere of vocal interactions with non-human agents, such as virtual assistants (Alexa, Siri, Google Assistant) or voice bots integrated into messaging applications and social platforms. These systems are not limited to providing functional responses but are progressively assuming a conversational and relational role, offering companionship, emotional support, and even forms of care. In this sense, the so-called companion AIs or vocal social bots (for example, Replika or Character.AI) open up unprecedented scenarios in which voice becomes the primary vehicle for constructing affective bonds and attributing agency to artificial entities [130,131]. What I define here as post-orality does not refer solely to communication between humans mediated by technological tools or to the digital transposition of human interaction but also encompasses forms of oral interaction between humans and non-humans, that is, between any entity, whether living or non-living, natural or artificial, endowed with the capacity to act. These are new forms of hybrid communication in which the boundaries between humans and artificial intelligence become porous. Speaking with a voice bot means not only using the voice as a technical interface but also projecting onto it social, cultural, and emotional expectations, a phenomenon that compels us to reflect on the ethical, epistemological, and political implications of such emerging vocal relations.

Second, the decline in ASR and diarization accuracy in multi-speaker settings reveals how platform affordances directly shape the methodological reliability of research, constraining what can be captured and understood. Third, the uneven performance across languages and accents illustrates how voice technologies risk reproducing linguistic hierarchies, privileging dominant groups while marginalizing minority varieties. Together, these findings substantiate the claim that post-orality cannot be studied using frameworks designed for text-centered environments.

Voice-based social media platforms represent a dynamic and evolving frontier for sociological research, offering valuable insights into digital identity, power dynamics, and community formation. Yet, their analysis requires not only innovative tools but also epistemological and ethical vigilance. As Bucholtz (2000) incisively argues in The Politics of Transcription [117], transcription is not a neutral act, it is politically charged, interpretative, and shaped by socio-theoretical positioning. Even the insertion of a pause or punctuation mark can alter meaning, reinforcing the need for reflexivity in every stage of the methodological pipeline.

This becomes even more urgent in the context of machine-driven processes. Current technologies, such as ASR and NLP, enable large-scale data processing but lack the human capacity for contextual or "diagonal" reading, which is crucial for interpreting ambiguity, irony, emotional tone, and implicit power dynamics. The limits of speaker diarization tools, especially when the number or identity of speakers is unknown, further compromise the reliability of analysis. Despite advances, there is no shared theoretical framework underpinning these technologies. Benchmarks are often arbitrary, datasets are inconsistent and claims of "fairness" are frequently unsupported by rigorous social inquiry.

Moreover, the ability to build and deploy these tools remains concentrated in the hands of a few dominant actors, big Tech firms with the financial and computational capital to shape the rules of engagement. As a result, the current ecosystem of voice analysis reproduces asymmetries in access to knowledge, infrastructure, and interpretive authority. The process of transcription, once a human act, is now algorithmically governed, yet still lacks accountability, transparency, and methodological coherence.

At the same time, future investigations should pursue hybrid methodological designs that combine computational scalability with qualitative and interpretive depth. Integrating large-scale pipelines with ethnographic inquiry, discourse analysis, and participatory

approaches would ensure that machine-driven analyses remain anchored in the lived social meanings of voice, rather than being reduced to decontextualized data traces. Equally important is the need for longitudinal and cross-platform studies. The ephemeral and real-time character of voice communication makes it a privileged site for observing processes of community formation, political discourse, and identity negotiation over time. Comparative analyses across different platforms, such as Clubhouse, X Spaces, Discord, Little Red Book, or dating applications, could shed light on how distinct affordances and governance structures shape the trajectories of post-orality in diverse sociocultural contexts.

Also, the growing recognition of voice as biometric and affective data demands closer engagement with legal and institutional frameworks. Future research should examine how existing regulations, such as the GDPR, might be adapted to address the specificities of vocal data, while also considering the development of fairness, transparency, and accountability standards in AI systems that process speech. Advancing this agenda will require sustained interdisciplinary collaboration, drawing together expertise from sociology, linguistics, computer science, law, and ethics to build methodological and regulatory approaches that are both technically robust and socially responsible.

As outlined so far, only a multidimensional pipeline, capable of combining qualitative depth with computational scalability in a hybrid framework that integrates digital methods with ethical rigor, can foster a genuine consolidation of the field. While it cannot eliminate its structural limitations, such an approach moves closer to a more responsible and inclusive way of studying voice in digital environments.

Ultimately, addressing the epistemic, technical, and economic asymmetries in this domain requires not only better tools but also deeper critical reflection. Voice is not just a medium of communication, it is a site of power, identity, and vulnerability. Developing a truly voice-centered methodology will depend on our capacity to situate technical processes within their broader social, political, and economic contexts. This is not merely a technical challenge, it is a conceptual and ethical imperative.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data collected in this study are available upon request from the corresponding author.

Conflicts of Interest: The author declares no conflicts of interest.

Notes

[1] Several digital platforms now incorporate synchronous oral communication as a core or complementary feature, including X Spaces, Clubhouse, Discord, Telegram, LinkedIn Audio Events, TikTok LIVE (primarily video-based but often centered on audio interaction), and Twitch (originally for video streaming, yet frequently hosting live oral exchanges such as talk shows and Q&As).

[2] In Discipline and Punish (1977), lt introduced the concept of the "calculated management of life," an idea that can now be applied to surveillance capitalism and biometric technologies such as those used in airports or security systems [110]. The analysis of vocal interactions, with their capacity to reveal emotional traits and power dynamics that remain invisible in textual analysis, represents a new frontier in social monitoring.

References

1. Scutari, M.; Malvestio, M. *The Pragmatic Programmer for Machine Learning: Engineering Analytics and Data Science Solutions*, 1st ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2025; p. 356.
2. Kitchin, R. Big data, new epistemologies and paradigm shifts. *Big Data Soc.* **2014**, *1*, 1–12. [CrossRef]
3. Kitchin, R. *Data Lives: How Data Are Made and Shape Our World*; Policy Press: Bristol, UK, 2021.
4. Striphas, T. Algorithmic culture. *Eur. J. Cult. Stud.* **2015**, *18*, 395–412. [CrossRef]

5. Burrell, J.; Fourcade, M. The society of algorithms. *Annu. Rev. Sociol.* **2021**, *47*, 213–237. [CrossRef]
6. Castells, M. *The Rise of the Network Society*; Blackwell: Oxford, UK, 1996.
7. Thompson, J.B. *The Media and Modernity. A Social Theory of the Media*; Blackwell: Cambridge, UK, 1995.
8. Crary, J. *24/7: Late Capitalism and the Ends of Sleep*; Verso Books: London, UK, 2013.
9. Chun, W.H.K. *Updating to Remain the Same: Habitual New Media*; MIT Press: Cambridge, MA, USA, 2016.
10. Floridi, L. *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*; Oxford University Press: Oxford, UK, 2014.
11. Floridi, L. *The Onlife Manifesto: Being Human in a Hyperconnected Era*; Springer Nature: Cham, Switzerland, 2015.
12. van Dijck, J. *The Culture of Connectivity: A Critical History of Social Media*; Oxford University Press: Oxford, UK, 2013.
13. van Dijck, J. Datafiction, dataism and dataveillance: Big Data between scientific paradigm and secular belief. *Surveill. Soc.* **2014**, *12*, 197–208. [CrossRef]
14. van Dijck, J.; Poell, T.; de Waal, M. *The Platform Society: Public Values in a Connective World*; Oxford University Press: Oxford, UK, 2018.
15. Gillespie, T. The politics of "platforms". *New Media Soc.* **2010**, *12*, 347–364. [CrossRef]
16. Gillespie, T. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*; Yale University Press: New Haven, CT, USA, 2018.
17. Galloway, A.R. *The Interface Effect*; Polity: Cambridge, UK, 2012.
18. Noble, S.U. *Algorithms of Oppression: How Search Engines Reinforce Racism*; NYU Press: New York, NY, USA, 2018.
19. Nakamura, L. Feeling good about feeling bad: Virtuous virtual reality and the automation of racial empathy. *J. Vis. Cult.* **2020**, *19*, 47–64. [CrossRef]
20. Benjamin, R. *Race After Technology: Abolitionist Tools for the New Jim Code*; Polity Press: Cambridge, UK, 2019.
21. D'Ignazio, C.; Klein, L.F. *Data Feminism*; MIT Press: Cambridge, MA, USA, 2020.
22. Han, B.C. *The Burnout Society*; Stanford University Press (Stanford Briefs): Stanford, CA, USA, 2015.
23. Han, B.C. *The Transparency Society*; Stanford University Press (Stanford Briefs): Stanford, CA, USA, 2015.
24. Han, B.C. *Psychopolitics: Neoliberalism and New Technologies of Power*; Verso Books: London, UK; New York, NY, USA, 2017.
25. Han, B.C. *In the Swarm: Digital Prospects*; MIT Press (Untimely Meditations): Cambridge, MA, USA, 2017.
26. Lovink, G. *Zero Comments: Blogging and Critical Internet Culture*; Routledge: New York, NY, USA, 2007.
27. Lovink, G. *Networks Without a Cause: A Critique of Social Media*; Polity Press: Cambridge, UK, 2012.
28. Lovink, G. *Sad by Design: On Platform Nihilism*; Pluto Press: London, UK, 2019.
29. Turkle, S. *Alone Together: Why We Expect More from Technology and Less from Each Other*; Basic Books: New York, NY, USA, 2011.
30. Morozov, E. *The Net Delusion: The Dark Side of Internet Freedom*; PublicAffairs: New York, NY, USA, 2011.
31. Morozov, E. *To Save Everything, Click Here: The Folly of Technological Solutionism*; PublicAffairs: New York, NY, USA, 2013.
32. Shirky, C. *Here Comes Everybody: The Power of Organizing Without Organizations*; Penguin Press (Penguin Group): New York, NY, USA, 2008.
33. Fuchs, C. *Social Media: A Critical Introduction*, 2nd ed.; SAGE Publications Ltd.: London, UK, 2014.
34. Fuchs, C. *Digital Humanism*; Emerald Publishing Limited: Leeds, UK, 2022.
35. Fuchs, C. *Media, Economy and Society: A Critical Introduction*; Routledge: London, UK, 2024.
36. Lévy, P. *Collective Intelligence: Mankind's Emerging World in Cyberspace*; Basic Books: New York, NY, USA, 1999.
37. Lévy, P. *Becoming Virtual: Reality in the Digital Age*; Plenum Trade: New York, NY, USA, 1997.
38. Balbi, G. *L'Ultima Ideologia*; La Terza Bari: Bari, Italy, 2022.
39. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*; PublicAffairs: New York, NY, USA, 2019.
40. Lyon, D. *The Culture of Surveillance: Watching as a Way of Life*; Polity Press: Cambridge, UK, 2018.
41. Deuze, M. *Media Work*; Polity Press: Cambridge, UK, 2007.
42. Deuze, M. Media life. *Media Cult. Soc.* **2011**, *33*, 137–148. [CrossRef]
43. Marwick, A.E. The public domain: Social surveillance in everyday life. *Surveill. Soc.* **2012**, *9*, 378–393. [CrossRef]
44. Zuboff, S. Surveillance Capitalism or Democracy? The Death Match of Institutional Orders and the Politics of Knowledge in Our Information Civilization. *Organ. Theory* **2022**, *3*, 1–79. [CrossRef]
45. Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*; Harvard University Press: Cambridge, MA, USA, 2015.
46. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 1997.
47. Tao, J.; Tan, T. Affective Computing: A Review. In *Affective Computing and Intelligent Interaction*; Tao, J., Tan, T., Picard, R.W., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3784, pp. 981–995. [CrossRef]
48. Crawford, K. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*; Yale University Press: New Haven, CT, USA, 2021.

49. Couldry, N.; Mejias, U.A. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*; Stanford University Press: Stanford, CA, USA, 2019.
50. Blodgett, S.L.; Barocas, S.; Daumé, H., III; Wallach, H. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5454–5476. [CrossRef]
51. Hovy, D.; Spruit, S.L. The social impact of natural language processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 591–598.
52. Brewer, R.; Ankenbauer, S.; Hashmi, M.; Upadhyay, P. Examining voice community use. *ACM Trans. Comput.-Hum. Interact.* **2024**, *31*, 1–29. [CrossRef]
53. Caroleo, L.; Maiello, G. Understanding polarization effects on voice-based social media: A Clubhouse analysis. *Ital. Sociol. Rev.* **2022**, *12*, 749. [CrossRef]
54. Corposanto, C.; Caroleo, L. Raccontami una storia. L'uso curativo dei social media orali nella popolazione anziana. *Salute Soc.* **2023**, *22*, 92–105. [CrossRef]
55. Kramer, N.C.; Winter, S. Impression management 2.0: The relationship of self-esteem, extraversion, self-efficacy, and self-presentation within social networking sites. *J. Media Psychol.* **2008**, *20*, 106–116. [CrossRef]
56. Cooley, C.H. *Human Nature and the Social Order*; Scribner's Sons: New York, NY, USA, 1902.
57. Zhao, S.; Grasmuck, S.; Martin, J. Identity construction on Facebook: Digital empowerment in anchored relationships. *Comput. Hum. Behav.* **2008**, *24*, 1816–1836. [CrossRef]
58. Nuyen, B.; Kandathil, C.; McDonald, D.; Thomas, J.; Most, S.P. The impact of living with transfeminine vocal gender dysphoria: Health utility outcomes assessment. *Int. J. Transgend. Health* **2023**, *24*, 99–107. [CrossRef] [PubMed] [PubMed Central]
59. Junior, C.N.V.; de Medeiros, A.M. Voice And Gender Incongruence: Relationship Between Vocal Self-Perception And Mental Health Of Trans Women. *J. Voice* **2022**, *36*, 808–813. [CrossRef]
60. Davies, S.; Papp, V.G.; Antoni, C. Voice and Communication Change for Gender Nonconforming Individuals: Giving Voice to the Person Inside. *Int. J. Transgenderism* **2015**, *16*, 117–159. [CrossRef]
61. Gillespie, T. Content Moderation, AI, and the Question of Scale. *Big Data Soc.* **2020**, *7*, 2053951720943234. [CrossRef]
62. Székely, É.; Miniota, J.; Hejná, M. Will AI Shape the Way We Speak? The Emerging Sociolinguistic Influence of Synthetic Voices. *arXiv* **2025**, arXiv:2504.10650. [CrossRef]
63. Behravan, M.; Mohammadrezaei, E.; Azab, M.; Gračanin, D. Multilingual Standalone Trustworthy Voice-Based Social Network for Disaster Situations. *arXiv* **2024**, arXiv:2411.08889.
64. Jones, J.M. The looking glass lens: Self-concept changes due to social media practices. *J. Soc. Media Soc.* **2015**, *4*, 100–124.
65. Gibson, J.J. *The Ecological Approach to Visual Perception*; Houghton Mifflin: Boston, MA, USA, 1979.
66. Norman, D.A. *The Psychology of Everyday Things*; Basic Books: New York, NY, USA, 1988.
67. Jung, K.; Park, Y.; Kim, H.; Lee, J. Let's Talk @Clubhouse: Exploring voice-centered social media platform and its opportunities, challenges, and design guidelines. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 30 April–5 May 2022; Association for Computing Machinery: New York, NY, USA, 2022. [CrossRef]
68. Jenkins, H. *Convergence Culture: Where Old and New Media Collide*; New York University Press: New York, NY, USA, 2006.
69. Kraus, M.W. Voice-only communication enhances empathic accuracy. *Am. Psychol.* **2017**, *72*, 644–654. [CrossRef]
70. Schroeder, J.; Kardas, M.; Epley, N. The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. *Psychol. Sci.* **2017**, *28*, 1745–1762. [CrossRef]
71. Kamiloğlu, R.G.; Boateng, G.; Balabanova, A.; Cao, C.; Sauter, D.A. Superior communication of positive emotions through nonverbal vocalisations compared to speech prosody. *J. Nonverbal Behav.* **2021**, *45*, 419–454. [CrossRef] [PubMed]
72. Daft, R.L.; Lengel, R.H. Information richness: A new approach to managerial behavior and organization design. *Manag. Sci.* **1986**, *32*, 554–571. [CrossRef]
73. Valacich, J.S.; Paranka, D.; George, J.F.; Nunamaker, J.F. Communication concurrency and the new media: A new dimension for media richness. *Commun. Res.* **1993**, *20*, 249–276. [CrossRef]
74. Dennis, A.R.; Valacich, J.S.; Speier, C.; Morris, M.G. Beyond media richness: An empirical test of media synchronicity theory. In Proceedings of the 31st Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 6–9 January 1998; IEEE: Piscataway, NJ, USA, 1998; Volume 1, pp. 48–57. [CrossRef]
75. Turkle, S. *Life on the Screen: Identity in the Age of the Internet*; Simon & Schuster: New York, NY, USA, 1995.
76. Taylor, T.L. *Play Between Worlds: Exploring Online Game Culture*; MIT Press: Cambridge, MA, USA, 2006.
77. Chen, M.G. Communication, coordination, and camaraderie in World of Warcraft. *Games Cult.* **2009**, *4*, 47–73. [CrossRef]
78. Kendall, L. *Hanging Out in the Virtual Pub: Masculinities and Relationships Online*; University of California Press: Berkeley, CA, USA, 2002.

79. Burnham, V. *Supercade: A Visual History of the Videogame Age, 1971–1984*; MIT Press: Cambridge, MA, USA, 2001.
80. Pearce, C. *Communities of Play: Emergent Cultures in Multiplayer Games and Virtual Worlds*; MIT Press (HTI Books): Cambridge, MA, USA, 2011.
81. Williams, D.; Ducheneaut, N.; Xiong, L.; Zhang, Y.; Yee, N.; Nickell, E. From tree house to barracks: The social life of guilds in World of Warcraft. *Games Cult.* **2006**, *1*, 338–361. [CrossRef]
82. Wadley, G.; Carter, M.; Gibbs, M. Voice in virtual worlds: The design, use, and influence of voice chat in online play. *Hum. Comput. Interact.* **2015**, *30*, 336–365. [CrossRef]
83. Fine, G.A. *Shared Fantasy: Role Playing Games as Social Worlds*; University of Chicago Press: Chicago, IL, USA, 2002.
84. Gee, J.P. *What Video Games Have to Teach Us about Learning and Literacy*, 2nd ed.; Palgrave Macmillan: New York, NY, USA, 2014.
85. Squire, K. *Video Games and Learning: Teaching and Participatory Culture in the Digital Age*; Teachers College Press: New York, NY, USA, 2011.
86. Slater, M.; Sanchez-Vives, M.V. Enhancing our lives with immersive virtual reality. *Front. Robot. AI* **2016**, *3*, 74. [CrossRef]
87. Boellstorff, T.; Nardi, B.; Pearce, C.; Taylor, T.L. *Ethnography and Virtual Worlds: A Handbook of Method*; Princeton University Press: Princeton, NJ, USA, 2012.
88. de Freitas, S.; Liarokapis, F. Serious games: A new paradigm for education? In *Serious Games and Edutainment Applications*; Ma, M., Oikonomou, A., Jain, L.C., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 9–23.
89. Steinkuehler, C.; Squire, K. Videogames and learning. In *The Cambridge Handbook of the Learning Sciences*, 2nd ed.; Sawyer, K., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 377–394.
90. Gee, J.P. *Games for Learning: An Ecological Perspective on the Potential of Games*; Routledge: London, UK, 2014.
91. Nardi, B.A. *My Life as a Night Elf Priest: An Anthropological Account of World of Warcraft*; University of Michigan Press: Ann Arbor, MI, USA, 2010.
92. Wiles, A.M.; Simmons, S.L. Establishment of an engaged and active learning community in the biology classroom and lab with Discord. *J. Microbiol. Biol. Educ.* **2022**, *23*, e00334-21. [CrossRef] [PubMed]
93. Saldanha, L.; da Silva, S.M.; Ferreira, P.D. "Community" in Video Game Communities. *Games Cult.* **2023**, *18*, 1004–1022. [CrossRef]
94. Jiang, J.A.; Middler, J.; Fiesler, C. Characterizing community guidelines on social media platforms. In Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing, Virtual, 8–22 November 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 287–291. [CrossRef]
95. Silverstone, R.; Haddon, L. Design and the domestication of ICTs: Technical change and everyday life. In *Communicating by Design: The Politics of Information and Communication Technologies*; Silverstone, R., Mansell, R., Eds.; Oxford University Press: Oxford, UK, 1996; pp. 44–74.
96. Rogers, R. *Digital Methods*; MIT Press: Cambridge, MA, USA, 2013.
97. Rogers, R. *Doing Digital Methods*; SAGE Publications: London, UK, 2019.
98. Venturini, T.; Rogers, R. *Digital Methods: A Short Introduction*; Polity: Cambridge, UK, 2025.
99. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 12449–12460.
100. Horiguchi, S.; Fujita, Y.; Watanabe, S.; Rudnicky, A.I.; Hershey, J.R. End-to-end joint speaker diarization and speech recognition for meeting transcription. *IEEE Trans. Audio Speech Lang. Process.* **2023**, *31*, 1234–1245.
101. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv* **2020**, arXiv:2006.11477.
102. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv* **2022**, arXiv:2212.04356. [CrossRef]
103. Bain, M.; Huh, J.; Han, T.; Zisserman, A. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *arXiv* **2023**, arXiv:2303.00747.
104. Communication, S.; Barrault, L.; Chung, Y.-A.; Cora Meglioli, M.; Dale, D.; Dong, N.; Duquenne, P.-A.; Elsahar, H.; Gong, H.; Heffernan, K.; et al. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation. *arXiv* **2023**, arXiv:2308.11596.
105. Bourdieu, P. *Language and Symbolic Power*; Polity Press: Cambridge, UK, 1991.
106. Giddens, A. *The Constitution of Society: Outline of the Theory of Structuration*; University of California Press: Berkeley, CA, USA, 1984.
107. Baym, N.K.; Boyd, D. Socially mediated publicness: An introduction. *J. Broadc. Electron. Media* **2012**, *56*, 320–329. [CrossRef]
108. Hirschberg, J.; Manning, C.D. Advances in natural language processing. *Science* **2015**, *349*, 261–266. [CrossRef]
109. Mansourifar, H.; Shi, W.; Wang, W. Conversational flow analysis in multi-party dialogues: Challenges and opportunities. *IEEE Trans. Multimed.* **2021**, *23*, 1567–1578.
110. Foucault, M. *Discipline and Punish: The Birth of the Prison*; Pantheon Books: New York, NY, USA, 1977.

111. Park, T.J.; Kanda, N.; Dimitriadis, D.; Han, K.J.; Watanabe, S.; Narayanan, S. A review of speaker diarization: Recent advances with deep learning. *Comput. Speech Lang.* **2022**, *72*, 101317. [CrossRef]
112. Kynych, F.; Cerva, P.; Zdansky, J.; Svendsen, T.; Salvi, G. A lightweight approach to real-time speaker diarization: From audio toward audio-visual data streams. *J. Audio Speech Music Process.* **2024**, 62. [CrossRef]
113. Lin, Y.; Cheng, M.; Li, Z.; Tang, B.; Li, M. Diarization-aware multi-speaker automatic speech recognition via large language models. *arXiv* **2025**, arXiv:2506.05796.
114. O'Shaughnessy, D. Speaker diarization: A review of objectives and methods. *Appl. Sci.* **2025**, *15*, 2002. [CrossRef]
115. Wu, X.; Nguyen, T.; Zhang, D.C.; Wang, W.Y.; Luu, A.T. FASTopic: Pretrained Transformer is a Fast, Adaptive, Stable, and Transferable Topic Model. *arXiv* **2024**, arXiv:2405.17978.
116. Mantelero, A. *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI*; Information Technology and Law Series; Springer: Cham, Switzerland, 2022; Volume 39. [CrossRef]
117. Bucholtz, M. The politics of transcription. *J. Pragmat.* **2000**, *32*, 1439–1465. [CrossRef]
118. Koenecke, A.; Nam, E.; Lake, J.; Nudell, M.; Quartey, Z.; Mengesha, C.; Toups, J.R.; Rickford, D.; Jurafsky, D.; Goel, S. Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 7684–7689. [CrossRef]
119. Hepburn, A.; Bolden, G.B. *Transcribing for Social Research*; SAGE Publications Ltd.: London, UK, 2017. [CrossRef]
120. Boersma, P.; Weenink, D. Praat: Doing Phonetics by Computer (Version 6.3) [Computer Program]. 2023. Available online: https://www.fon.hum.uva.nl/praat/ (accessed on 13 August 2025).
121. Microsoft. Azure Speech Service Documentation. 2024. Available online: https://learn.microsoft.com/azure/ai-services/speech-service/ (accessed on 13 August 2025).
122. Amazon Web Services. Amazon Transcribe Developer Guide. Available online: https://docs.aws.amazon.com/transcribe (accessed on 13 August 2025).
123. IBM. IBM Watson Speech to Text Documentation. 2024. Available online: https://cloud.ibm.com/apidocs/speech-to-text (accessed on 13 August 2025).
124. Caroleo, L. The Politics of Voice: The Evolution of Political Campaigns. In *(ECREA 2024): Communication & Social (Dis) Order, Proceedings of the 10th European Communication Conference, Ljubljana, Slovenia, 24–27 September 2024*; ECREA: Ljubljana, Slovenia, 2024; pp. 747–748.
125. McLuhan, M. *The Gutenberg Galaxy: The Making of Typographic Man*; University of Toronto Press: Toronto, ON, Canada, 1962.
126. McLuhan, M. *Understanding Media: The Extensions of Man*; McGraw-Hill: New York, NY, USA, 1964.
127. McLuhan, M.; Powers, B.R. *The Global Village: Transformations in World Life and Media in the 21st Century*; Oxford University Press: Oxford, UK, 1992.
128. Ong, W.J. *Orality and Literacy: The Technologizing of the Word*; Methuen: London, UK, 1982.
129. Buffardi, A.; de Kerckhove, D. *Il Sapere Digitale. Pensiero Ipertestuale E Conoscenza Connettiva*; Liguori Editore: Napoli, Italy, 2011.
130. Character.ai. Available online: https://character.ai (accessed on 20 August 2025).
131. Replika. Available online: https://replika.com (accessed on 20 August 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.