ISSN: 2473-4901
v12 n6702

Data Should Speak: AI-Driven Spoken Narratives and the Next Frontier in Data Storytelling

Michael O. Ojo
michaelojo.mailbox@gmail.com

Sushma Mishra
mishra@rmu.edu

Robert Morris University
Moon Township, PA 15108, USA

Abstract

Most organizations deliver analytical insights the same way they have for thirty years: charts, graphs, and dashboards on screens. That works — until the screen isn't available, the user can't see it, or the situation demands something different, faster, and more natural.

This paper argues that spoken data narrative — AI-generated, speech-synthesized spoken language that carries the full weight of organizational analytical insight — is a distinct and largely unexplored data storytelling modality. Grounded in dual coding theory and multiple resource theory, which together explain when spoken narrative complements visual analytics and when it outperforms visual delivery outright, the paper makes four contributions: it defines spoken narrative analytics (SNA) as a modality in its own right, distinct from data sonification, oral narrative analytics (ONA), and voice search; maps the AI technology stack — multimodal large language models, natural language generation, and neural text-to-speech synthesis — that makes SNA organizationally feasible now; identifies the use cases — mobile analytics, accessibility, high-frequency monitoring, and analytics democratization — where SNA offers a comparative advantage over visual delivery; and examines the remaining deployment challenges, including hallucination risk, information density constraints, and infrastructure integration. For IS educators, the paper highlights a practical gap: the engineering and governance skills required to build and govern SNA pipelines are not included in current IS curricula, and programs that address this first will be ahead of an emerging demand.

Keywords: data storytelling, spoken narrative, audio analytics, business intelligence, voice-driven analytics, natural language generation, artificial intelligence, analytics education, data communication.

ISSN: 2473-4901
v12 n6702

Data Should Speak: AI-Driven Spoken Narratives and the Next Frontier in Data Storytelling

Michael O. Ojo and Sushma Mishra

1. INTRODUCTION

Every dominant paradigm eventually reveals weaknesses that its very success obscures. The printing press democratized the written word and created a world legible only to those who could read. The dashboard democratized data and created an analytics landscape accessible only to those with a screen. For decades, the dominant model for delivering analytical insight has been unambiguously visual: data transformed into charts, graphs, and images, embedded in reports and presentations, and interpreted on screens. This paradigm has produced extraordinary tools and genuine advances. Yet its completeness has obscured a significant structural gap—one that is technically solvable, even now, but largely ignored.

Consider a few scenarios: the stakeholder who receives a performance dashboard on a smartphone while driving to a client meeting (Van der Heiden et al., 2022); the field worker who cannot monitor operational metrics because pulling out a device is impractical; the executive whose organization serves a visually impaired employee population, for whom the chart-and-graph paradigm is structurally exclusionary (Keilers et al., 2023; Erdemli, 2025); or simply the professional who processes information more effectively by listening. In each scenario, the visual paradigm fails not because of inaccuracy but because it delivers insights in the wrong modality.

This paper proposes that spoken narrative—the delivery of data-driven analytical insight through structured, AI-generated spoken language—is a distinct data storytelling modality that has largely been ignored by researchers and practitioners. Throughout, the paper refers to this as Spoken Narrative Analytics (SNA), defined as the NLG-driven transformation of analytical data into coherent spoken language that conveys insight, context, and narrative meaning. It distinguishes SNA from data sonification (non-linguistic audio signals), ONA, and voice analytics/voice search (speech as input rather than as output). While visual analytics still wins for exploration, comparison, and dense information tasks, and SNA complements rather than replacing it wholesale, the paper's central position is that SNA becomes the primary delivery mechanism in the specific contexts described in Section 5.

Three technological developments have converged to make SNA achievable at scale: advances in LLM-driven NLG; production-grade Text-to-Speech (TTS) technologies that render synthetic speech nearly indistinguishable from human speech (Tan et al., 2021; Hussain et al., 2025); and the proliferation of voice-first devices—smart speakers, in-vehicle systems, wearables, and mobile assistants—as delivery infrastructure for spoken analytics.

Despite these developments, spoken data storytelling remains nascent. The academic literature is dominated by work on visualization (Segel & Heer, 2010; Lee et al., 2015; Stolper et al., 2016), with comparatively little attention to audio or spoken modalities. Enterprise deployment remains sparse—the most prominent implementations have emerged from media organizations, most notably the Associated Press's use of Automated Insights' Wordsmith platform (Clerwall, 2014; Graefe, 2016)—and spoken data storytelling is essentially absent from IS curricula and competency frameworks.

This paper is guided by two research questions: (1) what defines Spoken Narrative Analytics as a distinct data storytelling modality, and (2) when does it outperform visual analytics as a delivery mechanism rather than merely complementing it? The paper proceeds as follows: Section 2 reviews the relevant literature; Section 3 outlines the methodology; Section 4 presents the conceptual framework and technology landscape; Section 5 covers use cases, implications for IS education, and deployment challenges; and Section 6 concludes.

2. LITERATURE REVIEW

2.1 The Data Storytelling Tradition

Data storytelling—the deliberate combination of data, narrative, and visual elements to communicate analytical insight—has drawn increasing scholarly attention over the past two decades. Segel and Heer (2010) proposed an early taxonomy of narrative visualization, spanning author-driven to reader-driven approaches. Lee et al. (2015) found that narrative framing significantly improves recall of data-driven information compared with unframed presentation. Stolper et al. (2016) introduced progressive disclosure—the incremental layering and revealing of insight—as a principle of narrative design.

The practitioner literature has been equally active. Knaflic (2015) developed data communication principles centered on the visual layer, while Dykes (2019) proposed a three-part framework in which data, narrative, and visuals are interdependent—each necessary but insufficient alone—arguing that their deliberate integration distinguishes genuine data storytelling from mere presentation.

2.2 Audio Analytics and Voice-Driven Interfaces

A modest yet growing body of literature examines the delivery of analytics through audio channels. Research on in-vehicle information systems has explored how spoken data summaries can provide drivers with situational awareness without diverting visual attention from the road (Strayer et al., 2017). Work in accessibility computing has similarly examined non-visual data presentation for blind and visually impaired users, identifying both the potential and the significant design challenges of translating visual data structures into audio equivalents (Ferres et al., 2013; Goncu & Marriott, 2011).

Consumer technology has seen widespread adoption of voice-driven analytics through platform providers (Amazon, Google, Apple) offering devices that support bidirectional voice analytics. Research shows users develop distinct mental models for voice-based information retrieval that differ significantly from those for visual search and dashboard interaction, with implications for how spoken analytics content should be structured and delivered (Porcheron et al., 2018; Lopatovska & Williams, 2018). Documented organizational deployments of SNA are examined in Section 3.3.

2.3 Distinguishing Spoken Narrative from Adjacent Concepts

Several related concepts in the literature are distinct from SNA as defined here. Data sonification (DS) translates data into non-linguistic audio signals—pitch, rhythm, and volume—for analytical purposes (Kramer, 1994; Hermann et al., 2011), as in EKG monitors, Geiger counters (Zanella et al., 2022), and NASA's sonification of telescope data (NASA, 2022). While DS serves important functions in scientific analysis and accessibility (Zanella et al., 2022), it produces no coherent semantic content—no insight, context, or story—unlike SNA. It differs from SNA in its medium of delivery, message precision, and literalness.

Oral Narrative Analytics (ONA), used in speech-language pathology and computational discourse analysis, systematically analyzes existing spoken narrative to derive structural or thematic insights (e.g., story-grammar coding, Labovian discourse parsing, narrative-arc extraction) (Labov & Waletzky, 1967). Like sonification, ONA is an analytic, input-side discipline: it takes extant spoken narrative and produces structured descriptions of it rather than generating spoken narrative as output.

Voice analytics, as typically defined in enterprise software, analyzes voice recordings to extract sentiment, intent, and behavioral patterns. Because voice is input rather than output, it is the reverse of SNA. It shares this input-side orientation with ONA but differs in its object: voice analytics extracts business signals from an interaction, while ONA extracts narrative structure from a story; neither generates spoken output from data. Conversational AI and chatbot interfaces likewise use natural language bidirectionally, but their design metaphor is dialogue and query-response, not the structured, narrative delivery of analytical insights that defines SNA.

2.4 Natural Language Generation and Automated Narrative

The technical foundation of SNA is natural language generation (NLG)—the computational production of human-readable text from structured data. The field has a substantial history (Reiter & Dale, 2000), but its relevance to organizational data communication accelerated with the emergence of LLMs. Early template-based systems worked well for narrow, repetitive tasks but required extensive manual

configuration and produced formulaic output with limited communicative range (Gatt & Krahmer, 2018). The shift to neural and transformer-based architectures changed that, enabling varied, contextually sensitive narratives at scale and making NLG a practical engine for spoken data storytelling.

Today's NLG systems produce narrative text that experts routinely judge comparable to human writing (Brown et al., 2020). Converting that text to speech requires Text-to-Speech (TTS) synthesis, which has crossed a similar threshold: neural TTS systems—ElevenLabs, Murf AI, Amazon Polly Neural, Google Cloud TTS—now produce synthetic speech that human evaluators find highly natural and difficult to distinguish from a human voice (Tan et al., 2021; Hussain et al., 2025). Since listener trust depends on perceived naturalness, these systems have cleared that bar in most organizational delivery contexts.

2.5 Theoretical Foundation: Dual Coding, Multiple Resource Theory, and Cognitive Load

The theoretical case for SNA rests on two lines of cognitive theory: one explaining why spoken narrative enhances visual analytics, and another explaining why it *outperforms* visual analytics alone under visual constraint.

Paivio's (1986) dual coding theory holds that cognition operates through two distinct, interconnected systems—verbal and nonverbal—and that information encoded in both is retained more robustly than when encoded in either alone (Clark & Paivio, 1991). Mayer's (2021) cognitive theory of multimedia learning extends this to instructional design: multimodal presentation deepens learning and reduces extraneous cognitive load relative to single-channel delivery, consistent with Reiter et al. (2016). This explains why pairing narrative with visuals encodes insight more deeply than either alone—not that narrative should replace visuals.

That comparative claim rests on Wickens' multiple resource theory (MRT) (Wickens, 1984, 2002, 2008): attention draws on separable resources—critically, the visual versus auditory modalities—rather than a single pool. Tasks that share a resource interfere; tasks that draw on different resources proceed concurrently with far less interference. This is why SNA gains its edge when the visual-spatial channel is already occupied—driving, walking, operating equipment—by routing insight through the auditory-verbal channel instead of competing for the same one, consistent with the driving-distraction findings reviewed in Section 2.2 (Strayer et al., 2014, 2015, 2017; Van der Heiden et al., 2022). The situationally-induced impairments and disabilities (SIID) framework (Sears et al., 2003) extends this to accessibility: a driver, a field worker with occupied hands and eyes, and someone permanently blind are functionally in the same state with respect to the visual channel, and MRT predicts the same outcome for all three—auditory-verbal delivery as mechanically superior, not merely an accommodation, whenever the visual-spatial resource is constrained or unavailable. Information foraging theory (Pirolli & Card, 1999) adds a smaller point: SNA lowers the effort cost of accessing insight when visual attention is unavailable, extending the effect to brief, low-stakes needs.

Together: dual coding and multimedia learning justify SNA as a complement when both channels are available; MRT and SIID justify it as the preferred primary channel when the visual-spatial resource is constrained, occupied, or absent—momentarily or persistently. SNA suits sequential, decision-relevant delivery under visual constraints; visual analytics suits exploratory tasks requiring simultaneous access to many data points.

The audio dimension of data storytelling is typically treated as a supplementary channel—the narrator accompanying a presentation—rather than a primary modality. SNA offers a theoretically grounded alternative to the assumption that data storytelling is inherently visual.

Recent work confirms that the written narrative layer of AI-driven analytics is maturing: Islam et al. (2024) demonstrate automated generation of data-driven narratives that integrate visualizations and text via LLMs, while adaptive gaze analytics in AI companions (Deng, 2025) and conversational image-recognition interfaces (Fathima et al., 2026) show both progress and persistent limitations in audio-enhanced analytics. In each case, audio is layered onto a system that remains screen-anchored—confirming that independent spoken delivery remains largely unaddressed. That gap—spoken narrative as a primary, standalone modality, independent of any screen, neither theorized nor studied as its own phenomenon—is what this paper addresses.

3. METHODOLOGY

3.1 Research Design

This paper employs conceptual analysis—appropriate for synthesizing fragmented knowledge, defining a new construct, and establishing a theoretical framework rather than testing empirical hypotheses (Jaakkola, 2020; Gregor, 2006). SNA spans adjacent literatures—NLG, audio/video analytics, data storytelling, accessibility—yet lacks a unifying framework; a proposed comparative framework emerging from this analysis appears in Section 4.1 (Table 2). Following Jaakkola's (2020) typology of conceptual-article designs, this paper adopts an integrative review approach: the literature search in Section 3.2 is structured and reproducible, but supports conceptual framework development rather than meeting formal Systematic Literature Review (SLR) completeness and appraisal standards, and shouldn't be evaluated against SLR reporting norms (e.g., PRISMA). Its purpose is simply to show that a unifying treatment of SNA does not yet exist and to identify the adjacent literature sources that inform the framework in Section 4.

3.2 Literature Collection

The literature review was conducted in April 2026 across four peer-reviewed databases—ACM Digital Library, IEEE Xplore, JSTOR, and Google Scholar—supplemented by targeted searches of Gartner and BI/data analytics practitioner publications. Records were included if peer-reviewed, conference-published, or from a recognized analyst firm (e.g., Gartner), and directly addressed spoken/audio delivery of analytics content, NLG applied to data narratives, TTS in an analytics context, or data communication through non-visual channels; excluded if they treated voice interfaces purely as input (voice search, ONA, command-and-control) without a narrative-generation component, or addressed sonification without a linguistic layer. No prior work using the term "Spoken Narrative Analytics" was found, so search terms were deliberately broad and adjacent-concept based (Table 1)—that breadth, and the resulting low direct-match yield, is itself evidence of the gap this paper identifies.

Across the four databases, initial searches returned more than 750 records. Applying inclusion/exclusion criteria at the title/abstract level, then full-text, eliminated off-topic and duplicate records, leaving 17 topically relevant—none a direct match to SNA as defined here, consistent with the paper's motivating gap. A second full-text pass for conceptual centrality left 7 surviving records, each synthesized for its contribution to technology capability, use-case/accessibility evidence, or theoretical grounding, and mapped onto the construct dimensions in Section 4.1; those records are cited throughout Sections 2 and 4.

Source/Journal		**Search Term 1**: ("Spoken Narrative" OR "speech synthesis" OR "Text-to-Speech") AND (“Analytics” OR "Business Intelligence") AND (AI OR "natural language generation” OR LLM)	**Search Term 2**: ("Spoken Narrative" OR "speech synthesis") AND (“Analytics” OR "Business Intelligence") AND (AI OR "natural language generation” OR LLM)	**Search Term 3**: ("Spoken Narrative") AND (“Analytics” OR "Business Intelligence") AND (AI OR "natural language generation” OR LLM)
ACM Digital Library	Details	725 records (filtered for context/relevance, recency).	725 records (filtered for context/relevance, recency).	5 records. Focus: communication, voice/video processing, AI, ML.
	Relevance/Match	N/A	N/A	2 relevant.
JSTOR	Details	53 records. Broad and varied; none on Data Analytics.	16 records. Broad and varied; none on Data Analytics.	0 records.
	Relevance/Match	0	0	0
IEEE Xplorer	Details	25 records. Video/text-language focus; 5 docs on Text/Video analysis delivering TTS via AI/NLG/ML.	13 records. Video/text-language focus; 5 docs on Text/Video analysis delivering TTS via AI/NLG/ML.	0 records.
	Relevance/Match	5 relevant	5 relevant	0 match.
Google Scholar	Details	15.8k records (filtered).	6,820 records (filtered).	108 records. Broad and varied; none on Data Analytics.
	Relevance/Match	Top 150 scanned; 10 relevant.	Top 150 scanned; 10 relevant.	2 relevant, sorted by relevance.

Table 1: Details of Literature searches and results

3.3 Organizational Practice Analysis

To ground the conceptual analysis in practice, the paper examined documented NLG-driven narrative implementations, drawing on published case studies, academic analyses, company announcements, and industry reports. Given the nascent state of organizational SNA deployments, the cases span NLG-to-text through NLG-to-speech implementations—reflecting the literature rather than a mature deployment population: the Associated Press's use of Automated Insights' Wordsmith (Clerwall, 2014; Graefe, 2016), The Washington Post's Heliograf (The Washington Post, 2016), and Arria NLG's deployment at EY (Arria NLG, n.d.). The first two are documented in peer-reviewed and journalistic sources independent of the vendor; the Arria/EY case, by contrast, is a vendor-published account included only as evidence that NLG-to-text reporting has reached production deployment, not as independent verification of outcomes.

ISSN: 2473-4901
v12 n6702

3.4 Limitations of Methodology

This work has three methodological limitations. No controlled studies have compared SNA with visual analytics in organizational settings, so claims about organizational value rely on extrapolation from related empirical literature. The practice analysis relies exclusively on publicly documented cases and practitioner-sourced observations, as proprietary deployments are absent from both academic and gray literature. The underlying technology landscape is evolving rapidly—specific technical claims reflect the field's state at the time of writing, though the conceptual framework is designed to outlast any particular instantiation.

4. THE PROPOSED CONCEPTUAL FRAMEWORK

4.1 Spoken Narrative as a Distinct Data Storytelling Modality

SNA constitutes a distinct data storytelling modality—differentiated from visualization, ONA, voice analytics, and data sonification alike—each with its own design requirements, cognitive mechanisms, and performance characteristics.

Four characteristics define it. First, linguistic structure: SNA delivers insight through coherent natural language, not through pitch or rhythm as in sonification. Second, algorithmic generation: the pipeline transforms structured data into narrative text and then into speech, enabling scale, real-time delivery, and personalization without human intervention. Third, primacy of delivery: SNA serves as the primary carrier of insight rather than a chart's supplement, particularly where visual presentation is unavailable or impractical. Fourth, conversational register: effective SNA approximates natural speech rather than a report read aloud.

The underlying pipeline, illustrated in Figure 1, works as follows.

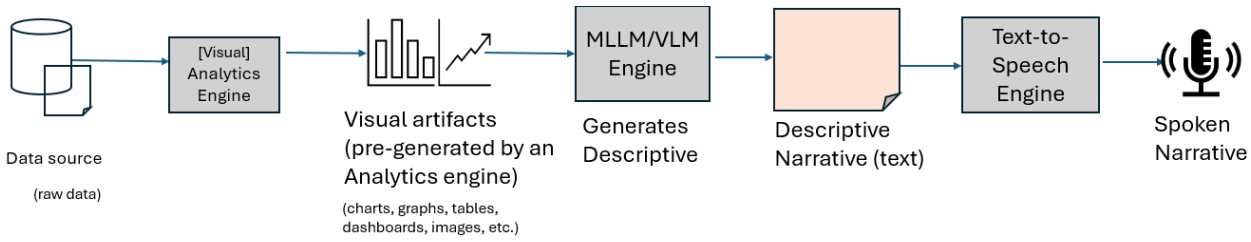

Figure 1: SNA process pipeline/workflow

An analytics engine preprocesses data into visual artifacts—charts, dashboards, tabular summaries, and images—though this step can be skipped if the data is already a visual-analytics output or if a meaningful narrative can be generated directly from raw data. A Multimodal Large Language Model (MLLM) or Vision-Language Model (VLM)—such as those powering Claude.ai and ChatGPT—reads these artifacts as a human analyst would, generating a contextually sensitive narrative that a Text-to-Speech (TTS) engine renders as audio. This supports voice-triggered drill-downs—dynamically generated spoken responses drawn from the same analytical context—a capability template-based NLG systems cannot provide. This pipeline above is presumed to be underpinned by data engineering best practices, including data quality, data governance, security, and compliance.

These characteristics have direct design implications. Unlike a dashboard, which presents dozens of data points simultaneously for the viewer to selectively scan, spoken narrative unfolds linearly: listeners cannot scan ahead or return, so SNA must sequence insights deliberately, anchor context at each step, and keep delivery within roughly 130–160 words per minute, since comprehension measurably degrades beyond that range (Reynolds & Givens, 2001). Voice characteristics—gender, accent, pitch, intonation—independently affect listener engagement and perceived trustworthiness (Jiang et al., 2021; Niculescu et al., 2024), a factor especially sensitive in multicultural settings where accent shapes credibility judgments of AI-delivered information. MLLM-driven generation mitigates these constraints through dynamic, context-aware construction rather than fixed scripts; prosodic design—rate, pause structure, emphasis, register—adds a further layer with no visual parallel, independently shaping comprehension and credibility.

Table 2 summarizes the proposed framework, comparing the existing visual analytics paradigm with SNA across dimensions relevant to organizational deployment. It also serves as a reference for the use-case analysis in Section 5.1 and the deployment challenges in Section 5.3. It outlines typical tendencies and tradeoffs rather than asserting either modality's categorical superiority: visual analytics retains clear advantages for exploratory analysis, multi-variable comparison, and dense information scanning, while the two modalities are frequently complementary, and SNA retains advantages where visuals are limited.

Dimension	Existing Visual Analytics Paradigm	Proposed SNA Framework
Primary output modality	Charts, dashboards, graphs, infographics	Spoken (audio) narratives
Generation mechanism	BI and Analytics tools (Tableau, Power BI, etc.)	Data pre-processing and analytics (optional)→MLLM/VLM→ NLG →Descriptive Narrative→TTS pipeline→Spoken Narrative
Audience requirement	Screen access, visual literacy, data literacy	Natural language comprehension only; possible multilingual
Delivery context	Desktop/screen-bound	Mobile, hands-free, anywhere audio, screen-limited environments/scenarios
Information structure	Simultaneous/spatial (multiple) data points — viewer selects	Sequential/linear — system prioritizes
Personalization	Static or pre-configured & deterministic views	Dynamic — MLLM can adjust register, depth, and emphasis per context
Accessibility	Typically requires fine motor skills (mouse/touch interaction to access tooltips and drill-downs). Excludes visually impaired and reading-limited users, as well as scenarios where visual presentation is unavailable or impractical.	Hands-free and screen-free by default — no separate tooling required for visually impaired users. Creates barriers for hearing-impaired users, users in noisy environments, and users (or scenarios) who need a persistent visual reference to strengthen the audio.
Update/delivery frequency	Scheduled refresh or manual	On-demand or real-time spoken briefing. Could be refreshable/repeatable on-demand.
Governance requirement	Data accuracy pre-validation before generating visuals	Data accuracy + narrative factual validation layer
Primary organizational use cases	Exploration, analysis, reporting	Monitoring, alerting, mobile delivery, analytics democratization, and real-time decision making.
IS Education Competency Model	Competency models exist to address the training of professionals.	Competency models do not exist to address the training of professionals

Table 2: SNA vs. Visual Analytics—Proposed Framework Comparison

4.2 The Technology Landscape: Current State and Trajectory

The SNA technology stack comprises four layers: data processing and analytics, MLLM/VLM narrative generation, speech synthesis, and delivery infrastructure (see Figure 1). Each layer has reached meaningful commercial maturity, but the defining development is the MLLM/VLM layer emerging as the intelligent bridge between visual analytics and spoken output. For an IS audience, the central question is whether the pipeline as a whole is accurate, auditable, governable, and trusted—properties that depend on the validation architecture in Section 5.3, not speech-synthesis quality alone. The vendor landscape below is illustrative, not the paper's central claim.

Earlier platforms—Arria NLG, Wordsmith, Yseop—offered governance and auditability suited to regulated industries (Gatt & Krahmer, 2018; Dale, 2020), but required significant per-domain setup and produced formulaic output. Contemporary MLLMs are more flexible: Claude (Anthropic, 2024), GPT-4o (OpenAI, 2024), and Gemini (Gemini Team, 2023) can read a chart or dashboard directly and generate spoken

narrative without templates, restructuring, or manual configuration, enabling conversational follow-up—a substantive advance over template-based approaches. For deployments requiring strict auditability, the practical answer is hybrid: MLLM narrative quality paired with a structured validation layer before delivery.

The TTS layer has matured as well. ElevenLabs, Murf AI, Amazon Polly Neural, Google Cloud TTS, and Microsoft Azure Cognitive Services all produce synthetic speech that human evaluators find difficult to distinguish from a human voice, with leading systems now broadcast-grade (Tan et al., 2021). ElevenLabs and Murf AI also offer voice customization for branded content. Remaining challenges include prosodic control—emphasis, pacing, and pauses for data narratives—and multilingual performance, which still degrades for lower-resource languages (Hussain et al., 2025).

The full SNA stack is commercially viable, as demonstrated by Google's NotebookLM Audio Overview, which uses Gemini's multimodal capabilities to generate spoken analytical discussions from uploaded content (Google, 2024). Key frontiers for organizational adoption at scale remain governance, auditability, BI platform integration, and enterprise security.

5. DISCUSSION

5.1 Organizational Use Cases and Comparative Advantages

Building on the maturing technology landscape in Section 4.2, four organizational scenarios stand out where SNA has a clear edge over visual analytics—delivering pre-interpreted, decision-relevant insight as the last mile of the analytical process.

The first is mobile, hands-free analytics: field workers, sales professionals, logistics personnel, healthcare providers, and executives in transit need performance data where screen engagement is impractical or unsafe. Strayer et al. (2014, 2015) show passive audio delivery imposes substantially less cognitive load than visual engagement with a screen while driving; SNA, as a pre-rendered spoken narrative, uses that low-load channel—a sales manager gets a spoken pipeline summary between client visits, a technician hears a site briefing while walking to a job. These aren't edge cases; they're daily reality for a large share of the workforce the dashboard was never designed for.

The second is accessibility and inclusion. Analytics platforms routinely fail employees with visual impairments or reading difficulties, a population whose accessible data representation has lagged well behind visualization's growth (Fan et al., 2023; Chundury et al., 2022; Erdemli, 2025). The WHO (2019) estimates 253 million people globally live with visual impairment—before counting reading difficulties, low visual literacy, or screen-access constraints more broadly. SNA delivers sighted stakeholders' insights through a fully accessible modality preserving narrative context and interpretive guidance; the same MLLM generating a mobile executive's briefing generates an equally accessible narrative for a screen-reader user, at little added engineering cost once the pipeline runs. As Table 2 notes, though, SNA broadens rather than solves accessible delivery universally—it can introduce its own barriers for hearing-impaired users, noisy environments, and users needing a persistent visual reference.

The third is high-frequency monitoring, alerting, and emergency decision support. Trading floors, logistics, and clinical environments demand sustained situational awareness of rapidly changing metrics, where continuous visual monitoring is costly and missed signals are consequential (Strayer et al., 2015). SNA offloads this to the auditory channel: a spoken alert that an SLA has been breached, a patient metric has crossed a threshold, or a landing should be delayed reaches the responsible person fastest, regardless of screen state, without dashboard navigation or color-code interpretation.

The fourth is analytics democratization. Visual analytics stays concentrated among technically proficient users; Gartner's finding that BI adoption sits around 30% despite decades of self-service investment reflects a modality problem, not a data-quality one (Gartner, 2017). The remaining 70% need a different interface to the same insight, not better charts—and SNA, delivered through voice-first devices already in their pockets, closes that gap by changing delivery mode rather than simplifying the data.

These four scenarios are conceptual claims—grounded in the Section 2.5 framework and related empirical literature—not controlled studies of SNA itself, which don't yet exist (Section 3.4). They're testable propositions motivating the research agenda in Section 6, not demonstrated outcomes.

5.2 Implications for IS Education and Curriculum Design

SNA poses a direct question for IS education: if organizations are going to build and deploy spoken analytics pipelines, where do the professionals who can do so come from? Right now, they're not being trained. IS2020—the ACM/AIS competency model for undergraduate IS programs (Leidig & Salmela, 2022)—includes no competencies for NLG-to-TTS pipeline design, MLLM configuration, spoken narrative design, or voice-output governance; the curriculum for the Data Architects and Data Analytics Engineers who will own these systems hasn't kept pace.

The core competency gap spans pipeline engineering, narrative design, and governance. On the engineering side: configuring and prompting MLLMs for data-driven narrative generation (Gatt & Krahmer, 2018; Anthropic, 2024; OpenAI, 2024); integrating neural TTS with appropriate prosodic control (Tan et al., 2021; Reynolds & Givens, 2001); and building the source-to-output validation layers that make spoken analytics trustworthy, not merely fluent. On the design side: structuring analytical insight for linear audio delivery—fundamentally different from building a dashboard, where sequencing, density, and register matter in ways visual design doesn't prepare graduates for. Both domains also need accessibility principles for screen-limited audiences (Fan et al., 2023; Chundury et al., 2022) and the AI governance and disclosure standards required for organizational deployment. None of these gaps is exotic—most extend what IS programs already teach, just not yet applied to the SNA modality.

5.3 Challenges and Barriers to Deployment of SNA

Earlier NLG-driven SNA systems required manually specifying content and pre-generating versions for each listener segment, but MLLMs and VLMs have largely overcome these constraints (Gatt & Krahmer, 2018; Dale, 2020): Claude, GPT-4o, and Gemini read visual analytical artifacts directly and generate audience-aware narratives that adjust register, depth, and emphasis without relying on static templates or editorial configuration (Anthropic, 2024; OpenAI, 2024; Gemini Team, 2023). Some challenges remain, though.

The first is trust and hallucination risk—a governance concern that should be treated as first-class engineering, not an afterthought. LLM-driven NLG carries the well-documented risk of confabulation: fluent, plausible, but factually incorrect output (Ji et al., 2023; Huang et al., 2025). This differs qualitatively from a chart displaying incorrect data—a bad dashboard number is visible and checkable, whereas a confidently stated incorrect insight may be accepted uncritically by a listener lacking the context to challenge it, since the auditory channel removes the visual verification cues chart readers use to catch anomalies. Production SNA deployments therefore need a validation layer that checks generated claims—stated figures, directional assertions, comparisons—against source data before audio is rendered; without it, organizations risk delivering confident-sounding but unverified claims at scale.

The second is information density. Spoken narrative delivers at 130–160 words per minute, a hard constraint with no visual equivalent (Reynolds & Givens, 2001): a dashboard showing 20 KPIs simultaneously could require 10–15 minutes of narrative to match that depth, forcing prioritization decisions that visual tools handle effortlessly. The solution is tiered delivery—a concise 90–120-second primary narrative covering the three to five most decision-relevant signals, with conversational drill-down for listeners wanting more. Since audio is otherwise linear, this requires non-sequential navigation via short voice commands ("skip," "next metric," "repeat," "explain further") and an underlying MLLM that maintains conversational state rather than treating each briefing as a one-shot narration—mirroring how an effective analyst briefs an executive and often improving on dense dashboards' "show everything" tendency.

The third is infrastructure integration. SNA requires integrating MLLM and TTS pipelines with existing data engineering and BI stacks, governance frameworks, and audio delivery endpoints. Most enterprise infrastructure wasn't designed to support them (Gartner, 2017)—though that's more manageable on modern cloud-native stacks, where MLLM APIs are managed services, neural TTS is a commodity call, and mobile/smart-speaker delivery is a solvable distribution problem. The harder challenge is upstream: the data quality, lineage, and governance issues that have long plagued visual analytics don't disappear when the output modality shifts to SNA. They simply become more audible.

6. CONCLUSION

The convergence of MLLM, NLG, and TTS technologies has substantially altered what is organizationally achievable, and this paper makes four contributions toward establishing the discipline around it. First, it establishes SNA as a distinct data-storytelling modality—precisely defined, bounded from adjacent concepts, and characterized by prosodic, cognitive, and editorial design requirements with no equivalent in visual analytics. Second, it maps the enabling technology landscape, providing the first systematic account of how the convergence of MLLM/VLM capabilities, neural TTS, and ambient computing infrastructure composes an organizational SNA pipeline. Third, it identifies four use cases—mobile analytics, accessibility, high-frequency monitoring, and analytics democratization—where SNA delivers a comparative advantage, while clarifying where visual tools remain superior. Fourth, it assesses the remaining deployment challenges—hallucination risk, information density, and infrastructure integration—showing that earlier NLG systems' template-based design and personalization constraints have been substantially resolved by contemporary MLLMs, leaving governance, validation, and integration as the primary engineering frontier.

Three research directions follow. First, empirical validation: the theoretical predictions in Section 2.5 remain untested in organizational settings. Second, governance and trust architecture: validating SNA outputs, quantifying confidence, and disclosing AI-generated content to listeners constitute a technically and ethically significant open problem at the intersection of AI governance and IS ethics. Third, SNA-specific design principles: the editorial, prosodic, and pipeline choices that determine whether a spoken analytics system delivers actionable insight or noise remain uncodified in the IS or HCI literature—representing one of the field's most pressing design challenges.

REFERENCES

Anthropic. (2024). *The Claude 3 model family: Opus, Sonnet, Haiku* [Model card]. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

Arria NLG. (n.d.). *EY: Ernst & Young* [Vendor case study]. Arria NLG. Retrieved July 21, 2026, from https://www.arria.com/cs-case-study-ey/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chundury, P., Patnaik, B., Reyazuddin, Y., Tang, C., Lazar, J., & Elmqvist, N. (2022). Towards understanding sensory substitution for accessible visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 1084–1094. https://doi.org/10.1109/TVCG.2021.3114829

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. https://doi.org/10.1007/BF01320076

Clerwall, C. (2014). Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice*, *8*(5), 519–531. https://doi.org/10.1080/17512786.2014.883116

Dale, R. (2020). Natural language generation: The commercial state of the art in 2020. *Natural Language Engineering*, *26*(4), 481–487. https://doi.org/10.1017/S135132492000025X

Deng, E. Y. (2025). AI-driven virtual companion: Advanced gaze analytics and adaptive interaction for enhanced engagement. In *Proceedings of the 2025 IEEE 3rd International Conference on Control, Electronics and Computer Technology* (ICCECT), pp. 837–845. IEEE. https://doi.org/10.1109/ICCECT64621.2025.11339672

Dykes, B. (2019). *Effective data storytelling: How to drive change with data, narrative, and visuals*. Wiley.

Erdemli, M. (2025). Designing accessible calendar tools for blind and low-vision users. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '25), Article 174, pp. 1–9. ACM. https://doi.org/10.1145/3663547.3759423

Fan, D., Siu, A. F., Rao, H., Kim, G. S., Vazquez, X., Greco, L., O'Modhrain, S., & Follmer, S. (2023). The accessibility of data visualizations on the web for screen reader users: Practices and experiences during COVID-19. *ACM Transactions on Accessible Computing*, *16*(1), Article 4. https://doi.org/10.1145/3557899

Fathima, N., Raman, P., Harini, P., Priya, S., Vijay, B., & Gowtham, C. (2026). Conversational image recognition system to comprehend an image. In *Proceedings of the 2026 International Conference on Visual Analytics and Data Visualization* (ICVADV), pp. 589–593. IEEE. https://doi.org/10.1109/ICVADV67766.2026.11470367

Ferres, L., Lindgaard, G., Sumegi, L., & Tsuji, B. (2013). Evaluating a tool for improving accessibility to charts and graphs. *ACM Transactions on Computer-Human Interaction*, *20*(5), 1–32.

Gartner. (2017). *Survey analysis: Why BI and analytics adoption remains low and how to expand its reach*. Gartner Research.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*(1), 65–170. https://doi.org/10.1613/jair.5477

Gemini Team, Google. (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. https://arxiv.org/abs/2312.11805

Goncu, C., & Marriott, K. (2011). GraVVITAS: Generic multi-touch presentation of accessible graphics. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction* (INTERACT 2011) (pp. 30–48). Springer.

Google. (2024, September 11). *NotebookLM now lets you listen to a conversation about your sources* [Blog post]. https://blog.google/technology/ai/notebooklm-audio-overviews/

Graefe, A. (2016). *Guide to automated journalism* [Tow Report]. Tow Center for Digital Journalism, Columbia University. https://doi.org/10.7916/D80G3XDJ

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, *30*(3), 611–642. https://doi.org/10.2307/25148742

Hermann, T., Hunt, A., & Neuhoff, J. G. (Eds.). (2011). *The sonification handbook*. Logos Publishing House. https://sonification.de/handbook/

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), Article 42. https://doi.org/10.1145/3703155

Islam, M., Hassan, N., & Bhatt, U. (2024). Automated generation of data-driven narratives using large language models. *arXiv preprint arXiv:2406.01107*. https://arxiv.org/abs/2406.01107

Hussain, I., Latif, S., & Qayyum, A. (2025). The state of TTS: A case study with human fooling rates. *arXiv preprint arXiv:2508.04179*. https://arxiv.org/abs/2508.04179

Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, *10*(1–2), 18–26. https://doi.org/10.1007/s13162-020-00161-0

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), Article 248. https://doi.org/10.1145/3571730

Jiang, X., Gossack-Keenan, K., & Pell, M. D. (2021). To believe or not to believe? How voice and accent information in speech alter listener impressions of trust. *Quarterly Journal of Experimental Psychology*, *74*(1), 56–76. https://doi.org/10.1177/1747021820939736

Keilers, C., Tigwell, G. W., & Peiris, R. L. (2023). Data visualization accessibility for blind and low vision audiences. In *Proceedings of INTERACT 2023*. Springer.

Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. Wiley.

Kramer, G. (Ed.). (1994). *Auditory display: Sonification, audification, and auditory interfaces*. Addison-Wesley.

Labov, W., & Waletzky, J. (1967). Narrative analysis: Oral versions of personal experience. In J. Helm (Ed.), *Essays on the Verbal and Visual Arts* (pp. 12–44). University of Washington Press.

Lee, B., Riche, N. H., Isenberg, P., & Carpendale, S. (2015). More than telling a story: Transforming data into visually shared stories. *IEEE Computer Graphics and Applications*, *35*(5), 84–90.

Leidig, P. M., & Salmela, H. (2022). The ACM/AIS IS2020 competency model for undergraduate programs in information systems: A joint ACM/AIS task force report. *Communications of the Association for Information Systems*, *50*, Article 25. https://doi.org/10.17705/1CAIS.05021

Lopatovska, I., & Williams, H. (2018). Personification of the Amazon Alexa: BFF or a master? In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 265–268). ACM.

Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781316941355

NASA Science. (2022). *Data sonifications: Black holes*. National Aeronautics and Space Administration. https://science.nasa.gov/science-research/astrophysics/data-sonifications/

Niculescu, A., Yeo, K. H., & D'Haro, L. F. (2024). The influence of accent and device usage on perceived credibility during interactions with voice-AI assistants. *Frontiers in Computer Science*, *6*, Article 1411414. https://doi.org/10.3389/fcomp.2024.1411414

OpenAI. (2024). *GPT-4o system card* [Technical report]. arXiv. https://arxiv.org/abs/2410.21276

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford University Press.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4), 643–675.

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). ACM.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press. https://doi.org/10.1017/CBO9780511519857

Reiter, E., Sripada, S., Hunter, J., Yu, J., & Davy, I. (2016). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, *167*(1–2), 137–169.

Reynolds, M. E., & Givens, J. (2001). Presentation rate in comprehension of natural and synthesized speech. *Perceptual and Motor Skills*, *92*(3c), 958–962. https://doi.org/10.2466/pms.2001.92.3c.958

Sears, A., Lin, M., Jacko, J., & Xiao, Y. (2003). When computers fade… Pervasive computing and situationally-induced impairments and disabilities. In *Proceedings of the 10th International Conference on Human-Computer Interaction* (HCI International '03) (Vol. 2, pp. 1298–1302). Lawrence Erlbaum Associates.

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1139–1148.

Stolper, C. D., Lee, B., Henry Riche, N., & Stasko, J. (2016). *Emerging and recurring data-driven storytelling techniques: Analysis of a curated collection of recent stories* [Technical Report]. Microsoft Research.

Strayer, D. L., Turrill, J., Coleman, J. R., Ortiz, E. V., & Cooper, J. M. (2014). *Measuring cognitive distraction in the automobile II: Assessing in-vehicle voice-based interactive technologies* [Technical Report]. AAA Foundation for Traffic Safety. https://aaafoundation.org/measuring-cognitive-distraction-automobile-ii-assessing-vehicle-voice-based-interactive-technologies/

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2015). *Measuring cognitive distraction in the automobile III: A comparison of ten 2015 in-vehicle information systems* [Technical Report]. AAA Foundation for Traffic Safety. https://aaafoundation.org/measuring-cognitive-distraction-automobile-iii-comparison-ten-2015-vehicle-information-systems/

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2017). The smartphone and the driver's cognitive workload: A comparison of Apple, Google, and Microsoft's intelligent personal assistants. *Canadian Journal of Experimental Psychology*, *71*(2), 93–110.

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*. https://arxiv.org/abs/2106.15561

The Washington Post. (2016, August 5). *The Washington Post experiments with automated storytelling to help power 2016 Rio Olympics coverage* [Press release]. https://www.washingtonpost.com/pr/wp/2016/08/05/the-washington-post-experiments-with-automated-storytelling-to-help-power-2016-rio-olympics-coverage/

Van der Heiden, R. M. A., Janssen, C. P., & Donker, S. F. (2022). The effect of cognitive load on auditory susceptibility during automated driving. *Human Factors*, *64*(7), 1195–1209.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 63–101). Academic Press.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

World Health Organization. (2019). *World report on vision*. WHO Press. https://www.who.int/publications/i/item/9789241516570

Zanella, A., Harrison, C. M., Lenzi, S., Cooke, J., Damsma, P., & Fleming, S. W. (2022). Sonification and sound design for astronomy research, education and public engagement. *Nature Astronomy*, *6*, 1241–1248. https://doi.org/10.1038/s41550-022-01721-z