Scilight

Journal of Educational Technology and Innovation
https://jeti.thewsu.org/index.php/cieti

Article

Bridging the Affective Gap: A Pedagogical Framework for Critical Voice Artistry in the AI Era

Jialin Ye [1] and Jingying Yang [2,*]

[1] School of Cultural Creativity and Media, Hangzhou Normal University, Hangzhou 310030, China
[2] School of Education, City University of Macau, Macau 999078, China
* Correspondence: m25092100206@cityu.edu.mo

How To Cite: Ye, J., & Yang, J. (2026). Bridging the Affective Gap: A Pedagogical Framework for Critical Voice Artistry in the AI Era. *Journal of Educational Technology and Innovation*, *8*(1), 12–24. https://doi.org/10.61414/wmee7j93

Received: 26 December 2025
Revised: 6 February 2026
Accepted: 10 March 2026
Published: 31 March 2026

Abstract: The rise of AI speech synthesis, while achieving impressive naturalness, has revealed a profound educational challenge: its failure to convey complex human emotions and contextual nuance—termed the "affective gap"—threatens to undermine the ecology of voice artistry and societal aesthetic discernment. This paper first diagnoses this gap by examining its key manifestations (compound emotion flattening, contextual deafness, the prosodic uncanny valley) and tracing its root cause to the epistemological divide between AI's data-driven pattern recognition and human embodied experience. It then analyzes the consequent structural disruption to the voice-acting industry's traditional "pyramid" training model and the broader risk of cultural aesthetic deskilling. In response, the paper's central contribution is to propose a novel pedagogical framework designed to bridge this gap. This framework advocates a decisive shift in voice education from skill transmission towards critical voice artistry, centered on cultivating students' capacities for deep textual/contextual analysis, empathetic and embodied sense-making, and the critical evaluation and direction of AI-generated speech. The paper argues that by integrating this critical pedagogical approach with strategic technology use, educators can empower future artists to navigate and shape a hybrid human-AI creative landscape. Ultimately, this work provides a theoretically grounded and actionable roadmap for innovating performing arts education in the AI era, positioning educational technology as vital steward of uniquely human expressive intelligence.

Keywords: AI speech synthesis; affective gap; pedagogical framework; critical voice artistry; performing arts education

1. Introduction

1.1. The Global Rise of Synthetic Voice and Its Latent Educational Challenge

The global landscape of synthetic voice generation is undergoing a seismic shift, moving from a niche research field to a cornerstone of the digital economy. Internationally, advancements in deep learning architectures and the availability of massive, curated datasets have propelled AI-generated speech from robotic monotones to a level of acoustic naturalness that challenges human perception (Tan et al., 2021). Pioneering models, such as those developed by OpenAI and Google, demonstrate capabilities in zero-shot voice cloning and expressive prosody control, fueling a vision of hyper-personalized audio interfaces and automated content creation at scale (Casanova et al., 2022; OpenAI, 2024). This technological trajectory is predominantly measured and celebrated through the lens of technical fidelity. This fidelity is quantified by metrics like the Mean Opinion Score (MOS), which prioritize the elimination of artificial noise and the attainment of fluent, human-like pronunciation (Chan & Kuang, 2025). This global trend is mirrored and intensified within specific national contexts, such as China's rapidly

expanding AI sector. Here, synthetic voices have seen widespread adoption in commercial applications, including virtual idols, audiobook production, customer service, and online education platforms. This growth is driven by both technological investment and strong market demand for scalable digital content solutions (China Academy of Information and Communications Technology, 2023). The discourse surrounding this adoption often emphasizes gains in efficiency, cost reduction, and accessibility, thereby accelerating the integration of these tools into the very fabric of media production and daily life.

However, this overwhelming focus on surface-level perfection and utilitarian application obscures a deeper, more pedagogically significant deficit. Beneath the technical sheen of global and local progress lies what we term the "affective gap": the systemic inability of contemporary AI to convey the complex, situated, and authentically human dimensions of emotional and contextual expression. While AI can simulate a palette of basic emotions, it fundamentally falters at rendering compound or contradictory feelings (e.g., bittersweet nostalgia), adapting to intricate social subtext and power dynamics, or producing the organic, intention-driven prosody that arises from lived experience (Cho et al., 2025; Łajszczak et al., 2024). This gap is not a transient technical bug but a persistent feature stemming from the data-driven paradigm of current AI.

Consequently, what appears as a success story of innovation transforms into an urgent and multifaceted educational challenge. For the field of voice artistry education—a domain dedicated to cultivating nuanced human expression—the proliferation of affectively limited yet technically proficient AI tools creates a profound paradigm crisis. It calls into question the foundational goals, curricular design, and pedagogical methods of a discipline whose traditional value proposition is being eroded by machines that expertly replicate the form of speech while remaining incapable of its substance. This disconnect between technological capability and expressive depth forms the core problematique that this paper seeks to address, arguing that educational response, not just technical refinement, is the critical imperative.

1.2. From Epistemological Divide to Educational Crisis

The "affective gap" is not merely an engineering shortfall. It reflects a deeper epistemological divide. At its core, state-of-the-art AI speech systems operate through indirect and decontextualized information processing. They learn by detecting statistical patterns across large datasets and modelling how language sounds in different contexts. What they cannot access is the lived "why" behind expression: the embodied sensations, socio-cultural meanings, and intentional states that give human speech its force and authenticity (Bender & Koller, 2020). Human expression, by contrast, is grounded and enactive. It emerges from lived experience within a body and a social world (Varela et al., 1991). For this reason, AI can simulate vocal expression with increasing sophistication, yet it remains limited in its capacity for situated understanding and genuinely felt response (Dreyfus, 2014). It can generate plausible affective forms, but it has no emotions of its own to express.

This limitation has consequences that extend well beyond technical performance. One immediate effect is the destabilization of traditional professional training. In fields such as voice acting, expertise has historically developed through a gradual apprenticeship structure, often resembling a pyramid. Novices begin with high-volume, standardized work and, over time, build technique, judgment, and professional experience (Lave & Wenger, 1991). The problem is that AI is especially effective at automating precisely these entry-level, pattern-based tasks. By contrast, the more interpretive and context-sensitive work at the top of the profession remains, at least for now, more resistant to automation. The result is an increasingly fragile training ecology. The experiential pipeline from novice to expert is weakened, producing a "missing middle" and raising concerns about the long-term renewal of skilled practitioners (Bakhshi et al., 2015).

A second consequence appears at the broader sociocultural level. As technically polished but affectively flattened synthetic voices become more common in media, interfaces, and entertainment, they may begin to reshape the normative soundscape. Audiences may gradually become accustomed to valuing clarity, smoothness, and consistency over expressive depth, subtlety, and the textured imperfections of human vocal performance (Woodruff et al., 2024). This is not simply a change in taste. It signals a broader process of aesthetic deskilling, in which collective sensitivity to vocal nuance is gradually diminished. Over time, this may weaken both public demand for sophisticated vocal art and the cultural value attached to the forms of expertise that arts education seeks to cultivate (Sennett, 2008).

Taken together, these shifts create a profound educational challenge. Traditional voice pedagogy was built for the transmission of relatively stable skills within a predictable professional hierarchy. That model is now under pressure. It risks preparing students for entry-level roles that are shrinking, while giving insufficient attention to the higher-order capacities that are becoming more important in hybrid human-AI environments. What now matters more are forms of learning centred on deep interpretation, contextual intelligence, critical judgment, and

creative direction (Xiong et al., 2025). The crisis, then, is not simply about whether technology should be used in the classroom. It is about rethinking what should be taught, why it matters, and what kind of future professional education ought to serve.

1.3. Thesis and Roadmap: Proposing a Framework for Critical Voice Artistry

This paper is grounded in the central argument that the "affective gap" in AI speech synthesis necessitates a paradigmatic shift in performing arts education. Rather than viewing AI as merely a new tool for skill acquisition or a threat to be resisted, we contend it must be recognized as a catalyst that fundamentally redefines the goals of voice artistry training. The primary educational response, therefore, cannot be to refine pedagogical methods for technical mimicry—a race that overlooks AI's inherent strengths and human core competencies. Instead, education must deliberately cultivate the uniquely human capabilities that reside beyond the reach of data-driven pattern recognition. These capabilities include hermeneutic depth for interpreting complex texts and contexts, embodied emotional intelligence for authentic expression, socio-cultural discernment, and the critical capacity to analyze, evaluate, and ethically direct synthetic media. This leads to our core research question: How can voice artistry education be systematically redesigned to bridge the affective gap, thereby empowering future practitioners to thrive as essential contributors in a hybrid human-AI creative ecosystem?

In response, this paper's central contribution is the development and elaboration of a Pedagogical Framework for Critical Voice Artistry. This framework proposes a decisive transition from the legacy model of skill transmission—focused on reproducing established techniques for a stable industry—to a new paradigm of critical artistry (Giroux, 2020). This paradigm positions the artist as a critical interpreter, a collaborative designer, and a strategic director. Their value lies in the very capacities that constitute the affective gap: contextual understanding, intentional creativity, and ethical curation (Yasuda & Maruyama, 2025). The framework is not anti-technology; rather, it advocates for the strategic integration of educational technology (EdTech) not as a substitute for human teaching, but as a means to deepen experiential learning, facilitate critical analysis of AI outputs, and simulate complex performance ecologies (Selwyn, 2022).

This article is a conceptual and theoretical paper. It does not report original empirical data; rather, it develops a pedagogical framework through theoretical analysis and synthesis of prior scholarship. In doing so, it provides a theoretical foundation for future empirical investigation and curriculum design. Accordingly, the paper unfolds as follows: It begins by examining the dual role of AI in performing arts education, analyzing its affordances and limitations. Building on this analysis, the paper then develops the core contribution—the Pedagogical Framework for Critical Voice Artistry—detailing its four foundational pillars. The manuscript concludes by discussing the framework's theoretical and practical implications and outlining avenues for future evaluation.

2. Manifestations of the Affective Gap: Defining the Educational Frontier

The "affective gap" is not an abstract concern but manifests in concrete, observable failures that demarcate the boundary between computational reproduction and human expression. These manifestations are not merely technical shortcomings; they delineate the precise terrain where human intelligence remains indispensable and, therefore, define the new competencies that must become the focus of voice artistry education. Diagnosing these symptoms is the first step in constructing a pedagogical response that targets the root of AI's limitations.

2.1. The Flatness of Compound Emotions: The Limits of Categorical Logic

Contemporary AI voice synthesis has mastered the simulation of a basic emotional palette—happiness, sadness, anger—treating emotions as discrete, categorical states that can be algorithmically selected and applied to speech output (Zhou et al., 2024). This engineering paradigm, however, collapses when confronted with compound emotions, such as bittersweet nostalgia, resentful gratitude, or awestruck fear. These states are not simple arithmetic sums of basic emotions. Instead, they are unique, phenomenologically rich experiences arising from complex cognitive appraisals, personal histories, and social contexts (Scherer, 2009). Their vocal expression requires a synthesis where one emotion modulates another, creating a singular, often contradictory, tonal quality—a tender warmth undercut by a tremor of loss, for instance. AI systems, operating on statistical correlations within decontextualized training data, lack the internal model of subjective experience necessary for this synthesis. When prompted to generate such complexity, the output typically defaults to a dominant basic emotion, flattens into ambiguous neutrality, or produces an implausible, segmented alternation between emotional cues (Cho et al., 2025; Łajszczak et al., 2024). The failure is epistemic. The system can process the label "bittersweet," but it cannot comprehend its intentional object—the reason why something is bittersweet. Furthermore, it cannot access the embodied memory that gives the feeling its authentic somatic and vocal signature. This emotional monoculture in

AI output starkly contrasts with the fluid, layered, and dynamic nature of human affective experience. Consequently, a core educational objective must shift. The goal is no longer to teach actors how to portray basic emotions, but to train them in the hermeneutic depth and intrapersonal awareness required to inhabit and express complex emotional realities. This represents a foundational pillar of the critical voice artistry we propose.

2.2. Contextual Deafness: The Missing "Scene" in AI Performance

A second, critical manifestation is a profound contextual deafness. While AI can process words and syntax, it remains critically blind to the rich, multifaceted "scene" or context that governs appropriate human communication (Goffman, 1981). This context includes the sociolinguistic fabric of an interaction (e.g., intimacy, power dynamics, unspoken subtext) and the physical-narrative environment (e.g., a library, a battlefield, a state of exhaustion).

This blindness leads to pragmatic failures. An AI cannot reliably adjust its vocal register for such nuanced differences. For example, it struggles to distinguish between a boss's reprimand, a lover's complaint, or a friend's joke—even when they all use the same words: "You're late" (Que & Ragni, 2025). It misses sarcasm, veiled threats, or diplomatic indirectness, as these depend on understanding shared knowledge and social goals beyond the literal text (Hovy, 2015). Furthermore, while audio filters can simulate shouting or whispering, AI lacks the embodied simulation of being in a situation, unable to organically generate the breathlessness of exertion or the tense restraint of a hidden fear (Huang et al., 2025).

In essence, AI performs decontextualized signal mapping. It excels at linking text to sound but fails at the situated understanding required for genuine performance. This failure means it cannot answer the fundamental questions of context: Who is speaking, to whom, under what conditions, and to what unspoken end? This deficit underscores that future voice artists must be trained not just as vocal technicians, but as dramaturges and contextual analysts. Education must therefore prioritize situational intelligence and socio-pragmatic understanding. This equips students to build and perform within the full contextual "field" of a scene, developing a competency that AI cannot automate.

2.3. The Uncanny Valley of Prosody: The Intentionality Gap

The affective gap culminates in a subtle yet unsettling phenomenon: an uncanny valley of prosody (Ross et al., 2024). As synthetic speech nears human-like fluency, its imperfections in rhythm, intonation, and pausing become peculiarly salient, triggering unease. This disquiet stems from micro-failures in the organic variability that signals a conscious, feeling mind. AI-generated intonation often exhibits "over-regularization". It follows statistically learned, smoothed contours that lack the moment-to-moment, semantically motivated adjustments of human speech (Chuenwattanapranithi et al., 2009). Similarly, AI-modeled pauses are mechanistically accurate but emotionally void, missing the cognitive planning, emotional weight, or filled hesitations ("um", "uh") that characterize natural, intentional discourse (Clark & Fox Tree, 2002). The result is a delivery that can sound correct yet feels hollow, pressured, or oddly inhuman. These prosodic micro-deficits are perceptually powerful because they directly index an intentionality gap. Human prosody is an audible trace of embodied cognition—the interplay of breath, laryngeal tension, and real-time cognitive-emotional processing (Huang et al., 2025). AI's prosody is a post-hoc acoustic mask, not the emergent signature of a lived moment. This reveals that the most elusive yet fundamental aspect of performance is presence—the sonic evidence of a mind actively making meaning. Therefore, a pivotal goal of education must be to cultivate this performative presence and intentionality. Through embodied practice and critical self-listening, students can develop an authentic, in-the-moment expressivity. This is a quality that defines artistic truth and lies fundamentally beyond algorithmic reach.

3. The Ontological Roots of the Gap: Data vs. Lived Experience as an Educational Fulcrum

The manifestations of the affective gap are not random errors but systematic outputs arising from a fundamental ontological divide between how AI "knows" and how humans understand. This divide is not merely a technical specification; it forms the epistemological fulcrum upon which any effective educational response must be leveraged. To design pedagogy that bridges the gap, we must first understand the nature of the chasm it seeks to span.

3.1. The AI's Epistemology: The Universe of Secondhand Data and Decontextualized Correlation

State-of-the-art AI speech systems operate within an epistemology of correlation. Their "knowledge" and "expression" are exclusively derived from identifying and replicating statistical patterns within colossal, pre-existing datasets of text and speech (Bender & Koller, 2020). This process is one of decontextualized signal mapping. For instance, an AI can learn that the acoustic features of slower speech rate and lower pitch frequently

co-occur with the textual label "sad" in its training corpus. It becomes extraordinarily adept at generating this correlation on command. However, this mastery is fundamentally indirect and derivative. The AI processes millions of descriptions of a "sunset" but has never felt its warmth; it analyzes thousands of recordings tagged "angry" but has never experienced the flush of adrenaline or the relational rupture that precipitates rage. Its world is a high-dimensional statistical shadow of human reality, a map of surface correlations stripped of their lived context (Harnad, 1990). A crucial educational insight emerges here: an AI learns the acoustic signature of an emotion but is constitutively blind to its intentional object and situated cause. It can mimic the sound of "heartbreak" but cannot comprehend the loss that gives the sound meaning, nor the personal history that shapes its unique vocal texture. Consequently, its output, however polished, is generic and combinatorial, an echo of past utterances rather than a novel expression born of present, felt experience. For educators, this clarifies that teaching students to merely replicate emotional "sounds" is to train them for a task at which AI already excels and is improving exponentially. The educational frontier must lie elsewhere.

3.2. The Human Foundation: Embodied, Situated, and Social Sense-Making

In stark ontological contrast, human cognition and expression are grounded, enactive, and embodied (Varela et al., 1991). We do not process abstract symbols detached from the world; we understand and communicate through our bodies, within specific situations, and against a backdrop of socio-cultural practices. Emotional expression originates not as a label but as a biological event—a racing heart, contracted muscles, altered breath—that is then shaped into culturally recognizable vocal and linguistic forms (Damasio, 1999). The voice is not a separate instrument but an extension of the autonomic nervous system; its prosody is a bio-acoustic readout of internal state (Porges, 2011). This is why a genuine tremor of fear is perceptibly different from a simulated one. Furthermore, human understanding is irreducibly social and narrative. We learn the meaning of "trust", "betrayal", or "sarcasm" not from definitions but through cumulative, lived interactions within a community (Fuchs, 2017). Our communicative acts are situated performances, deeply sensitive to the unspoken rules of power, intimacy, and shared history (Clark, 1996). This contextual intelligence is what allows a human actor to decide how to say a line based on who they are speaking to and why. This rich, embodied, and social matrix is the wellspring of authenticity, nuance, and creative meaning—the very qualities missing in AI speech. Therefore, education must pivot to become the dedicated cultivator of this matrix.

3.3. The Consequential Divide: Algorithmic Recombination vs. Authentic Creation and Affect

The ontological chasm yields two definitive consequences that directly inform pedagogical goals. First, it results in a categorical difference in creativity. AI exhibits combinatorial prowess, brilliantly interpolating and recombining elements from its training data to produce novel-seeming variations (Boden, 2009). However, this is pastiche, not poetry. Human creativity, especially in artistic interpretation, often involves transformational insight—a novel perception of a situation, character, or emotion that springs from a personal, embodied perspective and results in a genuinely new expression of meaning (Boden, 2003). An AI has no "self" from which such an original perspective can emerge. Its "choices" are probabilistic selections. Thus, education must foster not just the skill of variation, but the capacity for generative interpretation and authentic artistic voice.

Second, it creates a fundamental rift in the nature of emotional affect. AI-generated emotion is the product of offline computation: a deliberative retrieval and synthesis of stored patterns based on an input label. Human emotional expression, in contrast, is often the online, emergent output of a psychobiological state, unfolding in real-time as an integrated mind-body response to a situation (Scherer, 2009). The spontaneity of a gasp, the involuntary crack in a voice, the micro-adjustment in tone during a conversation—these are not "chosen" but experienced and expressed. AI cannot "feel with" another (empathy) and therefore cannot produce this quality of instantaneous, embodied emergence. Its affect is always a representation, never a present event. This underscores a critical educational objective: moving beyond teaching students to "indicate" emotion, towards training them to access, channel, and express genuine emotional truth from a place of embodied awareness.

In sum, the affective gap originates in the difference between processing data and living an experience. This analysis provides the theoretical blueprint for our pedagogical framework: education must systematically develop the embodied, contextual, and authentically creative capacities that define the human side of this ontological divide, equipping students not to imitate AI, but to master and direct the domain it cannot enter.

4. The Pedagogical Imperative: Industry Disruption and the Collapse of the Apprenticeship Model

The affective gap is not merely a technical limitation but a powerful disruptive force reshaping the professional ecology of voice performance. This disruption exposes a critical vulnerability in traditional talent

development pathways, creating an urgent imperative for educational reform. To understand the scale of the pedagogical challenge, we analyze the industry's structural transformation through the lens of the "Pyramid Model". Traditionally, the voice-acting profession has functioned as a pyramid: a broad base of practitioners engaged in high-volume, standardized work (e.g., telephony, basic narration); a midsection involving more varied and demanding projects; and a narrow apex occupied by elite artists capable of supreme interpretive and emotional performance. This structure was sustained by an implicit apprenticeship ecology, where novices entered at the base, honing foundational skills through repetition before progressing upward (Lave & Wenger, 1991). The affective gap, however, enables asymmetric automation: AI excels at automating the pattern-based, lower-complexity tasks at the pyramid's base while being incapable of the high-complexity work at the apex. This does not simply erase the base layer; it dismantles the entire developmental pipeline, collapsing the pyramid from the foundation up and creating a direct crisis for education.

4.1. The Erosion of the Foundational Base: Automating the Apprenticeship

The most immediate impact is the rapid automation of the standardized, bulk vocal work that constitutes the pyramid's base. AI synthesis offers near-zero marginal cost and instant scalability for tasks like IVR systems, e-learning modules, and routine announcements—precisely the domain where unique emotional depth is not a priority (Brynjolfsson & McAfee, 2014). From an educational standpoint, this automation represents more than job displacement; it signifies the systematic removal of the profession's primary apprenticeship platform.

For generations, these entry-level roles served as an indispensable, immersive training ground. It was here that aspiring actors built non-negotiable craft fundamentals—microphone technique, script sight-reading, vocal stamina, and professional discipline—through the sheer volume and variety of real-world work (Ciccarelli & Ciccarelli, 2013). This process constituted a form of situated, experiential learning that studio-based education struggles to replicate. The erosion of this base severs the traditional pathway at its origin, leaving educational institutions as the sole, yet insufficient, gateway to the profession.

4.2. The (Temporarily) Defended Apex: The Sanctuary of Human Artistry

Conversely, the work at the pyramid's apex—demanding the synthesis of complex emotions, deep narrative understanding, and unique artistic signature—remains a defensive height against direct AI replacement. This work requires the very capacities defined by the affective gap: embodied empathy, contextual intelligence, and authentic creative interpretation (Brade et al., 2025). Elite artists function as emotional architects, making nuanced choices drawn from lived experience—a qualitative reasoning process beyond current AI's reach. Thus, while AI may serve as a tool for ideation or placeholder audio, the final, authoritative performance that carries cultural and economic premium remains firmly human (du Sautoy, 2019). This defense, however, is not permanent but contingent on the continued cultivation of the expertise it requires. The apex cannot sustain itself in isolation. Its resilience depends on a healthy, reproducing ecosystem beneath it—the very ecosystem now under threat.

4.3. The Broken Pipeline: A Crisis of Succession and the Educational Vacuum

The most severe consequence is the structural collapse of the developmental pipeline. The pyramid model relied on graduated experiential learning, where skill and artistry were accrued through progressive challenges (Lave & Wenger, 1991). The automation of the base ruptures this continuum of experience.

Without the foundational platform of commercial work, future generations face a "missing middle" in their professional development (Bakhshi et al., 2015). They may graduate with theoretical knowledge and basic studio technique but will lack the polished craft, resilience, and pragmatic problem-solving skills forged in the crucible of entry-level work. Educational institutions, no matter how well-equipped, cannot fully simulate the diversity of client demands, tight deadlines, and iterative feedback that constituted the traditional apprenticeship.

Consequently, the industry risks a disastrous bifurcation: a shrinking cohort of aging masters at the apex, disconnected from a growing pool of under-prepared aspirants with no viable bridge between them. The pyramid does not merely shrink; its structural integrity fails. This broken pipeline is the ultimate pedagogical manifestation of the affective gap. A technology that can mimic outputs not only displaces tasks but actively jeopardizes the embodied, experiential process through which human artistic excellence is cultivated and transmitted across generations. It creates an educational vacuum, demanding nothing less than a wholesale re-invention of how voice artists are trained, for a future where the entry-level journeyman phase has vanished, and the path to mastery must begin anew in the classroom. This structural transformation of the traditional apprenticeship model is summarized in Table 1.

Table 1. Traditional vs. AI-Disrupted Voice Acting "Pyramid Model".

Tier	Traditional Pyramid (Human-Centric)	Post-AI Disruption
Base	High-volume, standardized work (e.g., telephony, basic narration)	Largely automated by AI
Middle	Varied, moderate-complexity projects	Reduced demand; hybrid roles emerge
Apex	High-complexity, emotional, interpretive performances	Still human-dominated, but pipeline broken

5. The Sociocultural Dimension: Aesthetic Deskilling as an Educational Frontier

The implications of the affective gap transcend the economics of the voice-acting profession, extending into the foundational realm of cultural perception and collective sensibility. The pervasive integration of technically polished but affectively flattened synthetic voices into media, interfaces, and entertainment poses a profound, long-term sociocultural risk: aesthetic deskilling. Adapted from theories of workplace deskilling, this concept describes the gradual erosion of a population's capacity to perceive, appreciate, and critically evaluate nuanced sensory and emotional information due to diminished exposure and practice (Sennett, 2008). As AI-mediated speech becomes the dominant sonic backdrop, it threatens to recalibrate human auditory expectation, with dire consequences for empathy, cultural diversity, and the public's capacity to engage with complex artistic expression. For education, this shifts the mission from merely training professionals to cultivating a discerning public and stewarding cultural literacy.

5.1. Recalibrating the "Educated Ear": When Flawlessness Becomes the Norm

Human perception is not static but is dynamically shaped by environmental inputs through neuroplastic adaptation (Irvine, 2018). The "ear" is a culturally formed organ of discernment. When the primary vocal models encountered daily—from virtual assistants and automated announcements to synthetic narrators in children's media—are characterized by narrow, predictable, and emotionally simplified prosody, the public's auditory palate is subtly trained to accept this as the normative benchmark for speech (Kassabian, 2016). This process constitutes a powerful form of aesthetic socialization. A generation reared in an AI-saturated soundscape may internalize a value system for "good" voice that prioritizes technical flawlessness, clarity, and consistency over expressive depth, unpredictable nuance, or the evocative "imperfections" that signify authentic human presence and effort (Woodruff et al., 2024). The raw vulnerability in a blues singer's voice, the hesitant, searching quality of a profound soliloquy, or the complex tonal shifts of a conflicted character may increasingly register not as emotionally rich but as deviant, inefficient, or unpolished. This recalibration of the "educated ear" creates a public less equipped to value the very artistry that advanced training seeks to produce, thereby undermining the ecosystem that supports high-level voice art. Consequently, a key educational objective expands to include fostering critical listening skills and sophisticated aesthetic judgment in students, preparing them to be both creators and connoisseurs in a hybrid media landscape.

5.2. The Blunting of Nuance: Eroding Emotional and Cultural Discernment

The direct corollary of this recalibrated expectation is a collective blunting of nuance appreciation. Discriminating subtle emotional cues—a strained warmth, a sarcasm laced with affection, a tremor of doubt beneath confidence—is a learned, cultural skill requiring exposure and interpretive frameworks (Gendlin, 1997). A soundscape dominated by broad-stroke or affectively neutral AI speech deprives the public of this essential practice, causing the cognitive and empathetic "muscles" for fine-grained interpretation to atrophy.

This deskilling operates on multiple fronts:

Cognitively, listeners may become less adept at parsing complex prosodic information, defaulting to simpler, categorical interpretations.

Emotionally, the capacity for empathetic resonance, often triggered by subconsciously detected vocal cues, may weaken (Juslin & Laukka, 2003). Culturally, shared competences for understanding genre-specific or culturally coded vocal performances (e.g., the cadences of classical theatre, the understatement in certain film traditions) erode if such models are absent from mainstream, AI-influenced media.

The result is a flattened auditory-emotional intelligence, where layers of communicative intent are lost on their audience. This creates a challenging environment for graduates of voice programs, who may find a market with a diminished appetite for the subtlety they have been trained to deliver. Education, therefore, must assume a role in public cultural education, not only honing students' expressive skills but also teaching them to articulate and advocate for the value of nuance, thereby educating their future audiences.

5.3. The Homogenization of Expression: Efficiency Versus Ecological Diversity

The ultimate cultural risk is a systemic move towards the homogenization of expressive culture. The economics of AI favor standardization: a single, highly adaptable model can generate globally comprehensible, inoffensive, and commercially optimized speech for infinite purposes. This industrial logic stands in direct opposition to the ecology of human expression, which thrives on idiosyncrasy, local inflection, historical particularity, and spontaneous innovation (Fischer, 1999).

As synthetic voices become the default for cost-effective content production, the vibrant diversity of human vocal styles—regional accents, sociolects, idiosyncratic timbres, and unique artistic signatures—risks marginalization in favor of a limited palette of "optimal" synthetic voices. This creates a vicious cycle: homogenized outputs train homogenized expectations, which in turn drive demand for more homogenized content (Horkheimer et al., 2002). The challenging, the acquired-taste, the locally specific, and the deeply personal in vocal art may be pushed to the cultural periphery.

In this context, the voice educator's role transforms into that of an ecologist of expression. The pedagogical framework must consciously counter this homogenizing pressure by valuing and cultivating idiosyncratic voice, cultural specificity, and personal artistic signature. It must frame diversity of expression not as a market inefficiency but as a core cultural and educational value. By doing so, education can help preserve the "wilderness" of human expression against the encroaching "managed garden" of synthetic uniformity, ensuring that the affective gap does not become a conduit for a profound cultural flattening.

6. A Pedagogical Framework for Critical Voice Artistry

The analysis thus far presents not merely a critique but a clear mandate: the "affective gap" defines the new frontier for voice artistry education. This gap, rooted in an ontological divide between data processing and lived experience, has catalyzed a dual crisis—the collapse of traditional apprenticeship pipelines and the risk of widespread aesthetic deskilling. In response, this section moves from diagnosis to prescription, proposing a comprehensive Pedagogical Framework for Critical Voice Artistry. This framework is designed not to compete with AI on its own terms but to systematically cultivate the uniquely human capacities that AI lacks, thereby bridging the gap by empowering a new generation of artists as essential interpreters, directors, and ethical guides in a hybrid creative ecosystem.

6.1. Theoretical Underpinnings: Beyond Skill Transmission

The proposed framework is rooted in a broader rethinking of what voice artistry education is meant to do. Rather than treating training as the efficient transmission of stable techniques, it approaches artistic formation as a process that is embodied, interpretive, and critically situated. This shift draws on three complementary lines of thought.

The first is embodied and enactive cognition. From this perspective, understanding and expression do not originate in abstract cognition alone, but emerge through sensorimotor experience and situated action (Varela et al., 1991). For voice training, this means that pedagogy cannot be reduced to vocal control or technical imitation. Students need opportunities to develop somatic awareness and to work from lived, felt experience as a central resource for artistic expression.

The second strand comes from critical pedagogy. If students are to work meaningfully in an environment shaped by synthetic media, they must be able to do more than perform within it. They must also learn to interpret and question it. Critical pedagogy offers a way of developing that capacity by encouraging students to examine power relations, cultural codes, and the ideological assumptions embedded in media systems, including AI-generated speech (Giroux, 2020). In this sense, criticality becomes an artistic as well as an ethical competence.

The third strand is technology-enhanced learning. Here, technology is not positioned as a substitute for the teacher or as a shortcut to artistic mastery. Its value lies in how it can extend experience, support complex simulation, and create new spaces for analysis, experimentation, and reflection (Selwyn, 2022). Used in this way, technology does not displace human formation; it can deepen it. Taken together, these perspectives move the framework beyond a narrow model of skill delivery. The aim is to cultivate critical voice artists who possess not only technical craft, but also hermeneutic depth, contextual intelligence, and the capacity to shape creative processes involving both human and artificial agents.

6.2. Curricular Pillars: A Four-Dimensional Model

The framework translates these theoretical commitments into four interrelated curricular pillars. Together, they move from inward awareness to outward collaboration, and from individual expression to critical engagement with hybrid creative systems.

The first pillar is Hermeneutic Depth & Contextual Intelligence. This module moves beyond textual surface to immersive analysis. Students engage with dramatic theory, socio-linguistics, and character psychology to build rich, contextualized "worlds" for performance. Techniques include granular script analysis, research into historical and social contexts, and exercises in adapting performance for radically different scenarios (e.g., the same speech for a courtroom vs. a confessional).

The second pillar, the embodied instrument and emotional archaeology, turns attention inward. Drawing from methodologies in phenomenology, somatics, and aspects of method acting, students learn to connect emotional truth to physical expression. Practices may include sensory and memory work, breath and movement studies, and biofeedback training to create direct links between internal state and vocal output, building a personal, accessible reservoir of authentic expression.

The third pillar is critical engagement with synthetic media, which becomes essential in an AI-mediated creative environment. Students are asked not only to use AI tools, but to listen to them critically and work with them deliberately. This involves learning how to identify markers of the affective gap in AI-generated speech, how to design prompts and refine outputs iteratively, and how to address questions of authorship, bias, and labour in AI-assisted production. A central goal is to develop competence in what might be called AI voice direction: the ability to analyze synthetic output, intervene in it, and shape it toward more contextually appropriate and ethically considered forms.

The fourth pillar, collaborative creation and professional agency, prepares students for a field in which voice work increasingly intersects with other creative and technical domains. Interdisciplinary collaboration, simulated client interaction, and project-based work with areas such as game design, immersive media, or virtual production can all be part of this training. Just as importantly, students are encouraged to articulate the distinctive value of their own artistic practice in a changing labour market. Resilience, entrepreneurship, and collaborative leadership are therefore treated not as peripheral skills, but as central elements of professional formation.

6.3. Integrative Pedagogy and Technology as an Augmenting Tool

These four pillars are not intended to function as isolated curriculum blocks. Their value lies in how they intersect within a studio-based and project-driven pedagogy. Interpretation, embodiment, critique, and collaboration should develop in relation to one another, rather than as separate competencies taught in parallel.

Within this model, educational technology is integrated strategically and selectively. One important use is immersive simulation. Virtual and augmented reality, for example, can place students inside the kinds of narrative and spatial contexts they are being asked to interpret, allowing them to experience how environment shapes voice, intention, and audience relation. Performing a Shakespearean soliloquy in a virtual Globe Theatre is pedagogically different from merely imagining the setting on the page.

Technology can also support more precise forms of analysis and visualization. Software that displays vocal features such as pitch, timing, and intensity gives students a way of examining performance with greater analytical clarity. This does not reduce expression to data. Rather, it offers a vocabulary for comparing human and synthetic performances and for identifying where nuance, tension, or emotional credibility may be gained or lost.

A further use lies in treating AI tools as a collaborative sandbox. Instead of positioning them as engines of final output, the framework treats them as spaces for experimentation. Students can use synthetic voice systems to test alternative interpretations, explore tonal variation, and probe the limits of machinic expression. In doing so, they are not surrendering artistic agency. They are refining it. The pedagogical value of the tool lies precisely in how it sharpens the student's own judgment as a director, interpreter, and critical maker.

6.4. Assessment Redesign: Valuing Process, Critique, and Curation

Traditional evaluation in voice training has often privileged the finished performance as the primary object of judgment. That model is too narrow for the present purpose. What now matters is not only how well students perform, but how well they interpret, reflect, compare, curate, and make creative decisions within hybrid environments.

One useful format is the critical curation portfolio. Such a portfolio might include a student's own performance of a piece, an AI-generated rendering of the same material, and a comparative critical commentary that identifies the differences between them. The task is not simply to declare one version better than the other, but to analyze where the affective gap appears, how it functions, and what kinds of directorial intervention might improve the synthetic output.

A second format is the directorial brief or project pitch. Documents and presentations where students conceive a project (e.g., an audio drama, a character for an immersive game) that strategically integrates human and AI voices, justifying every creative and ethical choice.

A third component is the reflective practice journal. Because the framework values embodied learning and critical self-awareness, students need space to document their evolving process. Journals can capture how they work through interpretive problems, how they respond to synthetic media, and how their understanding of their own artistic development changes over time. In this sense, reflection is not an optional supplement to performance. It is part of the evidence of learning itself. These differences between traditional and critical approaches to assessment are summarized in Table 2.

Table 2. Comparison of Traditional vs. Critical Assessment Methods.

Aspect	Traditional Assessment	Critical Pedagogy Assessment
Focus	Technical accuracy, imitation	Interpretation, authenticity, AI critique
Format	Solo performance grading	Portfolio, director's brief, reflective journal
Tech Role	Recording tool	Analytical & co-creative platform

6.5. The Metaverse as an Ultimate Test Case

The urgency and relevance of this framework are crystallized in the emerging paradigm of immersive virtual environments, often described as the metaverse. In such spaces, intelligent agents are expected to produce dialogue that is real-time, dynamic, and highly sensitive to context. Under these conditions, the affective gap becomes a central obstacle to believability and social presence (Oh et al., 2018). Voices must do more than sound fluent; they must convey layered emotions, respond credibly to unpredictable interaction, and sustain a sense of relational authenticity.

This future does not render the human voice artist obsolete. On the contrary, it makes the critical voice artist increasingly important. As synthetic expression becomes more pervasive, the human professional is likely to take on a more complex role as designer, director, and curator of believable vocal performance. The proposed framework is intended to prepare students for precisely this kind of work, equipping them to shape the emotional and interpretive texture of hybrid human-AI communicative environments.

The metaverse therefore functions not only as a futuristic application scenario, but also as a demanding test case for the pedagogical value of Critical Voice Artistry. At the same time, if the framework is to move beyond conceptual promise, it must also be examined in relation to the practical conditions of implementation and the ways its educational effectiveness might be evaluated.

6.6. Implementation Considerations and Future Evaluation

The proposed framework responds conceptually to the affective gap in AI-generated speech, but its educational significance will ultimately depend on how it can be enacted in practice. This requires attention not only to pedagogical ideals, but also to the institutional conditions that make those ideals workable.

One immediate challenge concerns the role of the teacher. Voice artistry educators are now being asked to do more than preserve established traditions of performance training. They must also help students engage critically with synthetic voice, AI-mediated production processes, and increasingly hybrid creative environments. That shift cannot be assumed. It calls for sustained professional development in AI literacy, digital pedagogy, and interdisciplinary curriculum design, so that instructors are equipped to guide reflective and ethically informed practice rather than simply demonstrate technique.

Implementation is also shaped by material conditions. The framework presumes access to a range of technologies, including voice synthesis tools, recording and editing platforms, nashi immersive simulation environments, and interfaces that support comparison between human and synthetic performance. For many institutions, such access is uneven. Software costs, hardware requirements, and ongoing technical support may all become practical constraints. In this sense, the question is not merely whether the framework is pedagogically desirable, but whether institutions can support it in ways that are sustainable over time.

A further issue lies in how the framework is introduced into the curriculum. Its value will not come from appending a few isolated AI-related activities to otherwise unchanged courses. What matters is deeper integration. Courses in dubbing, narration, vocal performance, or voice direction may need to be redesigned so that students are trained not only to perform well, but also to interpret context, evaluate synthetic outputs, and make deliberate artistic and ethical decisions within hybrid workflows. The difficulty, then, is one of coherence: technological engagement should extend and enrich artistic formation, not fragment it.

These practical considerations also point to the importance of future evaluation. As a conceptual and theoretical paper, this study does not claim to validate the framework empirically. Rather, it offers a foundation for further testing, adaptation, and refinement. One promising direction would be qualitative case studies of courses

in voice acting, dubbing, audiobook narration, or AI-assisted voice direction. Classroom observation, student reflection, and interviews with instructors could illuminate how the framework operates under real teaching conditions. At the same time, quasi-experimental or mixed-methods designs could compare courses adopting this model with those following more traditional approaches. Such studies could examine changes in students' contextual interpretation, embodied expressivity, critical AI literacy, ethical judgment, and sense of professional agency within human-AI creative ecologies. Evidence of this kind would be essential not only for assessing the framework's effectiveness, but also for identifying its limits and improving its future application.

7. Conclusions

The "affective gap" in AI-generated speech is not simply a technical shortcoming that will disappear with better models. It points to a deeper boundary between computational reproduction and human experience. AI can learn patterns, mimic vocal textures, and generate increasingly persuasive outputs. Yet it still operates through indirect and decontextualized data. Human expression does not. It emerges from embodied perception, lived experience, social context, and intentional meaning-making.

The consequences of this gap are already visible in the structure of professional training and cultural production. Through the analytical lens of the "Pyramid Model", we have seen how the affective gap drives an asymmetric disruption of the voice-acting industry. By automating the entry-level, pattern-based work that historically formed the foundation of professional apprenticeship, AI does not merely displace jobs—it actively dismantles the traditional pipeline for cultivating artistic mastery, threatening a future crisis of talent succession. At the same time, the growing normalization of affectively flattened synthetic media risks contributing to a broader aesthetic deskilling, gradually reducing public sensitivity to vocal nuance and expressive complexity.

Against this backdrop, the key educational question is not how to help students outperform AI at imitation. That is the wrong contest. A more meaningful response is to reconsider what voice artistry education should now prioritize. What follows from this, educationally, is a shift in priorities. The proposed framework for Critical Voice Artistry is built on the premise that education must deliberately cultivate those capacities that remain distinctly human: hermeneutic depth, embodied emotional intelligence, socio-cultural discernment, and the critical ability to analyze, direct, and ethically work with synthetic media.

Seen in this light, the future professional is not only a performer, but also an interpreter, curator, collaborator, and creative decision-maker within hybrid human-AI environments. This becomes especially clear in emerging spaces such as immersive media and the metaverse, where synthetic dialogue must be not only intelligible, but context-sensitive, emotionally credible, and socially meaningful. Here, human value lies less in repetition and more in judgment: the ability to shape tone, intention, and expressive direction in ways machines cannot autonomously determine.

The affective gap therefore should not be read only as a limit of AI. It is also a clarifying concept for education. It helps identify what must be protected, deepened, and reimagined in the training of future practitioners. Voice artistry education will remain relevant not by retreating from technological change, nor by surrendering to it, but by redefining its core around the human capacities that synthetic systems still cannot possess. From this perspective, AI does not simply pose a challenge; it compels the field to clarify what expressive education is for, and why it continues to matter.

Author Contributions

Conceptualization, J.Y. (Jialin Ye) and J.Y. (Jingying Yang); Methodology, J.Y. (Jialin Ye) and J.Y. (Jingying Yang); Formal Analysis, J.Y. (Jialin Ye); Investigation, J.Y. (Jialin Ye); Data Curation, J.Y. (Jialin Ye); Visualization, J.Y. (Jialin Ye); Writing—Original Draft Preparation, J.Y. (Jialin Ye); Resources, J.Y. (Jingying Yang); Supervision, J.Y. (Jingying Yang); Project Administration, J.Y. (Jialin Ye); Validation, J.Y. (Jingying Yang); Writing—Review & Editing, J.Y. (Jingying Yang). All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable. This study is a conceptual/theoretical paper and does not involve human participants.

Informed Consent Statement

Not applicable.

Data Availability Statement

No new data were created or analyzed in this study. This article is a conceptual and theoretical paper based on the analysis and synthesis of existing literature. Data sharing is therefore not applicable to this study.

Acknowledgments

The authors are grateful to the editors and the anonymous reviewers for their constructive comments.

Conflicts of Interest

The authors declare no conflict of interest.

References

Bakhshi, H., Benedikt Frey, H., & Osborne, M. (2015). *Creativity vs robots: The creative economy and the future of employment*. Nesta. https://media.nesta.org.uk/documents/creativity_vs._robots_wv.pdf

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Boden, M. A. (2003). *The creative mind: Myths and mechanisms*. Routledge. https://doi.org/10.4324/9780203508527

Boden, M. A. (2009). Computer models of creativity. *AI Magazine*, *30*(3), 23–34. https://doi.org/10.1609/AIMAG.V30I3.2254

Brade, S., Anderson, S., Kumar, R., Jin, Z., & Truong, A. (2025). SpeakEasy: Enhancing text-to-speech interactions for expressive content creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 756, pp. 1–19). Association for Computing Machinery. https://doi.org/10.1145/3706598.3714263

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., & Ponti, M. A. (2022). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 2709–2720). PMLR. https://proceedings.mlr.press/v162/casanova22a.html

Chan, C., & Kuang, J. (2025). *Toward objective and interpretable prosody evaluation in text-to-speech: A linguistically motivated approach.* arXiv preprint. https://doi.org/10.48550/arXiv.2511.02104

China Academy of Information and Communications Technology. (2023). *Artificial intelligence-generated content (AIGC) white paper*. https://interpret.csis.org/translations/artificial-intelligence-generated-content-aigc-white-paper-excerpt/

Cho, D. H., Oh, H. S., Kim, S. B., & Lee, S. W. (2025). EmoSphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Transactions on Affective Computing*, *16*(3), 2365–2380. https://doi.org/10.1109/TAFFC.2025.3561267

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., & Maneewongvatana, S. (2009). Encoding emotions in speech with the size code: A perceptual investigation. *Phonetica*, *65*(4), 210–230. https://doi.org/10.1159/000192793

Ciccarelli, D., & Ciccarelli, S. (2013). *Voice acting for dummies*. John Wiley & Sons.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Clark, W. L. (1996). *Being there: Putting brain, body, and world together again*. The MIT Press. https://doi.org/10.7551/mitpress/1552.001.0001

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Harcourt Brace.

Dreyfus, H. L. (2014). *Skillful coping: Essays on the phenomenology of everyday perception and action.* Oxford University Press. https://doi.org/10.1093/ACPROF:OSO/9780199654703.001.0001

du Sautoy, M. (2019). Can AI ever be truly creative? *New Scientist*, *242*(3229), 38–41. https://doi.org/10.1016/S0262-4079(19)30840-1

Fischer, E. (1999). *The necessity of art: A Marxist approach*. Verso.

Fuchs, T. (2017). *Ecology of the brain: The phenomenology and biology of the embodied mind*. Oxford University Press. https://doi.org/10.1093/MED/9780199646883.001.0001

Gendlin, E. T. (1997). *Experiencing and the creation of meaning: A philosophical and psychological approach to the subjective*. Northwestern University Press.

Giroux, H. A. (2020). *On critical pedagogy*. Bloomsbury Academic. https://doi.org/10.5040/9781350145016

Goffman, E. (1981). *Forms of talk*. University of Pennsylvania Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Horkheimer, M., Adorno, T. W., & Noeri, G. (2002). *Dialectic of enlightenment*. Stanford University Press.

Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 752–762). Association for Computational Linguistics (ACL). https://doi.org/10.3115/v1/p15-1073

Huang, K., Tu, Q., Fan, L., Yang, C., Zhang, D., Li, S., Fei, Z., Cheng, Q., & Qiu, X. (2025). *InstructTTSEval: Benchmarking complex natural-language instruction following in text-to-speech systems.* arXiv preprint. https://doi.org/10.48550/arXiv.2506.16381

Irvine, D. R. F. (2018). Plasticity in the auditory system. *Hearing Research*, *362*, 61–73. https://doi.org/10.1016/j.heares.2017.10.011

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814. https://doi.org/10.1037/0033-2909.129.5.770

Kassabian, A. (2016). *Ubiquitous listening: Affect*, *attention*, *and distributed subjectivity*. University of California Press. https://doi.org/10.1525/CALIFORNIA/9780520275157.001.0001

Łajszczak, M., Cámbara, G., Li, Y., Beyhan, F., Van Korlaar, A., Yang, F., Joly, A., Martín-Cortinas, Á., Abbas, A., & Michalski, A. (2024). *BASE TTS: Lessons from building a billion-parameter text-to-speech model on 100k hours of data*. arXiv preprint. https://doi.org/10.48550/arXiv.2402.08093

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815355

Oh, C. S., Bailenson, J. N., & Welch, G. F. (2018). A systematic review of social presence: Definition, antecedents, and implications. *Frontiers Robotics AI*, *5*, 114. https://doi.org/10.3389/FROBT.2018.00114

OpenAI. (2024, March 29). *Navigating the challenges and opportunities of synthetic voices: Insights from voice engine*. https://openai.com/index/navigating-the-challenges-and-opportunities-of-synthetic-voices/

Porges, S. W. (2011). *The polyvagal theory: Neurophysiological foundations of emotions*, *attachment*, *communication*, *and self-regulation*. W. W. Norton & Company.

Que, S., & Ragni, A. (2025). *VisualSpeech: Enhance prosody with visual context in TTS.* arXiv preprint. https://doi.org/10.48550/arXiv.2501.19258

Ross, A., Corley, M., & Lai, C. (2024). Is there an uncanny valley for speech? Investigating listeners' evaluations of realistic synthesised voices. In Y. Chen, A. Chen, & A. Arvaniti (Eds.), *Speech prosody 2024* (pp. 1115–1119). International Speech Communication Association.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, *23*(7), 1307–1351. https://doi.org/10.1080/02699930902928969

Selwyn, N. (2022). *Education and technology: Key issues and debates* (4th ed.). Bloomsbury Academic. https://doi.org/10.5040/9781350145573

Sennett, R. (2008). *The craftsman*. Yale University Press.

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). *A survey on neural speech synthesis*. arXiv preprint. https://doi.org/10.48550/arXiv.2106.15561

Varela, F. J., Rosch, E., & Thompson, E. (1991). *The embodied mind: Cognitive science and human experience*. The MIT Press. https://doi.org/10.7551/MITPRESS/6730.001.0001

Woodruff, A., Shelby, R., Kelley, P. G., Rousso-Schindler, S., Smith-Loud, J., & Wilcox, L. (2024). How knowledge workers think generative AI will (not) transform their industries. *In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 641, pp. 1-26). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642700

Xiong, F., Yu, X., Wai, H., & Ma, A. (2025). AI-driven research ecosystem: Unifying human-AI collaboration models and new research thinking paradigms. *Journal of Educational Technology and Innovation*, *7*(1), 39–53. https://doi.org/10.61414/n0n76c97

Yasuda, A., & Maruyama, Y. (2025). Creativity in the age of generative AI. *AI and Ethics*, *6*(1), 46. https://doi.org/10.1007/s43681-025-00848-9

Zhou, K., Zhang, Y., Zhao, S., Wang, H., Pan, Z., Ng, D., Zhang, C., Ni, C., Ma, Y., & Nguyen, T. H. (2024). *Emotional dimension control in language model-based text-to-speech: Spanning a broad spectrum of human emotions*. arXiv preprint. https://doi.org/10.48550/arXiv.2409.16681