CREATIUM

Research

Human or Artificial:
Comparing Perceptions of AI and Human Voices

Dr. Kelly Puzio
Director of Research, Creatium

Abstract

As artificial intelligence (AI) becomes increasingly prevalent, understanding how humans perceive and evaluate AI voices is crucial. This study investigated three dimensions of AI voice perception: technical quality, emotional expressiveness, and the ability to distinguish between AI voices and human voices. Two online experiments were conducted with college-educated U.S.-based participants ($N_1 = 60$, $N_2 = 120$) who evaluated audio clips generated by AI or human beings. Using a mixed design, participants rated the quality, emotional expressiveness, and human-likeness of each voice on a 7-point scale. Results revealed a compelling paradox: while participants could readily distinguish between AI and human voices, they consistently rated AI voices as superior in both technical quality and emotional expressiveness. In Study 1, AI voices received significantly higher ratings for voice quality ($M = 5.54$ vs. 4.64, $p < .001$) and emotional expression ($M = 4.83$ vs. 3.97, $p < .001$), with medium effect sizes. Study 2 replicated these findings with a larger sample, though with smaller effect sizes. Both studies showed that participants could accurately identify human voices, rating them significantly more human-like than AI voices. These findings challenge traditional assumptions about the relationship between naturalness and effectiveness in synthetic speech, suggesting that recognizing a voice as artificial does not impair its perceived quality or emotional impact. This research has important implications for the development and deployment of AI voices across educational, commercial, and interactive media.

Background

The perception and evaluation of artificially generated (AI) voices represent a critical area of investigation as synthetic speech becomes increasingly prevalent in daily life. Recent neuroscientific evidence reveals a fascinating paradox that directly informs the present study: while humans may struggle to consciously differentiate between AI and human voices, their brains may process these voices through distinct neural pathways (Fischer, Niebuhr, Jensen, & Bodenhagen, 2019; Skjegstad & Frühholz, 2024). This finding aligns with our investigation's focus on how listeners evaluate voice quality while simultaneously attempting to distinguish between human and artificial sources.

The technical evolution of AI voice synthesis has achieved unprecedented levels of naturalness and emotional expressivity, raising important questions about how listeners perceive and evaluate these voices. While earlier research highlighted synthetic speech limitations (e.g., Foster, 2007), recent studies suggest that state-of-the-art AI voices can achieve performance comparable to or superior to human voices in certain contexts (Moriuchi, 2021). Our study builds on this work by directly comparing perceptions of voice quality and emotional expressiveness between AI and human voices, while also examining listeners' ability to identify the voice source.

Research in human-computer interaction has demonstrated that voice quality significantly influences trust formation and social presence, with more natural-sounding AI voices generating higher levels of both compared to obviously synthetic voices (Craig & Schroeder, 2017; Kim, Merrill, Xu, & Kelly, 2022). These findings inform our investigation's focus on voice quality assessment and suggest potential explanations for why listeners might rate AI voices differently from human voices, even when they can identify them as artificial.

Research in advertising and consumer behavior indicates that while consumers can often distinguish between human and AI voices, this recognition does not necessarily diminish effectiveness. In fact, AI voices sometimes achieve superior results in message recall and persuasion (Rodero, 2017). This paradox—where artificial voices can outperform human voices despite being identifiable as synthetic—directly parallels our study's investigation of voice quality and emotional expressiveness ratings.

The phenomenon of anthropomorphism provides a theoretical framework for understanding our study's findings. Research shows that users often attribute human-like qualities to AI voices while consciously recognizing their artificial nature (Underwood, 2017). This dual processing aligns with the Computers Are Social Actors (CASA) paradigm, which explains why humans might respond positively to AI voices while maintaining awareness of their artificial nature (Edwards et al., 2019). These theoretical approaches help contextualize our investigation of how listeners can simultaneously recognize AI voices as artificial while rating them highly on quality and emotional expressiveness.

The present study extends this body of research by examining three critical dimensions: voice quality perception, emotional expressiveness evaluation, and human-AI voice discrimination. By investigating these aspects simultaneously, we aim to better understand the complex relationship between technical voice quality, emotional communication, and source recognition in the context of AI-generated speech.

Research Questions

This investigation examined how listeners perceive and distinguish between AI-generated and human voices. Specifically, we investigated three primary questions: How do listeners evaluate the quality of AI-generated voices compared to human voices? How do they perceive the emotional expressiveness of AI versus human voices? And can listeners accurately distinguish between AI-generated and human voices? Each participant rated multiple audio clips on three dimensions: voice quality, emotional expressiveness, and their perception of whether the voice was human or artificial. This design allowed us to examine both the technical and perceptual aspects of voice evaluation.

Research Methods – Study 1

The online survey was distributed to 120 Prolific participants via SurveyMonkey, targeting U.S.-based native English speakers aged 21-29 with college degrees. Each participant was asked to evaluate four short (30-second) audio clips (one each of humor, science, technology, and travel content). For each content type, participants were randomly assigned to hear either an AI-generated or human-voiced version of the clip. Due to a limitation in the SurveyMonkey platform, the source of the audio file (AI vs. human) could only be tracked when participants provided written text responses in a non-required field adjacent to each audio clip. Of the 120 initial participants, only 60 provided trackable data, yielding 236 audio file ratings for analysis (some ratings were missing due to incomplete responses). This constraint, while reducing our usable sample size, ensured that our analysis included only verified audio file ratings where we could definitively determine whether participants were evaluating AI-generated or human voices. The final

dataset included 103 ratings of AI-generated voices and 133 ratings of human voices across the four content categories.

Table 1
Descriptive Statistics, by Condition

Measure	AI Voices	Human Voices	Effect Size (r)
Voice Quality	5.54 (1.56)	4.64 (1.68)	0.28
Emotional Expressiveness	4.83 (1.57)	3.97 (1.69)	0.26
Human-likeness	4.09 (2.04)	5.70 (1.40)	0.39

Note. The total *N* was 236, with 103 AI voice ratings and 133 human voice ratings.

A Mann-Whitney U test was conducted to examine differences between AI and human voices across three dimensions ($N = 236$; AI Voices $n = 103$, Human Voices $n = 133$). AI voices were rated significantly higher than human voices on both voice quality (AI Voices: $M = 5.54$, $SD = 1.56$; Human Voices: $M = 4.64$, $SD = 1.68$; $U = 4660.00$, $p < .001$, $r = 0.28$) and emotional expression (AI: $M = 4.83$, $SD = 1.57$; Human: $M = 3.97$, $SD = 1.69$; $U = 4771.00$, $p < .001$, $r = 0.26$), with medium effect sizes. However, when participants were asked to rate how human-like the voices sounded, human voices received significantly higher ratings than AI voices (Human: $M = 5.70$, $SD = 1.40$; AI: $M = 4.09$, $SD = 2.04$; $U = 3840.00$, $p < .001$, $r = 0.39$), with a medium-to-large effect size.

Research Methods – Study 2

The online survey was distributed to 120 Prolific participants via SurveyMonkey, targeting U.S.-based native English speakers aged 21-29 with college degrees. Each participant was asked to evaluate four short (30-second) audio clips (one each of humor, science, technology, and travel content). For each content type, participants were randomly assigned to hear either an AI-generated or human-voiced version of the clip. Learning from the technical limitations encountered in Study 1, we modified the survey platform configuration to ensure reliable tracking of audio source data for all participants. All 120 participants completed the survey successfully, yielding 473 usable audio file ratings for analysis (a small number of individual ratings were missing due to participants skipping occasional items). The final dataset included 248 ratings of AI-generated voices and 225 ratings of human voices across the four content categories, providing a substantially larger and more complete sample than Study 1.

Table 2
Descriptive Statistics, by Condition

Measure	AI Voices	Human Voices	Effect Size (r)
Voice Quality	5.48 (1.34)	5.08 (1.58)	0.12
Emotional Expressiveness	4.46 (1.49)	4.14 (1.54)	0.10
Human-likeness	3.84 (2.14)	5.64 (1.51)	0.41

Note. The total *N* was 473, with 248 AI voice ratings and 225 human voice ratings.

A Mann-Whitney U test was conducted to examine differences between AI and human voices in a larger sample ($N = 473$; AI Voices $n = 248$, Human Voices $n = 225$). The analysis revealed that AI voices received slightly, but significantly, higher ratings than human voices for both voice quality (AI Voices: $M = 5.48$, $SD = 1.34$; Human Voices: $M = 5.08$, $SD = 1.58$; $U = 24094.00$, $p = .010$, $r = 0.12$) and emotional expression (AI Voices: $M = 4.46$, $SD = 1.49$; Human Voices: $M = 4.14$, $SD = 1.54$; $U = 24808.50$, $p = .03$, $r = 0.10$), though both effect sizes were small. For the human-likeness measure, human voices were rated significantly more human-like than AI voices (Human Voices: $M = 5.64$, $SD = 1.51$; AI Voices: $M = 3.84$, $SD = 2.14$; $U = 14867.50$, $p < .001$, $r = 0.41$), with a medium-to-large effect size.

This larger study confirms the pattern found in Study 1, with some notable differences in magnitude. While AI voices still outperformed human voices in both technical quality and emotional expressiveness, the advantages were more modest in this larger sample (about 0.4 points difference on a 7-point scale, compared to nearly 1 point difference in Study 1). However, the ability of listeners to distinguish between AI and human voices remained strong and consistent across both studies, with human voices receiving substantially higher ratings on the human-likeness measure. This suggests that while AI voices maintain a slight edge in technical aspects, their artificial nature remains clearly detectable to listeners, a finding that held true even with a larger and more diverse sample of ratings.

Results

Across two studies with different sample sizes and participant engagement levels, a compelling pattern emerged in how listeners perceive AI-generated versus human voices. Both studies demonstrated that even when participants were aware they might be hearing AI voices, they still rated AI voices as superior in both technical quality and emotional expressiveness, though the magnitude of this advantage varied notably between studies.

These findings reveal an interesting paradox at the heart of AI voice perception: even though participants could readily distinguish between AI and human voices (consistently rating human voices as more human-sounding), they nevertheless judged the AI voices to be of higher quality and better at expressing emotions. This suggests that while current AI technology has not fully mastered sounding completely human-like, it has already surpassed human voices in terms of perceived technical and emotional performance. The consistency of this pattern across both studies, despite their different methodological approaches, indicates that AI's superior performance is not just a statistical fluke, but represents a meaningful advancement in voice technology.

The magnitude of AI's advantage varied between studies in important ways. In Study 1 ($N = 236$), where participants demonstrated higher engagement by completing optional text responses, AI voices showed superiority on both quality (AI Voices: $M = 5.54$ vs. Human Voices: $M = 4.64$) and emotional expressiveness (AI Voices: $M = 4.83$ vs. Human Voices: $M = 3.97$), with medium effect sizes ($r = 0.28$ and 0.26 respectively). Study 2 ($N = 473$) replicated these findings but with smaller effect sizes ($r = 0.12$ and 0.10) and more modest differences in ratings (0.4 points), possibly reflecting increased noise in the data from a less selectively engaged participant pool.

However, both studies revealed a robust and consistent finding regarding human detection: participants could readily distinguish between AI and human voices, as evidenced by substantially higher human-like ratings for actual human voices. This effect was strong and consistent across both studies (Study 1: $r = 0.39$; Study 2: $r = 0.41$), with human voices receiving two points higher on the human-likeness measure. These results suggest that while current AI technology has achieved superior technical performance in voice generation, its artificial nature remains distinctly recognizable to listeners. The stronger effect sizes in Study 1 suggest an intriguing direction for future research: investigating whether the more robust effects could be replicated with more engaged or attentive participants, as the difference in effect sizes between studies might reflect the impact of participant engagement on rating precision rather than true differences in voice perception.

Limitations

Several methodological constraints should be considered when interpreting these findings. Our sample was restricted to college-educated U.S.-based native English speakers aged 21-29, limiting generalizability across different age groups, educational backgrounds, and cultural contexts. The study's online format, while allowing for larger sample sizes, prevented control over participants' listening environments and equipment quality, potentially introducing variability in voice quality assessments. Furthermore, our study's focus on short (30-second) audio clips across four content categories may not fully capture the complexities of AI voice perception in real-world applications where longer exposure is common. The variation in participant engagement levels between studies, as evidenced by the stronger effect sizes in Study 1's more engaged participants, suggests that attention levels may significantly influence voice perception and evaluation. Future research should address these limitations by examining more diverse populations, implementing controlled listening conditions, and investigating perceptions of AI voices across longer exposure periods and broader content domains.

Discussion

Our findings reveal a compelling paradox in AI voice perception that extends current theoretical understanding in several important ways. First, the consistent superior ratings of AI voices for both technical quality and emotional expressiveness, despite clear recognition of their artificial nature, challenges traditional assumptions about the relationship between naturalness and effectiveness in synthetic speech. This finding aligns with recent neuroscientific research showing distinct processing pathways for AI and human voices (Skjegstad & Frühholz, 2024), while suggesting that conscious recognition of artificial sources does not necessarily impair performance evaluations.

The results also extend the CASA paradigm (Edwards et al., 2019) by demonstrating that humans can simultaneously maintain awareness of a voice's artificial nature while rating it more favorably than human voices. This suggests a more nuanced relationship between anthropomorphism and performance evaluation than previously theorized. The consistent pattern across both studies, despite different sample sizes and engagement levels, indicates a robust phenomenon rather than a statistical artifact.

The variation in effect sizes between studies provides valuable insights about participant engagement and valid assessment of voice perception. The stronger effects were observed in Study 1's self-selected sample; that is, those who took time to provide optional written feedback likely represent more ecologically valid findings, as these participants mirror real-world users who actively engage with voice content. While traditional methodological perspectives might view this self-selection as a limitation, we argue it actually enhanced external validity by capturing participants who approach voice evaluation similarly to more natural contexts. This finding has important implications for both research methodology and practical applications. In research, it suggests that screening for engaged participants, rather than maximizing sample size, may provide more accurate insights into voice perception. For applications like educational content or long-form media, it indicates that AI voices can effectively serve audiences who are naturally motivated to engage with the content, rather than those participating purely for monetary compensation or with minimal attention. The smaller effects in Study 2 likely reflect dilution from participants approaching the task with minimal engagement, rather than representing true population effects.

Future research should investigate whether these findings extend to different demographic groups, particularly across age ranges and cultural contexts (e.g., Zellou, Cohn, & Ferenc Segedin, 2021). Additionally, examining how these perceptions might change with longer exposure to AI voices could provide valuable insights for applications requiring extended interaction. Finally, investigating the specific technical characteristics that contribute to AI voices' superior ratings could inform the development of more effective synthetic speech systems.

These findings have significant practical implications for voice technology deployment across various sectors. The ability of AI voices to outperform human voices in perceived quality and emotional expressiveness, even while being recognized as artificial, suggests that developers should focus on optimizing these performance aspects rather than solely pursuing perfect human mimicry. This insight could particularly benefit applications in education, advertising, and human-computer interaction where voice quality and emotional expressiveness are crucial for effectiveness.

References

Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers & Education*, 114, 193-205.

Edwards, C., Edwards, A., Stoll, B., Lin, X., & Massey, N. (2019). Evaluations of an artificial intelligence instructor's voice: Social identity theory in human-robot interactions. *Computers in Human Behavior*, 90, 357-362.

Fischer, K., Niebuhr, O., Jensen, L. C., & Bodenhagen, L. (2019). Speech melody matters—How robots profit from using charismatic speech. *ACM Transactions on Human-Robot Interaction*, 7(1), 1-21.

Foster, M. E. (2007). Enhancing human–computer interaction with embodied conversational agents. In C. Stephanidis (Ed.), *Universal access in human-computer interaction: Ambient interaction*

(Lecture Notes in Computer Science, Vol. 4555, pp. 828–837). Springer.

Kim, J., Merrill, K., Jr., Xu, K., & Kelly, S. (2022). Perceived credibility of an AI instructor in online education: The role of social presence and voice features. *Computers in Human Behavior*, 134, 107305.

Moriuchi, E. (2021). An empirical study on anthropomorphism and engagement with disembodied AIs and consumers' re-use behavior. *Psychology & Marketing*, 38(1), 21-42.

Rodero, E. (2017). Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices. *Computers in Human Behavior*, 77, 336-346.

Skjegstad, C., & Frühholz, S. (2024). Neural dynamics of processing natural and digital emotional vocalizations [Conference Presentation]. *Federation of European Neuroscience Societies (FENS) Forum* 2024, Vienna, Austria.

Underwood, J. (2017). Exploring AI language assistants with primary EFL students. In K. Borthwick, L. Bradley & S. Thouësny (Eds), *CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp. 317-321). Research-publishing.net.

Zellou, G., Cohn, M., & Ferenc Segedin, B. (2021). Age- and gender-related differences in speech alignment toward humans and voice-AI. *Frontiers in Communication*, 6, 1-14.

Author Note

Kelly Puzio, PhD, is the Director of Research at Creatium. Dr. Puzio is a learning scientist whose work focuses on personalized learning and artificial intelligence. He earned a PhD in Learning, Teaching, and Diversity from Vanderbilt University, a master's degree in education from DePaul University, and a bachelor's degree from the University of Notre Dame. He has been a certified teacher in New Zealand and the United States and brings experience as a faculty member in the Department of Teaching and Learning at Washington State University. Dr. Puzio has served as Principal Investigator on more than $5 million in funding from the National Institutes of Health and the National Science Foundation, and he has been recognized as a Young Scholar by the International Literacy Association.