When Machines Speak with Feeling: Investigating Emotional Prosody, Authenticity, and Trust in AI vs. Human Voices

Guangrui Fan (fgr@tyust.edu.cn)
College of Computer Science and Technology, Taiyuan University of Science and Technology,
WaLiu Road, WanBaiLin, Taiyuan, 030024, Shanxi, China

Dandan Liu (s2134717@siswa.um.edu.my)
Department of Media and Communication Studies, Faculty of Arts and Social Sciences, Universiti Malaya,
50603 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia

Abstract

Emotional prosody—vocal cues that convey affect—profoundly influences how listeners interpret a speaker's intentions. We conducted two studies comparing AI- and human-generated emotional speech. In Study 1 ($N = 38$), participants categorized five emotions (happy, sad, angry, neutral, fear) expressed by human voices and by an advanced text-to-speech (TTS) system. Human recordings exhibited higher overall accuracy (79.82% vs. 72.65%) and were rated significantly more natural, an effect partially explained by micro-perturbations (e.g., jitter, shimmer) that enhanced perceived authenticity. In Study 2 ($N = 53$), these validated stimuli were incorporated into short scenarios, with each speaker labeled as either "human" or "AI." Even when participants heard identical clips, those informed that the speaker was human exhibited greater trust and empathy, resulting in higher donation and advice-following rates. Although contemporary TTS systems effectively convey broad affective states, explicit AI labeling reduces perceived credibility and social engagement, underscoring the critical role that preexisting expectations play in human–AI communication.

Keywords: Emotional Prosody; Text-to-Speech (TTS); Authenticity; Trust; Empathy

Introduction

Emotional prosody encompasses the nuanced modifications in intonation, rhythm, stress, and other vocal parameters that enable speakers to convey affective states beyond the literal content of words (Maltezou-Papastylianou, Russo, Wallace, Harmsworth, & Paulmann, 2022; Jungers, Hupp, Rardon, McDonald, & Song, 2024). These vocal cues help listeners interpret a speaker's intent and detect underlying emotions, thereby influencing interpersonal rapport and shaping communication outcomes (Lausen & Hammerschmidt, 2020). For instance, research on atypical prosodic processing underscores the centrality of these cues in social interactions, as difficulties in recognizing or expressing emotional prosody can give rise to miscommunication (Leipold, Abrams, Karraker, Phillips, & Menon, 2023). Similarly, studies with cross-cultural samples illustrate how prosodic variation—such as an "interested-sounding" tone—promotes both accurate emotion detection and enhanced trust or willingness to disclose personal information (Baharin, Lamarche, Weinstein, & Paulmann, 2024). Seemingly minor modulations in pitch, amplitude, and temporal patterns significantly influence perceptions of the speaker's credibility and relational warmth, suggesting a powerful role for prosody in shaping persuasion, conflict resolution, and joint decision-making (Niebuhr, Thumm, & Michalsky, 2018).

Recent advancements in text-to-speech (TTS) technology have propelled synthetic voices far beyond their early monotone incarnations, enabling the generation of sophisticated vocal expressions that convey nuanced emotional prosody (Bott, Lux, & Vu, 2024). Key innovations include neural network architectures that dynamically model prosodic dimensions—such as pitch, duration, and energy—thereby enhancing the naturalness and expressiveness of AI-generated speech (Cho, Oh, Kim, Lee, & Lee, 2024). As a result, systems integrating these features have found broad applicability in voice assistants, service robots, and educational software, marking a shift toward more socially and emotionally aware human–machine interactions (Pande & Mishra, 2024). Widely used digital assistants such as Amazon's Alexa and Google Assistant reflect this progress; nonetheless, the proliferation of quasi-human AI voices raises questions about ethical implications, the fidelity of synthesized affect, and the potential for unintended social consequences (Han, Kelly, Nikou, & Svee, 2022).

In examining how emotional expression is processed and interpreted—whether in human or AI-generated voices—two theoretical perspectives emerge as paramount. The basic emotions model identifies discrete categories such as happiness, sadness, anger, and fear as universally recognizable, whereas dimensional approaches posit that emotions can be mapped along continuous axes such as valence and arousal (Laukka, Juslin, & Bresin, 2005; DE CAROLIS, Ferilli, Palestra, Redavid, et al., 2017). In AI speech synthesis, researchers have applied these frameworks to analyze how listeners decode vocal features as emotional signals (Bott et al., 2024). However, the media equation framework complicates this picture by asserting that individuals often respond to computer-generated stimuli as though they were human, attributing social presence to synthetic voices that convey emotional cues (Lee & Nass, 2005; Nass & Moon, 2000). This ascription of "human-like" qualities, though conducive to engagement, may also arouse skepticism or discomfort if the emotional expression seems inauthentic, paralleling the "uncanny valley" phenomenon (Wang, Lilienfeld, & Rochat, 2015; Chandra, Shirish, & Srivastava, 2022). Within this dynamic, perceived authenticity becomes a crucial determinant of trust. Some studies propose that the disclosure of an AI's artificial nature undermines users' sense of sincerity, whereas others suggest that transparency can alleviate suspicions of

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

manipulation (Glikson & Woolley, 2020).

Despite the clear significance of emotional prosody for user perceptions and behaviors, relatively few empirical investigations have explored how contextual factors—such as varying scenarios or explicit speaker labeling—affect judgments of authenticity, trust, empathy, and compliance. Labeling a voice as "AI," for example, may alter listeners' willingness to rely on or empathize with the speaker, yet the processes that underlie these shifts remain insufficiently understood.

The present research addresses these gaps through two interconnected studies designed to investigate: (a) the accuracy with which listeners detect and label emotional prosody in AI-generated speech relative to human speech, (b) how emotional prosody from AI influences perceptions of authenticity, empathy, and trust, and (c) how contextual elements such as speaker identity disclosure might moderate these relationships. In Study 1, a set of emotional speech stimuli—both AI- and human-generated—will be validated through controlled listening tasks, enabling an assessment of emotion identification accuracy, perceived intensity, and naturalness. Study 2 extends these validated stimuli into more realistic contexts, manipulating speaker identity disclosure and measuring trust, empathy, and behavioral compliance. By comparing outcomes across the two studies, the work aims to offer insights into the interplay between emotional prosody, authenticity, and trust, thereby informing theoretical perspectives on vocal emotion and guiding the ethical and design considerations of emotionally expressive AI.

Study 1

Method

Participants Thirty-eight individuals ($N = 38$; 19 women, 19 men; $M_{\text{age}} = 24.7$, $SD = 4.2$) were recruited from a large public university in China. All reported normal hearing and fluent command of Mandarin Chinese. To minimize confounds related to professional speech training, individuals who self-identified as voice actors or who had received extensive vocal coaching were excluded. Written informed consent was obtained prior to data collection, and the institutional review board of the host university approved all procedures.

Stimuli A set of 20 neutral Mandarin sentences (e.g., "快递预计今天到达。" ["The package will arrive today."]) was selected as the base content. Two trained native Mandarin speakers (one male, one female) recorded each sentence in five targeted emotions: happy, sad, angry, neutral, and fear. Recordings were made in a sound-attenuated studio using a professional-grade microphone (44.1 kHz sampling rate). Acoustic checks (e.g., pitch contour, amplitude) ensured that the intended emotional tone was reflected in each take.

Each of the same 20 Mandarin sentences and five emotions was synthesized using Seed-TTS, a versatile high-quality speech generation model based on an autoregressive Transformer architecture (Anastassiou et al., 2024). Seed-TTS includes a speech tokenizer, token language model, token diffusion model, and acoustic vocoder modules, making it particularly efficient and precise for Chinese speech synthesis. A pilot script introduced variations in punctuation and interjections to elicit emotive delivery. Final files were output as MP3 at 24 kHz and amplitude-normalized for consistency. All AI-generated stimuli included a disclosure that they were AI voices, though participants were not explicitly told which individual clips were AI-generated vs. human.

To confirm that each emotional category was perceptually distinct, all files underwent acoustic analysis for key prosodic features (mean $F0$, $F0$ range, duration, amplitude (Owren & Bachorowski, 2007)). Six pilot participants verified that each category was reasonably distinguishable, yielding a final pool of 200 total files (20 sentences $\times$ 5 emotions $\times$ 2 speaker types). Given potential time constraints in the main experiment, participants were randomly assigned a subset (20) of these stimuli to ensure each session remained under 30 minutes.

Procedure Data collection took place individually in a quiet laboratory setting. After consenting, participants were seated in front of a desktop computer equipped with over-ear headphones. They received instructions to listen carefully to each utterance and to rate what they heard on several dimensions. Each audio clip was presented once in random order, followed immediately by on-screen rating prompts, which includes: *Emotion Classification* (Forced-choice labeling of the perceived emotion (happy, sad, angry, neutral, fear)); *Confidence Rating* (A 5-point scale (1 = not at all confident, 5 = extremely confident) reflecting how certain participants were about their classification.) *Perceived Emotional Intensity* (A 7-point scale (1 = very low intensity, 7 = very high intensity) assessing how strongly the emotion was expressed) and *Naturalness* (A 7-point scale (1 = highly synthetic, 7 = very natural or human-like)).

Data Analysis Recognition accuracy was evaluated using confusion matrices (participant's chosen emotion vs. the intended emotion). To examine potential differences across speaker type (human vs. AI) and emotion category, mixed-effects models were employed, specifying Speaker Type and Emotion as fixed effects and Participant as a random intercept. Dependent variables included accuracy, intensity, and naturalness ratings; confidence was analyzed similarly to test whether listeners were more or less certain in their emotional judgments for AI vs. human speech. Finally, acoustic features were regressed against perceived intensity and accuracy scores to determine which prosodic parameters best predicted successful emotion recognition.

Results

As shown in Table 1, human-recorded stimuli yielded higher overall recognition accuracy (M = 79.82%, SD = 40.15%) compared to AI-generated stimuli (M = 72.65%, SD = 44.60%), $t(1898) = 3.672$, $p < .001$, Cohen's $d = 0.169$. When broken down by specific emotions, human-recorded

Table 1: Overall Recognition Accuracy (%)

Metric	AI	Human
Mean Accuracy	72.65	79.82
Standard Deviation	44.60	40.15
t-test	3.672	
p-value	0.0002	
Cohen's d	0.169	

speech led to higher accuracy in every category. Particularly pronounced differences arose for fear (61.08% vs. 69.47%) and sad (69.85% vs. 81.87%), suggesting that subtle emotional tones remain more challenging for AI systems to convey convincingly. By contrast, angry and happy approached similar recognition rates in both conditions.

Participants rated the naturalness of each sample. A mixed-effects model confirmed a significant effect of speaker type, $t = 65.039$, $p < .001$. Human speech was perceived as considerably more natural ($M \approx 5.50$) than AI speech ($M \approx 4.00$), across all five emotions. Emotion-specific effects were relatively minor in comparison, although neutral and fear stimuli displayed slightly larger gaps between AI and human recordings. A two-sample t-test indicated slightly higher confidence for human recordings, $t = 2.893$, $p = 0.0039$, though this difference appeared relatively small in practical terms. No significant difference emerged in intensity ratings ($t = -0.238$, $p = 0.812$), suggesting that participants perceived the "strength" of the expressed emotion similarly across AI and human stimuli.

An acoustic analysis compared key parameters (mean $F0$, $F0$-range, intensity, speech rate, jitter, shimmer, and harmonics-to-noise ratio). Notably, jitter and shimmer showed strong positive correlations with naturalness ($r \approx 0.73$ and $r \approx 0.76$, respectively, both $p < .001$). These findings imply that minor perturbations typically present in human speech may enhance the perceived realism of synthesized audio. Pitch range also emerged as an important predictor of intensity ratings ($r \approx 0.39$), underscoring the value of dynamic prosodic variation in successful emotion portrayal.

Discussion

Study 1 revealed that while participants generally recognized emotional prosody in both AI- and human-generated speech, human recordings achieved consistently higher accuracy and naturalness ratings. This advantage was particularly pronounced for fear and sadness—emotions known to rely on subtler variations in pitch and timing (DE CAROLIS et al., 2017). The difficulty TTS systems face in capturing these nuances indicates that despite significant advancements, replicating the acoustic complexity of human emotional expression remains a formidable challenge. Notably, listeners rated AI-generated fear and sadness as both less accurate and less natural, highlighting the importance of micro-perturbations (e.g., jitter, shimmer) in conveying authenticity. This aligns with prior research suggesting that a certain "roughness" in human vocal signals can paradoxically enhance perceived realism; in contrast, perfectly "clean" AI outputs can sound mechanical or uncanny (Bott et al., 2024).

These findings underscore both the promise and current limitations of TTS in conveying emotional speech. On one hand, AI voices displayed near-human performance for relatively high-arousal emotions such as anger or happiness, suggesting their readiness for many real-world applications where broad emotional cues suffice. On the other hand, the consistent underperformance on sadness and fear points to the need for more sophisticated modeling that integrates subtle prosodic cues and minor vocal fluctuations. This gap also aligns with dimensional models of emotion, which stress the role of continuously varying valence and arousal in shaping how listeners interpret affect (Jin & Youn, 2023). Moreover, the modest difference in participants' confidence ratings between AI and human voices suggests that while listeners sense something "off," they are not entirely dismissive of synthetic speech. In light of these results, an important question emerges regarding how these perception gaps translate into broader social judgments: If AI voices can convey recognizable emotions but still sound somewhat inauthentic, does this affect how listeners respond in terms of trust, empathy, or willingness to act on the speaker's requests? Study 2 addresses this question by examining user trust and compliance under different identity labels, thus probing whether contextual factors—such as explicitly calling a voice "AI" vs. "human"—exacerbate or diminish the perceptual shortcomings observed in this first study.

Study 2

Method

Participants and Setting Fifty-three adults (N = 53; 25 men, 28 women; M age = 25.9, SD = 3.8) were recruited from the same public university in China involved in Study 1. None of these individuals had participated in the previous study, ensuring that all were naïve to the stimuli. All reported normal hearing and fluent Mandarin proficiency. Ethical clearance was granted by the university's human research ethics committee, and written informed consent was obtained from each participant prior to the onset of the experiment.

Design and Conditions The study employed a 2×4 mixed design, with disclosed identity (AI-labeled vs. Human-labeled) as a between-participant factor and emotion (angry, happy, neutral, sad) as a within-participant factor. Twenty-six participants were told all voice clips were human-recorded (Human-labeled), while twenty-seven were informed that all clips were AI-generated (AI-labeled). In actuality, each participant was exposed to both four AI clips and four human clips. However, this true source (AI vs. human) was only disclosed in the final debriefing to facilitate partial deception regarding the speaker's identity.

Stimuli and Scenarios The stimuli consisted of eight short Mandarin recordings validated in Study 1—two clips each for angry, happy, neutral, and sad. (Because fear recognition was especially challenging in Study 1, we excluded it here to focus on the four most reliably recognized categories.) Four clips were recorded by human voice actors, and four were synthesized using Seed-TTS (Anastassiou et al., 2024). All audio files were amplitude-normalized for consistent loudness.

Each clip was embedded in a brief on-screen scenario, designed to simulate a common conversational context. The specific scenarios and associated behavioral measures were: *Schedule Change (Neutral or Sad)*: The speaker announces a sudden change in timetable and asks for feedback; *Project Feedback (Angry or Happy)*: The speaker critiques or praises a work project, followed by "Would you follow this advice?"; *Personal Advice (Sad or Happy)*: The speaker offers personal guidance, followed by "Would you follow this advice?"; *Donation Request (Neutral or Angry)*: The speaker asks for a contribution to a cause, followed by "How likely are you to donate?"

Randomization Scheme and Distribution To ensure balanced exposure to both human and AI clips within each labeling condition, we adopted the following procedure:

- **Stimulus Pool:** Eight total recordings (4 AI, 4 human). Each emotion (angry, happy, neutral, sad) had two potential clips (one AI, one human).
- **Scenario Assignment:** For each of the four emotions, we created a pair of identical scenario texts. One scenario version was linked to the human recording, and one was linked to the AI recording of the same emotional content.
- **Participant Allocation:** Participants were first randomly assigned to a labeling condition: AI-labeled (n = 27) or Human-labeled (n = 26). Each participant then randomly drew one scenario–clip pairing per emotion (angry, happy, neutral, sad) from the pool. For instance, if a participant drew the AI–angry clip, the other participant in the same label condition might draw the human–angry clip. Over the course of the data collection, this ensured that each participant encountered exactly one clip for each of the four emotions, two of which happened to be AI-based and two of which were human-based (though the exact distribution could vary across individuals to maintain randomization). Crucially, participants in the AI-labeled condition were told (falsely) that all of these clips originated from an AI speaker, while those in the Human-labeled condition were told that all came from a single human speaker. The actual mixture (2 AI + 2 human) was hidden from them.
- By rotating the AI vs. human clip assignment for each emotion among participants, we prevented any single clip from being over-represented in one label condition and preserved a balanced sampling of both AI and human stimuli overall.

Procedure Participants were tested individually in a quiet laboratory setting. Each participant was assigned to one of the two labeling conditions (AI-labeled or Human-labeled) based on a pre-planned rotation. After a short introduction explaining that they would hear "voice messages in Mandarin from one speaker expressing various emotions or making requests," participants encountered four scenario–audio pairs in random order. Each scenario appeared on screen, followed by a single audio clip played through over-ear headphones. Immediately afterwards, participants rated the speaker on trust, empathy, and naturalness. Each item used a 7-point scale. Depending on the scenario, participants also answered a behavioral or compliance-related question using a Likert-type response format. To capture broader attitudes, participants completed a brief Trust Toward AI scale (five items) and the short-form empathy questionnaire (short-form Interpersonal Reactivity Index, five items), which served as a potential moderator for emotion recognition. Following completion of all tasks, the experimenter debriefed each participant, explaining that the voice clips were in fact drawn from both human and Seed-TTS sources, and providing further details regarding the study's aims and hypotheses.

Data Analysis Each participant contributed four within-subject ratings (angry, happy, neutral, sad), while Disclosed Identity (AI-labeled vs. Human-labeled) served as a between-participant factor. Ratings of trust (1–7) and empathy (1–7) were analyzed via mixed-effects linear models specifying participant as a random intercept, with Emotion (angry, happy, neutral, sad) and Disclosed Identity as fixed effects. Post-hoc comparisons used Tukey corrections where needed. Behavioral measures were analysed by nonparametric tests (Mann-Whitney U) that allowed for repeated measures across scenarios.

Results

A mixed-effects linear model examined how disclosed identity (AI-labeled vs. Human-labeled) and emotion (angry, happy, neutral, sad) affected trust (1-7 scale). Participants who believed the speaker was human reported significantly higher trust overall ($b = 1.16$, $p < .001$), as demonstrated in Figure 1. Notably, trust for angry speech exhibited the largest labeling effect (Cohen's $d = 1.365$), whereas the effect was less pronounced for sad (Cohen's $d = 0.733$) and happy (Cohen's $d = 0.572$) clips. The model also revealed small main effects for emotion (e.g., both happy and neutral were rated more trustworthy than angry), though these effects sometimes interacted negatively with labeling, implying a partial attenuation of gains in the Human-labeled condition when the speaker expressed certain emotions.

A parallel mixed-effects model for empathy likewise indicated a substantial main effect of disclosed identity (AI-labeled vs. Human-labeled). Participants in the Human-labeled condition reported higher empathy overall ($b = 1.25$, $p < .001$). The model found that sad speech elicited more empathy than angry speech ($b = 0.397$, $p = 0.009$), although

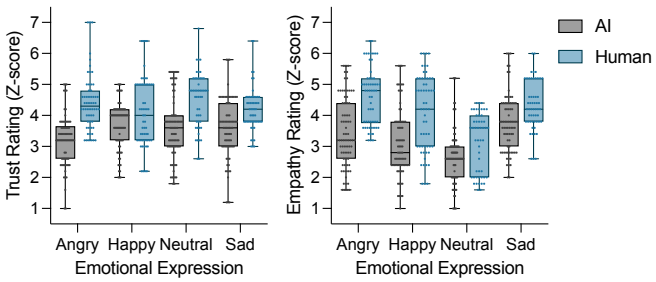

Figure 1: Trust and empathy ratings by speaker identity (AI vs. Human) and emotional expression. Box plots show distribution of participant ratings on 1-7 scales, with individual data points overlaid.

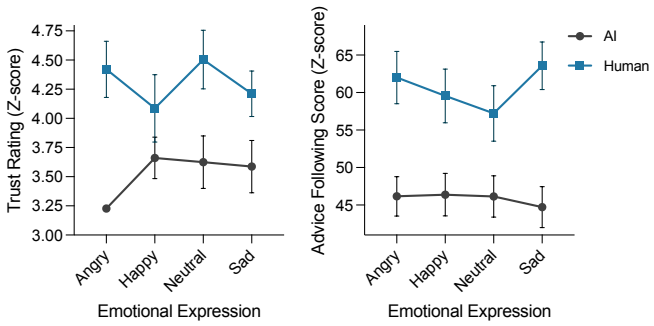

Figure 2: Mean trust ratings (left) and advice following scores (right) as a function of emotional expression and speaker identity. Error bars represent ± 1 standard error of the mean.

certain emotional effects were dampened by labeling interactions (Figure 1). For example, the empathy advantage of sad expressions was partially offset in the Human-labeled group (interaction term: $b = -0.779$, $p = 0.001$). However, the *"happy + human-labeled = lower empathy gain"* interaction emerged as a somewhat surprising result. While one might expect a cheerful human speaker to evoke stronger empathy, the data indicated that happy emotion combined with human labeling actually diminished participants' empathy scores relative to what one might predict from the additive effects alone.

A plausible explanation is that extremely positive or "overly cheerful" tones from a human coworker can be interpreted as inauthentic or even irritating, thus undercutting an empathetic response. By contrast, a happy-sounding AI voice might be perceived as more benign or "programmed," leading participants to respond neutrally or slightly more empathically in the absence of perceived social incongruity. This interplay between content (emotional tone) and context (speaker identity) underscores the complexity of user perceptions in realistic communication settings.

Two behavioral measures were examined: donation intent and advice following (scaled 0–100). A nonparametric Mann-Whitney U test for donation showed a statistically significant difference favoring the Human-labeled condition ($p < .001$, $r = 0.801$). Participants were substantially more willing to donate when they believed the speaker was human. Similarly, a mixed-effects model of advice following indicated a substantial identity effect (b = 14.9, $p < .001$), with listeners in the Human-labeled group assigning an average of 14.93 points more on a 0–100 scale than those in the AI-labeled condition (Figure 2). No significant emotion main effects or interactions emerged for advice following, suggesting that identity labeling accounted for the largest variance in compliance.

Discussion

Study 2 demonstrates that labeling a voice as "human" versus "AI" can decisively influence how listeners perceive and respond to emotionally expressive speech. Despite using the same stimuli for both groups, participants told they were hearing a human speaker reported markedly higher trust, empathy, and willingness to comply. This outcome echoes the results of Study 1—which showed that AI voices, though recognizable across various emotions, were perceived as less natural—by underscoring that social and cognitive biases about a speaker's identity can overshadow purely acoustic cues. Even if AI-generated speech can approximate certain prosodic features convincingly, the label "AI" appears to trigger skepticism or diminished affective engagement.

Interestingly, the effect of emotional tone itself (angry, happy, neutral, sad) was less pronounced than the effect of labeling. Although participants generally empathized more with sad-sounding speakers and found angry ones less trustworthy, label-based expectations had a stronger overall impact on donation intentions, advice following, and subjective ratings of trust. The nuance observed with "happy" prosody—where listeners expressed somewhat lower empathy toward a cheerful human coworker compared to a cheerful AI—suggests that emotional appropriateness may interact with identity expectations in context-specific ways. In other words, a positive emotional display from a purportedly human speaker can be interpreted as artificial or overenthusiastic under certain conditions, aligning with sociolinguistic evidence that excessive positivity can spark suspicion or irritation rather than understanding (Baharin et al., 2024). Such findings highlight the complexity of human–AI interactions, where not only the vocal signal but also the presumed "personhood" of the speaker shapes listeners' reactions.

From an applied standpoint, these findings point to potential strategies for designing and deploying emotionally expressive AI systems. Improved prosodic modeling, as demonstrated by Study 1, may help reduce listeners' sense that something is "missing" in AI voices. However, the robust labeling effect observed here suggests that such technical enhancements must be paired with thoughtful disclosure practices. While concealing AI identity could bolster trust, it raises ethical concerns regarding user autonomy and informed consent. Future work might explore "partial disclosure" approaches—for instance, clarifying that content

is "AI-assisted" rather than fully AI-generated—to preserve transparency while reducing knee-jerk skepticism. Longitudinal studies, in which users repeatedly interact with a consistently empathetic AI-labeled speaker, could further reveal whether these initial biases diminish over time or whether they persist in extended engagements. Taken together, Study 2's results highlight the interplay of prosodic cues and social framing, suggesting that even highly advanced TTS will face challenges establishing rapport if listeners remain anchored in preconceived notions about artificiality.

General discussion

Taken together, our two studies illustrate both the technical strides and persisting limitations of text-to-speech (TTS) technology in conveying emotional prosody, as well as the strong influence of social labeling on listener perceptions. Study 1 showed that while AI-generated voices can approach human performance in portraying anger or happiness (Bott et al., 2024; Chang & Chien, 2024), they struggle to capture subtler emotions such as sadness and fear—particularly due to the absence of micro-perturbations like jitter and shimmer. This gap highlights an "authenticity deficit" in which perfectly smooth synthesis paradoxically sounds artificial.

Study 2 underscored how beliefs about speaker identity can override acoustic realism. Participants labeled as hearing a "human" speaker reported significantly higher trust, empathy, and compliance than those told they were listening to an "AI" voice, even when the exact same recording was used. These results extend the "media equation" framework (Nass, 2005; Zhao, 2006) by showing that labeling can weaken the natural inclination to treat computers as social actors: if listeners are primed to view the speaker as non-human, they may withhold the empathy and cooperation they otherwise exhibit. At the same time, the difficulty AI voices face in emulating subtle affect resonates with the "uncanny valley" concept (Mori, 1970), suggesting that minor—but perceptible—discrepancies from human speech can evoke discomfort or mistrust.

From an ethical standpoint, our findings raise questions about whether and how to disclose a speaker's artificiality. Full transparency safeguards user autonomy and informed consent, ensuring that people know they are interacting with AI. Yet, as shown here, such disclosure can reduce trust and engagement. Conversely, partial or nonexistent disclosure might enhance rapport but risks deceiving users, compromising ethical guidelines and potentially eroding trust if the deception is later revealed. Future studies should examine disclosure strategies that inform users about AI involvement without triggering reflexive skepticism, such as labeling a system as "AI-assisted" rather than entirely synthetic.

A promising avenue for subsequent research is to investigate how repeated interactions might reshape these perceptions. While the "labeling effect" was pronounced in a single-session setting, prolonged exposure to an AI-labeled voice that consistently delivers empathetic and reliable responses could erode initial biases over time. Longitudinal studies or field deployments (e.g., in customer service or healthcare) might illuminate whether repeated, positive encounters gradually offset the negative impact of labeling, thereby revealing the extent to which trust can be built or restored in human–AI communication.

On the practical side, these insights urge developers to refine prosodic algorithms for more subtle emotional states while acknowledging that simply improving acoustic fidelity may not suffice. Transparent yet nuanced disclosure practices—for instance, labeling a system as "AI-assisted"—may help mitigate reflexive mistrust without compromising user autonomy. Ethically, balancing authenticity and transparency becomes paramount. If AI systems that convey rich emotional prosody are deployed without clear disclaimers, users may feel deceived; yet if the label "AI" is too salient, they may be predisposed to doubt the speaker's intentions or sincerity.

Conclusion

In sum, the present research demonstrates that while advanced text-to-speech (TTS) systems can approximate certain discrete emotions—particularly high-arousal states such as anger or happiness—they remain less adept at conveying more subtle affective cues. Acoustic and perceptual data indicate that vocal irregularities (e.g., jitter and shimmer) are pivotal for conveying authenticity, and current synthesis algorithms struggle to reproduce these nuances fully. In addition, the findings from Study 2 underscore how social and cognitive biases can overshadow acoustic information. Even when exposed to identical speech content, participants who believed they were listening to a human reported higher trust, empathy, and compliance than those who were told the speaker was AI-generated.

These outcomes hold significant implications for both theory and practice. From a theoretical perspective, they highlight the interplay between bottom-up auditory signals and top-down expectations in shaping emotion perception and interpersonal judgments. Practically, the results suggest that further refinements in prosodic modeling—especially for low-intensity emotions—must be accompanied by careful consideration of how AI identity is disclosed. Transparency in labeling and opportunities for repeated interaction with AI voices may help mitigate skepticism while still respecting users' need for informed engagement. By addressing both the technical and social dimensions of emotional speech synthesis, researchers and developers can foster AI systems that are not only more natural-sounding but also more likely to be trusted and embraced in real-world communicative contexts.

References

Anastassiou, P., Chen, J., Chen, J., Chen, Y., Chen, Z., Chen, Z., … others (2024). Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Baharin, S. A. S., Lamarche, V., Weinstein, N., & Paulmann, S. (2024). Does my tone of voice affect your disclosure? a cross-cultural comparison of how showing an interest through vocal cues affects listeners'well-being and self-disclosure.

Bott, T., Lux, F., & Vu, N. T. (2024). Controlling emotion in text-to-speech with natural language prompts. *arXiv preprint arXiv:2406.06406*.

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To be or not to be... human? theorizing the role of human-like competencies in conversational artificial intelligence agents. *Journal of Management Information Systems*, *39*(4), 969–1005.

Chang, C.-J., & Chien, W.-C. (2024). Towards a positive thinking about deepfakes: Evaluating the experience of deepfake voices in the emotional and rational scenarios. In *International conference on human-computer interaction* (pp. 311–325).

Cho, D.-H., Oh, H.-S., Kim, S.-B., Lee, S.-H., & Lee, S.-W. (2024). Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. *arXiv preprint arXiv:2406.07803*.

DE CAROLIS, B., Ferilli, S., Palestra, G. C., Redavid, D., et al. (2017). Emotion-recognition from speech-based interaction in aal environment. In *Ceur workshop proceedings* (Vol. 1803, pp. 92–104).

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Han, S., Kelly, E., Nikou, S., & Svee, E.-O. (2022). Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI & SOCIETY*, 1–13.

Jin, S. V., & Youn, S. (2023). Social presence and imagery processing as predictors of chatbot continuance intention in human-ai-interaction. *International Journal of Human–Computer Interaction*, *39*(9), 1874–1886.

Jungers, M. K., Hupp, J. M., Rardon, J. A., McDonald, S. A., & Song, Y. (2024). The effect of emotional prosody and referent characteristics on novel noun learning. *Language and Cognition*, *16*(4), 1881–1898.

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, *19*(5), 633–653.

Lausen, A., & Hammerschmidt, K. (2020). Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications*, *7*(1), 1–17.

Lee, K.-M., & Nass, C. (2005). Social-psychological origins of feelings of presence: Creating social presence with machine-generated voices. *Media Psychology*, *7*(1), 31–45.

Leipold, S., Abrams, D. A., Karraker, S., Phillips, J. M., & Menon, V. (2023). Aberrant emotional prosody circuitry predicts social communication impairments in children with autism. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(5), 531–541.

Maltezou-Papastylianou, C., Russo, R., Wallace, D., Harmsworth, C., & Paulmann, S. (2022). Different stages of emotional prosody processing in healthy ageing–evidence from behavioural responses, erps, tdcs, and trns. *Plos one*, *17*(7), e0270934.

Mori, M. (1970). The uncanny valley: the original essay by masahiro mori. *Ieee Spectrum*, *6*(1), 6.

Nass, C. (2005). *Wired for speech: How voice activates and advances the human-computer relationship.* MIT press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, *56*(1), 81–103.

Niebuhr, O., Thumm, J., & Michalsky, J. (2018). Shapes and timing in charismatic speech-evidence from sounds and melodies. In *9th international conference on speech prosody 2018* (pp. 582–586).

Owren, M. J., & Bachorowski, J.-A. (2007). Measuring emotion-related vocal acoustics. *Handbook of emotion elicitation and assessment*, 239–266.

Pande, A., & Mishra, D. (2024). Humanoid robot as an educational assistant–insights of speech recognition for online and offline mode of teaching. *Behaviour & Information Technology*, 1–18.

Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology*, *19*(4), 393–407.

Zhao, S. (2006). Humanoid social robots as a medium of communication. *New Media & Society*, *8*(3), 401–419.