Computers in Human Behavior Reports 22 (2026) 101142

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Does speech prosody shape social perception equally for AI and human voices? A 16-attribute rating study

Wenjun Chen [a,b], Marc D. Pell [b], Xiaoming Jiang [a,c,*]

[a] Institute of Language Sciences, Shanghai International Studies University, Shanghai, 201620, China
[b] School of Communication Sciences and Disorders, McGill University, Montréal, H3G 1A8, Canada
[c] Key Laboratory of Language Science and Multilingual Artificial Intelligence, Shanghai International Studies University, Shanghai, 201620, China

ARTICLE INFO

Keywords:
Social perception
Voice cloning
Group categorization
Prosody
Voice assistant
Affective speech

ABSTRACT

AI-generated speech can now replicate humanlike prosodic patterns, yet whether these cues shape social perception equivalently to human voices remains unknown. We used voice cloning to closely match speaker identity and transfer prosodic style across human and AI voices. We had 40 native Mandarin Chinese speakers rate 320 utterances (half human-produced) in two prosodic styles, confident (greater F0 variability, and faster speech rate) and doubtful (higher mean F0), across 16 perceptual attributes spanning acoustic properties and social impressions. Listeners rated human and AI voices differently: human voices received higher ratings on such attributes as humanlikeness, animateness, and emotional richness, whereas AI voices scored higher on speaking rate and nasality. Confident prosody shifted ratings relative to doubtful prosody across all 15 attributes for both sources, with AI voices showing greater gains on most of these. Principal component analysis revealed two core perceptual dimensions: social appeal (capturing attractiveness, animateness, and pleasantness) and vocal expressiveness (capturing speech rate, loudness, and confidence-related features), with AI voices scoring substantially lower on social appeal. Meanwhile, vocal expressiveness strongly predicted social appeal for human voices, but only weakly for AI voices, and AI-human differences in social appeal widened with increasing expressiveness levels. As participants were not informed of the voice source, our findings suggest that in-group/out-group categorization of human vs. AI voice source implicitly constrains how listeners process short-term prosodic cues in speech. We propose that commercialized AI voices should remain perceptibly identifiable as AI-generated, and that highly humanlike prosody should be used with caution in interpersonal contexts.

1. Introduction

As AI-generated speech becomes increasingly integral to modern digital life, from virtual assistants to audiobooks and navigation systems (Bérubé et al., 2024), questions arise about whether it conveys prosodic patterns as effectively as human speech. Human speech is inherently expressive, with prosodic patterns carrying distinct pragmatic meanings that profoundly influence social impressions (Jiang & Pell, 2024; Larrouy-Maestri et al., 2025; Pinheiro, 2025). For example, the same statement can elicit different levels of listener trust depending on the tone of voice, or prosody: a confident delivery (firm and assertive, as if stating a fact) vs. a doubtful delivery (hesitant and question-like, as if uncertain about the content) (Jiang & Pell, 2017). Klüber et al. (2025) recently proposed that prosodic elements can significantly enhance the attribution of human-like experience to artificial agents, raising questions about whether established prosodic processes operate equivalently in AI-generated voices.

To address this question, the present study employs voice cloning technology to closely match speaker identity across human and AI voices (Lavan et al., 2025; Rosi et al., 2025), while systematically varying prosodic style between confident and doubtful expressions (Chen et al., 2026b), allowing us to isolate the effect of voice source on prosodic perception. Through a 16-attribute perceptual rating study, we investigate whether AI-generated prosody achieves functional equivalence to human vocal expression.

1.1. AI voice disadvantage stems from reduced prosodic expressiveness and categorical expectations

Available evidence suggests that differential perceptions between AI

* Corresponding author. 1550 Wenxiang Road, Shanghai 201620, China.
E-mail address: xiaoming.jiang@shisu.edu.cn (X. Jiang).

https://doi.org/10.1016/j.chbr.2026.101142
Received 17 December 2025; Received in revised form 26 May 2026; Accepted 29 May 2026
Available online 1 June 2026

and human voices may reflect both the prosodic disadvantages of AI-generated speech and listeners' categorical expectations regarding artificial vs. human sources.

On the one hand, compared to human voices, AI voices are perceived as less emotionally expressive (Cohn et al., 2020) and more monotonous in vocal quality (Kühne et al., 2020), indicating a prosodic expressiveness disadvantage in AI voices. These perceptual disadvantages may explain why, in applied contexts such as narrative advertising, human voices led to higher perceived communicative effectiveness, greater attention, and better recall compared to synthetic voices (Rodero, 2017; Rodero & Lucas, 2023). Likewise, Kühne et al. (2020) demonstrated that human voices were rated more favorably than synthetic voices across multiple attributes, including intelligibility, naturalness of prosody, trustworthiness, confidence, enthusiasm, pleasantness, likability, appealingness, credibility, and human-likeness. Similar human voice advantages have been widely reported in terms of trust (Becker et al., 2025; Noah et al., 2021), attractiveness (Bruder et al., 2025), and evaluative ratings across multiple social attributes (Stern et al., 2006), with a systematic review of voice in human-agent interaction further corroborating this pattern (Seaborn et al., 2021).

On the other hand, as Stern et al. (2006) suggested, knowing that a voice is artificial may reduce listeners' preference for human over synthetic voices, suggesting that categorical source knowledge (Li et al., 2021), not acoustics alone, shapes these perceptions. This categorization does not, however, uniformly disadvantage AI voices. While listeners predominantly prefer human voices, analogous to in-group preferences for native-accented over foreign-accented speakers (Domínguez-Arriola et al., 2025; Jiang et al., 2018; Lam et al., 2025; Mauchand & Pell, 2022; Paladino & Mazzurega, 2020), AI voices can nonetheless be favored in specific scenarios: functional tasks (Im et al., 2023), technical contexts requiring consistency (Abdulrahman & Richards, 2022), and long-form content (Cambre et al., 2020). This latter pattern reflects AI voices' utilitarian advantages, where task efficiency and consistency are prioritized over social engagement (Gong & Lai, 2003; Zellou & Cohn, 2020).

However, disentangling the relative contributions of prosodic disadvantages (Cohn et al., 2020; Klüber et al., 2025; Kühne et al., 2020) and categorical expectations (Li et al., 2021; Stern et al., 2006) to these perceptual differences remains challenging, as both factors likely operate simultaneously in natural listening conditions.

1.2. Neural evidence suggests underlying acoustic differences between human and AI voices

Current neuroimaging evidence suggests that this in-group/out-group distinction (Li et al., 2021; Stern et al., 2006) is rooted in differential social-emotional processing of AI vs. human voices in the brain. Tamura et al. (2015) found that the left posterior insula, involved in auditory feature processing and directing attention to salient social-emotional events (Koch et al., 2020; Zhang et al., 2022), responds more strongly to human than artificial voices. In addition, the left dorso-central insula selectively activates for human speech that conveys both vocal affect and communicative intent, but neither for robotic voices nor for non-social human speech, reflecting its role in integrating affective speech quality with communicative intention (Di Cesare et al., 2022). Recently, Roswandowitz et al. (2024) found that the nucleus accumbens, a key node in the mesolimbic reward pathway for processing social vocal signals (Abrams et al., 2016), shows evolved sensitivity to authentic human voices while AI voices trigger compensatory auditory cortex activation.

Note that the engagement of the above-listed brain regions beyond the primary auditory cortex does not imply that acoustic input is irrelevant. Rather, it indicates that voice perception may involve hierarchical processing where acoustic features are integrated with higher-order categorization processes such as social identity detection and communicative intent recognition (Andics et al., 2010; Brück et al., 2011; Jiang & Pell, 2024; Latinus et al., 2013; Pinheiro, 2025). Collectively, these findings suggest that listeners' brain responses to human and AI voices diverge across multiple regions involved in social-emotional processing.

Such differential brain responses imply that underlying acoustic differences may drive the distinct social-emotional processing of human vs. AI voices. Computational studies have identified mel-frequency cepstral coefficients (MFCC), a composite measure capturing spectral envelope, formant structure, and coarticulation patterns, as a critical marker distinguishing human from AI voices (Bisogni et al., 2024). Other studies have reported more explainable acoustic differences: AI voices often lack micro-perturbations such as jitter and shimmer that characterize natural voice quality (Fan & Liu, 2025; Rozsypal & Millar, 1979), exhibit flatter pitch contours with reduced F0 variation (Kim et al., 2025), and show more uniform temporal patterns with less natural rhythm (Yoon et al., 2026).

Chen et al. (2026a) recently leveraged the temporal resolution of electroencephalography (EEG) to characterize the rapid neural distinction between human and AI voices. In that study, 40 listeners completed a speaker identity learning task while neural responses were recorded, using a subset of 192 stimuli (human vs. AI voices × confident vs. doubtful prosody), with participants' attention directed toward speaker identity rather than voice source or prosodic features. Multivariate pattern analysis (MVPA) revealed that neural discrimination of human vs. AI voice source emerged rapidly (~200 ms post-onset), substantially preceding prosodic decoding (emerging only after 1300 ms).

These acoustic differences (e.g., MFCC; Bisogni et al., 2024) may thus underpin both the rapid perceptual discrimination and the differential social-emotional brain responses to human vs. AI voices. The present study employs the same voice conditions and AI cloning service as Chen et al. (2026a), assuming that listeners can implicitly detect voice source differences based on acoustic features alone. Accordingly, participants were not informed of the voice source during the experiment.

1.3. Listeners partially apply human voice perception patterns to AI voices

Despite the above perceptual differences and categorical processing patterns, available evidence also suggests that AI and human speech perception share substantial latent dimensional structures underlying social-cognitive and perceptual processing. Y. Li et al. (2025), based on a study of 2187 participants, suggest that warmth and competence are the two most prominent factors affecting trust in AI, which corresponds with the general social impression formation framework of warmth (intentional benevolence) and competence (capability to enact intentions) (Cuddy et al., 2008). Likewise, just as valence and dominance have been identified as the two core dimensions underlying social impressions formed from human voices (McAleer et al., 2014), these same two dimensions emerge when rating synthetic voices on 17 social traits (Shiramizu et al., 2022).

Beyond these broad dimensional similarities, more specific perceptual parallels have also been documented across human and synthetic voices. These include findings that listeners recognized personality cues in computer-synthesized speech and demonstrated similarity-attraction effects parallel to human voices (Nass & Lee, 2000), that emotional prosody conveyed through synthesized and recorded voices similarly influenced content perceptions (Nass et al., 2001), that gender-based power perceptions and stereotyping operated comparably across voice types (Mullennix et al., 2003), that accent-based preferences mirrored human voice perception patterns (Tamagawa et al., 2011), that participants showed comparable emotional alignment toward human and voice-AI interlocutors (Cohn et al., 2020), and that acoustic features affected trustworthiness ratings equivalently across voice types (Maltezou-Papastylianou et al., 2025).

These convergent findings suggest that listeners apply established perceptual processes from human voice research when evaluating synthetic voices, consistent with the Computers Are Social Actors (CASA)

framework, which posits that people automatically and unconsciously exhibit social responses to technology in ways similar to human–human interaction (Reeves & Nass, 1996). However, the explanatory power of CASA is not without limitations. Recent studies suggest that these automatic and unconscious social responses may have weakened as people have accumulated experience with technology (Heyselaar, 2023), consistent with the view that social responses to technological artifacts are not automatic or universal but emerge contextually through lived experience (Kaptelinin & Dalli, 2025).

Namely, contemporary research challenges the original assumption of automatic and unconscious social responding: rather than unconsciously applying human-human interaction rules, people develop distinct human-media scripts, specialized norms and expectations tailored specifically to AI agents as a social category (Gambino et al., 2020). Evidence shows that people exhibit reduced sensitivity to face-threatening criticism from AI compared to humans (Elsinger & Hörner, 2024), display faster erosion and lower overall politeness toward AI agents (Lazebnik et al., 2025), experience less apprehension about AI-delivered negative feedback due to face-saving opportunities (Pei et al., 2024), rate chatbots lower on benevolence despite polite language (Brummernhenrich et al., 2025), and use profanity toward chatbots that would be inappropriate in human interaction (Park et al., 2021).

Against this backdrop, while listeners apply shared perceptual patterns when processing human and AI voices at the individual level, categorical distinctions between voice sources may substantially constrain how these patterns operate. The present study, therefore, asks whether prosodic patterns that modulate social impressions in human voices, specifically the contrast between confident and doubtful prosody (Jiang et al., 2020), transfer equivalently to AI-generated speech, or whether categorical voice source distinctions limit their effectiveness.

1.4. Voice cloning and style transfer: the present study's approach

Unlike human speech, which has a biological basis in the vocal apparatus (Ghazanfar & Rendall, 2008), artificially generated speech relies on engineered systems. Current methods include formant synthesis, concatenative synthesis, statistical parametric synthesis, and neural text-to-speech (TTS), with neural TTS representing the current state-of-the-art (Kaur & Singh, 2023; Muhamad et al., 2024; Tan et al., 2021). Formant synthesis uses rule-based algorithms to simulate vocal tract acoustics, producing intelligible but robotic-sounding output (Klatt, 1982; Schroeter, 2008). Concatenative synthesis stitches together pre-recorded speech segments, achieving natural sound but requiring large databases and limiting prosodic flexibility (Hunt & Black, 1996). Statistical parametric synthesis trains statistical models to predict acoustic parameters, enabling flexible manipulation but producing over-smoothed output (Ling et al., 2015; Zen et al., 2009). Neural TTS employs deep learning to learn end-to-end text-to-speech mappings, with recent neural voice cloning systems replicating specific speaker identities from limited samples while maintaining prosodic control (Arık et al., 2018; Arık et al., 2017; Hinterleitner, 2017; Kumar et al., 2023; Li et al., 2019; Tan et al., 2021).

Despite the widespread adoption of neural TTS in commercial voice assistants nowadays, most studies examining perceptual differences between human and AI voices employ pre-configured commercial voice assistants with fixed, non-customizable speaker identities (Cohn et al., 2020; Fortunati et al., 2022; Shiramizu et al., 2022). This design potentially confounds voice source with speaker identity, as observed differences could stem from artificiality, from comparing different individuals, or from both. To address this confound, voice cloning technology can be employed to hold speaker identity constant (Lavan et al., 2025; Rosi et al., 2025). The present study employs a similar commercially available voice technology but with customizable speaker identity through neural voice cloning. In essence, this allows us to obtain the original human version of what would otherwise be a fixed AI assistant voice.

Contemporary neural voice cloning systems are capable of preserving not only speaker identity but also prosodic style from the input recordings, enabling systematic manipulation of prosodic characteristics while maintaining speaker identity (Du et al., 2021; Ji et al., 2025; Li et al., 2024; Y. A. Li et al., 2025; Neekhara et al., 2021). In the present study, we exploited this capability by recording human audio samples in two prosodic styles (confident vs. doubtful) and inputting these recordings to the voice cloning system to generate corresponding AI voices.

We selected confident and doubtful prosodies because previous research has established their acoustic and perceptual distinctiveness (Jiang & Pell, 2017), with confident expressions characterized by a higher F0 range and greater amplitude, and doubtful expressions by higher mean F0, slower speaking rate, and more pauses. Validation studies confirmed that cloning speaker identity successfully preserved the intended prosodic characteristics: listeners rated confident AI voices significantly higher on confidence than doubtful voices, and acoustic analyses revealed corresponding systematic F0 differences (Chen et al., 2026b; Chen & Jiang, 2023), demonstrating the viability of this approach for replicating confident and doubtful prosody in AI-generated speech. Still, our approach represents an attempt to maximally match rather than fully replicate speaker identity. Some residual acoustic differences will always remain between any two utterances of the same content, even when they are not perceptible to listeners.

1.5. The present study

Throughout this paper, "AI voices" refers to AI-cloned voices generated from human recordings; voice cloning was employed as a methodological tool to hold speaker identity constant, ensuring comparability between human and AI voices while isolating the effect of prosodic variation. We use the "AI voices" or "AI-generated" throughout to maintain focus on the human vs. AI comparison as the primary object of inquiry, rather than on the cloning process itself.

The present study examined whether prosodic patterns shaping social impressions of human voices transfer equivalently to AI-generated speech, using a 2 (voice source: human vs. AI) × 2 (prosody: confident vs. doubtful) design with 16 perceptual attributes spanning acoustic properties and social impressions. These attributes span those with documented human advantages, those established in early synthesis research, and those examined individually in prior work but not yet studied together within a single comprehensive rating study.

The tested attributes include speech quality-related cues, such as loudness, pitch/squeakiness, speaking rate, nasality, monotonicity (Mullennix et al., 2003), accent (Tamagawa et al., 2011), and naturalness (Kuriki et al., 2016; Nussbaum et al., 2025). Explicit social inferences that require holistic evaluation of the voice were collected, including emotion richness, animateness (Kuriki et al., 2016), confidence level (Eyssel & Hegel, 2012; Jiang & Pell, 2017), attractiveness (Brave et al., 2005; Bruder et al., 2025; Sundar et al., 2017), pleasantness (Dou et al., 2021), human-likeness (Kuriki et al., 2016), dominance (Law et al., 2021; Shiramizu et al., 2022), trustworthiness (Law et al., 2021), and friendliness (Dou et al., 2021). See Table S1 for the original Chinese scale items and their English translations. Our rating experiment addressed two research questions.

- RQ1: How do human and AI voices differ in listeners' perceptual ratings across 16 attributes?
- RQ2: Does confident vs. doubtful prosody shape social perception equivalently for human and AI voices?

2. Methods

2.1. Participants

Forty-five native Mandarin Chinese speakers participated in the rating study, among whom data from five participants were excluded due to incomplete participation, resulting in 40 valid participant data sets (20 males: $M = 24.6$, $SD = 2.09$ years; 20 females: $M = 22.3$, $SD = 2.54$ years). The mean years of education were $M = 18.6$ ($SD = 1.79$) for males and $M = 18.2$ ($SD = 2.59$) for females. None of the participants reported having a speech or hearing impairment or a psychiatric history. Participants provided written informed consent prior to participation and were compensated at 50 RMB/hour. The study was approved by the Ethics Committee of the Institute of Language Sciences at Shanghai International Studies University.

2.2. Stimuli

Eight native standard Mandarin university students (4 males: M age $= 22.5 \pm 3.4$ years, M education $= 19.0 \pm 3.4$ years; 4 females: M age $= 22.2 \pm 1.9$ years, M education $= 19.2 \pm 1.3$ years) with Putonghua Proficiency Test scores above 87/100 and backgrounds in acting, speech, or music training served as speakers.

Each speaker produced 30 utterances (15 geographical, 15 trivia statements) in three prosodic conditions (neutral, doubtful, confident) using the established elicitation paradigm (Jiang & Pell, 2017). As trivia statements were particularly suited for eliciting doubtful prosody, these 15 trivia sentences were input to Huawei's *Xiaoyi* (Celia) voice cloning service to generate AI speakers replicating each human speaker's identity across three prosodic conditions; each AI speaker then produced all 30 sentences. This input size (15 utterances per speaker) follows the requirement set by the voice cloning service, which is calibrated by the provider to ensure reliable speaker-identity reproduction.

For the present study, five geographical statements ($M = 12.2 \pm 0.84$ characters) and five trivia statements ($M = 11.0 \pm 1.0$ characters) were selected from the original pool based on three criteria: absence of internal commas, moderate sentence length, and balanced topic coverage (see Table S2). A total of 320 sentences were used in the present study: 2 confidence levels (confident vs. doubtful) × 2 speaker types (human vs. AI-cloned) × 8 speakers × 10 statements. All audio clips were normalized to −30 dBFS with a sampling rate of 44,100 Hz to ensure comparable acoustic properties across human and AI voices.

The present stimuli were drawn from the corpus reported in Chen and Jiang (2023). That study confirmed that confident and doubtful prosody were effectively replicated by the AI voice cloning system.

2.3. Experiment procedures

Participants wore Bose QuietComfort 45 noise-cancelling headphones and sat comfortably in front of a computer screen. Before the experiment began, participants were asked to adjust the volume to a comfortable level. The experiment was programmed and conducted using *PsychoPy* v2022.2.5 (Peirce et al., 2019). The 16 rating scales were divided into four groups of four scales, with each group displayed on a separate display page.

Participants heard an audio clip and saw four scales on a screen, with the option to replay the audio clip by pressing the play icon. Each audio clip was presented across four screens, with questions fully randomized in sequence within each screen across trials. The four screens maintained fixed content and sequence, with earlier screens focusing on

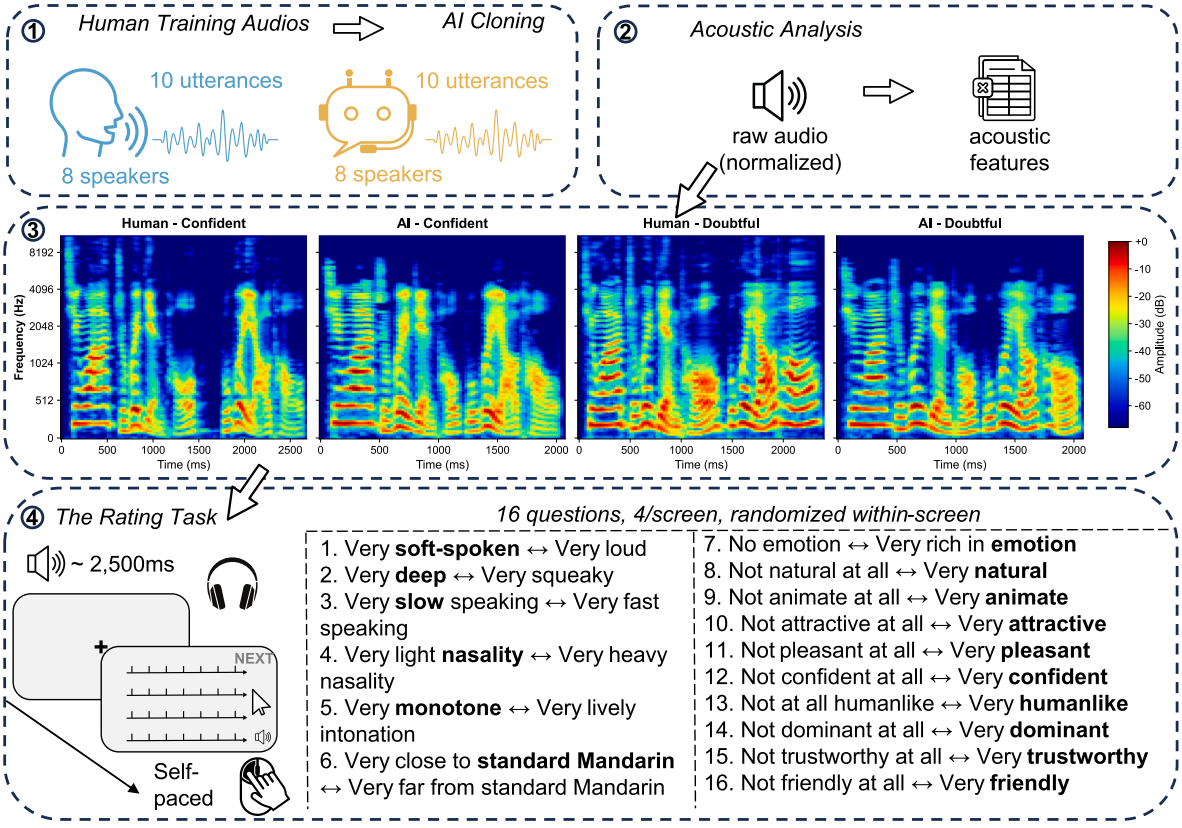

Fig. 1. ***Experimental pipeline.*** **(1)** Eight speakers each produced 10 selected utterances in confident and doubtful prosody; recordings were input to voice cloning service to generate corresponding AI-cloned audio. **(2)** All audio clips were normalized to −30 dBFS, then used for acoustic analysis and participant ratings. **(3)** Representative mel spectrograms are shown for one male speaker producing the same statement across four conditions ("青蛙只会点头不会摇头"[Frogs can only nod their heads but cannot shake them]). Acoustic analyses of the full utterances were conducted to validate prosodic manipulation across human and AI-cloned voices. **(4)** The full set of 320 audio clips was presented to 40 participants in a within-subjects rating task, with each clip evaluated across 16 perceptual attributes.

speech quality attributes and later screens targeting social impression measures. Participants registered their impressions using 7-point Likert scales via mouse clicks (see Fig. 1).

Participants were not informed whether each clip was human or AI-generated, to avoid label-induced biases. Any differentiation between human and AI voices therefore relied on acoustic cues alone (Chen et al., 2026a). The experiment consisted of four blocks (80 trials each), totaling 320 sentence ratings. Participants completed two blocks per laboratory visit, with a recommended 10-min break between blocks (though actual break duration varied based on participant needs). Each block lasted approximately 60-90 min, with variability due to individual pacing. Block sequence was counterbalanced using a Latin square design to control for order effects.

2.4. Data analysis

Acoustic analyses were first conducted to validate prosodic manipulation across human and AI voices (Section 2.4.1). Inter-rater reliability was assessed to ensure data integrity (Section 2.4.3).

To address RQ1, LMMs compared human and AI voice ratings across 16 attributes (Section 2.4.4), with PCA examining the underlying perceptual structure (Section 2.4.6). To address RQ2, voice source × confidence level interactions were tested using LMMs (Section 2.4.5), followed by a post hoc analysis of whether vocal expressiveness moderated AI-human differences in social appeal (Section 2.4.7).

2.4.1. Acoustic analysis

Four acoustic measures were extracted from all 320 audio files using the *parselmouth* library (Jadoul et al., 2018) in Python: F0 mean, F0 range, and F0 SD (all in semitones), and speech rate (characters per second). F0 was estimated using Praat's autocorrelation method with sex-specific pitch ranges (75–300 Hz for males, 100–500 Hz for females; Boersma & Weenink, 2021) and converted to semitones ($12 \times \log_2[\text{Hz}/1]$) to account for the nonlinear frequency-pitch relationship. F0 range and SD were computed across all voiced frames; speech rate was calculated as character count divided by utterance duration. LMMs were fitted in R (v4.4.0) (R-Core-Team, 2024) using *lme4* and lmerTest (Bates et al., 2015; Kuznetsova et al., 2017): *measure* ~ *source* × *confidence* + *(1|speaker)* + *(1|item)*. Bonferroni correction was applied across 12 tests ($\alpha = .0042$).

2.4.2. Perceptual rating data pre-processing

The rating data comprised 12,800 observations (40 participants × 320 utterances, excluding practice trials), with each observation containing ratings across 16 perceptual attributes. After listwise deletion of cases with missing values on any attribute, the final dataset comprised 12,749 observations.

2.4.3. Post-hoc quality assessments

Inter-rater reliability was assessed for all 16 perceptual attributes using two-way mixed intraclass correlation coefficients (ICC(2,1) for single raters and ICC(2,k) for average measures) computed with *psych::ICC()* (Revelle, 2025).

ICC(2,1) values ranged from 0.11 to 0.37 across attributes, indicating moderate single-rater reliability consistent with expected individual variability in perceptual judgments. ICC(2,k) values ranged from 0.83 to 0.96, indicating excellent reliability for the group mean ratings used in subsequent analyses. Human-likeness showed the highest reliability (ICC(2,k) = 0.96), while nasality and accent showed the lowest (ICC(2,k) = 0.83), reflecting their greater subjectivity. These results support the use of mean ratings across participants as reliable estimates of group-level voice perception (Table S3).

A complementary simulation-based power sensitivity analysis (simr) (Green & MacLeod, 2016) further confirmed that all 16 attribute-level linear mixed-effects models achieved simulated power of 1.00 (95% CI lower bound = 0.982) under the Bonferroni-corrected threshold ($\alpha = .003125$). That is, across 200 Monte Carlo simulations per model, every simulated dataset successfully recovered the observed source effect, indicating ample statistical resolution to detect the human–AI differences reported here (see Table S7).

2.4.4. Comparing human and AI voice ratings across 16 attributes

To address RQ1, perceptual differences between human and AI voices were examined across each of the 16 rating attributes using LMMs fitted in R version 4.4.0 (R-Core-Team, 2024) with the lme4 and lmerTest packages (Bates et al., 2015; Kuznetsova et al., 2017). Each model followed the formula: *attribute* ~ *source* + *ai_voice_exposure* + *(1|participant_id)* + *(1|speaker)* + *(1|item)*, where voice source (human vs. AI) was included as a fixed effect, participants' self-reported AI voice exposure frequency ("How frequently do you encounter AI-generated voices in your daily life?", 7-point scale) as a covariate, and random intercepts for participant, speaker, and item to account for the nested data structure. Bonferroni correction was applied with an adjusted significance threshold of $p < .003125$ (0.05/16 attributes).

2.4.5. Voice source × confidence level interaction effects on perceptual ratings

To address RQ2, we examined whether the effect of confidence level on voice perception differed between human and AI voices by fitting LMMs for each of the 15 perceptual attributes (excluding confidence_level) using R version 4.4.0 (R-Core-Team, 2024) with the *lme4* and *lmerTest* packages (Bates et al., 2015; Kuznetsova et al., 2017). Each model followed the formula: *attribute* ~ *source* × *confidence_std* + *ai_voice_exposure* + *(1|participant_id)* + *(1|speaker)* + *(1|item)*, where voice source and standardized confidence ratings (*confidence_std*, derived from participants' 7-point Likert scale ratings of perceived confidence) were included as fixed effects along with their interaction, AI voice exposure frequency as a covariate, and random intercepts for participant, speaker, and item. Conditional slopes for each voice source were derived from model coefficients: the human slope corresponds to the *confidence_std* main effect, and the AI slope to the sum of the *confidence_std* main effect and the interaction term. Bonferroni correction was applied across 45 tests (15 attributes × 3 effects; $\alpha = .001111$).

2.4.6. Principal component analysis of voice perception ratings

To further address RQ1, principal component analysis (PCA) was conducted to examine the dimensional structure underlying voice perception across the 16 rating attributes (McAleer et al., 2014; Yu, 2022), using R's *prcomp()* function on standardized variables (scale = TRUE, center = TRUE). Component extraction employed Kaiser's criterion (eigenvalue >1.0) combined with scree plot examination to determine the optimal number of components to retain. Variables with absolute loadings ≥0.30 were considered meaningful contributors to each component. To respect the within-subjects design, component scores were first averaged per participant per voice source, and group differences between AI and human voices were assessed using paired-samples t-tests with Bonferroni correction ($\alpha = .017$ for three components) and Cohen's *d* effect sizes. Hotelling's T^2 test assessed overall multivariate differences between voice source groups in the principal component space.

2.4.7. Post hoc analysis: vocal expressiveness as a moderator of AI–human differences in social appeal

Building on the PCA results, a post hoc moderation analysis was conducted to examine whether vocal expressiveness (PC2) moderates the relationship between voice source and social appeal (PC1), further addressing RQ2. Although PC1 and PC2 are orthogonal by construction, this analysis examines whether the PC2–PC1 relationship differs as a function of voice source within each group, following the precedent of using PCA-derived component scores in subsequent regression and interaction analyses (McAleer et al., 2014; Yu, 2022). PC2 scores were standardized (M = 0, SD = 1) and voice source was dummy-coded (AI =

1, Human = 0). The moderation model followed the formula: *PC1 ~ source × PC2_standardized + (1|participant_id)*, with a random intercept for participant to account for the within-subjects design. To probe significant interactions, conditional effects analysis was conducted by dividing PC2 into terciles (Low, Medium, High) and calculating AI-Human differences in PC1 within each level using paired-samples t-tests on participant-level means, with Cohen's *d* effect sizes. Simple slopes analysis examined the relationship between PC2 and PC1 separately for AI and human voices using linear mixed-effects models with random intercepts for participants.

3. Results

3.1. Acoustic validation of prosodic manipulation in human and AI voices

Significant confidence main effects were found across all four measures (all *ps* < 0.001, Bonferroni-corrected): relative to confident prosody, doubtful prosody showed higher mean F0 (human $\beta = 1.74$, AI $\beta = 2.69$), narrower F0 range (−4.03, −4.68), smaller F0 SD (−1.09, −1.12), and slower speech rate (−0.56, −0.16), confirming successful prosodic manipulation in both human and AI-cloned voices.

Source × confidence interactions did not survive Bonferroni correction for F0 mean, F0 range, or F0 SD, indicating that the confidence-related F0 changes were largely comparable across human and AI voices; the F0 mean interaction was marginal ($\beta = 0.95$, $p = .00419$, just above the corrected $\alpha = .0042$). The interaction was significant only for speech rate ($\beta = 0.40$, $p < .001$): human voices showed a larger speech-rate difference between prosodic conditions ($\beta = -0.56$) than AI voices ($\beta = -0.16$). See Fig. 2 and Table S4.

3.2. Perceptual advantages of human over AI voices across 16 attributes

The results of LMMs with Bonferroni correction are presented in Fig. 3 and Table S5. Human voices were rated as significantly more human-like ($\beta = 2.30$), animate ($\beta = 1.73$), emotionally rich ($\beta = 1.54$), vivid in vocal expression (monotonicity scale; $\beta = 1.49$), attractive ($\beta = 1.23$), natural ($\beta = 1.15$), pleasant ($\beta = 0.98$), trustworthy ($\beta = 0.87$), friendly ($\beta = 0.77$), dominant ($\beta = 0.74$), and confident ($\beta = 0.74$) than AI voices. Human voices were also perceived as carrying stronger regional accents diverging from standard Mandarin ($\beta = 0.58$), louder ($\beta = 0.37$), and higher in pitch/squeakiness ($\beta = 0.34$).

Conversely, AI voices received significantly higher ratings on nasality ($\beta = -0.65$) and speaking rate ($\beta = -0.66$), indicating that AI voices were perceived as more nasal and faster than human voices.

AI voice exposure frequency ($M = 3.62$, $SD = 1.50$) was included as a covariate but did not significantly predict ratings on any attribute (all *ps* > 0.14), indicating that prior familiarity with AI voices did not moderate the observed effects.

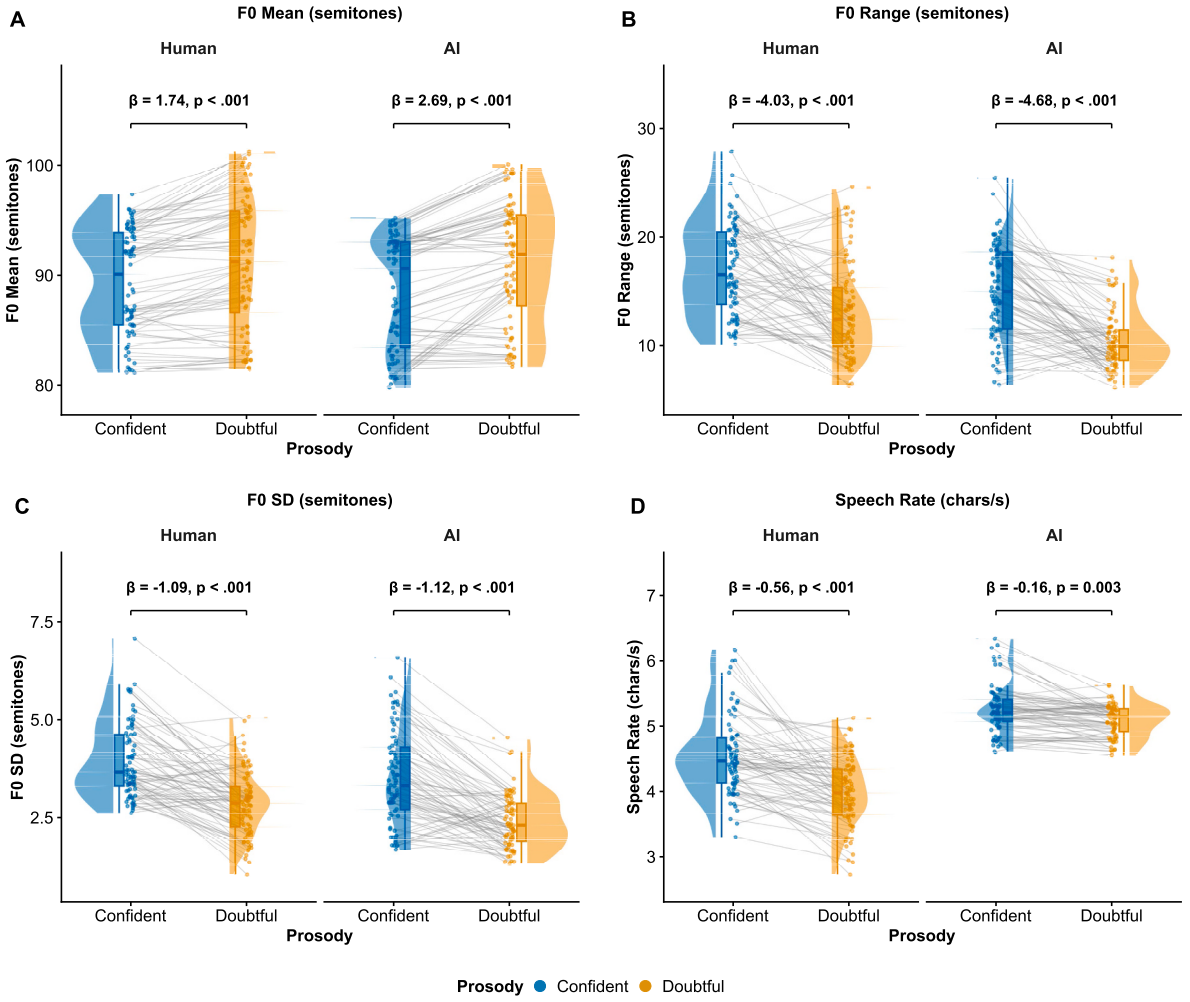

Fig. 2. ***Acoustic differences between confident and doubtful prosody in human and AI-cloned voices.*** Violin plots with individual data points and connecting lines for paired observations (same speaker and item): **(A)** F0 Mean, **(B)** F0 Range, **(C)** F0 SD, **(D)** Speech Rate. Source × confidence interactions were non-significant for A–C, indicating largely comparable prosodic F0 patterns across human and AI-cloned voices, and significant for D ($p < .001$). β = conditional confidence slope (Doubtful − Confident) shown separately for human and AI voices; Bonferroni-corrected $\alpha = .0042$.

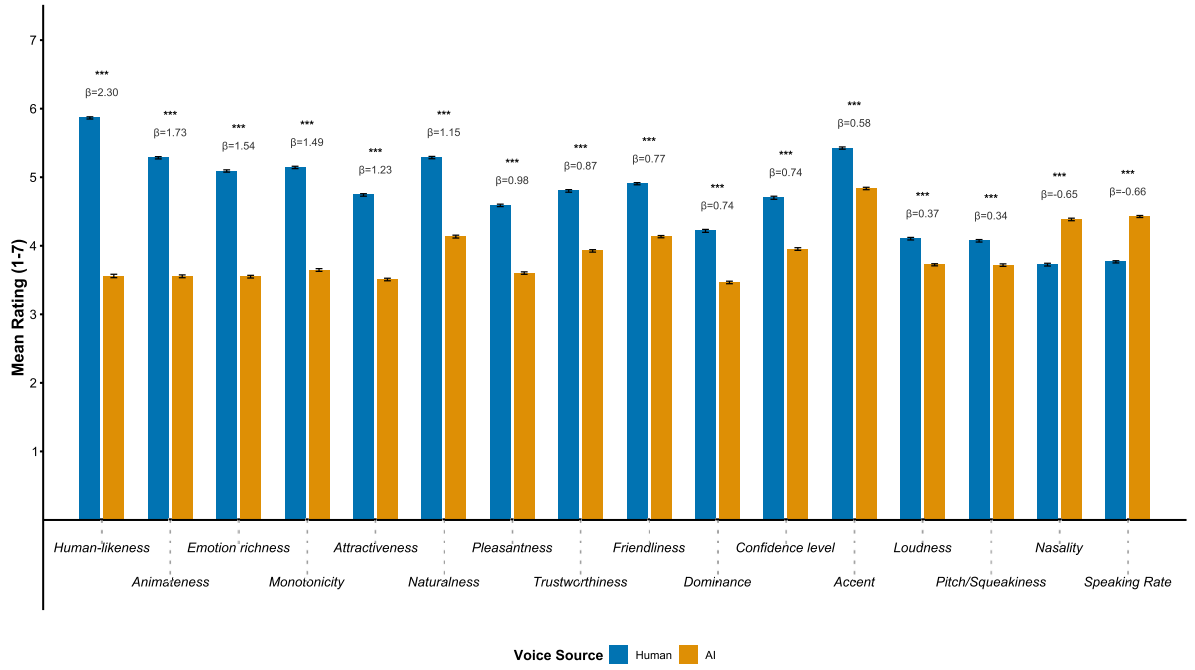

Fig. 3. ***Mean perceptual ratings of human and AI-cloned voices across 16 attributes.*** Mean ratings (±SE) on 7-point scales. β = unstandardized coefficient for voice source (Human vs. AI) from LMMs. ***$p < .003125$ (Bonferroni-corrected). Attributes ordered by effect magnitude (Human − AI).

3.3. Confident prosody modulates perceptual ratings for both voice sources, but more so for AI voices

To examine how confidence level affects voice perception and whether these effects differ between voice sources, we fitted LMMs with voice source × confidence level interactions for each of the 15 perceptual attributes (Fig. 4 and Table S6).

Confident prosody significantly affected ratings for human voices across all 15 attributes (15/15 surviving Bonferroni correction). Specifically, compared to doubtful prosody, confident human voices were rated as more human-like ($\beta = 0.24$), friendly ($\beta = 0.16$), natural ($\beta = 0.37$), emotionally rich ($\beta = 0.18$), animate ($\beta = 0.36$), attractive ($\beta = 0.53$), pleasant ($\beta = 0.51$), trustworthy ($\beta = 0.94$), and dominant ($\beta = 0.99$). On speech quality attributes, confident human voices were also perceived as more vivid (monotonicity; $\beta = 0.17$), farther from standard Mandarin (accent; $\beta = 0.37$), louder ($\beta = 0.35$), rated higher on the pitch/squeakiness scale ($\beta = 0.15$), and faster in speaking rate ($\beta = 0.36$), while sounding less nasal ($\beta = -0.16$).

The same pattern held for AI voices: confident prosody significantly shifted ratings on all 15 attributes (15/15 surviving Bonferroni correction). Confident AI voices were rated as more human-like ($\beta = 0.64$), friendly ($\beta = 0.42$), natural ($\beta = 0.66$), emotionally rich ($\beta = 0.39$), animate ($\beta = 0.66$), attractive ($\beta = 0.69$), pleasant ($\beta = 0.65$), trustworthy ($\beta = 0.95$), and dominant ($\beta = 0.90$). On speech quality attributes, confident AI voices were perceived as more vivid (monotonicity; $\beta = 0.37$), farther from standard Mandarin (accent; $\beta = 0.50$), louder ($\beta = 0.24$), rated higher on the pitch/squeakiness scale ($\beta = 0.14$), and faster in speaking rate ($\beta = 0.14$), while sounding less nasal ($\beta = -0.22$).

However, voice source × confidence interactions were significant for 12 of 15 attributes, indicating that the magnitude of prosodic modulation differed between voice sources. AI voices showed larger rating gains from confident prosody in social perception domains, including human-likeness (AI: $\beta = 0.64$ vs. Human: $\beta = 0.24$), friendliness (0.42 vs. 0.16), monotonicity (0.37 vs. 0.17), emotion richness (0.39 vs. 0.18), animateness (0.66 vs. 0.36), and naturalness (0.66 vs. 0.37), as well as moderate advantages for accent (0.50 vs. 0.37), attractiveness (0.69 vs. 0.53), and pleasantness (0.65 vs. 0.51). Conversely, human voices showed larger gains for speaking rate (0.36 vs. 0.14), loudness (0.35 vs. 0.24), and dominance (0.99 vs. 0.90). Trustworthiness, pitch/squeakiness, and nasality showed non-significant interactions. These findings demonstrate that confident prosody produces greater perceptual change in AI than human voices, particularly across social attributes.

3.4. Underlying dimensional structure of voice perception (PCA analysis)

The LMM results revealed significant perceptual differences between AI and human voices across individual attributes. PCA was conducted to understand the underlying structure organizing these differences. The analysis revealed that voice perception is organized along two primary orthogonal dimensions of social appeal and vocal expressiveness, with AI and human voices occupying distinct regions in this reduced perceptual space.

The scree plot (Fig. 5A) identified three principal components that met the Kaiser criterion (eigenvalues >1), collectively explaining 61.8% of the total variance. PC1 accounted for 42.7% of the variance, with strong negative loadings on attractiveness (−0.33), animateness (−0.33), and pleasantness (−0.31), indicating that higher PC1 scores correspond to lower ratings on these positive attributes. PC2 explained 11.1% of variance, characterized by positive loadings on speaking rate (0.53), loudness (0.49), pitch/squeakiness (0.36), dominance (0.33), and confidence level (0.31). PC3 explained an additional 8.0% of variance with mixed loadings that partially overlap the PC2 expressiveness dimension, indicating it likely captures residual prosodic variance not fully explained by the primary two-factor structure.

The variable loadings plot (Fig. 5B) revealed distinct clustering patterns: attributes related to expressiveness clustered in the first quadrant with positive loadings on PC2, representing short-term prosodic cues, including speaking rate, loudness, pitch/squeakiness, and confidence level. In contrast, social appeal-related attributes clustered in the third quadrant with negative loadings on both components, including attractiveness, pleasantness, animateness, and naturalness. This pattern indicates that vocal expressiveness and social appeal represent orthogonal dimensions of voice perception.

Factor score profiles (Fig. 5C) showed significant group differences across all three components (PC1, PC2, and PC3; all $ps < 0.001$, Bonferroni-corrected paired t-tests). The largest difference occurred in

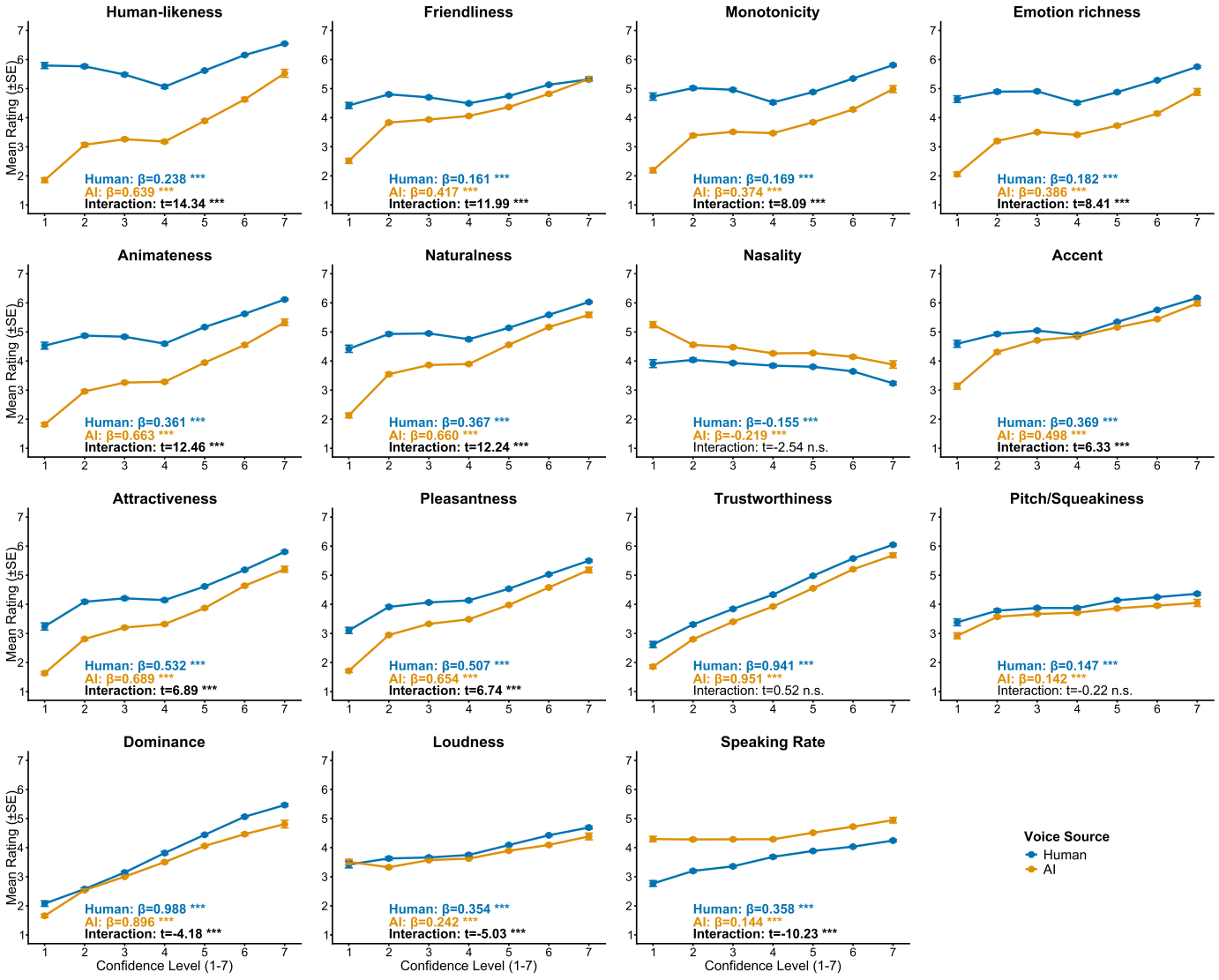

Fig. 4. ***How varying prosody from doubtful to confident shapes perceptual ratings differently for human and AI voices across 15 attributes.*** Mean ratings (±SE) for human and AI voices across confidence levels (1 = Not confident at all, 7 = Very confident). β = conditional slope derived from LMMs with voice source × confidence level interaction, AI voice exposure frequency as covariate, and random intercepts for participant, speaker, and item. A steeper slope indicates greater sensitivity to prosodic variation. Panels ordered by strength of AI prosodic sensitivity (slope ratio = AI slope/Human slope). ***$p < .001111$ (Bonferroni-corrected); n.s. = not significant.

PC1, where AI voices scored significantly higher than human voices ($d = 1.45$, large effect), indicating that AI voices were perceived as less attractive, animate, and pleasant. Large effects were also observed in PC2 ($d = 1.05$) and PC3 ($d = 0.91$). This pattern demonstrates that the primary distinction between AI and human voices lies in the social appeal dimension (PC1), with secondary differences in vocal expressiveness (PC2).

The bivariate distribution plot (Fig. 5D) demonstrated clear separation between AI and human voice clusters in the PC1–PC2 space, with AI voices centred at higher PC1 values ($M = 1.28$) and human voices at lower PC1 values ($M = -1.31$). The 68% and 95% confidence ellipses showed minimal overlap between groups, with AI voices distributed more variably in PC1 ($SD = 2.45$) compared to human voices ($SD = 2.07$). Extreme cases (defined as exceeding the 90th percentile on either |PC1| or |PC2|) comprised 18.1% of AI voices and 19.7% of human voices, indicating comparable perceptual variability between AI and human voices. Multivariate analysis confirmed significant overall group differences (Hotelling's $T^2 = 2421.88$, $p < .001$), indicating that AI and human voices occupy distinct regions in this reduced perceptual space.

3.5. Post hoc analysis based on PCA: vocal expressiveness moderates AI–human differences in social appeal

To examine whether prosodic effects also manifest in principal component space, we tested PC2 (vocal expressiveness) as a moderator of voice source effects on PC1 (social appeal), building on the PCA results reported above. Fig. 6A shows a statistically significant but small voice source × PC2 interaction ($\beta = 0.40$, $t = 10.70$, $p < .001$). Human voices demonstrated a stronger negative expressiveness-to-appeal relationship ($\beta = -0.50$, $t = -22.54$, $p < .001$), indicating that more expressive human voices were rated as substantially more socially appealing. AI voices also showed a significant but markedly weaker relationship ($\beta = -0.14$, $t = -4.94$, $p < .001$), suggesting that expressiveness gains translated into only modest social appeal improvements for AI voices.

Fig. 6B shows that the raw AI-human difference in social appeal grew across expressiveness levels ($\Delta = 2.09$, 2.80, and 2.91 for low, medium, and high), although the standardized effect size remained large and stable throughout ($d = 1.36$, 1.40, and 1.42). This pattern suggests that

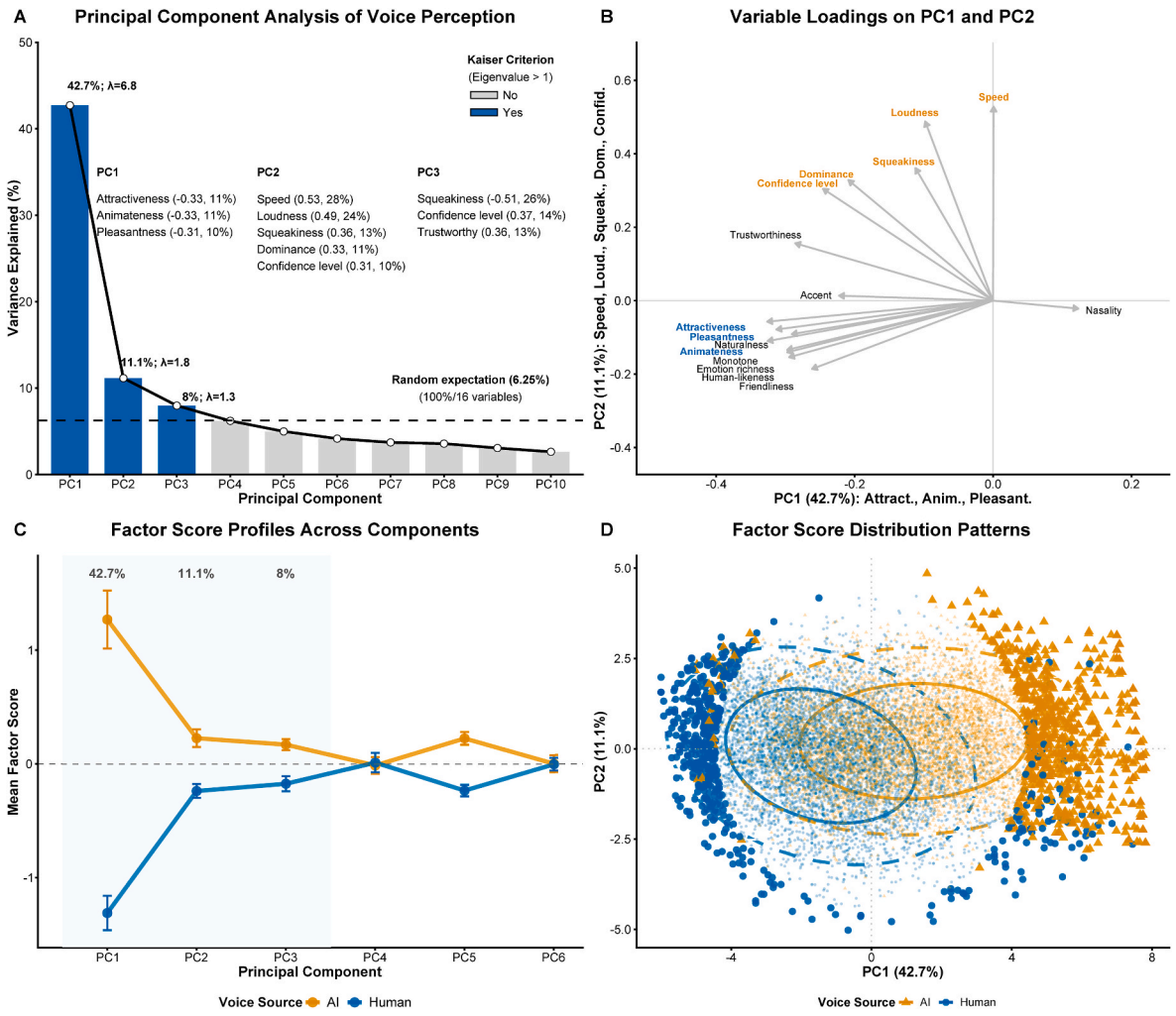

Fig. 5. ***Principal component analysis of voice perception ratings.*** **(A)** Scree plot; the first three components explain 61.8% of variance (Kaiser criterion: eigenvalues >1). **(B)** Variable loadings on PC1 (42.7%: attractiveness, animateness, pleasantness; note that all three load negatively, such that higher PC1 scores indicate lower social appeal) and PC2 (11.1%: speaking rate, loudness, pitch/squeakiness, dominance, confidence level). **(C)** Mean factor scores (±SE) for AI and human voices across components; group differences were significant across all three components (paired *t*-tests, all *ps* < 0.001, Bonferroni-corrected). **(D)** Factor score distributions in PC1–PC2 space; ellipses represent 68% (solid) and 95% (dashed) confidence intervals. Enlarged points mark extreme cases (radial distance $\sqrt{(PC1^2 + PC2^2)}$ above the 90th percentile), a display-only criterion distinct from the marginal one in Section 3.4 (90th percentile on either |PC1| or |PC2|). Multivariate group differences were significant (Hotelling's $T^2 = 2421.88$, $p < .001$).

while prosodic cues can modulate AI voice perception at individual dimensional levels, categorical voice source distinctions constrain processing at more fundamental perceptual levels, with categorical boundaries remaining resilient even when prosodic features vary.

4. Discussion

Our study directly compared human and AI-cloned voices across 16 perceptual attributes, using voice cloning to hold speaker identity constant while systematically varying prosodic style between confident and doubtful expressions. For RQ1, human voices received higher ratings on most attributes, replicating previous literature, with AI voices perceived as faster and more nasal. For RQ2, while confident prosody reliably modulated ratings for both voice sources, only human voices showed expressiveness-to-appeal mappings in overall social impressions, suggesting that categorical processing constrains prosodic benefits for AI voices.

4.1. Human voice advantages across perceptual attributes: replications, small contradictions, and social role implications

We observed human voice advantages across 11 attributes, replicating previous literature. Human voices were rated as more human-like (Kühne et al., 2020; Schreibelmayr & Mara, 2022), trustworthy, confident, and pleasant (Kühne et al., 2020), more animate and emotionally rich (Kuriki et al., 2016), more attractive (Bruder et al., 2025), more natural (Fan & Liu, 2025), and more vivid in vocal expression (monotonicity) (Mullennix et al., 2003). AI voices were conversely perceived as faster in speaking rate and more nasal than human voices (Mullennix et al., 2003).

The present study also provides direct evidence for human voice advantages on friendliness and dominance, extending existing literature on social perception of artificial voices (Im et al., 2023; Shiramizu et al., 2022). Among these, research on human-human communication suggests that lower F0 variability is associated with perceived physical dominance (Hodges-Simeon et al., 2010). Our acoustic analyses revealed that AI voices exhibited a narrower F0 range ($\beta = -2.01$) and smaller F0 variability ($\beta = -0.42$) than human voices (see **Table S4**), which might have been expected to yield higher dominance ratings for

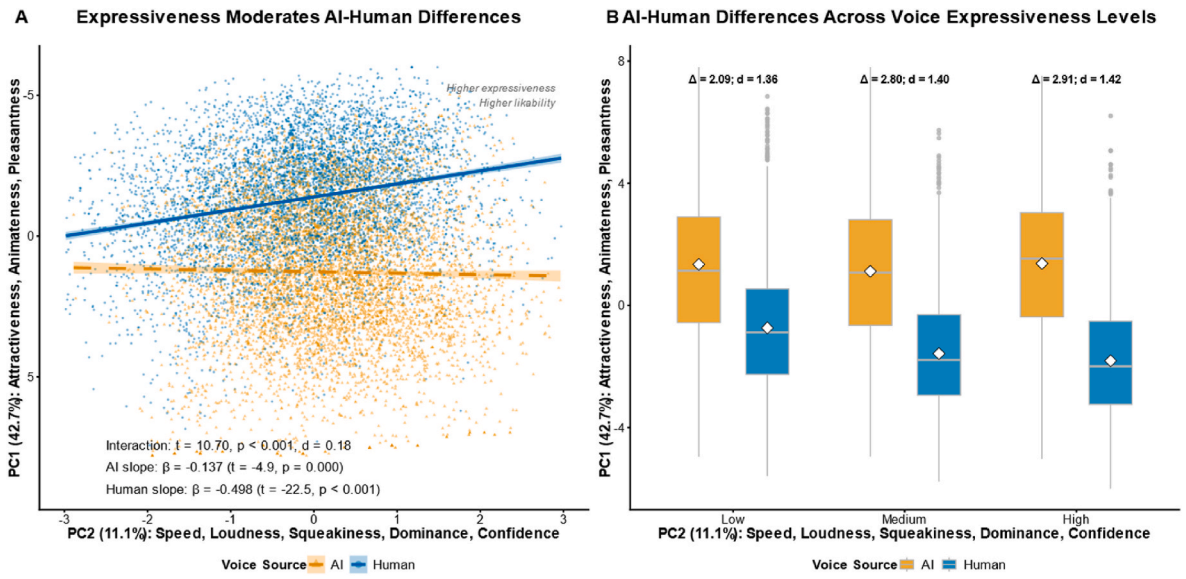

Fig. 6. ***Vocal expressiveness moderates AI–human differences in social appeal.*** Panel **(A)** shows the interaction between voice source and vocal expressiveness (PC2) on social appeal (PC1); a significant interaction ($\beta = 0.40$, $t = 10.70$, $p < .001$) indicates that the expressiveness-to-appeal relationship is stronger for human than AI voices. Panel **(B)** shows AI-human differences in social appeal across low, medium, and high expressiveness levels (tercile split), with Cohen's *d* effect sizes.

AI voices. Yet the opposite was found. Notably, our prosody data show the reverse: confident voices, with greater F0 variability, were rated as more dominant (Section 3.3). The same acoustic feature thus predicted dominance in opposite directions across the two comparisons, so acoustics alone cannot explain the human advantage. This pattern may instead reflect categorical processing, consistent with the possibility that CASA-based social attributions are not universally applied (Heyselaar, 2023; Kaptelinin & Dalli, 2025) to AI voices: even when acoustic cues might otherwise support social trait attributions such as dominance, listeners' categorical recognition of the voice as artificial may override these cues, limiting the transfer of human social perception patterns to AI-generated speech. A similar logic may apply to friendliness: categorical expectations that AI voices lack social warmth may constrain friendliness attributions regardless of acoustic input.

We also report three findings that contradict previous literature: despite normalization to −30 dBFS at a sampling rate of 44,100 Hz across all audio files, human voices were perceived as louder, rated higher on the pitch/squeakiness scale, and more regionally accented than AI voices, contrary to Mullennix et al. (2003). Regarding loudness ($\beta = 0.37$) and pitch/squeakiness ($\beta = 0.34$), both effects were small and may partly reflect the substantial technological advances in speech synthesis over the past two decades: Mullennix et al. (2003) employed the DECtalk text-to-speech system, a formant synthesis technology originally developed in the 1980s (Klatt, 1982), whereas our study used modern neural voice cloning technology. The pitch/squeakiness finding may additionally reflect the greater tonal variation of naturally produced Mandarin compared to the smoother, more uniform output of AI-generated speech, though this remains speculative. Regarding regional accent, the difference likely reflects acoustic smoothing introduced by the voice cloning process, which may attenuate subtle regional prosodic features even when speaker identity is preserved, and may additionally reflect categorical expectations that AI-generated speech sounds more standardized and accent-neutral.

4.2. Naturalness may not be the most direct measure for capturing AI–human voice distinctions

We also note that human-likeness ($\beta = 2.30$) (Kuriki et al., 2016), animateness ($\beta = 1.73$) (Kuriki et al., 2016), and naturalness ($\beta = 1.15$) (Fan & Liu, 2025) all effectively discriminated between voice sources, though with varying effect magnitudes. Among these, naturalness may be a less sensitive measure as technological advances render AI voices increasingly natural-sounding, whereas human-likeness, which more directly targets the machine-like vs. human-like distinction, may provide a more sensitive measure for future research examining listeners' categorical voice source perception.

This is consistent with the conceptual framework proposed by Nussbaum et al. (2025), which distinguishes two types of naturalness: deviation-based naturalness (how much a voice deviates from a normal reference) and human-likeness-based naturalness (how closely a voice resembles a real human voice). Our findings support the latter as a more appropriate conceptualization for AI-human voice research, as human-likeness emerged as the strongest discriminator, suggesting that listeners' categorical distinctions between AI and human voices are better captured by direct human-likeness judgments than by broader naturalness impressions.

4.3. Categorical processing constrains prosodic benefits: AI voices show attenuated expressiveness-to-appeal mappings

The attribute-level advantages discussed above hint at categorical processing of human vs. AI voices. The PCA results further illuminate this by revealing how vocal expressiveness and social appeal organize differently across voice sources. Importantly, this categorical processing did not arise from explicit source labeling: participants were not informed whether each stimulus was human or AI-generated, and categorical distinctions are instead assumed to emerge implicitly from underlying acoustic differences detectable within ~200 ms of voice onset (Chen et al., 2026a).

Previous research found that both human (McAleer et al., 2014) and synthetic (Shiramizu et al., 2022) voice evaluations independently organize along valence and dominance dimensions, suggesting parallel perceptual structures across voice types. In the current study, we manipulated both group identity (AI vs. human) and prosodic states (confident vs. doubtful) to examine how these factors jointly structure voice perception. Our PCA revealed two key dimensions: social appeal (PC1, capturing attractiveness, animateness, and pleasantness) and vocal expressiveness (PC2, capturing speed, loudness, pitch/squeakiness, confidence level, and dominance). The inclusion of dominance in the expressiveness dimension likely reflects that more

confident-sounding voices inherently elicit greater dominance impressions (Hodges-Simeon et al., 2010), as confident prosody involves acoustic features such as increased intensity and pitch variation that signal assertiveness and control. This two-component structure reveals the core perceptual dimensions listeners use when evaluating AI vs. human voices. We quantitatively demonstrate that social appeal and vocal expressiveness represent two distinct perceptual dimensions, where expressiveness may function as a diagnostic marker (Kühne et al., 2020; Rodero & Lucas, 2023) for voice source detection, while social appeal (Cuciniello et al., 2022; Noah et al., 2021; Rodero, 2017; Rodero & Lucas, 2023; Romportl, 2014; Seaborn et al., 2021; Stern et al., 2006) reflects distinct attitudes toward AI vs. human speakers.

Meanwhile, our findings regarding how confident prosody alters perceptual ratings revealed a dissociation reminiscent of Simpson's Paradox between individual attribute and principal component analyses. Simpson's Paradox occurs when a trend that appears in several groups of data disappears or reverses when the groups are combined, often due to confounding variables or differences in group composition (Kievit et al., 2013; Simpson, 1951). We use the term by analogy: rather than a single association reversing upon aggregation, the AI advantage seen at the attribute level (perceived confidence predicting individual ratings) reverses at the integrated component level (expressiveness predicting social appeal). At the individual attribute level, all 15 attributes shifted significantly with prosody for both voice sources (more positive ratings on 14 attributes, lower nasality ratings), with AI voices showing greater prosodic sensitivity than human voices across most of the 15 attributes. However, our PCA results revealed the opposite pattern: human voices demonstrated substantially stronger expressiveness-to-appeal mappings, while AI voices showed a significantly weaker mapping that did not meaningfully translate expressiveness into overall social appeal gains. This dissociation suggests that while prosodic confidence enhances individual perceptual dimensions more strongly for AI voices, these improvements are substantially attenuated at the level of integrated social perception.

Fig. 7A offers a preliminary conceptual illustration of this pattern based on the current findings, though it should be noted that this is not a formal theoretical model but rather an initial attempt to summarize the observed results. This pattern suggests that while listeners do apply some human voice perception patterns to AI voices at the individual attribute level, consistent with CASA-based predictions (Reeves & Nass, 1996), these responses are constrained at the level of integrated social perception by categorical voice source distinctions, analogous to accent-based in-group/out-group processing (Paladino & Mazzurega, 2020). Rather than automatically and unconsciously applying human social rules (Heyselaar, 2023; Kaptelinin & Dalli, 2025), listeners appear to maintain categorical boundaries between AI and human voices that limit the extent to which prosodic expressiveness translates into social appeal. As per the Stereotype Content Model (Cuddy et al., 2008), AI and human voices may be perceived as distinct social groups, with categorical recognition of AI voice source preceding and constraining the social-emotional impact of prosodic cues. Such categorical constraints may be rooted in the rapid perceptual detection of voice source: neural discrimination of human vs. AI voices emerges as early as ~200 ms post-onset, substantially preceding prosodic decoding (Chen et al., 2026a), and voice naturalness assessments, including human-likeness judgments, are proposed to occur at early stages of voice object analysis prior to the processing of social-communicative content (Nussbaum et al., 2025).

The present study controlled for speaker identity (who is saying it) (Kuhl, 2011) through voice cloning, held linguistic content (what is said) constant across stimuli, and systematically varied prosodic style (how things are said) (Hellbernd & Sammler, 2016; Pinheiro, 2025) between confident and doubtful conditions. As illustrated in Fig. 7B, AI and human voices differed along both axes.

On the long-term traits axis, voice source was rapidly detectable from underlying acoustic features within ~200 ms of voice onset (Chen et al., 2026a), indicating that AI and human voices occupy distinct categorical positions even when speaker identity is nominally held constant. We propose that AI-human differences function as sociocultural primitives within long-term trait processing (Schuller & Batliner, 2013), manifesting through acoustic markers such as speech rate, F0 variability, and distributed spectral properties such as MFCC (Bisogni et al., 2024). On the short-term states axis, prosodic realization was not fully equivalent across voice sources, as reflected in the acoustic and perceptual differences reported above.

Accordingly, long-term categorical representations of voice source may constrain how short-term prosodic cues are interpreted and weighted in social perception, akin to how accent-based in-group/out-group distinctions constrain prosodic processing (Jiang et al., 2018, 2020).

4.4. Limitations and considerations

The present study employed Mandarin Chinese stimuli rated by university students, limiting the generalizability of our findings to other languages, accent backgrounds, and user populations. We used a single voice cloning technology, and results may vary with different synthesis approaches. We must note that, although our cloning procedure was designed to maximally match speaker identity across human and AI voices, some residual acoustic differences inevitably remain between

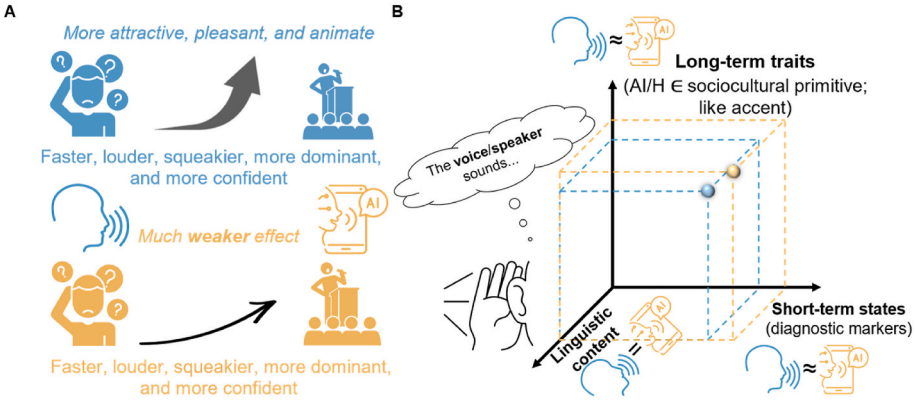

Fig. 7. ***Preliminary interpretive illustration of expressiveness-to-appeal mappings in human and AI voices.*** Panel **(A)** illustrates that vocal expressiveness was associated with substantially stronger social appeal gains for human voices than for AI voices, consistent with categorical processing accounts. Panel **(B)** depicts the proposed relationships between linguistic content (what is said), prosodic short-term states (how things are said), and voice source as a long-term trait (who is saying it); data points represent perceived rather than objective speaker identity, as voice cloning produces an approximation of the human voice. Both panels are preliminary interpretations based on the current sample and AI voice technology, and their generalizability awaits more empirical validation.

any two utterances of the same content, even when they fall below the threshold of conscious perception; our human–AI comparisons should therefore be understood as comparisons between a human voice and its closest currently-achievable AI counterpart, rather than between two acoustically identical speakers. Stimuli were restricted to factual statements with confident and doubtful prosody, and the controlled laboratory setting may not capture real-world human-computer interactions where contextual factors and extended exposure could influence voice perception. Additionally, perceptual ratings based on audio alone may not directly predict behavioral outcomes in actual human-robot collaboration, where visual cues, embodiment, task demands, and extended interaction experience may modulate voice perception.

A special note is that the interpretive illustrations presented in Fig. 7 are preliminary, based on a sample of 40 listeners using a single AI voice cloning technology in a single language. The proposed framework should be regarded as a speculative account grounded in the current findings rather than an established theoretical model; replication with larger and more diverse samples, different AI synthesis systems, and other languages is needed before broader conclusions can be drawn. We also note that perceptual ratings alone cannot directly adjudicate mechanism-level differences but can only indicate possible ones; combining explicit ratings with time-resolved measures (e.g., EEG) or functional neuroimaging approaches (e.g., fMRI) will be needed to test this in future work.

In addition, unlike designs in which each listener rates only a single trait to minimize halo effects (McAleer et al., 2014), our participants rated all 16 attributes for every stimulus. Because the confidence predictor (*confidence_std*) and the 15 outcome attributes were therefore all provided by the same listeners for the same stimuli, the conditional slopes may partly reflect shared-rater (common-method) variance, including halo effects (e.g., a voice rated high on one positive trait being difficult to rate low on a related one), rather than purely manipulation-driven effects.

Finally, we note that the social appeal and vocal expressiveness dimensions identified by PCA are likely tied to the prosodic manipulation employed here. Studies manipulating other prosodic cues such as engaging vs. neutral tone (Domínguez-Arriola & Pell, 2026) may reveal comparable expressiveness-appeal dissociations, while those without prosodic manipulation may instead replicate the valence and dominance structure identified in prior voice perception research (McAleer et al., 2014; Shiramizu et al., 2022).

5. Conclusions

Our study offers a preliminary explanatory framework for future research examining how individuals perceive and interact with AI agents through speech. The human voice advantages observed in currently available studies may be explained by three contributing factors: (1) listeners rapidly detect AI-generated voices based on underlying acoustic features (Bisogni et al., 2024; Nussbaum et al., 2025; Chen et al., 2026a); (2) AI-generated speech is inherently less prosodically expressive than genuine human vocal expression (Cohn et al., 2020; Klüber et al., 2025; Kühne et al., 2020); and (3) the integration of available cues to identify voice source invokes out-group categorical expectations (Gampe et al., 2023; Heyselaar, 2023; Kaptelinin & Dalli, 2025; Li et al., 2021; Stern et al., 2006), leading listeners to withhold social attributions such as dominance and friendliness from AI voices regardless of acoustic input (the current study). Together, these three factors may serve as useful reference points for future research on the social perception of AI voices.

As highlighted in the American Psychological Association 2026 Trends Report (Andoh, 2026), AI companions are increasingly engineered to evoke human-like attachment through natural-sounding speech that mimics human cadence and tone, simulated empathy (Brandtzaeg et al., 2022; Liu-Thompkins et al., 2022; Mari et al., 2024), and personalized interaction (Adewale & Muhammad, 2025). Users can develop genuine emotional bonds with such agents, with some research suggesting that AI companionship can alleviate loneliness to a degree comparable to human interaction (De Freitas et al., 2025; Montag et al., 2025; Xie & Pentina, 2022). Moreover, specific populations such as elderly users (Jones et al., 2021; Montag et al., 2025) and individuals with autism spectrum disorder (Kuriki et al., 2016) may process AI voice cues differently, underscoring the need for inclusive design considerations. It is for these reasons that we propose AI voices should remain perceptibly identifiable as AI-generated, and that designing prosodically expressive AI voices that closely mimic human emotional richness should be approached with particular caution in interpersonal settings. Furthermore, given that AI voices are increasingly deployed alongside embodied facial expressions in social robots and virtual agents, we propose that prosodic expressiveness should be designed to match the artificiality level of facial expressions where technically feasible, as multimodal coherence between voice and facial expression may enhance communicative effectiveness (Cihodaru-Ștefanache & Podina, 2025; Hu et al., 2024).

CRediT authorship contribution statement

Wenjun Chen: Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation, Conceptualization. **Marc D. Pell:** Writing – review & editing, Supervision, Resources. **Xiaoming Jiang:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization.

Data statement

The code and data supporting the findings of this study are openly available in the Open Science Framework repository at https://osf.io/qgrva/overview?view_only=2a4d21db2419426299af9229e8945675.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Claude (Anthropic) to assist with highly customized data analysis and visualizations, as well as to refine wording in the manuscript. Grammarly was also used to check grammar and improve sentence clarity. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

This research was supported by the National Natural Science Foundation of China (Grant No. 32471109), awarded to X. Jiang. The PhD studentship of W. Chen was supported by a McGill-CSC (China Scholarship Council) Joint Scholarship, part of which is sourced from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2022-04363) awarded to M. D. Pell.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbr.2026.101142.

References

Abdulrahman, A., & Richards, D. (2022). Is natural necessary? Human voice versus synthetic voice for intelligent virtual agents. *Multimodal Technologies and Interaction, 6*(7), 51. https://doi.org/10.3390/mti6070051

Abrams, D. A., Chen, T., Odriozola, P., Cheng, K. M., Baker, A. E., Padmanabhan, A., Ryali, S., Kochalka, J., Feinstein, C., & Menon, V. (2016). Neural circuits underlying mother's voice perception predict social communication abilities in children. *Proceedings of the National Academy of Sciences, 113*(22), 6295–6300. https://doi.org/10.1073/pnas.1602948113

Adewale, M. D., & Muhammad, U. I. (2025). From virtual companions to forbidden attractions: The seductive rise of artificial intelligence love, loneliness, and intimacy—A systematic review. *Journal of Technology in Behavioral Science*. https://doi.org/10.1007/s41347-025-00549-4

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *NeuroImage, 52*(4), 1528–1540. https://doi.org/10.1016/j.neuroimage.2010.05.048

Andoh, E. (2026). *AI chatbots and digital companions are reshaping emotional connection*. https://www.apa.org/monitor/2026/01-02/trends-digital-ai-relationships-emotional-connection.

Arık, S.Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). *Deep voice: Real-time neural text-to-speech* proceedings of the 34th international conference on machine learning. *Proceedings of Machine Learning Research*. https://proceedings.mlr.press/v70/arik17a.html.

Arık, S.Ö., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural voice cloning with a few samples. *Advances in Neural Information Processing Systems, 31*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Becker, D., Braach, L., Clasmeier, L., Kaufmann, T., Ong, O., Ahrens, K., Gäde, C., Strahl, E., Fu, D., & Wermter, S. (2025). Influence of robots' voice naturalness on trust and compliance. *ACM Transactions on Human-Robot Interaction, 14*(2). https://doi.org/10.1145/3706066. Article 29.

Bérubé, C., Nißen, M., Vinay, R., Geiger, A., Budig, T., Bhandari, A., Pe Benito, C. R., Ibarcena, N., Pistolese, O., Li, P., Sawad, A. B., Fleisch, E., Stettler, C., Hemsley, B., Berkovsky, S., Kowatsch, T., & Kocaballi, A. B. (2024). Proactive behavior in voice assistants: A systematic review and conceptual model. *Computers in Human Behavior Reports, 14*, Article 100411. https://doi.org/10.1016/j.chbr.2024.100411

Bisogni, C., Loia, V., Nappi, M., & Pero, C. (2024). Acoustic features analysis for explainable machine learning-based audio spoofing detection. *Computer Vision and Image Understanding, 249*, Article 104145. https://doi.org/10.1016/j.cviu.2024.104145

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer. In (Version Version 6.2.09) Praat. https://www.praat.org/ER.

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI friend: How users of a social chatbot understand their Human–AI friendship. *Human Communication Research, 48*(3), 404–429. https://doi.org/10.1093/hcr/hqac008

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies, 62*(2), 161–178. https://doi.org/10.1016/j.ijhcs.2004.11.002

Brück, C., Kreifelts, B., & Wildgruber, D. (2011). Emotional voices in context: A neurobiological model of multimodal affective information processing. *Physics of Life Reviews, 8*(4), 383–403. https://doi.org/10.1016/j.plrev.2011.10.002

Bruder, C., Breda, P., & Larrouy-Maestri, P. (2025). Attractive synthetic voices. *Computers in Human Behavior: Artificial Humans, 6*, Article 100211. https://doi.org/10.1016/j.chbah.2025.100211

Brummernhenrich, B., Paulus, C. L., & Jucks, R. (2025). Applying social cognition to feedback chatbots: Enhancing trustworthiness through politeness. *British Journal of Educational Technology, 56*(6), 2321–2340. https://doi.org/10.1111/bjet.13569

Cambre, J., Colnago, J., Maddock, J., Tsai, J., & Kaye, J. (2020). *Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content* proceedings of the 2020 CHI conference on. *Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376789. Honolulu, HI, USA.

Chen, W., & Jiang, X. (2023). Voice-cloning artificial-intelligence speakers can also mimic human-specific vocal expression. *Preprints*. https://doi.org/10.20944/preprints202312.0807.v1

Chen, W., Pell, M. D., & Jiang, X. (2026a). Human brains implicitly and rapidly distinguish AI from human voices before decoding prosodic meaning. *bioRxiv, 2026*. https://doi.org/10.64898/2026.04.08.716483, 2004.2008.716483.

Chen, W., Pell, M. D., & Jiang, X. (2026b). Prosodic cues strengthen human-AI voice boundaries: Listeners do not easily perceive human speakers and AI clones as the same person. *Computers in Human Behavior: Artificial Humans, 7*, Article 100261. https://doi.org/10.1016/j.chbah.2026.100261

Cihodaru-Ștefanache, Ș., & Podina, I. R. (2025). Exploring the uncanny valley effect: A critical systematic review of attractiveness, anthropomorphism, and uncanniness. *Frontiers in Psychology, 16*, Article 1625984. https://doi.org/10.3389/fpsyg.2025.1625984

Cohn, M., Raveh, E., Predeck, K., Gessinger, I., Möbius, B., & Zellou, G. (2020). Differences in gradient emotion perception: Human vs. Alexa voices. *Proceedings of Interspeech*. https://doi.org/10.21437/Interspeech.2020-1938.

Cuciniello, M., Amorese, T., Cordasco, G., Marrone, S., Marulli, F., Cavallo, F., Gordeeva, O., Carrión, Z. C., & Esposito, A. (2022). Identifying synthetic voices' qualities for conversational agents. *Applied Intelligence and Informatics, Reggio Calabria, Italy*. https://doi.org/10.1007/978-3-031-24801-6_24, 2022//.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. (2025). AI companions reduce loneliness. *Journal of Consumer Research, 52*(6), 1126–1148. https://doi.org/10.1093/jcr/ucaf040

Di Cesare, G., Cuccio, V., Marchi, M., Sciutti, A., & Rizzolatti, G. (2022). Communicative and affective components in processing auditory vitality forms: An fMRI study. *Cerebral Cortex, 32*(5), 909–918. https://doi.org/10.1093/cercor/bhab255

Domínguez-Arriola, M. E., Bazzi, L., Mauchand, M., Foucart, A., & Pell, M. D. (2025). Does criticism in a foreign accent hurt less? *Language, Cognition and Neuroscience*, 1–20. https://doi.org/10.1080/23273798.2025.2547350

Domínguez-Arriola, M. E., & Pell, M. D. (2026). Not worth my time! understanding factors that make speech socially engaging. *Journal of Experimental Psychology: Human Perception and Performance, 52*(5), 672–692. https://doi.org/10.1037/xhp0001410

Dou, X., Wu, C.-F., Lin, K.-C., Gan, S., & Tseng, T.-M. (2021). Effects of different types of social robot voices on affective evaluations in different application fields. *International Journal of Social Robotics, 13*(4), 615–628. https://doi.org/10.1007/s12369-020-00654-9

Du, Z., Sisman, B., Zhou, K., & Li, H. (2021). Disentanglement of emotional style and speaker identity for expressive voice conversion. arXiv preprint arXiv:2110.10326 . https://doi.org/10.48550/arXiv.2110.10326.

Elsinger, F., & Hörner, C. (2024). I am no human after all: Receiving negative feedback from AI versus humans. *SSRN, 4946407*, SSRN. https://doi.org/10.2139/ssrn.4946407.

Eyssel, F., & Hegel, F. (2012). (S)he's got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology, 42*(9), 2213–2230. https://doi.org/10.1111/j.1559-1816.2012.00937.x

Fan, G., & Liu, D. (2025). When machines speak with feeling: Investigating emotional prosody, authenticity, and trust in AI vs. human voices. *Proceedings of the Annual Meeting of the Cognitive Science Society*. https://escholarship.org/uc/item/8vr8s6h8.

Fortunati, L., Edwards, A., Edwards, C., Manganelli, A. M., & de Luca, F. (2022). Is Alexa female, male, or neutral? A cross-national and cross-gender comparison of perceptions of Alexa's gender and status as a communicator. *Computers in Human Behavior, 137*, Article 107426. https://doi.org/10.1016/j.chb.2022.107426

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–85. https://doi.org/10.3316/INFORMIT.097034846749023

Gampe, A., Zahner-Ritter, K., Müller, J. J., & Schmid, S. R. (2023). How children speak with their voice assistant Sila depends on what they think about her. *Computers in Human Behavior, 143*, Article 107693. https://doi.org/10.1016/j.chb.2023.107693

Ghazanfar, A. A., & Rendall, D. (2008). Evolution of human vocal production. *Current Biology, 18*(11), R457–R460. https://doi.org/10.1016/j.cub.2008.03.030

Gong, L., & Lai, J. (2003). To mix or not to mix synthetic speech and human speech? Contrasting impact on judge-rated task performance versus self-rated performance and attitudinal responses. *International Journal of Speech Technology, 6*(2), 123–131. https://doi.org/10.1023/A:1022382413579

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language, 88*, 70–86. https://doi.org/10.1016/j.jml.2016.01.001

Heyselaar, E. (2023). The CASA theory no longer applies to desktop computers. *Scientific Reports, 13*(1), Article 19693. https://doi.org/10.1038/s41598-023-46527-9

Hinterleitner, F. (2017). Speech synthesis. In *Quality of synthetic speech: Perceptual dimensions, influencing factors, and instrumental assessment* (pp. 5–18). Singapore: Springer. https://doi.org/10.1007/978-981-10-3734-4_2.

Hodges-Simeon, C. R., Gaulin, S. J. C., & Puts, D. A. (2010). Different vocal parameters predict perceptions of dominance and attractiveness. *Human Nature, 21*(4), 406–427. https://doi.org/10.1007/s12110-010-9101-5

Hu, Y., Chen, B., Lin, J., Wang, Y., Wang, Y., Mehlman, C., & Lipson, H. (2024). Human-robot facial coexpression. *Science Robotics, 9*(88). https://doi.org/10.1126/scirobotics.adi4724. eadi4724.

Hunt, A. J., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*. https://doi.org/10.1109/ICASSP.1996.541110

Im, H., Sung, B., Lee, G., & Xian Kok, K. Q. (2023). Let voice assistants sound like a machine: Voice and task type effects on perceived fluency, competence, and consumer attitude. *Computers in Human Behavior, 145*, Article 107791. https://doi.org/10.1016/j.chb.2023.107791

Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics, 71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Ji, S., Chen, Q., Wang, W., Zuo, J., Fang, M., Jiang, Z., Huang, H., Wang, Z., Cheng, X., Zheng, S., & Zhao, Z. (2025). ControlSpeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers) Vienna, Austria*. https://doi.org/10.18653/v1/2025.acl-long.346

Jiang, X., Gossack-Keenan, K., & Pell, M. D. (2020). To believe or not to believe? How voice and accent information in speech alter listener impressions of trust. *Quarterly*

Journal of Experimental Psychology, 73(1), 55–79. https://doi.org/10.1177/1747021819865833
Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication, 88*, 106–126. https://doi.org/10.1016/j.specom.2017.01.011
Jiang, X., & Pell, M. D. (2024). Tracking dynamic social impressions from multidimensional voice representation. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2024.08.005
Jiang, X., Sanford, R., & Pell, M. D. (2018). Neural architecture underlying person perception from in-group and out-group voices. *NeuroImage, 181*, 582–597. https://doi.org/10.1016/j.neuroimage.2018.07.042
Jones, V. K., Hanus, M., Yan, C., Shade, M. Y., Blaskewicz Boron, J., & Maschieri Bicudo, R. (2021). Reducing loneliness among aging adults: The roles of personal voice assistants and anthropomorphic interactions. *Frontiers in Public Health, 9*, Article 750736. https://doi.org/10.3389/fpubh.2021.750736
Kaptelinin, V., & Dalli, K. C. (2025). Understanding contextual framing: A nonessentialist perspective on social interactions with technological artifacts. *20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. https://doi.org/10.1109/HRI61500.2025.10974062, 2025.
Kaur, N., & Singh, P. (2023). Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review, 56*(7), 5837–5880. https://doi.org/10.1007/s10462-022-10315-0
Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology, 4*, 513. https://doi.org/10.3389/fpsyg.2013.00513
Kim, M., Park, J., Jeong, M., & Song, J. (2025). What determines personality impressions of synthetic and natural voices? The effects of voice quality and intonation. *Language and Speech, 0*(0), Article 00238309251389567. https://doi.org/10.1177/00238309251389567
Klatt, D. (1982). The klattalk text-to-speech conversion system. *ICASSP '82. IEEE international conference on acoustics, speech, and signal processing*. https://doi.org/10.1109/ICASSP.1982.1171431
Klüber, K., Schwaiger, K., & Onnasch, L. (2025). Affect-enhancing speech characteristics for robotic communication. *International Journal of Social Robotics, 17*(2), 315–333. https://doi.org/10.1007/s12369-025-01221-w
Koch, S. B. J., Galli, A., Volman, I., Kaldewaij, R., Toni, I., & Roelofs, K. (2020). Neural control of emotional actions in response to affective vocalizations. *Journal of Cognitive Neuroscience, 32*(5), 977–988. https://doi.org/10.1162/jocn_a_01523
Kuhl, P. K. (2011). Who's talking? *Science, 333*(6042), 529–530. https://doi.org/10.1126/science.1210277
Kühne, K., Fischer, M. H., & Zhou, Y. (2020). The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in Neurorobotics, 14*, Article 593732. https://doi.org/10.3389/fnbot.2020.593732
Kumar, Y., Koul, A., & Singh, C. (2023). A deep learning approaches in text-to-speech system: A systematic review and recent research perspective. *Multimedia Tools and Applications, 82*(10), 15171–15197. https://doi.org/10.1007/s11042-022-13943-4
Kuriki, S., Tamura, Y., Igarashi, M., Kato, N., & Nakano, T. (2016). Similar impressions of humanness for human and artificial singing voices in autism spectrum disorders. *Cognition, 153*, 1–5. https://doi.org/10.1016/j.cognition.2016.04.004
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13
Lam, P. C. H., Cui, H., & Pell, M. D. (2025). The influence of speaker accent on the neurocognitive processing of politeness. *Brain Research, 1865*, Article 149897. https://doi.org/10.1016/j.brainres.2025.149897
Larrouy-Maestri, P., Poeppel, D., & Pell, M. D. (2025). The sound of emotional prosody: Nearly 3 decades of research and future directions. *Perspectives on Psychological Science, 20*(4), 623–638. https://doi.org/10.1177/17456916231217722
Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology, 23*(12), 1075–1080. https://doi.org/10.1016/j.cub.2013.04.055
Lavan, N., Irvine, M., Rosi, V., & McGettigan, C. (2025). Voice clones sound realistic but not (yet) hyperrealistic. *PLoS One, 20*(9), Article e0332692. https://doi.org/10.1371/journal.pone.0332692
Law, T., Chita-Tegmark, M., & Scheutz, M. (2021). The interplay between emotional intelligence, trust, and gender in human–robot interaction. *International Journal of Social Robotics, 13*(2), 297–309. https://doi.org/10.1007/s12369-020-00624-1
Lazebnik, T., Zalmanson, L., & Mokryn, O. (2025). Mind your manners: The dynamics of politeness in Human-AI vs. human-human interactions. *Proceedings of the ACM on Human-Computer Interaction, 9*(7). https://doi.org/10.1145/3757631. Article CSCW450.
Li, Y., Baizhou, W., Yuqi, H., Jun, L., Junhui, W., & Luan, S. (2025). Warmth, competence, and the determinants of trust in artificial intelligence: A cross-sectional survey from China. *International Journal of Human-Computer Interaction, 41*(8), 5024–5038. https://doi.org/10.1080/10447318.2024.2356909
Li, Y. A., Han, C., & Mesgarani, N. (2025). StyleTTS: A style-based generative model for natural and diverse text-to-speech synthesis. *IEEE Journal of Selected Topics in Signal Processing, 19*(1), 283–296. https://doi.org/10.1109/JSTSP.2025.3530171
Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M.. Neural speech synthesis with transformer network. https://doi.org/10.1609/aaai.v33i01.33016706.
Li, T., Wang, Z., Zhu, X., Cong, J., Tian, Q., Wang, Y., & Xie, L. (2024). U-Style: Cascading U-Nets with multi-level speaker and style modeling for zero-shot voice cloning. *IEEE/ACM Transactions on Audio, Speech and Language Processing, 32*, 4026–4035. https://doi.org/10.1109/TASLP.2024.3453606
Li, W., Zhou, X., & Yang, Q. (2021). Designing medical artificial intelligence for in- and out-groups. *Computers in Human Behavior, 124*, Article 106929. https://doi.org/10.1016/j.chb.2021.106929
Ling, Z. H., Kang, S. Y., Zen, H., Senior, A., Schuster, M., Qian, X. J., Meng, H. M., & Deng, L. (2015). Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine, 32*(3), 35–52. https://doi.org/10.1109/MSP.2014.2359987
Liu-Thompkins, Y., Okazaki, S., & Li, H. (2022). Artificial empathy in marketing interactions: Bridging the human-AI gap in affective and social customer experience. *Journal of the Academy of Marketing Science, 50*(6), 1198–1218. https://doi.org/10.1007/s11747-022-00892-5
Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S. (2025). How do voice acoustics affect the perceived trustworthiness of a speaker? A systematic review. *Frontiers in Psychology, 16*, Article 1495456. https://doi.org/10.3389/fpsyg.2025.1495456
Mari, A., Mandelli, A., & Algesheimer, R. (2024). Empathic voice assistants: Enhancing consumer responses in voice commerce. *Journal of Business Research, 175*, Article 114566. https://doi.org/10.1016/j.jbusres.2024.114566
Mauchand, M., & Pell, M. D. (2022). Listen to my feelings! how prosody and accent drive the empathic relevance of complaining speech. *Neuropsychologia, 175*, Article 108356. https://doi.org/10.1016/j.neuropsychologia.2022.108356
McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PLoS One, 9*(3), Article e90779. https://doi.org/10.1371/journal.pone.0090779
Montag, C., Spapé, M., & Becker, B. (2025). Can AI really help solve the loneliness epidemic? *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.08.002
Muhamad, S. S., Veisi, H., Mahmudi, A., Abdullah, A. A., & Rahimi, F. (2024). Kurdish end-to-end speech synthesis using deep neural networks. *Natural Language Processing Journal, 8*, Article 100096. https://doi.org/10.1016/j.nlp.2024.100096
Mullennix, J. W., Stern, S. E., Wilson, S. J., & Dyson, C.-l. (2003). Social perception of male and female computer synthesized speech. *Computers in Human Behavior, 19*(4), 407–424. https://doi.org/10.1016/S0747-5632(02)00081-X
Nass, C., Foehr, U., Brave, S., & Somoza, M. (2001). The effects of emotion of voice in synthesized and recorded speech. *Proceedings of the AAAI symposium emotional and intelligent II: The tangled knot of social cognition*. MA: North Falmouth.
Nass, C., & Lee, K. M. (2000). Does computer-generated speech manifest personality? An experimental test of similarity-attraction. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. https://doi.org/10.1145/332040.332452. The Hague, The Netherlands.
Neekhara, P., Hussain, S., Dubnov, S., Koushanfar, F., & McAuley, J. (2021). *Expressive neural voice cloning* proceedings of the 13th Asian conference on machine learning. *Proceedings of Machine Learning Research*. https://proceedings.mlr.press/v157/neekhara21a.html.
Noah, B., Sethumadhavan, A., Lovejoy, J., & Mondello, D. (2021). Public perceptions towards synthetic voice technology. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting, 65*(1), 1448–1452. https://doi.org/10.1177/1071181321651128
Nussbaum, C., Frühholz, S., & Schweinberger, S. R. (2025). Understanding voice naturalness. *Trends in Cognitive Sciences, 29*(5), 467–480. https://doi.org/10.1016/j.tics.2025.01.010
Paladino, M. P., & Mazzurega, M. (2020). One of Us: On the role of accent and race in real-time In-Group categorization. *Journal of Language and Social Psychology, 39*(1), 22–39. https://doi.org/10.1177/0261927x19884090
Park, N., Jang, K., Cho, S., & Choi, J. (2021). Use of offensive language in human-artificial intelligence chatbot interaction: The effects of ethical ideology, social competence, and perceived humanlikeness. *Computers in Human Behavior, 121*, Article 106795. https://doi.org/10.1016/j.chb.2021.106795
Pei, J., Wang, H., Peng, Q., & Liu, S. (2024). Saving face: Leveraging artificial intelligence-based negative feedback to enhance employee job performance. *Human Resource Management, 63*(5), 775–790. https://doi.org/10.1002/hrm.22226
Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y
Pinheiro, A. P. (2025). Behind a voice there is a speaker: Why vocal emotion research needs to become 'personal. *Affective Science, 6*(3), 562–574. https://doi.org/10.1007/s42761-025-00317-w
R-Core-Team. (2024). *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.
Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people, 10*. Cambridge University Press.
Revelle, W. (2025). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University [R package] https://CRAN.R-project.org/package=psych.
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. https://doi.org/10.1016/j.chb.2017.08.044
Rodero, E., & Lucas, I. (2023). Synthetic versus human voices in audiobooks: The human emotional intimacy effect. *New Media & Society, 25*(7), 1746–1764. https://doi.org/10.1177/14614448211024142
Romportl, J. (2014). Speech synthesis and uncanny Valley. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, speech and dialogue text, speech and dialogue, brno, Czech Republic*. https://doi.org/10.1007/978-3-319-10816-2_72
Rosi, V., Soopramanien, E., & McGettigan, C. (2025). Perception and social evaluation of cloned and recorded voices: Effects of familiarity and self-relevance. *Computers in Human Behavior: Artificial Humans, 4*, Article 100143. https://doi.org/10.1016/j.chbah.2025.100143

Roswandowitz, C., Kathiresan, T., Pellegrino, E., Dellwo, V., & Frühholz, S. (2024). Cortical-striatal brain network distinguishes deepfake from real speaker identity. *Communications Biology, 7*(1), 711. https://doi.org/10.1038/s42003-024-06372-6

Rozsypal, A. J., & Millar, B. F. (1979). Perception of jitter and shimmer in synthetic vowels. *Journal of Phonetics, 7*(4), 343–355. https://doi.org/10.1016/S0095-4470(19)31069-1

Schreibelmayr, S., & Mara, M. (2022). Robot voices in daily life: Vocal human-likeness and application context as determinants of user acceptance. *Frontiers in Psychology, 13*, Article 787499. https://doi.org/10.3389/fpsyg.2022.787499

Schroeter, J. (2008). Basic principles of speech synthesis. In J. Benesty, M. M. Sondhi, & Y. A. Huang (Eds.), *Springer handbook of speech processing* (pp. 413–428). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-49127-9_19.

Schuller, B., & Batliner, A. (2013). Functional aspects. In *Computational paralinguistics* (pp. 107–158). https://doi.org/10.1002/9781118706664.ch5

Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2021). Voice in human–agent interaction: A survey. *ACM Computing Surveys, 54*(4), 1–43. https://doi.org/10.1145/3386867

Shiramizu, V. K. M., Lee, A. J., Altenburg, D., Feinberg, D. R., & Jones, B. C. (2022). The role of valence, dominance, and pitch in perceptions of artificial intelligence (AI) conversational agents' voices. *Scientific Reports, 12*(1), Article 22479. https://doi.org/10.1038/s41598-022-27124-8

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B, 13*(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

Stern, S. E., Mullennix, J. W., & Yaroslavsky, I. (2006). Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. *International Journal of Human-Computer Studies, 64*(1), 43–52. https://doi.org/10.1016/j.ijhcs.2005.07.002

Sundar, S. S., Jung, E. H., Waddell, T. F., & Kim, K. J. (2017). Cheery companions or serious assistants? Role and demeanor congruity as predictors of robot attraction and use intentions among senior citizens. *International Journal of Human-Computer Studies, 97*, 88–97. https://doi.org/10.1016/j.ijhcs.2016.08.006

Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A., & Broadbent, E. (2011). The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics, 3*, 253–262. https://doi.org/10.1007/s12369-011-0100-4

Tamura, Y., Kuriki, S., & Nakano, T. (2015). Involvement of the left insula in the ecological validity of the human voice. *Scientific Reports, 5*(1), 8799. https://doi.org/10.1038/srep08799

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

Xie, T., & Pentina, I. (2022). Attachment theory as a framework to understand relationships with social chatbots: A case study of Replika. *Proceedings of the 55th Hawaii international conference on system sciences*. https://doi.org/10.24251/HICSS.2022.258

Yoon, D., Oh, G. E., & Kent, R. (2026). Perceiving emotion in human and AI voices: Sensitivity to acoustic cues in Korean speech. *Lingua, 330*, Article 104083. https://doi.org/10.1016/j.lingua.2025.104083

Yu, A. C. L. (2022). Perceptual Cue Weighting Is Influenced by the Listener's Gender and Subjective Evaluations of the Speaker: The Case of English Stop Voicing. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.840291, 2022.

Zellou, G., & Cohn, M. (2020). Social and functional pressures in vocal alignment: Differences for human and voice-AI interlocutors. *Proceedings of InterspeechInterspeech*. https://doi.org/10.21437/Interspeech.2020-1335.

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication, 51*(11), 1039–1064. https://doi.org/10.1016/j.specom.2009.04.004

Zhang, Y., Zhou, W., Huang, J., Hong, B., & Wang, X. (2022). Neural correlates of perceived emotions in human insula and amygdala for auditory emotion recognition. *NeuroImage, 260*, Article 119502. https://doi.org/10.1016/j.neuroimage.2022.119502