Computers in Human Behavior: Artificial Humans 7 (2026) 100261

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

# Prosodic cues strengthen human-AI voice boundaries: Listeners do not easily perceive human speakers and AI clones as the same person

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

[a] Institute of Language Sciences, Shanghai International Studies University, Shanghai, 201620, China
[b] Key Laboratory of Language Science and Multilingual Artificial Intelligence, Shanghai International Studies University, Shanghai, 201620, China
[c] School of Communication Sciences and Disorders, McGill University, Montréal, H3A 1G1, Canada

ARTICLE INFO

*Keywords:*
Humanlikeness
Speech
Voice cloning
Voice identity
Prosody
Bayesian

ABSTRACT

Previous studies concluded that listeners struggle to discriminate AI from human voices, but these studies used monotone-like speech and did not examine prosodic expressiveness, a key advantage of human over AI speakers. This study explores whether prosodic expressiveness facilitates human-AI voice discrimination. We recorded human prosodic speech with confident and doubtful expressions, trained AI models to replicate these prosodic patterns, had AI models generate new sentences, and then had human speakers produce equivalent prosodic expressions for the same sentences. In Experiment 1, we had 48 listeners rate humanlikeness and perceived confidence in 11,808 audio samples, finding that AI speech was consistently rated as less humanlike regardless of prosody. We selected 768 audios (AI × human, confident × doubtful prosody) for Experiment 2, where 80 listeners completed an identity discrimination task, telling whether two sounds were from the same speaker. Bayesian modeling results revealed near-ceiling performance for human-human/AI-AI pairs, with inconsistent prosodies decreasing accuracy by ~7%, while listeners do not easily categorize AI and human as sharing the same identity (~54% accuracy when prosody matches, dropping to ~36% when inconsistent). We observed accuracy–reaction time synchronization; in human–AI/AI–human pairs only, however, listeners relied less on distance cues when the two voices' identities were distant beyond a certain threshold. Overall, we found that listeners perceive AI speech as lower in humanlikeness, and prosodic variation further promotes rejecting AI and human voices as sharing the same identity, indicating that human acceptance of AI voices as equivalent to human voices is limited.

## 1. Introduction

With AI-generated speech becoming increasingly common in daily life, both the general public and researchers are prompted to wonder whether AI-generated and human-produced speech are perceived similarly. Among many perceptual dimensions of interest is speaker identity, the social manifestation of oneself (Scott & McGettigan, 2016). Through several methodological improvements (introduced in later paragraphs), our current study presents evidence challenging recent concerns that (1) listeners are reportedly poor at identifying AI speech as deepfake (Mai et al., 2023) and (2) tend to believe that human and AI speech share the same speaker identity (Barrington et al., 2025). We report that: In Experiment 1, listeners are capable of detecting AI voices when assessed using Likert-scale ratings that capture graded perceptions of humanlikeness, rather than binary forced-choice judgments (Barrington et al., 2025). In Experiment 2, prosodic variation (*tone of voice*) (Xu, 2019), a dimension often absent from prior AI voice studies despite its perceptual importance (Kühne et al., 2020; Rodero, 2017; San Segundo et al., 2025), strengthens the human–AI voice boundary such that listeners do not predominantly attribute the same speaker identity to AI-cloned and human voices.

In our discussion section, we highlight and propose that future research prioritize psychological investigations of AI-human categorical perception using controlled experimental paradigms to uncover the perceptual mechanisms (e.g., listeners' expectations of AI voices) rather than focusing on demonstrating the technical challenge that humans struggle to detect increasingly sophisticated AI speech. This is because AI voices defined in existing research (and likely in soon-to-be-available studies as well) are mostly monotone or moderately prosodically rich, but still fall far short of game-changing systems like Sora 2 (OpenAI,

* Corresponding author. Institute of Language Sciences, Shanghai International Studies University, Shanghai, 201620, China.
*E-mail address:* xiaoming.jiang@shisu.edu.cn (X. Jiang).

https://doi.org/10.1016/j.chbah.2026.100261
Received 26 September 2025; Received in revised form 2 January 2026; Accepted 6 February 2026
Available online 7 February 2026

2025; YetiAF, 2025) already encountered in everyday contexts. We specifically recommend that businesses not deploy game-changer AI voice generators in daily AI voice agent applications to maintain clear AI-human boundaries detectable to listeners, especially when they attempt to make AI-generated speech prosodically rich.

### 1.1. AI voice perception parallels accent categorization but lacks understanding of detectability foundations

Humans readily form in-group and out-group distinctions through minimal cues (Gluszek & Dovidio, 2010; Tajfel et al., 1971). A well-established example is accent perception in sociolinguistics. As a basis, listeners' brains are sensitive to phonological deviations (Best et al., 2001; Goslin et al., 2012) as well as prosodic features such as intonation patterns, rhythm, and stress placement (Mauchand & Pell, 2022a; Pandey, 2015, pp. 301–319; Polyanskaya et al., 2017). This detection triggers increased cognitive load and a sense of "otherness" (Anderson, 2007; Wilmot et al., 2024), interacting with culturally learned accent associations: For instance, Birmingham accent is more likely to be associated with guilt attributions than RP accent (Dixon et al., 2002). These accent-based stereotypes thus emerge from two mechanisms: the immediate perceptual marking of out-group membership through acoustic deviations, and pre-existing social expectations linking specific accents to particular traits or behaviors (Fuertes et al., 2012; Hosoda et al., 2007; Rakić et al., 2011).

However, unlike accent perception, where both (1) detectability and (2) in-group/out-group mechanisms are well-established, AI voice perception research presents an incomplete picture. Evidence for out-group bias is abundant: human voices are consistently rated more favorably across multiple attributes, including intelligibility, prosody, trustworthiness, confidence, enthusiasm, pleasantness, human-likeness, likability, and naturalness (Kühne et al., 2020), with similar preferences reported across various human-computer interaction contexts (Cuciniello et al., 2022; Noah et al., 2021; Rodero & Lucas, 2023; Romportl, 2014; Seaborn et al., 2021; Stern et al., 2006).

What remains critically unresolved is AI voice detectability itself. Current evidence suggests humans struggle to distinguish high-quality synthetic speech from real voices (Diel et al., 2024; Mai et al., 2023), with approximately 80% of participants believing AI-cloned voices belonged to the original human speaker (Barrington et al., 2025). These blurred AI-human boundaries (Nussbaum et al., 2025) seemingly undermine the practical social significance of investigating in-group/out-group mechanisms in AI voice perception. The present study is positioned to address this gap by providing evidence that AI voices in everyday applications, such as AI voice agents deployed in smartphones, in our case, remain perceptually detectable. Critically, we aim to redirect research attention from mere AI detectability toward the in-group and out-group perceptual mechanisms that parallel accent perception frameworks.

### 1.2. Prosody and accent interactions are well-studied; prosody and AI-human distinctions are not

Still, referring to studies in accent perception, another gap emerges apart from (1) detectability and (2) in-group and out-group perception. Specifically, (3) the interaction between long-term social-cultural cues (i.e., accent) and short-term paralinguistic cues conveyed through prosodic variation, such as basic emotions and pragmatic intentions (Schuller & Batliner, 2013), remains underexplored in AI voice research.

For human accented speech, a well-established understanding exists regarding the interaction between accent and prosodic cues. A foundation for this understanding is the universality of prosodic perception: listeners can access prosodic information across different languages (Gussenhoven & Chen, 2021; Pell et al., 2009, Pell et al., 2026), across accents (Domínguez-Arriola et al., 2025; Squizzero, 2025), and even in the absence of semantic content (Liu & Pell, 2012; Paulmann & Uskul, 2014; Zhang & Pell, 2022). Building on this universality, research has demonstrated how in-group and out-group perception interacts with prosodically embedded pragmatic intentions. For example, event-related potentials (ERPs) evidence revealed that vocal confidence, which conveys a speaker's feeling of knowing through confident vs. doubtful prosodic expressions, is processed via a rapid "direct route" for in-group accents but requires an effortful "indirect route" for out-group accents, demonstrating distinct neural processing mechanisms depending on speaker accent (Jiang et al., 2020).

Given prosody's universality, and given that both accent and human-AI differences signal social group distinctions, prosody-by-source interactions may also emerge in AI voice perception. However, this parallel remains largely speculative due to limited empirical investigation. This gap partly stems from fundamental differences in production mechanisms: human prosodic variation arises from biological vocal tract modulation, with measurable laryngeal and respiratory adjustments corresponding to emotional prosodies such as happiness, sadness, and anger (Carey & McGettigan, 2017; Kim et al., 2020).

In contrast, AI-generated prosody results from computational manipulation rather than biological vocal production. In contemporary text-to-speech systems, emotional prosody is typically controlled through combinations of strategies, including discrete emotion labeling (e.g., happiness, sadness), continuous modulation along affective dimensions such as arousal and valence, and reference-based transfer of prosodic patterns extracted from exemplar speech, allowing models to learn statistical associations between linguistic content, emotion representations, and acoustic realizations from large-scale human speech corpora (Hassani & Kangavari, 2025; Ma et al., 2025). These processes rely on Mel-spectrograms, perceptually motivated acoustic representations widely used in neural speech synthesis, which encode pitch-related and temporal cues salient to human listeners (De et al., 2025). As a result, AI-generated prosody can elicit emotion-like perceptual responses in listeners, even though such prosodic variation is produced through computational control rather than biological affect (Bruder et al., 2025; Crumpton & Bethel, 2016).

Despite AI systems' capability to generate human-comparable prosodic variations, perceptual research has mostly operationalized "AI voice" as monotone synthetic speech (Roswandowitz et al. (2024), among others listed in the next section). Meanwhile, growing interest has emerged in integrating accent perception and human-AI voice perception, with "The HUM.AI.N-ACCENT project" starting to exam how accent-based perceptual mechanisms shape communication and social judgments in interactions with both human and AI interlocutors (European-Commission, 2025).

This raises a parallel question: if prosody interacts with accent perception, as demonstrated in human speech research, might it similarly interact with human-AI categorical distinctions, shaping how listeners process AI-generated emotional cues? The present study is situated within this framework.

### 1.3. AI voice research needs to control not only speaker identity and speech content, but also prosody

Despite potentially shared in- and out-group mechanisms, accent studies differ methodologically from AI-human voice research. Due to stimuli preparation challenges, it is less feasible to hold speaker identity constant in accent studies, as accent serves as a linguistic marker tied to speakers' backgrounds (Floccia et al., 2006; Kuhl et al., 2008). However, for human and AI voice perception, AI cloning is technically viable for controlling voice identity.

Consequently, we observe shifts in AI voice studies over recent years: from (1) using existing AI-generated speech from commercial tools alongside human speech from different speakers, without controlling for speaker identity (Abdulrahman & Richards, 2022; Kühne et al., 2020; Mullennix et al., 2003; Rodero & Lucas, 2023; Schreibelmayr & Mara, 2022), to (2) having participants produce sentences (or accessing

open-source corpora) and then applying AI voice cloning tools to create deepfake AI versions of either the same content (Barrington et al., 2025; Kirk, 2025; Mai et al., 2023; Rosi et al., 2025; Warren et al., 2024) or new content (Roswandowitz et al., 2024).

Despite these efforts, a critical control element remains missing: listeners report exploiting intonation and emotional prosodic cues to distinguish human from AI speech (Kühne et al., 2020; Rodero & Lucas, 2023). Prosody, the suprasegmental aspects of speech including pitch, duration, amplitude, and voice quality (Xu, 2019), conveys such paralinguistic information. For example, "This tastes like chicken" may differentially affect listener trust when spoken with confident (firm tone, steady pitch) vs. doubtful prosody (hesitant delivery, variable pitch) (Jiang & Pell, 2015, 2017). However, such richness of vocal expression, which is inherently advantageous for human voices, is often suppressed in experimental designs that use monotone or neutral prosody (Barrington et al., 2025; Roswandowitz et al., 2024), potentially disadvantaging human voices in AI detection tasks. This suppression of prosodic expressiveness may mainly lead to conclusions that AI and human voices are indistinguishable (Barrington et al., 2025; Mai et al., 2023). Hence, an important question remains: Would AI voices remain difficult to detect if studies employed prosodically expressive human speech?

Including prosodic control in experimental designs allows researchers to isolate whether listeners distinguish AI from human voices based on source alone or on prosodic cues. Studies that control for both audio content and prosodic variation (e.g., musical prosody) reveal that human voices are still consistently perceived as more humanlike and natural, with distinct neural activation in the left posterior insula (a region associated with social communication) for human voices only (Kuriki et al., 2016; Tamura et al., 2015).

Beyond the necessity of isolating prosodic cues' contributions, controlling for prosody in AI and human speech comparisons is now technically feasible, as AI speech technology has become increasingly capable of producing expressive and emotive speech (Kolekar et al., 2024). Moreover, established methodological workflows exist for constructing prosodically varied stimuli while holding speech content constant, such as manipulating confident vs. doubtful prosody to convey speakers' feeling of knowing regarding the subject being discussed (Jiang & Pell, 2015, 2017; Swerts & Krahmer, 2005).

### 1.4. Ecological validity: should studies compare human speech directly to its AI clone, or have AI generate new content?

Currently, many controlled perceptual studies that directly compare human speech with AI-generated or AI-cloned versions and explicitly ask participants to detect deepfakes consistently conclude that detection is difficult, although detection improves when the cloned voice belongs to a familiar person (Rosi et al., 2025). Still, accuracy ranges from approximately 60% (Barrington et al., 2025; San Segundo et al., 2025) to 73% (Mai et al., 2023), suggesting that even with awareness and focused attention, listeners struggle to reliably distinguish AI from human voices. However, these referenced studies typically have AI voices reproduce identical utterances as human recordings.

Because voice cloning systems train on speakers' acoustic patterns, reproducing the same sentences may favor AI by reducing natural variability and eliminating content-driven detection cues. In contrast, real-world AI voice applications (e.g., Google Maps) may generate novel utterances not in training data. Same-text paradigms may therefore overestimate AI-human indistinguishability. To address this ecological validity concern, we adopted AI-generated voices producing novel utterances rather than reproducing human speech content, following a rationale similar to Lavan et al. (2025).

### 1.5. Gaps also remain in speaker identity processing research

In parallel with the experimental design considerations above, literature on speaker identity processing and within-speaker variation also warrants the incorporation of prosodic cues. Just as we might have a mental image of a person's typical appearance, listeners form mental templates of how someone's voice should sound and use these templates to identify individual speakers (Latinus & Belin, 2011; Latinus et al., 2013; Papcun et al., 1989). However, existing explanations of prototype-based processing in voice perception have been challenged by both theoretical perspectives and empirical findings.

Theoretically, Lavan and McGettigan (2023) offered an alternative perspective on the functional necessity of it. They argued that the traditional two-step mechanism of calculating deviant features relative to a prototype and then comparing them to stored representations adds unnecessary complexity. Instead, they proposed a simplified mechanism that directly compares acoustic features to stored representations. For example, when hearing a new voice, rather than first asking how this voice differs from the average voice and then determining if these differences match any stored patterns, listeners could directly ask if this voice matches any stored voice patterns.

Meanwhile, a literature review on speaker identity processing has established that within-speaker variability can be harmful to successful voice identity perception (Lavan, Burton, et al., 2019), yet current empirical findings show conflicting results that warrant further investigation using controlled experimental paradigms. Within-speaker variability has been operationalized through two primary approaches: acoustic manipulation of physical voice parameters such as vocal tract length (VTL) and fundamental frequency (F0) (Lavan, Knight, & McGettigan, 2019), and natural prosodic variation conveyed through emotional expressions such as fearful vs. neutral speech (Xu & Armony, 2021). However, these studies yielded divergent findings. Lavan, Knight, and McGettigan (2019) found evidence supporting prototype-based processing, with listeners rating never-heard identity averages as most familiar despite training on acoustically variable examples only. In contrast, Xu and Armony (2021) observed chance-level recognition performance when prosody was mismatched between encoding and test phases, demonstrating the extremely harmful effects of prosodic variation on voice identity processing. However, we presume that the harmful effects of natural prosodic variation may be less severe, as Xu and Armony (2021) did not require explicit talker name memorization like Lavan, Knight, and McGettigan (2019) did, which would have promoted better encoding and deeper identity memory to facilitate later recognition.

### 1.6. Why speaker discrimination is an ideal paradigm

To address the above gaps in speaker identity processing, we employed an AX discrimination paradigm, where listeners judge whether two sequentially presented voice samples belong to the same or different speakers. This paradigm offers four methodological advantages.

First, AX discrimination tasks may access more shallow levels of identity processing compared to recognition tasks. Rather than requiring full identity extraction and conscious access to "who is speaking," discrimination tasks can operate on implicit acoustic similarity comparisons at more basic processing stages, making them less susceptible to ceiling effects and better suited for isolating the specific effects of prosodic variation as within-speaker variation (Chen & Jiang, 2024; Levi, 2019; Winters et al., 2008).

Second, this task is robust to conditions with substantial within-speaker variation, including challenging scenarios such as unfamiliar languages and time-reversed speech (Fleming et al., 2014). Since prosodic variation in pragmatically marked speech (Jiang & Pell, 2017) represents a relatively moderate form of acoustic change compared to highly expressive voices (Lavan, Burston, et al., 2019), it produces quantifiable effects on discrimination performance rather than completely disrupting identity processing.

Third, this discrimination paradigm allows us to examine whether

cognitive discrimination mechanisms mirror computational speaker verification approaches (Lavan & McGettigan, 2023). Computational systems extract acoustic embeddings and calculate similarity metrics (e.g., Euclidean distance) to determine whether audio samples belong to the same or different speakers. By computing acoustic distances between voice samples using Wav2Vec2 embeddings (Baevski et al., 2020), we can test whether listeners' discrimination performance systematically corresponds to computational distance metrics. If human discrimination mirrors computational verification, we would expect to observe systematic relationships between acoustic distance and both discrimination accuracy and reaction times.

Fourth, employing the same discrimination task used by Barrington et al. (2025) enables direct comparison of our identity discrimination results while simultaneously allowing us to examine how prosodic variation influences identity differentiation.

### 1.7. The current study

Our Experiment 1 aims to address current concerns that listeners are reportedly poorly equipped to detect AI voice clones from short audio clips. For Experiment 1, we went beyond existing approaches that directly compare human source speech with its AI-cloned version. We first obtained human audio and created AI-cloned voice avatars, then had the AI avatars generate entirely new sentences, with the original human speakers returning approximately one month later to produce the same sentences with confident and doubtful prosody. We analyzed acoustic features and recruited 48 listeners to rate humanlikeness and perceived confidence, addressing three research questions:

- RQ1: Do AI-cloned speakers and their source human speakers maintain comparable acoustic identity (F0) across prosodic contexts? We compared AI avatars' novel utterances against human speakers' newly produced speech to avoid circular comparisons.
- RQ2: Can listeners perceptually distinguish between human and AI voices based on humanlikeness ratings?
- RQ3: Can AI voices produce prosodic variations (confident vs. doubtful) perceptually comparable to human expressions, as measured by vocal confidence ratings?

Our Experiment 2 aims to address current concerns that listeners tend to perceive AI-cloned and human speech as sharing the same identity when both use neutral prosody (Barrington et al., 2025). We selected audios from Experiment 1 to form a validated stimulus set and asked participants to complete an AX discrimination task, judging whether two audio samples were from the same speaker. To examine whether listeners' voice identity discrimination mirrors computational speaker verification approaches, we modeled listeners' performance against Wav2Vec 2.0 acoustic distance representations of the audio pairs presented in each trial. This acoustic distance calculation method has demonstrated strong capacity for capturing speaker-specific identity information and emotional prosody (Baevski et al., 2020; Pepino et al., 2021). Our Experiment 2 employed a 2 (Source: Human vs. AI) × 2 (Speaker Identity: Same vs. Different) × 2 (Prosody Consistency: Consistent vs. Inconsistent) design. We aim to address three research questions:

- RQ4: How does prosodic consistency affect discrimination performance in within-source pairs (human-human, AI-AI) vs. cross-source pairs (human-AI, AI-human)?
- RQ5: Which type of within-speaker variation more significantly impacts identity discrimination: prosodic variation or AI-human source differences?
- RQ6: Do listeners apply computational logic similar to speaker verification systems? Specifically, does the acoustic distance between audio pairs predict listeners' discrimination performance?

An overview of the study procedure from Experiment 1 to Experiment 2 is illustrated in Fig. 1, which will be referenced throughout the Method section.

## 2. Experiment 1: Acoustic and perceptual comparison of expressive AI and human speech

Experiment 1 had four objectives: assessing whether AI-cloned speakers and their source human speakers maintain comparable acoustic identity (F0) across prosodic contexts (RQ1), obtaining humanlikeness ratings to test whether listeners can perceptually distinguish between human and AI voices (RQ2), obtaining vocal confidence ratings to test whether AI voices produce prosodic variations (confident vs. doubtful) perceptually comparable to human expressions (RQ3), and selecting stimuli for Experiment 2.

### 2.1. Materials and methods for experiment 1

The study was approved by the Ethics Committee of the Institute of Language Sciences at Shanghai International Studies University. Participants signed an informed consent form before participating in the experiments.

#### 2.1.1. Participants

**Speech production participants.** Participants consisted of 24 native Chinese speakers (12 females, 12 males) from Shanghai International Studies University. All speakers had experience in acting, speech, or music training and demonstrated Mandarin proficiency with Putonghua Proficiency Test scores ranging from 87 to 91 out of 100. Male participants ($M$ = 24.55 years, $SD$ = 2.09; education: $M$ = 18.55 years, $SD$ = 1.79; height: $M$ = 174.02 cm, $SD$ = 20.64) and female participants ($M$ = 22.30 years, $SD$ = 2.54; education: $M$ = 18.20 years, $SD$ = 2.59; height: $M$ = 165.24 cm, $SD$ = 11.42) reported no history of speech, hearing, neurological, or psychiatric impairments. Speakers received compensation of 60 RMB per hour.

**Perceptual rating participants.** Participants consisted of 48 independent listeners (24 males: $M$ = 21.30 years, $SD$ = 0.48; education: $M$ = 18.17 years, $SD$ = 0.44; 24 females: $M$ = 21.30 years, $SD$ = 0.48; education: $M$ = 17.00 years, $SD$ = 0.63). Participants were native Chinese speakers recruited from Shanghai International Studies University, Shanghai University of Engineering Science, and Donghua University. None reported any history of speech, hearing, neurological, or psychiatric impairments. Listeners received compensation of 50 RMB per hour.

#### 2.1.2. Stimuli preparation

**Textual content.** The text to be expressed consisted of 123 Mandarin Chinese sentences of controlled length ($M$ = 7.58, $SD$ = 1.45 characters). A key advantage of Mandarin Chinese is that confident and doubtful expressions can be achieved through prosodic modification alone without altering sentence structure. For example, "她准备结婚了" could express confident intention (She is preparing to get married.) or uncertain questioning (Is she preparing to get married?).

**Human recording and AI voice cloning.** This step followed established procedures for recording expressive confident and doubtful speech (Chen & Jiang, 2023; Jiang & Pell, 2017). Following Huawei Nova 9's *Celia* system requirements, 24 speakers were recruited to produce three versions of 15 sentences under confident, doubtful, and neutral prosody conditions. Each speaker's three prosodic versions were then input into the *Celia* system to generate AI voice avatars that captured the individual speaker's indentity across different prosodic conditions. Neutral prosody was included for avatar training but excluded from subsequent experiments, as confident and doubtful prosodies provide more distinct acoustic differences (particularly in F0 patterns) with greater perceptual discriminability for our experimental manipulation (Jiang & Pell, 2017).

**AI generation and human replication.** The AI voice models,

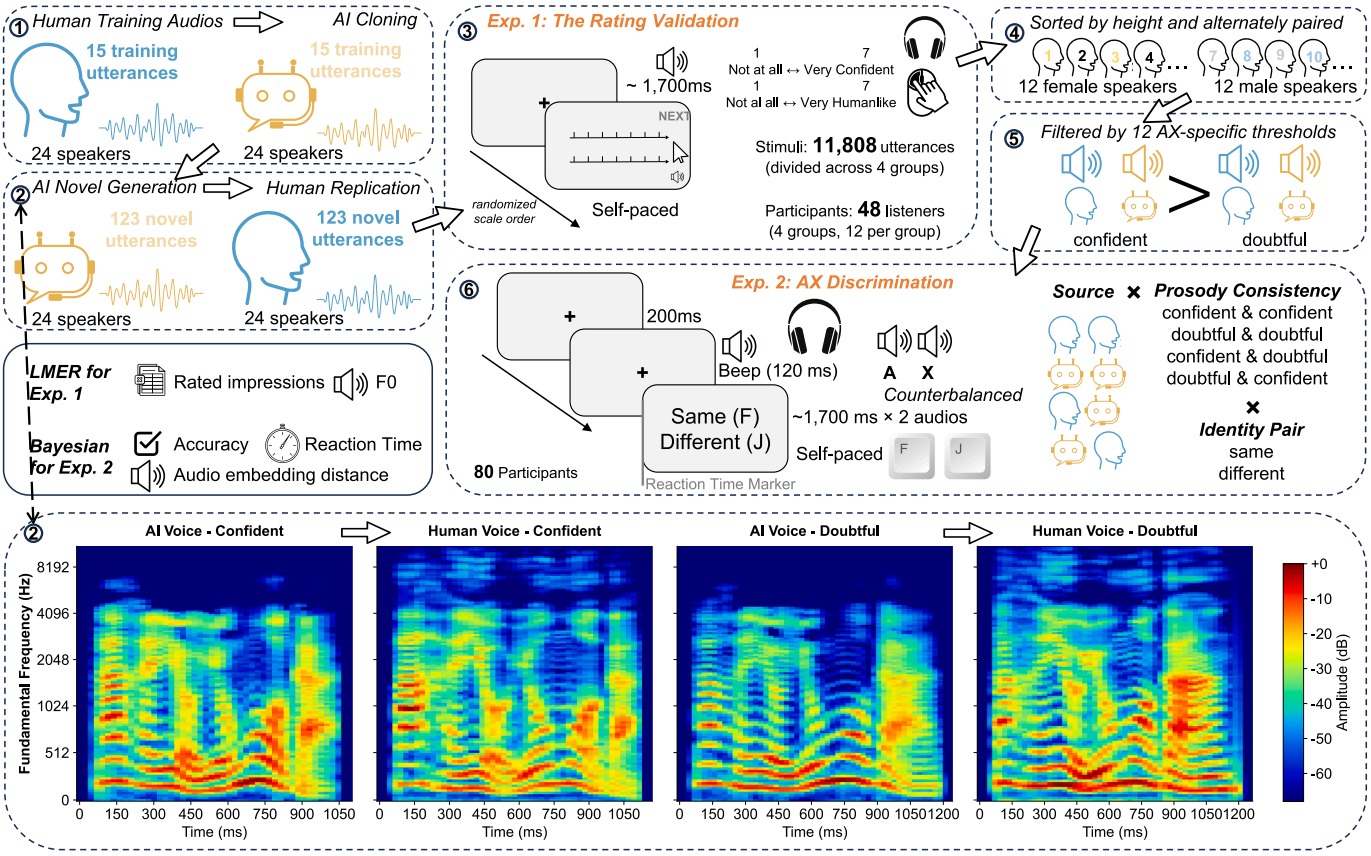


**Fig. 1.** ***Experimental Steps Diagram.*** Study methodology: **(1)** 24 speakers recorded 15 utterances in confident/doubtful/neutral prosody for AI model training. **(2)** AI models trained on confident/doubtful human speech generated 123 novel utterances. Original speakers then returned to the laboratory and, upon hearing each AI-produced confident/doubtful audio, were instructed to disregard any unnaturalness and produce the corresponding prescribed confident/doubtful sentences according to their human intuition. An example mel spectrogram for the sentence "她/他 (both pronounced "tā") 很有幽默感" [He/She has a great sense of humor] is shown. **(3)** 48 listeners rated 11,808 clips on perceived humanlikeness and confidence using 7-point scales. **(4)** and **(5)** 24 speakers were grouped into 12 pairs, and clips were filtered by AX-specific thresholds. **(6)** 80 participants completed AX discrimination tasks, judging whether paired clips came from the same speaker. For data analysis, linear mixed-effects models examined perceived ratings and F0 characteristics in Experiment 1, while Bayesian modeling analyzed accuracy and reaction time data using experimental factors and acoustic distance as predictors in Experiment 2.

comprising 24 speakers across three prosodic conditions (72 total models), were then used to generate readings of 123 novel sentences. Only confident and doubtful prosodic conditions were used for subsequent experiments to maximize prosodic contrast for the discrimination task. Audio files were segmented into individual recordings, resulting in a total of 5904 separate audio files (123 sentences × 2 prosodic conditions × 24 speakers). The original 24 speakers were invited back to the laboratory approximately one month later, using an identical recording setup as the initial session. Participants viewed sentence text on screen with prosodic condition prompts displayed in the upper left corner. Participants were informed that they would hear AI-generated versions of their own voices but were instructed to disregard any potential unnaturalness and instead produce the text naturally, according to the specified prosodic condition from their perspective as human speakers. The resulting 5904 human-produced sentences were manually segmented using *Praat 6.2.09* (Boersma & Weenink, 2021).

### 2.1.3. Independent rating

Participants wore Bose QuietComfort QS45 noise-cancelling headphones and completed the task using *PsychoPy* version 2022.2.4 (Peirce et al., 2019). All 11,808 audio clips were normalized to -30 dBFS with a sampling rate of 44,100 Hz. Each participant rated 2072 clips across three laboratory sessions (approximately 120 min each), with each clip rated by 12 listeners. Participants listened to individual audio clips and rated them on two 7-point Likert scales: humanlikeness (1 = *not at all humanlike*, 7 = *very much humanlike*) and confidence level (1 = *not at all confident*, 7 = *very confident*) (Jiang & Pell, 2017; Mullennix et al., 2003). Participants could replay each clip once in addition to the initial automatic playback.

To minimize fatigue, participants completed ratings across three laboratory visits (Sessions A, B, and C), with each session divided into four blocks. Between blocks, participants were given mandatory breaks around 10 min, though they could also self-pace their progress within blocks by taking breaks between individual trials as needed. Stimuli were presented in randomized order within each block, and the vertical positions of the two rating scales (humanlikeness above confidence, or vice versa) were randomized on each trial to prevent systematic order effects.

### 2.1.4. Post-hoc quality assessments

First, inter-rater reliability was assessed using intraclass correlation coefficients (ICC) with a two-way random effects model (`psych::ICC()`) (Revelle, 2025), calculated on a random sample of 1000 stimuli. We calculated ICC(2,1) for single raters and ICC(2,k) for average measures. Second, variance component analysis was conducted using linear mixed-effects models `lmer(rating ~ Source * Confidence_level + (1|Speaker) + (1|Participant_ID) + (1|Item)` to decompose total variance into participant-related, stimulus-related, and residual components. We examined participant-level rating patterns for insufficient variance (SD < 0.5) or excessive reliance on extreme values (>80% at the scale endpoints). Third, fatigue effects were tested by including block number (1-12) as a fixed-effect predictor in linear

mixed-effects models `lmer(rating ~ Source * Confidence_level + Overall_block + (1|Speaker) + (1|Participant_ID) + (1|Item)`, with session-level changes also examined descriptively.

First, inter-rater reliability was moderate for single raters (humanlikeness: ICC(2,1) = 0.55, 95% CI [0.52, 0.57]; confidence: ICC(2,1) = 0.54, 95% CI [0.52, 0.56]) but excellent for averaged ratings (ICC(2,k) = 0.98 for both scales). Second, variance component analysis showed that participant-related variance accounted for 8.7% (humanlikeness) and 4.2% (confidence) of total variance, with most variance attributable to stimulus differences (residual: 85.8% and 89.0%). All 48 participants demonstrated appropriate rating variance (SD range: 1.50-2.93 for humanlikeness, 1.33-2.71 for confidence). Third, fatigue analysis revealed negligible block effects (humanlikeness: $\beta = 0.02$, SE = 0.001; confidence: $\beta = 0.007$, SE = 0.001), with stable ratings across sessions (Session A to C change: 0.17 points for humanlikeness, 0.02 points for confidence on 7-point scales, representing less than 3% of the scale range). Overall, no systematic confounds were detected.

### 2.2. Data analysis

**Audio: F0 analysis.** To avoid potential interference from consonantal segments that can compromise F0 tracking accuracy in Mandarin Chinese vocal confidence expressions (Feng & Jiang, 2024), all vowel segments were extracted from each sentence and concatenated into vowel-only sequences using the *extractvowels* plugin from the *Praat Vocal Toolkit* (Corretgé, 2024). F0 analysis was then conducted on these vowel-only sequences using *Praat*'s autocorrelation method with sex-specific pitch ranges (75–300 Hz for males, 100–500 Hz for females) (Boersma & Weenink, 2021). F0 mean values from paired audio samples (where both human and AI-generated versions existed for the same speaker and item) were analyzed using mixed-effects linear models (Bates et al., 2015) in R version 4.3.3 (R-Core-Team, 2024) using RStudio (Build 402) (Posit-team, 2024): `f0_mean_hz ~ Source * Confidence_level * sex + (1|Speaker) + (1|Item)`. Post hoc comparisons used the emmeans package (Lenth et al., 2021), and Cohen's *d* effect sizes quantified practical significance.

**Audio: Representation visualization.** Using the pre-trained Wav2Vec2-base model (Baevski et al., 2020), contextual embeddings were extracted from each of the 768 audio samples by averaging the last hidden states across time steps. To reduce computational complexity and noise, principal component analysis (PCA) was first applied using scikit-learn 1.6.1 (Pedregosa et al., 2011) to reduce dimensionality from 768 to 50 components (explaining 88.7% of variance). Subsequently, t-distributed Stochastic Neighbor Embedding (t-SNE) with perplexity =

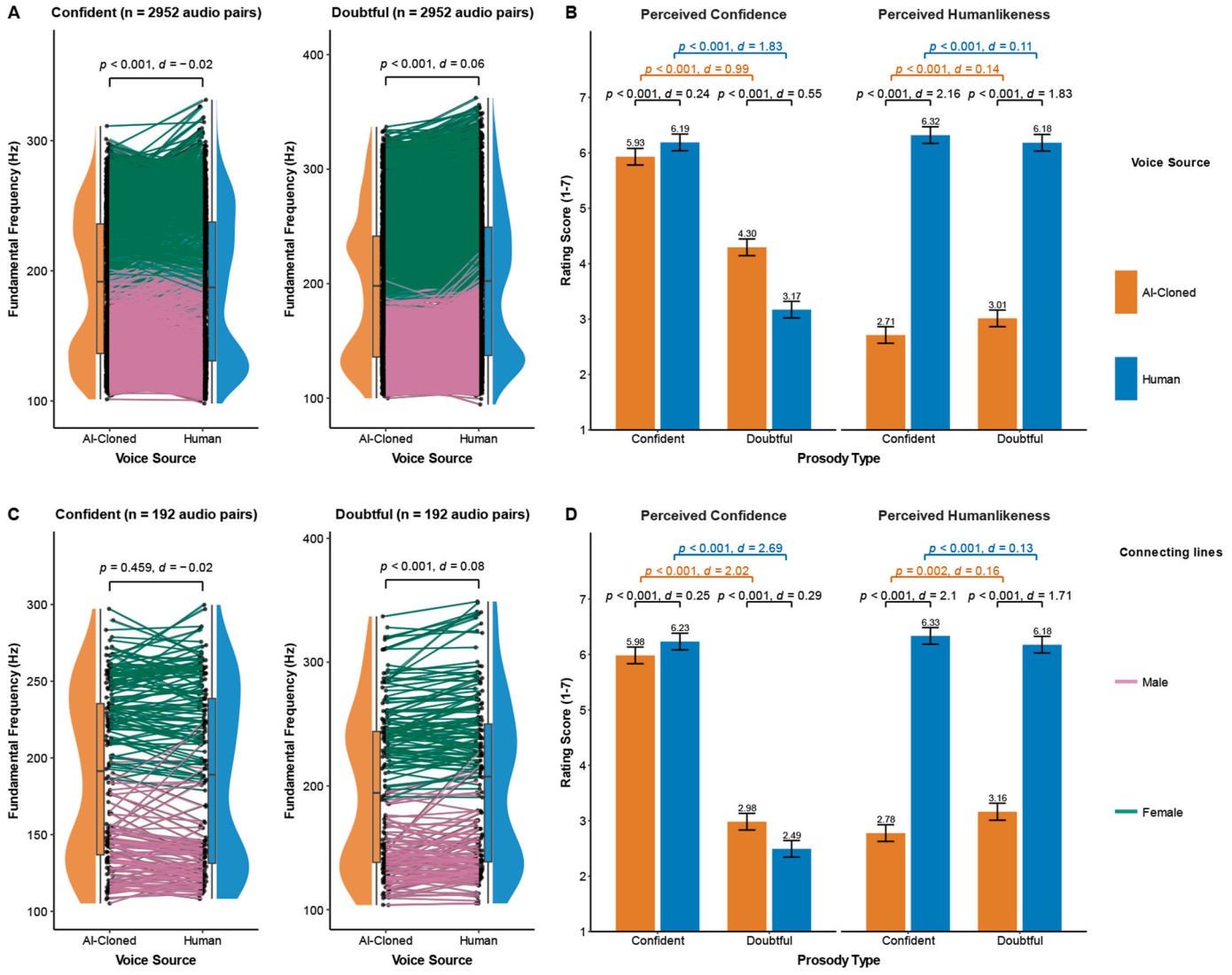


**Fig. 2.** ***Comparison of F0 Characteristics and Perceptual Ratings Between AI-Cloned and Human Voices.*** Panels **A** and **C** show F0 distributions, while Panels **B** and **D** show perceptual ratings. Panels **A** and **B** display results from the complete corpus of 11,808 audio clips, while Panels **C** and **D** show results from the filtered subset used in Experiment 2 (768 audio clips, selected based on prosodic rating thresholds).

50 was applied to project the PCA-reduced features into a 2D space for visualization (see Supplementary Analysis 1; Figure S1). This two-stage approach serves as a supplementary exploratory tool to illustrate clustering patterns across voice source, prosody, talker sex, and specific speakers, with the caveat that dimensionality reduction may introduce distance distortions that do not fully preserve the original high-dimensional relationships (Pepino et al., 2021).

**Ratings.** To examine main effects of voice source (AI-generated vs. human) and prosody type (confident vs. doubtful) as well as their interaction, we conducted statistical analyses using separate linear mixed-effects models for each rating type: `Perceived_humanlikeness ~ Source × Confidence_level + (1|Speaker) + (1|Participant_ID) + (1|Item)` and `Perceived_confidence ~ Source × Confidence_level + (1|Speaker) + (1|Participant_ID) + (1|Item)`, with post hoc comparisons performed using the aforementioned *emmeans* package (Lenth et al., 2021) and Cohen's *d* effect sizes calculated.

### 2.3. Results for experiment 1

In both datasets, F0 differences between AI-generated and human voices were minimal to negligible. The complete corpus (11,808 clips; Fig. 2A) showed very small differences in both confident ($p = .001$, $d = -0.02$) and doubtful conditions ($p < .001$, $d = 0.06$). Similarly, the experimental subset (768 clips; Fig. 2C) revealed no significant F0 differences in confident conditions ($p = .459$, $d = -0.02$) and only small differences in doubtful conditions ($p < .001$, $d = 0.08$). Effect sizes were interpreted by conventional standards, where 0.2-0.3 were small, 0.5 medium, and ≥0.8 large (Cohen, 2013).

**For confidence ratings**, both voice types demonstrated expected prosodic effects, with confident prosody consistently rated higher than doubtful prosody across datasets (complete corpus – Human: 6.19 vs. 3.18; AI-generated: 5.93 vs. 4.30; experimental subset – Human: 6.25 vs. 2.50; AI-generated: 6.00 vs. 3.01). Between voice sources, human voices received higher confidence ratings than AI-generated voices in confident conditions, while AI-generated voices were rated as more confident in doubtful conditions across both datasets, though with small to moderate effect sizes (confident conditions: $d = 0.24$-$0.25$; doubtful conditions: $d = 0.29$-$0.55$). See Fig. 2B and D.

**For humanlikeness ratings**, human voices were consistently perceived as substantially more humanlike than AI-generated voices across all conditions, despite the acoustic similarity in F0. Large effect sizes were observed in both datasets and prosodic conditions (complete corpus: confident $d = 2.16$, doubtful $d = 1.83$; experimental subset: confident $d = 2.10$, doubtful $d = 1.71$). Human voices maintained consistently high humanlikeness ratings across prosodic conditions (complete corpus: 6.33 confident, 6.19 doubtful; experimental subset: 6.33 confident, 6.18 doubtful), while AI-generated voices received markedly lower ratings (complete corpus: 2.71 confident, 3.00 doubtful; experimental subset: 2.77 confident, 3.14 doubtful). Interestingly, AI-generated voices were rated as slightly more humanlike when produced with doubtful prosody compared to confident prosody, though the prosodic condition had minimal overall impact on humanlikeness ratings within each voice type ($d < 0.16$ across all comparisons). This pattern suggests that humanlikeness perception is primarily driven by voice source, with doubtful prosody potentially masking some artificial qualities in AI-generated speech. See Fig. 2B and D.

### 2.4. Discussion of experiment 1 results

Experiment 1's three research questions yielded clear answers: AI-cloned speakers and their source human speakers maintained comparable acoustic identity (F0) across prosodic contexts, even when both produced novel content approximately one month after AI avatar training (RQ1); listeners consistently distinguished human from AI voices based on humanlikeness ratings (RQ2); and AI voices produced prosodic variations (confident vs. doubtful) that were perceptually comparable to human expressions, as evidenced by comparable confidence ratings (RQ3). Meanwhile, Experiment 1 revealed that prosody had dual effects: it enhanced listeners' ability to identify AI speech as less humanlike compared to past monotone studies (Barrington et al., 2025), yet within AI voices, doubtful prosody was paradoxically rated as more humanlike than confident prosody.

**Why did our study show enhanced AI detection compared to previous research?** We found that while AI-generated and human voices showed comparable F0 characteristics, listeners consistently rated human voices as substantially more humanlike, indicating that F0 is not necessarily a reliable cue for distinguishing AI from human voices. Listeners likely relied on more holistic evaluation cues, with prosody being one potential factor. Meanwhile, our humanlikeness finding is different from Barrington et al. (2025), who found modest accuracy in binary real/AI classification (69% for real voices, 63% for AI voices). Two methodological factors may contribute to this difference. First, stimuli design: participants in Barrington et al. (2025) reported minimal reliance on emotional cues, likely because their experimental design controlled for emotional encoding differences between AI-cloned and human voices from the outset, limiting access to the intonation and emotional cues that listeners typically use to identify AI voices (Kühne et al., 2020; Rodero & Lucas, 2023). Second, measurement approach: continuous ratings may capture discrimination abilities that binary classification obscures. Bakkouche et al. (2025) similarly demonstrated that continuous scales effectively captured naturalness differences between AI systems. Thus, expressive prosody and continuous ratings may explain enhanced AI source detection in our study.

**Listeners may not expect AI to produce certain prosodic sentences.** We also found that listeners perceived AI voices as slightly more humanlike in the doubtful than in the confident prosody condition, though effect sizes were small ($d = 0.13$-$0.16$). This tendency can be explained by listeners associating AI-generated speech with monotone delivery (Cohn et al., 2021; Di Cesare et al., 2022; Gampe et al., 2023; Mogali et al., 2024), while confident prosody may more closely resemble neutral, monotone speech patterns compared to doubtful prosody (Jiang & Pell, 2017). Listeners may have lower expectations for AI systems to produce complex prosodic variations characteristic of doubtful speech, making such expressions seem more authentically human when encountered. This may support the human-likeness-based naturalness framework by Nussbaum et al. (2025), which assumes perceptual boundaries between human and non-human voice categories. Our finding that doubtful prosody enhanced humanlikeness suggests that listeners may utilize expectation mechanisms when evaluating human-specific prosody, making AI voices appear more humanlike.

**Parallels with accent perception?** Just as listeners maintain categorical distinctions between accent groups regardless of prosodic variation (Jiang et al., 2020; Lam et al., 2025; Mauchand & Pell, 2022c), our participants consistently rated human voices as more humanlike than AI voices across all prosodic conditions. This suggests that AI-human distinctions function as categorical social boundaries (Nussbaum et al., 2025) similar to those observed in accent perception. However, a critical question remains unanswered in our current design: which is detected first, the human vs. AI distinction or the different prosodic levels? Since prosodic cues are decoded later in the utterance (after the immediate early milliseconds of acoustic processing), and given that similar temporal precedence is observed in accent detection where early acoustic cues influence subsequent prosodic interpretation (Jiang et al., 2020), we presume that AI-human categorization should emerge earlier than prosodic processing. So far, this temporal sequence has not been tested using electrophysiological methods. If AI-human categorization indeed precedes prosodic processing at early sensory stages, this would provide strong neurophysiological evidence supporting the hypothesis that human-AI distinctions operate through accent-like categorical mechanisms.

## 3. Experiment 2: effects of within-speaker variation on voice identity discrimination

Building on Experiment 1's validated stimuli, Experiment 2 addressed RQ4, RQ5, and RQ6 through an identity discrimination task. We examined how prosodic consistency affected discrimination performance in within-source vs. cross-source pairs (RQ4), which type of within-speaker variation (prosodic variation vs. AI-human source differences) more significantly impacted identity discrimination (RQ5), and whether listeners' discrimination performance mirrored computational speaker verification logic by testing relationships between acoustic distance and performance (RQ6). Experiment 2 followed the same ethical approval and consent protocols as Experiment 1.

### 3.1. Materials and methods for experiment 2

#### 3.1.1. Participants

A total of 80 native Chinese speakers (40 females, 40 males) participated in the perception experiment. Female participants ($M$ = 22.18 years, $SD$ = 2.37; education: $M$ = 17.88 years, $SD$ = 2.13) and male participants ($M$ = 23.68 years, $SD$ = 2.78; education: $M$ = 17.95 years, $SD$ = 2.91) were all university students. None of the participants reported any hearing, speech-motor, neurological, or psychiatric impairments. Participants received compensation of 50 RMB per hour.

We adopted a Bayesian sequential learning approach where sample adequacy is assessed through evidence accumulation and parameter estimation precision instead of predetermined power thresholds (Kruschke, 2015; Wagenmakers et al., 2018). Across 80 participants, we obtained 30,624 valid observations, which yielded stable posterior estimates and satisfactory convergence diagnostics (Vehtari et al., 2021) in our final Bayesian models (accuracy: $\widehat{R}$ = 1.006; reaction time: $\widehat{R}$ = 1.003). Posterior intervals were generally narrow and thus suitable for substantive interpretation, with the exception of spline effects, whose wider intervals are expected given their flexibility in modeling nonlinear relationships. See Tables S4 and S6.

#### 3.1.2. Stimuli

Stimuli were selected from Experiment 1 through a systematic filtering process (Step 4 in Fig. 1). First, the 24 speakers were divided into 12 pairs based on biological sex and height similarity to control for perceived vocal size, as height-related differences in vocal tract length could make speakers sound markedly different in body size (Pisanski et al., 2014). This pairing strategy balanced task difficulty by preventing overly obvious vocal size disparities while avoiding excessive similarity.

Second, audio clips were filtered based on prosodic rating thresholds. Items were retained only if, within each AX group, the minimum confident rating across 12 listeners exceeded the minimum confident threshold, and the maximum doubtful rating fell below the maximum doubtful threshold. Thresholds were defined independently for each group to ensure that confident prosody was consistently rated as more confident than doubtful prosody across both human and AI speech sources.

Following this filtering process, 12 AX groups were established, with each speaker pair assigned different text items to ensure no repetition across groups. Each speaker contributed 8 items, with items kept consistent within AX groups but varying across different group combinations. Some text items were shared across multiple groups, and the assignment process across all groups used 46 distinct sentences in total (see Table S1), which maintained controlled-length characteristics ($M$ = 7.50, $SD$ = 1.47 characters), comparable to the overall corpus ($M$ = 7.58, $SD$ = 1.45). This selection process yielded 768 total audio stimuli (24 speakers × 2 sources × 2 prosodic conditions × 8 items per speaker) distributed across 12 AX groups for the discrimination experiment.

#### 3.1.3. Procedures

Participants were seated in a sound-attenuated laboratory wearing Bose QuietComfort QS45 noise-cancelling headphones and instructed to make keyboard responses using designated keys. The task required participants to determine whether paired speech clips were produced by the same or different speakers. Prior to the experiment, participants were explicitly instructed: "*Please concentrate on identifying whether the speakers are the same, and ignore whether they sound natural. Even if the intonation varies or the audio sounds somewhat unnatural, if you feel the voice timbre belongs to the same person, you should still judge them as the same speaker.*" Each AX discrimination trial began with a 200ms beep, followed by two consecutive audio clips (Sound A and Sound X; Fig. 1). Participants indicated their judgments by pressing *F* or *J*, with key-response mappings counterbalanced across participants. Each participant completed 384 trials organized into four blocks of 96 trials each, with self-paced breaks between blocks.

#### 3.1.4. Data analysis

**Acoustic distance computation.** Feature extraction and distance calculations were performed in Python 3.13.2 (Rossum & Drake, 2009) using PyTorch 2.7.1 (Paszke et al., 2019) and the Transformers library 4.52.4 (Wolf et al., 2020). Using the Wav2Vec2-base model (Baevski et al., 2020), we extracted 768-dimensional feature vectors by averaging the final-layer outputs over time for each audio sample. Audio files were preprocessed to a 16 kHz mono format prior to feature extraction. Euclidean distances were calculated using SciPy 1.15.3 (Virtanen et al., 2020) between paired feature vectors to quantify acoustic dissimilarity, with larger distances indicating greater acoustic differences. These continuous distance measures served as predictors in Bayesian mixed-effects models.

**Data preprocessing.** Each participant completed 384 trials, except for one participant who had 288 trials due to data loss but was retained in the analysis. A total of 30,624 trials from 80 participants were included in the final dataset. Three variables were log-transformed due to skewness: reaction time (from 13.796 to −0.050), F0 difference (from 1.634 to −0.612), and Wav2Vec2 acoustic distance (from 2.202 to 1.308). Three experimental factors were coded for analysis: source pair (human-human, human-AI, AI-human, AI-AI, with human-human as reference category), speaker identity (same vs. different speakers), and prosodic consistency (consistent vs. inconsistent prosody). Acoustic distance showed low correlation with speaker identity ($r$ = −0.134, $t$ = −23.62, $p < 0.001$) and minimal variance inflation (*Variance Inflation Factor* = 1.018), confirming minimal multicollinearity concerns for this predictor. Speaker identity and prosodic consistency used Helmert coding (−1, +1), where regression coefficients represent the effect of a two-unit change from one category to the other.

**Bayesian modeling.** Mixed-effects models were fitted using a phase-wise approach, progressing from basic main effects through interactions to acoustic distance integration. Model selection used WAIC (Watanabe-Akaike Information Criterion) across phases (Tables S2–S5). The final models employed Bayesian mixed-effects regression with nonlinear acoustic distance effects modeled using smooth functions. The accuracy model used logistic regression: `response_correctness ~ source_pair_c × speaker_identity_c × prosodic_consistency_c + s(log_acoustic_distance) + (1|participant) + (1|AX_item_combination)`. The reaction time model used linear regression with log-transformed response times: `log_reaction_time ~ source_pair_c × speaker_identity_c × prosodic_consistency_c + s(log_acoustic_distance) + (1|participant) + (1|AX_item_combination)`. Both models were fitted using the *brms* package with default priors (Bürkner, 2017). MCMC sampling used 4000 iterations (2000 warmup) across four chains with convergence confirmed by $\widehat{R} < 1.1$.

**Fixed effects estimates.** Bayesian posterior distributions for all fixed effects were visualized using forest plots (Wirtz & Pfenninger,

2024). Both accuracy and reaction time models included experimental factors (source pairing, speaker identity, prosodic consistency), their interactions, and nonlinear acoustic distance effects using smooth functions. Effect credibility was assessed using 95% highest density intervals (HDI) and region of practical equivalence (ROPE) analysis, with ROPE ranges of [−0.18, 0.18] for accuracy (log-odds scale) and [−0.1, 0.1] for reaction time (log-RT scale). Effects were classified as positive (HDI above the upper ROPE boundary), negative (HDI below the lower ROPE boundary), or not credible (HDI overlapping with ROPE).

**Posthoc analysis.** For each pairwise comparison, we computed differences between posterior prediction distributions (generated using `posterior_epred` with population-level effects only), calculated raw differences (percentage points for accuracy, milliseconds for reaction time), and Bayesian standardized effect sizes (using pooled standard deviations from log-transformed scales) for practical and cross-variable comparisons. Directional probabilities indicated the proportion of posterior samples supporting a particular direction of difference, with statistical significance marked by asterisks. For example, a directional probability of 0.999 (***) indicates that 99.9% of posterior samples support one condition performing better than the other.

**Acoustic distance effects.** The nonlinear acoustic distance effects on accuracy and reaction time were analyzed through: (1) correlation analysis examining accuracy-RT relationships across all experimental conditions, (2) performance variability analysis quantifying condition-specific response patterns, and (3) cognitive conflict zone detection identifying systematic valleys and peaks at critical acoustic distances (see Supplementary Analyses 5-7, Tables S6-S8).

### 3.2. Results for experiment 2

**Fixed effects for accuracy.** First, we found significant reductions in accuracy for human-AI ($\beta = -1.97$) and AI-human pairings ($\beta = -2.03$) relative to human-human baselines, while AI-AI pairing showed no credible difference (ROPE = 39.3%). Second, different-speaker pairs showed enhanced accuracy ($\beta = 0.69$), and inconsistent prosody reduced accuracy ($\beta = -0.60$). Third, significant interactions emerged between speaker identity and source pairings, with human-AI and AI-human showing large negative interactions. Fourth, prosodic consistency positively interacted with human-AI and AI-human pairings. Fifth, three-way interactions showed mixed evidence: Human-AI and AI-

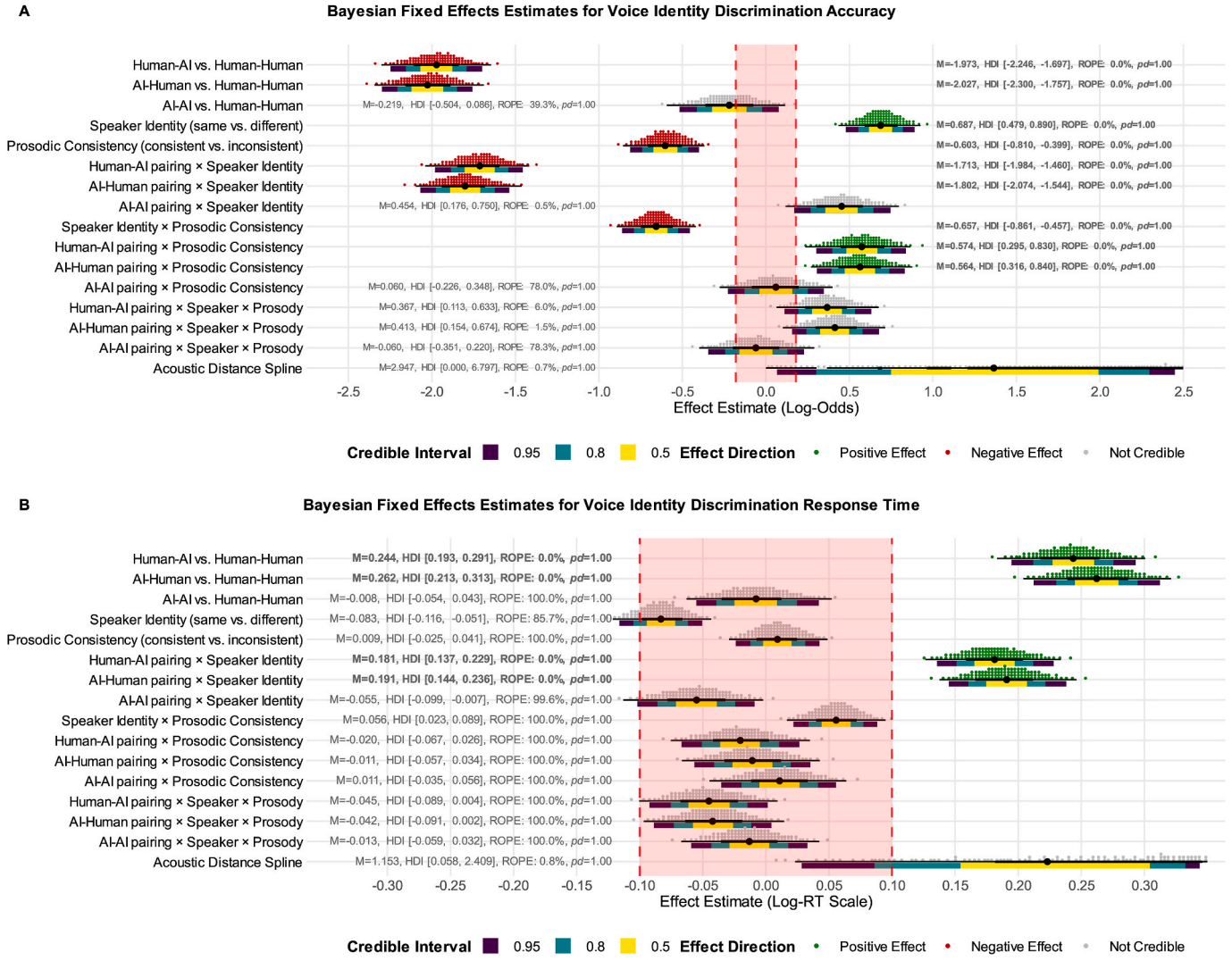


**Fig. 3.** ***Bayesian Fixed Effects Estimates for Voice Identity Discrimination Performance.*** Panel **A** shows fixed effects for discrimination accuracy; Panel **B** shows fixed effects for response time. All predictors were contrast-coded with Human-Human pairing, same speakers, and consistent prosody as baselines. Voice pairings: Human-AI = human first voice paired with AI second voice; AI-Human = AI first voice paired with human second voice; AI-AI = both voices generated by AI. The Acoustic Distance Splines capture nonlinear relationships between log-transformed acoustic similarity and performance. Positive effects (green) indicate higher performance than baseline; negative effects (red) indicate lower performance; grey effects are not credible. Effect estimates shown with 95% HDI. Red shaded regions indicate ROPE: ±0.18 log-odds for accuracy, ±0.1 log units for response time. *M* = posterior mean; *pd* = probability of direction.

human conditions demonstrated barely credible interactions (ROPE = 6.0% and 1.5%, respectively), while AI-AI conditions showed no credible interaction (ROPE = 78.3%). The acoustic distance spline effect showed substantial evidence (ROPE = 0.7%), with 99.3% of the HDI falling outside the equivalence region, indicating meaningful nonlinear acoustic similarity influences on discrimination accuracy despite technical overlap with the ROPE boundary. See Fig. 3A.

**Fixed effects for response time.** Human-AI and AI-human pairings significantly increased response times relative to human-human baselines ($\beta = 0.24$ and $\beta = 0.26$, respectively), and both showed significant positive interactions with speaker identity ($\beta = 0.18$ and $\beta = 0.19$, respectively), indicating that cross-source conditions create additional processing difficulty when speakers are the same. The acoustic distance spline effect showed substantial evidence (ROPE = 0.8%), indicating meaningful nonlinear relationships between acoustic distance and response time. See Fig. 3B.

**Prosodic consistency effects on performance.** For same-speaker discrimination, prosodic inconsistency significantly reduced accuracy across all pairings: Human-Human (−8.6%, $d = 7.04$, $pd = 1.000$), AI-AI (−6.9%, $d = 6.49$, $pd = 1.000$), Human-AI (−15.6%, $d = 3.65$, $pd = 0.996$), and AI-Human (−13.7%, $d = 3.25$, $pd = 0.993$). Within-source pairings showed significantly faster responses with inconsistency (Human-Human: 56ms faster, $d = -2.12$, $pd = 0.997$; AI-AI: 51ms faster, $d = -2.02$, $pd = 0.996$), while cross-source pairings showed minimal response time changes (Fig. 4A). For different-speaker discrimination, inconsistency effects varied substantially: Human-Human showed minimal accuracy change (−1.0%, $d = -0.56$, $pd = 0.665$) but significantly faster responses (47ms faster, $d = 1.54$, $pd = 0.975$). AI-AI showed improved accuracy with inconsistency (+5.2%, $d = -1.81$, $pd = 0.927$), while cross-source pairs showed significant improvements (Human-AI: +5.9%, $d = -2.74$, $pd = 0.985$; AI-Human: +4.5%, $d = -2.15$, $pd = 0.949$) with modest response acceleration (Fig. 4B).

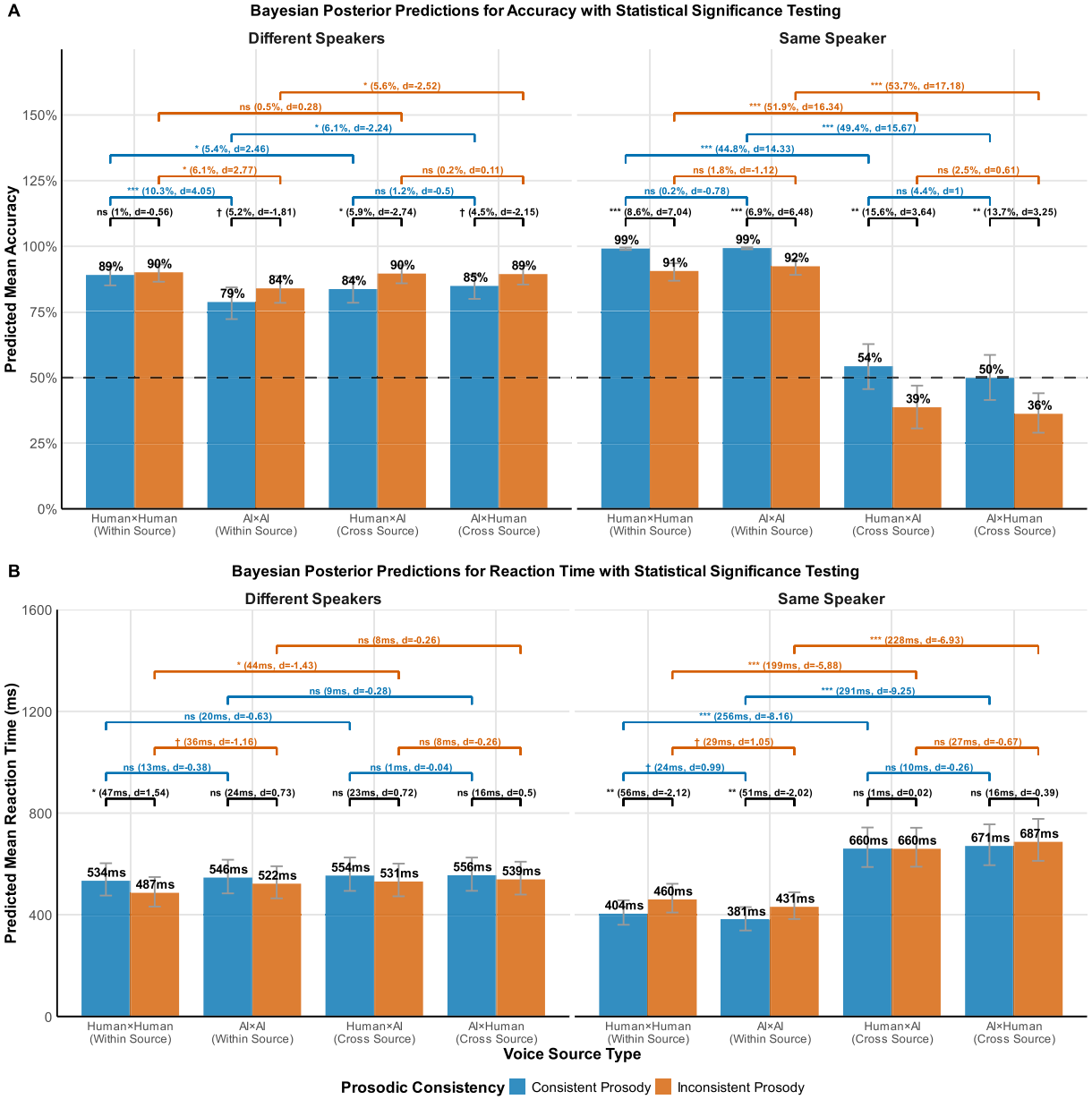


**Fig. 4.** ***Bayesian Posterior Predictions for Voice Identity Discrimination Performance.*** Panel **A** shows predicted accuracy rates; Panel **B** shows predicted reaction times across voice source pairings and speaker identity conditions. Significance levels: *** $pd > 0.999$, ** $pd > 0.99$, * $pd > 0.95$, † $pd > 0.90$, ns = nonsignificant. Parenthetical values show effect magnitude and standardized effect size (Cohen's $d$). Error bars = 95% credible intervals. The dashed line at 50% represents chance performance.

**Human-Human vs. AI-AI performance differences.** Same-speaker discrimination showed comparable accuracy (99.2% vs. 99.3%, $d = -0.78$, $pd = 0.726$) but Human-Human required significantly longer processing time (405ms vs. 381ms, $d = 0.99$, $pd = 0.903$). Different-speaker discrimination significantly favored Human-Human accuracy in both prosodic conditions (consistent: 89.1% vs. 78.8%, $d = 4.05$, $pd = 0.999$; inconsistent: 90.1% vs. 84.0%, $d = 2.77$, $pd = 0.987$) with comparable response times.

**Cross-source pairing effects and source substitution asymmetries.** Cross-source pairing order (Human-AI vs. AI-Human) showed no significant differences in accuracy or response times across conditions ($pd$ values $< 0.8$), indicating presentation order did not bias judgments. However, source substitution produced asymmetric effects: replacing human voices with AI in same-speaker conditions dramatically impaired performance (Human-Human to Human-AI: 99.2% to 54.3% accuracy, 405ms to 662ms response time), while AI-to-human substitution was less disruptive (AI-AI to AI-Human: 99.3% to 49.9% accuracy, 381ms to 672ms response time), suggesting human voice processing creates greater cognitive conflicts when substituted than AI voice processing.

We observed visual synchronization patterns between accuracy and reaction time across acoustic distance effects (Fig. 5), supported by the following three lines of evidence.

**Correlation analysis supported a unified cognitive mechanism.** Accuracy and reaction time showed strong negative correlations across all 16 experimental conditions ($r = -0.895$, 95% CI [−0.908, −0.880], $p < .001$), indicating that acoustic distance simultaneously influences both judgment speed and accuracy. This consistent negative relationship remained significant after multiple comparison corrections (see Table S6).

**Variability analysis revealed difficulty hierarchies.** The analysis demonstrated a three-tiered difficulty structure (see Table S7). Easy conditions (same-source, same-speaker) showed minimal variability (accuracy range: 3.3%), medium conditions (different speakers) exhibited intermediate variability (9.3%), while demanding conditions (cross-source, same-speaker) displayed maximum variability (22.2%). Accuracy and reaction time variability measures were highly correlated as well ($r = 0.962$, $p < .001$).

**Valley detection confirmed cognitive conflict zones with recovery patterns.** Cross-source same-speaker conditions showed accuracy valleys at acoustic distances 1.13 and 1.574-1.596, with reaction time peaks occurring at identical positions. The correlation between valley and peak positions was perfect ($r = 1.00$, $p < .001$) across eight valley-peak pairs, indicating systematic synchronization between discrimination difficulties and processing delays at specific acoustic distances (see Table S8). Beyond these conflict zones, performance showed visually apparent recovery patterns in Fig. 5, with accuracy increasing and reaction times decreasing at greater acoustic distances, suggesting that listeners shift from acoustic comparison strategies to alternative judgment mechanisms when acoustic differences become sufficiently distinct.

### 3.3. Discussion of experiment 2 results

Experiment 2 addressed RQ4 to RQ6. The answers provided were: First, prosodic variation facilitates human-AI voice discrimination, with the consequence that cross-source pairs (human-AI/AI-human) do not easily perceive speakers as the same person. Second, while both AI-human source differences and prosodic variation represent within-speaker variation, source differences had a significantly greater impact, possibly involving mechanisms similar to accent perception. Third, listeners do compute acoustic distances between audio samples to assist in identity judgments, similar to computational approaches. Still, acoustic distance is not the sole cue used when deciding whether two voices belong to the same talker.

**Within-source prosodic effects on telling speakers together.** Our findings using different-text pairs support evidence that voice identity

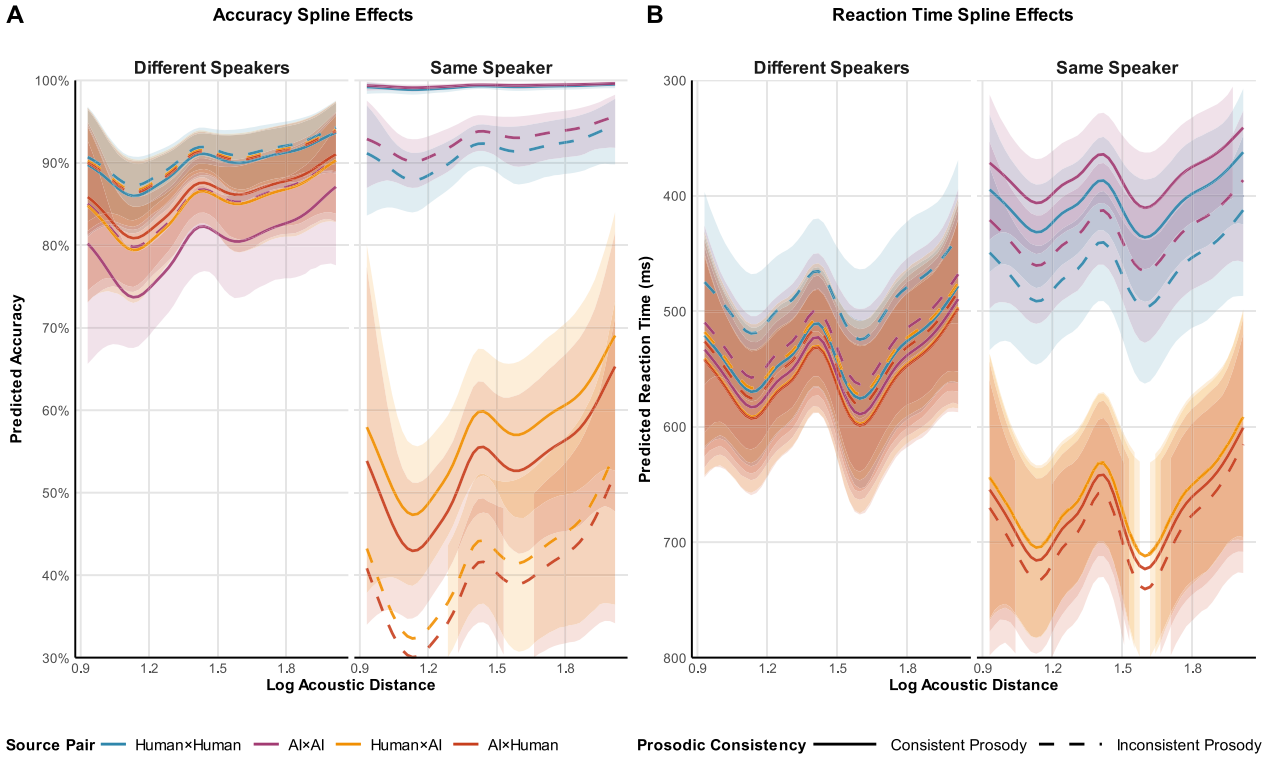


**Fig. 5.** ***Effects of Acoustic Distance on Voice Identity Discrimination Performance.*** Panel **A** shows nonlinear spline effects on accuracy across log-transformed acoustic distance values; Panel **B** shows linear effects on reaction time. Left facets show different speakers; right facets show the same speaker conditions. Colored lines represent voice source pairings. Solid lines indicate consistent prosody; dashed lines indicate inconsistent prosody. Lines represent posterior mean predictions from Bayesian; shaded ribbons show 90% credible intervals. Higher acoustic distance values indicate greater acoustic dissimilarity between paired voices.

extraction is independent of speech content (Zäske et al., 2017). Additionally, our within-source results address findings from Xu and Armony (2021), who reported chance-level performance when listeners heard prosodic mismatches between encoding and recognition phases. In contrast, our discrimination task achieved near-ceiling accuracy with prosodic pairs. This performance difference likely stems from the distinct cognitive demands of each task. Xu and Armony's recognition task imposed greater memory demands, requiring participants to encode speaker identities into long-term memory and subsequently retrieve them. Our discrimination task, instead, required participants to rapidly encode identity impressions for immediate comparison. This design bypasses the need for successful deep memory consolidation and long-term retention (Lavan, Knight, & McGettigan, 2019).

**AI voices show greater within-group homogeneity than human voices.** Our Bayesian analysis with 80 participants confirms the pattern observed in our preliminary study with 36 participants using mixed-effects logistic regression (Chen et al., 2024): while human-human and AI-AI pairs show comparable same-speaker discrimination accuracy, AI-AI pairs demonstrate greater accuracy decline for different-speaker discrimination under prosodic inconsistency. Additionally, human-human pairs required longer processing times than AI-AI pairs for same-speaker discrimination (405ms vs. 381ms), suggesting listeners process AI voices more readily as a homogeneous group with greater within-group similarity than human voices. This reflects that AI voices show greater within-group homogeneity than human voices, consistent with out-group homogeneity bias principles where members of an out-group (AI voices) are perceived as more similar to each other than members of an in-group (human voices) (Ackerman et al., 2006).

**Prosodic variation constitutes within-speaker identity variation.** Our previous acoustic analyses showed that confident prosody involves lower F0 and higher VTL value compared to doubtful prosody, and AI-cloned voices successfully replicated these acoustic patterns (Chen & Jiang, 2023; Jiang & Pell, 2017). The present study thus simulated VTL/F0 manipulations similar to Lavan, Knight, and McGettigan (2019), but using naturally produced prosodic variation rather than acoustic modifications. Our results similarly support that listeners can perceive speaker identity across within-speaker variation, and prosodic inconsistency did harm performance (Lavan, Burton, et al., 2019).

**AI-human identity sharing: Validity considerations.** Before analyzing cross-source effects, we must establish whether AI and human audio genuinely share speaker identity. Several lines of evidence support this validity: F0 analysis revealed no significant differences under confident prosody and minimal effects under doubtful prosody; cross-source discrimination achieved above-chance accuracy (>50%) under consistent prosody rather than well-below-chance performance under inconsistent prosody. While 2D clustering in Figure S1 showed apparent differences, this visualization cannot adequately reflect the actual computed distances between audio pairs, as computations relied on comprehensive 768-dimensional data demonstrating genuine comparability. Furthermore, the ultimate nonlinear utilization of acoustic distance to influence accuracy provides retrospective empirical validation for the design's effectiveness.

**AI voice as an accent-like long-term trait.** When forming social impressions from speech, voices contain long-term traits including sociocultural attributes, with accent being a particularly salient example, while short-term states include expressive prosodic cues (Schuller & Batliner, 2013). In our study, confident vs. doubtful prosody represents short-term states, while we propose that AI-human distinctions may be categorized as long-term traits with salience comparable to accent. This is supported by existing literature showing that accent as a long-term trait influences listeners' interpretation of vocal confidence and doubt and broader social impression formation (Jiang et al., 2018, 2020). We propose that AI voice perception may involve similar mechanisms to accent perception, given that in-group/out-group biases appear in both AI speech research (Kühne et al., 2020) and accent perception studies (Bestelmeyer et al., 2014; Paladino & Mazzurega, 2020). Future research on AI voice perception could benefit from adopting existing accent perception experimental paradigms and theoretical frameworks for understanding social impression formation.

**Listeners utilize acoustic distance as predicted, but cross-source pairs show unique patterns.** We hypothesized that listeners calculate acoustic distance between audio samples to assist identity discrimination. Our results confirm this through three key lines of evidence from accuracy and reaction time curves. However, findings do not universally support the prediction that greater acoustic distance leads to "different speaker" judgments. On the one hand, for same-speaker discrimination, three conditions showed expected patterns within thresholds (human-human and AI-AI pairs under inconsistent prosody, plus both cross-source pairs), where greater distances led to incorrect "different speaker" judgments. On the other hand, cross-source pairs showed unique recovery patterns beyond certain thresholds, in which listeners no longer followed the "greater distance = different speaker" rule but instead demonstrated faster reaction times and higher accuracy. We speculate this reflects cognitive strategy switching when long-term trait differences (AI vs. human) exist between audio samples, triggering alternative processing when simple distance calculations become insufficient.

**Acoustic distance utilization and cross-source patterns** support **the direct matching account.** Voice identity processing has been explained by prototype-based models, in which listeners first calculate acoustic deviations from average voice representations before comparing these patterns to stored speaker identities (Latinus & Belin, 2011; Latinus et al., 2013; Maguinness et al., 2018). Alternatively, direct matching accounts suggest that acoustic features are compared directly to stored voice representations without intermediate prototype calculations (Lavan & McGettigan, 2023). Our results support the direct matching account through the converging patterns described above, particularly the perfect synchronization between accuracy and reaction time, condition-specific variability, and adaptive strategy switching in cross-source conditions. However, it is important to acknowledge that our discrimination task design inherently favors direct matching mechanisms, as rapid sequential comparison is more conducive to immediate acoustic feature matching than extended prototype-based processing.

## 4. General discussion

Our study aims to respond to five gaps: three in AI voice perception design and two in speaker identity processing. For AI voice perception: First, most existing studies compared human recordings with AI clones of identical utterances (e.g., Roswandowitz et al. (2024)), but real-world AI generates novel content, and same-text cloning may artificially favor AI; we used entirely new sentences (Lavan et al., 2025) for both AI avatars and returning human speakers (our research design). Second, most prior studies used monotone prosody (e.g., Barrington et al. (2025)), suppressing human voices' natural expressiveness and potentially inflating AI-human similarity; we incorporated confident and doubtful prosodic variation (RQ1, RQ3). Third, binary detection tasks may miss graded perceptions (Barrington et al., 2025); we used Likert-scale humanlikeness ratings (RQ2). For speaker identity processing: First, conflicting findings on within-speaker variation (Lavan, Knight, & McGettigan, 2019; Xu & Armony, 2021) required systematic examination of prosody consistency (RQ4, RQ5). Second, direct acoustic matching mechanisms (Lavan & McGettigan, 2023) needed empirical testing via acoustic distance modeling of voice identities (RQ6).

Our findings do not support concerns that people are poorly equipped to detect AI-powered voice clones, demonstrating detection at two perceptual levels: (1) categorically, listeners distinguish AI from human voices based on humanlikeness despite expressive prosody comparable to human speech, and (2) at identity level, listeners do not readily perceive AI-cloned and human voices as the same speaker, with prosodic variation serving as diagnostic markers that reinforce categorical

distinctions (reducing from 54% to 36%). These findings extend (1) within-speaker variation research by treating natural prosodic variation as comparable to acoustic manipulations of VTL or F0 (Lavan, Knight, & McGettigan, 2019) and (2) carry theoretical, managerial, and practical implications for AI voice technologies.

### 4.1. Theoretical contributions

#### 4.1.1. Theoretical implications for identity perception from within-speaker variation

Previous work on within-speaker variation has primarily focused on identity perception across different vocalization types from the same person, such as speaking vs. laughing or whispering (Lavan, Burton, et al., 2019). Our study examines within-speaker variation through: (1) within-human voices, (2) within-AI voices, and (3) cross-source comparisons where the same person's identity is represented by both human speech and AI clones. Given the potential challenges posed by memory load if testing identity recognition across prosodic cues (Xu & Armony, 2021) and potentially from AI vs. human source differences for identity perception, we employed an AX discrimination paradigm (Fleming et al., 2014).

Our results reveal two key patterns. Within-category comparisons (human-human, AI-AI) remained tolerable despite prosodic variation, demonstrating listeners' ability to accommodate acoustic differences within perceptual categories. In contrast, cross-category comparisons (human-AI) proved substantially harmful, with performance near chance levels that declined further when prosodic variation was present. This pattern indicates that identity perception is constrained by categorical representations: prosodic cues are accommodated as within-speaker variation within categories but serve as diagnostic boundary markers across human and AI sources.

#### 4.1.2. Theoretical implications for categorical boundaries and expectation effects of AI speech

Beyond identity discrimination, these categorical distinctions manifest in listeners' perception of voice humanlikeness itself. Our results suggest that human speech possesses an inherent "human signature" (Roswandowitz et al., 2024; Tamura et al., 2015) that remains detectable even when AI systems attempt to replicate expressive prosody. Just as listeners demonstrate sensitivity to accent variations that signal social group membership (Caballero & Pell, 2020; Fiske et al., 2018; Mauchand & Pell, 2022b), our results indicate that listeners are similarly attuned to the subtle acoustic signatures that distinguish artificial from human vocal production.

Theoretically, these findings advance understanding of AI voice perception mechanisms. While humanlikeness perception may operate along a continuum, with voices rated as more or less human-like (Nussbaum et al., 2025), listeners' ultimate categorical decisions (AI vs. human) (Barrington et al., 2025; Lavan et al., 2025) follow social categorization frameworks analogous to in-group/out-group mechanisms. Although AI voices producing human-specific prosodic patterns (e.g., doubtful prosody) were rated as slightly more humanlike than those with confident prosody, this small effect may reflect expectation violations: listeners may not expect AI to produce expressive prosody, making such variability seem paradoxically more authentic. This possible expectation-based interpretation (Gampe et al., 2023) opens critical directions for future research. Specifically, violation-of-expectation paradigms could illuminate the cognitive mechanisms underlying AI voice perception at two levels.

At the semantic level, listeners may expect AI voices to produce functional instructions (e.g., "turn right in 200 m") rather than intimate expressions (e.g., "give me a hug"). For example, future studies may observe that AI-generated criticism statements are perceived as less hurtful than human criticism, as in the case of accent perception (Domínguez-Arriola et al., 2025). At the paralinguistic level, if listeners currently expect AI voices to sound monotone, presenting human-produced expressive prosody but labeling it as AI-generated (Gampe et al., 2023) could test whether expectation violations trigger aversion responses. Such paradigms would clarify whether categorical AI-human distinctions primarily reflect acoustic reality or expectation-driven interpretation.

Future research may reveal linguistic-paralinguistic interactions in AI voice perception analogous to accent perception. In accent research, mismatches between accent and stereotypical content elicit N400 potentials, indicating expectation violations when acoustic social cues conflict with semantic content (Pélissier & Ferragne, 2022). Similarly, if listeners can reliably detect AI speech, hearing AI produce unexpected human-associated content (e.g., warmth expressions) or human-typical expressive prosody may trigger comparable neural signatures of expectation violation.

### 4.2. Limitations and managerial/practical implications

#### 4.2.1. Current research uses simpler AI than real-world game-changer systems: Conservative interpretation needed

Achieving humanlike AI voices through text-appropriate prosodic cues appears already technically feasible, as demonstrated by multimodal AI systems like Sora 2 (OpenAI, 2025) and viral demonstration videos (YetiAF, 2025). The challenge is further compounded by Sora 2's multimodal integration, where visual content may distract from audio cues while background noise masks acoustic artifacts that would be detectable in audio-only contexts.

However, a critical gap exists between how AI voices are operationally defined in research and how the public increasingly perceives them in everyday contexts. Most AI voice perception studies have employed monotone AI speech as the operational definition of "AI voice" (e.g., Roswandowitz et al. (2024), among others cited earlier), and even the present study that introduces prosodic variation represents only an initial step toward the expressive capabilities that the public continuously encounters through systems like Sora 2. We therefore suggest that findings based on monotone or controlled expressive AI speech (including the present results) should be interpreted as conservative estimates of AI–human discriminability.

This gap suggests that existing findings showing listeners cannot perfectly identify AI voices as deepfakes in binary tasks, with approximately 40% of AI-generated audio being misclassified as human (Barrington et al., 2025; San Segundo et al., 2025) or 27% (Mai et al., 2023), may reflect genuine perceptual challenges. We suppose these difficulties could become even more pronounced if those studies employed prosodically rich AI speech comparable to Sora 2. While our Likert-scale humanlikeness ratings demonstrated significant AI-human distinctions, this does not ensure that listeners would consistently identify AI voices correctly in binary forced-choice tasks, particularly under time constraints.

#### 4.2.2. Research must expand beyond university students and laboratory contexts

As McGettigan et al. (2025) emphasize, experiments with healthy younger adults do not capture real-world experiences, requiring research expansion beyond laboratory samples. This is because, for example, after completing a speech-in-noise intelligibility task, only 26% of elderly individuals recognized that computer-generated speech had been presented, compared to 83% of younger adults (Herrmann, 2023). Likewise, individuals with autism spectrum disorder showed no differential ratings of humanlikeness between AI-generated and human singing voices, whereas neurotypical listeners clearly distinguished them (Kuriki et al., 2016).

By the same logic, many other populations and real-world contexts require investigation (see McGettigan et al. (2025) for review). Vulnerable groups, such as individuals with hearing loss, children, and augmentative and alternative communication (AAC) users, may perceive AI and human voices very differently than university students

do. Real-world settings, such as clinical voice banking, eldercare facilities, schools, and grief counseling, involve far more than just detecting whether a voice is AI: they involve trust, emotional attachment, and ongoing relationships with AI voices.

Given that individual differences in voice processing may exist within tested groups, including young healthy adults but more likely in at-risk populations, future studies may utilize standardized tools (e.g., recent ones like Humble et al. (2023) and Xu et al. (2025), among others) and use those results as predictors or grouping variables to investigate not only speaker identity discrimination ability but also human vs. AI voice distinction ability, provided participants are cognitively capable of such tasks.

#### 4.2.3. *Research priorities for everyday AI applications*

We recommend treating high-fidelity systems like Sora 2 as game-changers rather than as extensions of existing technologies. Future research should prioritize everyday AI applications where detection remains feasible, rather than state-of-the-art systems designed to be undetectable. Demonstrating listeners' inability to recognize cutting-edge AI voices is simply a demonstration of technical success.

The greater research value should lie in understanding how people perceive and interact with detectable AI voices through categorical perception mechanisms, particularly in-group and out-group processing (Gluszek & Dovidio, 2010; Tajfel et al., 1971), and how expressive prosody interacts with listeners' expectations when AI speech violates assumptions about artificial voices. This research focus enables investigation of psychologically and socially consequential outcomes predicted by the Computers Are Social Actors (CASA) framework (Reeves & Nass, 1996), such as anthropomorphization, expectation-driven trust, and interaction patterns with artificial agents, rather than merely documenting detection failure.

#### 4.2.4. *Deployment recommendations and responsible practices*

We recommend that AI voice deployment should prioritize AI voices that maintain perceptually transparent boundaries for listeners. This is because cloning familiar voices engages richer person representations, emotional associations, and social expectations than unfamiliar voices (McGettigan et al., 2025). We do not recommend unrestricted deployment of highly humanlike AI voices, especially when combined with voice cloning capabilities that replicate specific speaker identities and convey rich emotional expressions.

## CRediT authorship contribution statement

**Wenjun Chen:** Writing – original draft, Visualization, Methodology, 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 and code availability

Data, analysis code, and materials for both experiments are available at: https://osf.io/26yr5/?view_only=5b208c266b824f8eb105bed9dad4739ahttps://osf.io/26yr5/?view_only=5b208c266b824f8eb105bed9dad4739a.

## Declaration of generative AI

During the preparation of this work, the authors used Claude Sonnet 4 (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 these tools/services, the author(s) reviewed and edited the content as needed and take(s) 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.

## Acknowledgements

We thank the four anonymous reviewers for their valuable suggestions. We thank Dr. Morgan Sonderegger for valuable guidance and feedback on Bayesian modeling in the advanced statistics course W. Chen took. We thank Dr. Mason A. Wirtz for generously sharing customized code for visualizing Bayesian posterior summaries.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbah.2026.100261.

## References

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

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., Maner, J. K., & Schaller, M. (2006). They all look the same to me (unless they're angry) from out-group homogeneity to out-group heterogeneity. *Psychological Science, 17*(10), 836–840. https://doi.org/10.1111/j.1467-9280.2006.01790.x

Anderson, K. T. (2007). Constructing "otherness": Ideologies and differentiating speech styleAbstract. *International Journal of Applied Linguistics, 17*(2), 178–197. https://doi.org/10.1111/j.1473-4192.2007.00145.x

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th international conference on neural information processing systems*. https://doi.org/10.5555/3495724.3496768. Vancouver, BC, Canada.

Bakkouche, L., McGhee, C., Lau, E., Cooper, S., Luo, X., Rees, M., Alter, K., Post, B., & Schwarz, J. (2025). Finding the human voice in AI: Insights on the perception of AI-voice clones from naturalness and similarity ratings. In *Proceedings of interspeech 2025*. https://doi.org/10.21437/Interspeech.2025-947. Rotterdam, The Netherlands.

Barrington, S., Cooper, E. A., & Farid, H. (2025). People are poorly equipped to detect AI-powered voice clones. *Scientific Reports, 15*(1), Article 11004. https://doi.org/10.1038/s41598-025-94170-3

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

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America, 109*(2), 775–794. https://doi.org/10.1121/1.1332378

Bestelmeyer, P. E. G., Belin, P., & Ladd, D. R. (2014). A neural marker for social bias toward In-group accents. *Cerebral Cortex, 25*(10), 3953–3961. https://doi.org/10.1093/cercor/bhu282

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

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

Bürkner, P.-C. (2017). Brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Caballero, J. A., & Pell, M. D. (2020). Implicit effects of speaker accents and vocally-expressed confidence on decisions to trust. *Decision, 7*(4), 314. https://doi.org/10.1037/dec0000140

Carey, D., & McGettigan, C. (2017). Magnetic resonance imaging of the brain and vocal tract: Applications to the study of speech production and language learning. *Neuropsychologia, 98*, 201–211. https://doi.org/10.1016/j.neuropsychologia.2016.06.003

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., & Jiang, X. (2024). Memorization-based training and testing paradigm for robust vocal identity recognition in expressive speech using event-related potentials analysis. *Journal of Visualized Experiments*, (210), Article e66913. https://doi.org/10.3791/66913

Chen, W., Jiang, X., Ge, J., Shan, S., Zou, S., & Ding, Y. (2024). Inconsistent prosodies more severely impair speaker discrimination of artificial-intelligence-cloned than human talkers. In *Proc. Speech prosody 2024*. https://doi.org/10.21437/SpeechProsody.2024-171. Leiden, The Netherlands.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

Cohn, M., Predeck, K., Sarian, M., & Zellou, G. (2021). Prosodic alignment toward emotionally expressive speech: Comparing human and alexa model talkers. *Speech Communication, 135*, 66–75. https://doi.org/10.1016/j.specom.2021.10.003

Corretgé, R. (2024). Praat vocal toolkit. https://www.praatvocaltoolkit.com.

Crumpton, J., & Bethel, C. L. (2016). A survey of using vocal prosody to convey emotion in robot speech. *International Journal of Social Robotics, 8*(2), 271–285. https://doi.org/10.1007/s12369-015-0329-4

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

De, S., Bostan, I., & Sastry, N. (2025). Making social platforms accessible: Emotion-aware speech generation with integrated text analysis. In L. M. Aiello, T. Chakraborty, & S. Gaito (Eds.), *Social networks analysis and mining Cham*. https://doi.org/10.1007/978-3-031-78554-2_7

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

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports, 16*, Article 100538. https://doi.org/10.1016/j.chbr.2024.100538

Dixon, J. A., Mahoney, B., & Cocks, R. (2002). Accents of guilt?:effects of regional accent, race, and crime type on attributions of guilt. *Journal of Language and Social Psychology, 21*(2), 162–168. https://doi.org/10.1177/02627x02021002004

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

European-Commission. (2025). *Bridging communication gaps in human and Human-AI interactions: The role of accented speech on neurocognitive mechanisms and social dynamics*. Publications Office of the European Union. https://cordis.europa.eu/project/id/101226709.

Feng, S., & Jiang, X. (2024). Acoustic encoding of vocally expressed confidence and doubt in Chinese bidialectics. *Journal of the Acoustical Society of America, 156*(4), 2860–2876. https://doi.org/10.1121/10.0032400

Fiske, S. T., Lin, M., & Neuberg, S. L. (2018). The continuum model: Ten years later. *Social Cognition*, 41–75. https://doi.org/10.4324/9781315187280

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences, 111*(38), 13795–13798. https://doi.org/10.1073/pnas.1401383111

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1276–1293. https://doi.org/10.1037/0096-1523.32.5.1276

Fuertes, J. N., Gottdiener, W. H., Martin, H., Gilbert, T. C., & Giles, H. (2012). A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology, 42*(1), 120–133. https://doi.org/10.1002/ejsp.862

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

Gluszek, A., & Dovidio, J. F. (2010). Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging in the United States. *Journal of Language and Social Psychology, 29*(2), 224–234. https://doi.org/10.1177/0261927X09359590

Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language, 122*(2), 92–102. https://doi.org/10.1016/j.bandl.2012.04.017

Gussenhoven, C., & Chen, A. (2021). *The Oxford handbook of language prosody*. Oxford University Press.

Hassani, S. M., & Kangavari, M. R. (2025). Emotion-aware speech generation by utilizing prosody in artificial agents: A systematic review. *Circuits, Systems, and Signal Processing*. https://doi.org/10.1007/s00034-025-03336-x

Herrmann, B. (2023). The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults. *International Journal of Speech Technology, 26*(2), 395–415. https://doi.org/10.1007/s10772-023-10027-y

Hosoda, M., Stone-Romero, E. F., & Walter, J. N. (2007). Listeners' cognitive and affective reactions to English speakers with standard American English and Asian accents. *Perceptual and Motor Skills, 104*(1), 307–326. https://doi.org/10.2466/pms.104.1.307-326

Humble, D., Schweinberger, S. R., Mayer, A., Jesgarzewsky, T. L., Dobel, C., & Zäske, R. (2023). The jena voice learning and memory test (JVLMT): A standardized tool for assessing the ability to learn and recognize voices. *Behavior Research Methods, 55*(3), 1352–1371. https://doi.org/10.3758/s13428-022-01818-3

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/10.1177/1747021819865583

Jiang, X., & Pell, M. D. (2015). On how the brain decodes vocal cues about speaker confidence. *Cortex, 66*, 9–34. https://doi.org/10.1016/j.cortex.2015.02.002

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., 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

Kim, J., Toutios, A., Lee, S., & Narayanan, S. S. (2020). Vocal tract shaping of emotional speech. *Computer Speech & Language, 64*, Article 101100. https://doi.org/10.1016/j.csl.2020.101100

Kirk, N. W. (2025). "eh? Aye!": Categorisation bias for natural human vs AI-augmented voices is influenced by dialect. *Computers in Human Behavior: Artificial Humans, 4*, Article 100153. https://doi.org/10.1016/j.chbah.2025.100153

Kolekar, S. S., Richter, D. J., Bappi, M. I., & Kim, K. (2024). Advancing AI voice synthesis: Integrating emotional expression in multi-speaker voice generation. In *2024 international conference on artificial intelligence in information and communication (ICAIIC)*. https://doi.org/10.1109/ICAIIC60209.2024.10463204. Osaka, Japan.

Kruschke, J. K. (2015). Chapter 13 - Goals, power, and sample size. In J. K. Kruschke (Ed.), *Doing bayesian data analysis* (2nd ed., pp. 359–398). Academic Press. https://doi.org/10.1016/B978-0-12-405888-0.00013-1.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

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

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

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

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology, 2*, 175. https://doi.org/10.3389/fpsyg.2011.00175

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., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology, 72*(9), 2240–2248. https://doi.org/10.1177/1747021819836890

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review, 26*, 90–102. https://doi.org/10.3758/s13423-018-1497-7

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

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications, 10*(1), 1–9. https://doi.org/10.1038/s41467-019-10295-w

Lavan, N., & McGettigan, C. (2023). A model for person perception from familiar and unfamiliar voices. *Communications Psychology, 1*(1), 1. https://doi.org/10.1038/s44271-023-00001-4

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2021). Emmeans: Estimated marginal means, aka least-squares means. *The Comprehensive R Archive Network* [Computer software] R package version 1.5. 1.. https://cran.r-project.org/web/packages/emmeans/index.html.

Levi, S. V. (2019). Methodological considerations for interpreting the Language familiarity effect in talker processing. *WIREs Cognitive Science, 10*(2), Article e1483. https://doi.org/10.1002/wcs.1483

Liu, P., & Pell, M. D. (2012). Recognizing vocal emotions in mandarin Chinese: A validated database of Chinese vocal emotional stimuli. *Behavior Research Methods, 44*(4), 1042–1051. https://doi.org/10.3758/s13428-012-0203-3

Ma, F., Xie, Y. F., Ni, S. G., & Ma, F. (2025). A review of human emotion synthesis based on generative technology. *IEEE transactions on affective computing, 16*(4), 2579–2598. https://doi.org/10.1109/TAFFC.2025.3573878

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia, 116*, 179–193. https://doi.org/10.1016/j.neuropsychologia.2018.03.039

Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLoS One, 18*(8), Article e0285333. https://doi.org/10.1371/journal.pone.0285333

Mauchand, M., & Pell, M. D. (2022a). French or québécois? How speaker accents shape implicit and explicit intergroup attitudes among francophones in Montréal. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement, 54*(1), 1–8. https://doi.org/10.1037/cbs0000292

Mauchand, M., & Pell, M. D. (2022b). French or Québécois? How speaker accents shape implicit and explicit intergroup attitudes among francophones in Montréal. *Canadian*

*Journal of Behavioural Science/Revue canadienne des sciences du comportement, 54*(1), 1. https://doi.org/10.1037/cbs0000292
Mauchand, M., & Pell, M. D. (2022c). 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
McGettigan, C., Bloch, S., Bowles, C., Dinkar, T., Lavan, N., Reus, J. C., & Rosi, V. (2025). Voice conversion and cloning: Psychological and ethical implications of intentionally synthesising familiar voice identities. *Journal of the British Academy, 13*(3), a31. https://doi.org/10.5871/jba/013.a31
Mogali, S. R., Ng, O., Tan, J. X., San, T. H., & Ng, K. B. (2024). Voice-over anatomy lectures created by AI-voice cloning technology: A descriptive article. *Anatomical Sciences Education, 17*(9), 1686–1693. https://doi.org/10.1002/ase.2524
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
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
OpenAI. (2025). Sora 2 is here. *OpenAI*. https://openai.com/index/sora-2/.
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
Pandey, P. (2015). Indian English pronunciation. *The handbook of English pronunciation*.
Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America, 85*(2), 913–925. https://doi.org/10.1121/1.397564
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.
Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition & Emotion, 28*(2), 230–244. https://doi.org/10.1080/02699931.2013.812033
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.
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
Pélissier, M., & Ferragne, E. (2022). The N400 reveals implicit accent-induced prejudice. *Speech Communication, 137*, 114–126. https://doi.org/10.1016/j.specom.2021.10.004
Pell, M.D., Cui, H., Mori, Y., & Jiang, X. (2026). Speak or shout? Nonverbal vocalizations promote rapid detection of emotions in vocal communication. *PLoS One, 21*(1), Article e0327529. 10.1371/journal.pone.0327529.
Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior, 33*, 107–120. https://doi.org/10.1007/s10919-008-0065-7
Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint*. https://doi.org/10.48550/arXiv.2104.03502
Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., Fink, B., DeBruine, L. M., Jones, B. C., & Feinberg, D. R. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour, 95*, 89–99. https://doi.org/10.1016/j.anbehav.2014.06.011
Polyanskaya, L., Ordin, M., & Busa, M. G. (2017). Relative salience of speech rhythm and speech rate on perceived foreign accent in a second language. *Language and Speech, 60*(3), 333–355. https://doi.org/10.1177/0023830916648720
Posit-team. (2024). RStudio: Integrated development environment for R. In *Posit software*. PBC. http://www.posit.co/.
R-Core-Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.
Rakić, T., Steffens, M. C., & Mummendey, A. (2011). When it matters how you pronounce it: The influence of regional accents on job interview outcome. *British Journal of Psychology, 102*(4), 868–883. https://doi.org/10.1111/j.2044-8295.2011.02051.x
Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK, 10*(10), 19–36.
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
Rossum, G.v., & Drake, F. L. (2009). *Python 3 reference manual*.
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
San Segundo, E., López-Jareño, A., Wang, X., & Yamagishi, J. (2025). Human perception of audio deepfakes: The role of language and speaking style. *arXiv e-prints, arXiv-2512*. https://doi.org/10.48550/arXiv.2512.09221
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
Schuller, B., & Batliner, A. (2013). *Computational paralinguistics: Emotion, affect and personality in speech and language processing*. John Wiley & Sons. https://doi.org/10.1002/9781118706664
Scott, S., & McGettigan, C. (2016). The voice: From identity to interactions. In *APA handbook of nonverbal communication* (pp. 289–305). https://doi.org/10.1037/14669-011
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
Squizzero, R. (2025). The effects of perceived ethnicity and prosodic accuracy on intelligibility, comprehensibility, and accentedness in L2 mandarin Chinese. *Language and Speech, 0*(0), Article 00238309251361010. https://doi.org/10.1177/00238309251361010
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
Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language, 53*(1), 81–94. https://doi.org/10.1016/j.jml.2005.02.003
Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology, 1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202
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
Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian analysis, 16*(2), 667–718. https://doi.org/10.1214/20-BA1221
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., & Bright, J. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2
Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3
Warren, K., Tucker, T., Crowder, A., Olszewski, D., Lu, A., Fedele, C., Pasternak, M., Layton, S., Butler, K., & Gates, C. (2024). Better be computer or I'm dumb": A large-scale evaluation of humans as audio deepfake detectors. In *Proceedings of the 2024 on ACM SIGSAC conference on computer and communications security (CCS '24)*. Salt Lake City, UT, USA https://doi.org/10.1145/3658644.3670325.
Wilmot, N. V., Vigier, M., & Humonen, K. (2024). Language as a source of otherness. *International Journal of Cross Cultural Management, 24*(1), 59–80. https://doi.org/10.1177/14705958231216936
Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America, 123*(6), 4524–4538. https://doi.org/10.1121/1.2913046
Wirtz, M. A., & Pfenninger, S. E. (2024). Capturing thresholds and continuities: Individual differences as predictors of L2 sociolinguistic repertoires in adult migrant learners in Austria. *Applied Linguistics, 45*(2), 249–271. https://doi.org/10.1093/applin/amad055
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: state-of-the-art natural language processing. In Q. Liu, & D. Schlangen (Eds.), *Proceedings of the*

*2020 conference on empirical methods in natural language processing: System demonstrations online*. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xu, Y. (2019). Prosody, tone, and intonation. In *The routledge handbook of phonetics* (pp. 314–356). Routledge. https://doi.org/10.4324/9780429056253.

Xu, H., & Armony, J. L. (2021). Influence of emotional prosody, content, and repetition on memory recognition of speaker identity. *Quarterly Journal of Experimental Psychology, 74*(7), 1185–1201. https://doi.org/10.1177/1747021821998557

Xu, T., Jiang, X., Zhang, P., & Wang, A. (2025). Introducing the sisu voice matching test (SVMT): A novel tool for assessing voice discrimination in Chinese. *Behavior Research Methods, 57*(3), 86. https://doi.org/10.3758/s13428-025-02608-3

YetiAF. (2025). AI has COMPLETELY LOST ITS GRIP ? (Sora 2 clips that spiral out of Nowhere). https://www.youtube.com/watch?v=fO1spQ_DX50.

Zäske, R., Awwad Shiekh Hasan, B., & Belin, P. (2017). It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content. *Cortex, 94*, 100–112. https://doi.org/10.1016/j.cortex.2017.06.005

Zhang, S., & Pell, M. D. (2022). Cultural differences in vocal expression analysis: Effects of task, language, and stimulus-related factors. *PLoS One, 17*(10), Article e0275915. https://doi.org/10.1371/journal.pone.0275915