Neuropsychologia 229 (2026) 109493

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Human and AI voice identities evoke shared neural signatures during speaker recognition across changes in speech content and prosody

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

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

ARTICLE INFO

Keywords:
AI voice
Voice identity
Recognition
Prosody
Multivariate pattern analysis

ABSTRACT

Both biologically-produced human voices and algorithmically-generated AI speech manifest speaker identity. Critically, prosodic variations modulate the acoustic dimensions (e.g., fundamental frequency) that also shape individual speaker identity representations. So far, it remains unclear whether listeners process speaker identities in human and AI voices through neurologically equivalent mechanisms, nor how prosodic cues might influence these cognitive processes. We examined event-related potentials during old/new speaker discrimination after name-based identity learning, and further analyzed correctly recognized old speakers, comparing trials where prosody matched vs. mismatched between learning and testing. For old/new discrimination, multivariate pattern analysis (MVPA) revealed three significant late windows (662-1498 ms) with Pz as the primary contributor for AI voices, yet no clusters for human voices. Univariate analyses revealed that human voices showed earlier widespread discrimination (N250: 200-280 ms), while both voice types converged on Pz as the strongest contributor based on effect size rankings for late old/new effects (400-800 ms). These old/new effects emerged across completely different speech content between learning and testing, extending content-independent parietal ERP effects beyond syllabic stimuli. For speaker-specific prosodic expectation effects in the 500-900 ms window, unexpected prosody elicited late positivity for human voices compared to the prosody used during learning, whereas AI voices elicited late negativity. The late positivity resembles P600 components observed for communicative style expectancy violations, while the late negativity likely reflects effortful reprocessing of prosodic violations within atypical synthetic signals, analogous to accented speech processing. These findings advance understanding of voice identity processing and have implications for AI voices in human-computer interaction.

1. Introduction

Speaker identity is an essential manifestation of the self in social interaction (Scott and McGettigan, 2016) and our ability to rapidly identify speakers is deployed constantly: recognizing a friend's voice within seconds of answering a phone call, distinguishing between two speakers on the radio, or immediately identifying a celebrity's voice in an advertisement. Today, however, devices like smartphones also deliver AI-generated speech alongside human voices. Voice synthesis technology has advanced rapidly, with early work demonstrating that speaker-adaptive synthesis could reconstruct a personalized voice from as little as 5 min of speech recording (Yamagishi et al., 2012). Such capabilities are now widely available: Apple's *Personal Voice* allows users to record 150 utterances to clone their speaking style (Apple, 2023), while Huawei's *Xiaoyi* assistant in Chinese can clone both speaker identity and prosodic style from as few as 15 utterances (Chen and Jiang, 2023). When the same identity exists in both human and AI voices, do listeners rely on the same perceptual mechanisms to process identity features for each voice type?

One lens to examine this question is through speaker recognition, the ability to learn and later identify individual speakers, which has revealed distinct neural mechanisms for processing familiar vs. unfamiliar voices (Maguinness et al., 2018; Sidtis and Zäske, 2021). A key motivation for examining speaker recognition across familiarity is an existing gap in understanding how identity processing generalizes when speech content or prosody varies at the utterance level (Lavan et al.,

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

https://doi.org/10.1016/j.neuropsychologia.2026.109493
Received 23 January 2026; Received in revised form 21 April 2026; Accepted 11 May 2026
Available online 13 May 2026

2019a; Xu and Armony, 2021; Zäske et al., 2014). We address these two gaps while using this recognition framework to compare how listeners process human vs. AI-cloned voices. Our study provides an updated understanding of the intersection of voice identity perception and how speech prosody may influence it, and most importantly, whether AI-cloned voices engage the same perceptual mechanisms as human voices.

1.1. Speech-content-dependent or speech-independent voice identity processing?

Voice identity arises from anatomical, acoustic, and articulatory cues and functions as a relatively stable manifestation of oneself in social communication (Scott and McGettigan, 2016; Sidtis and Kreiman, 2012). This ability is evolutionarily ancient, with newborns already showing differential neural responses to maternal vs. unfamiliar voices (Adam-Darque et al., 2020). Unfamiliar voices undergo a gradual learning process in which listeners repeatedly analyze their acoustic structure and compare it to an internal prototype voice (Maguinness et al., 2018). Through such iterative perceptual cycles, stable identity features are extracted and consolidated into a reference pattern, ultimately giving rise to the neural signatures observed for familiar vs. unfamiliar voices. A meta-analysis of functional magnetic resonance imaging (fMRI) studies revealed systematic differences across familiarity levels (Sun et al., 2023). Unfamiliar voices mainly activate the bilateral superior temporal gyri for basic acoustic analysis. Familiar voices additionally recruit the right inferior and middle frontal gyri involved in identity processing. Higher familiarity further engages regions such as the fusiform gyrus, parahippocampal cortex, and insula, reflecting access to person-specific information.

Beyond spatial patterns revealed by fMRI, electroencephalography (EEG) studies show that voice-familiarity levels differ in their temporal dynamics. Plante-Hébert et al. (2021), who manipulated familiarity through frequency of exposure rather than explicit learning phases, found that trained-familiar voices elicited a distinct N250 (300–350 ms) compared to rarely heard unfamiliar voices, whereas intimately familiar voices showed different components compared to unfamiliar voices: a P2 (200–250 ms) and a sustained LPC (450–850 ms). However, when voices are explicitly learned and later tested with manipulated speech content, different patterns emerge. Zäske et al. (2014) reported a parietal old/new LPC effect (300–700 ms at Pz) for correctly recognized old vs. new voices when test utterances matched study utterances, but this effect was limited to the same-utterance condition. This seemingly suggests that the LPC primarily reflects speech-content-dependent episodic retrieval rather than abstract voice representations. Notably, however, Zäske et al. (2014) also identified a speech-independent effect in beta band oscillations (16–17 Hz, 290–370 ms) at central and right temporal sites for learned compared with novel voices, suggesting that content-invariant voice identity processing may be captured through oscillatory rather than purely time-domain ERP measures.

Meanwhile, fMRI evidence from the same research group points to speech-independent voice representations. Zäske et al. (2017) used an analogous recognition paradigm but presented German utterances to non-German speakers, isolating voice processing from semantic content. Learned voices showed reduced activation compared to novel voices in right-lateralized regions, including the posterior superior temporal gyrus and frontal areas. Critically, unlike the LPC effect in their 2014 ERP study, these fMRI novelty reductions persisted regardless of whether test utterances matched those heard during learning, indicating that voice identity representations can generalize across speech content. These findings suggest that while fMRI reveals speech-invariant voice identity representations, current ERP markers remain sensitive primarily to speech-content-dependent episodic retrieval.

It is notable that the above studies utilized emotionally neutral prosody (Plante-Hébert et al., 2021; Zäske et al., 2014, 2017), which minimizes cognitive load and avoids potential alterations to voice identity representations that prosodic variation might introduce (Lavan et al., 2019c; Xu and Armony, 2021). Although prior ERP studies minimized prosodic variation by using neutral prosody, they still did not detect neural old/new differences across changes in speech content (Zäske et al., 2014).

Meanwhile, there are reasons to anticipate such effects. This expectation is grounded in behavioral work demonstrating reliable extraction of speech-invariant identity cues across utterances. In memory-intensive old/new paradigms, cross-utterance voice recognition remains robust when prosody (suprasegmental vocal features; see next section for details) is held constant: Zäske et al. (2014) reported accuracies of ~62–70% across blocks using neutral prosody, and Xu and Armony (2021) likewise observed above-chance performance (~67%–72%) in same-prosody/different-content trials with either fearful or neutral prosody.

Arguably, the absence of ERP differences in Zäske et al. (2014) at the utterance level, despite above-chance behavioral accuracy, reflect insufficient strength of identity encoding. This speculation is supported by evidence from explicit identity learning tasks: when listeners receive name-label training accompanied by accuracy verification, they successfully recognize speakers despite substantial within-speaker variations in glottal pulse rate (GPR, perceived as fundamental frequency, F0) and vocal tract length (VTL) that exceed those typically associated with changes in speech content Lavan et al. (2019c). Crucially, by pairing voices with names and confirming successful learning before testing, Lavan et al. (2019) ensured that robust person-level representations were formed, enabling listeners to track identity-relevant acoustic variability while ignoring irrelevant within-speaker variation. By inference, implementing a similar name-labeling procedure and accuracy check should hypothetically strengthen identity encoding sufficiently to observe a parietal old/new effect at Pz, as reported by Zäske et al. (2014) even when speech content varies, rather than merely reflecting episodic retrieval.

Beyond the utterance-level evidence reviewed above, a more direct line of evidence comes from the syllabic level. Schweinberger et al. (2011) used an adaptation paradigm with vowel-consonant-vowel syllables (e.g., /aba/, /igi/) and highly familiar speakers, where participants adapted to one speaker's syllable before making a two-alternative forced-choice identity judgment on a subsequent test voice. Critically, parietal effects (P3, Pz, P4) reflecting voice identity persisted even when the adaptor and test stimuli contained different syllables, demonstrating content-independent voice identity processing at the parietal level. We therefore expect that analogous parietal effects should emerge at the utterance level in an old/new recognition paradigm, given sufficiently strong identity encoding through name-labeling as described above.

1.2. Prosody as more than acoustic variation: a speaker-specific style cue?

Beyond the linguistic content discussed above, the same utterance can be perceived quite differently depending on how it is said, that is, through speech prosody (suprasegmental features such as pitch, rhythm, and intonation patterns) (Xu, 2019). Prosodic variation conveys paralinguistic information about speakers' emotional states (e.g., happy vs. sad) (Pell and Skorup, 2008), epistemic certainty (e.g., feeling of knowing: confident vs. doubtful) (iang and Pell (Jiang and Pell, 2015, 2017), and communicative intentions (e.g., sincere vs. ironic) (Fish et al., 2017; Mauchand et al., 2020).

While variation in speech content with similar prosodic patterns likely induces only small within-speaker fluctuations in structural identity cues, such as F0 and VTL (Mathias and von Kriegstein, 2019), changes in paralinguistic cues can be far more disruptive (Anikin et al., 2021; Belyk et al., 2022; Pisanski et al., 2022). Real-time MRI studies show that emotional prosody systematically shifts vocal-tract configuration, with happy speech produced using shorter effective VTL and greater articulatory opening than angry, sad, or neutral speech (Kim et al., 2020). Meanwhile, F0 and VTL function in an inverse relationship

when listeners estimate vocal-tract dimensions (Darwin et al., 2003; Neuhaus et al., 2024). Prosodic variations across emotional, social, and pragmatic contexts are thus characterized by systematic modulations of both F0 and VTL (Larrouy-Maestri et al., 2025). For instance, confident prosody is produced with lower F0 and longer effective VTL than doubtful prosody in both English (Jiang and Pell, 2017) and Mandarin (Chen and Jiang, 2023). Most importantly, listeners are sensitive to such coordinated shifts in VTL and F0, influencing how they derive individual identity representations (Lavan et al., 2019a; 2019c).

Given that prosodic shifts systematically modulate the same acoustic dimensions that define speaker identity (Lavan et al., 2019b; Pinheiro, 2025), a critical question arises: do listeners treat prosody merely as acoustic variability to be normalized, or do they encode prosodic patterns as part of a speaker's identity representation? Evidence from syntactic processing suggests the latter may be plausible. showed that listeners internalize a speaker's habitual syntactic preferences (e.g., marked OSV vs. SVO word order), and violations of these speaker-specific patterns elicit P600 responses even when the utterance meaning remains unchanged. This demonstrates that listeners bind linguistic style to speaker identity.

We propose that prosodic style may function analogously. If listeners learn a talker under one prosodic style but later encounter a different style from the same speaker, this mismatch may elicit expectancy-violation responses. Such findings would carry two implications. First, prosodic style, like syntactic style, functions as a speaker-specific cue bound to identity (). Second, successful recognition despite prosodic shifts would indicate that listeners extract identity representations that generalize across acoustic variation, with P600-like prediction errors () signaling the comparison between stored and incoming identity cues (Lavan et al., 2019a; 2019c).

1.3. Are AI and human speaker identities represented similarly in the cognitive system?

Human voice production relies on physiological articulatory systems, whereas AI-synthesized speech is generated by deep learning models (Khanjani et al., 2023), with acoustic features derived from parametric structures rather than anatomical constraints. Thus, the frameworks discussed in the preceding subsections concerning how speech content and prosody influence identity recognition are grounded in human voice perception and cannot be automatically generalized to AI voices. Currently, there exists no systematic understanding of whether AI voices support identity memorization and recognition at behavioral and neural levels. Based on existing relevant evidence, we propose that AI voice identity learning and recognition mechanisms may share components with human voice perception.

One source of evidence comes from deepfake detection tasks, predominantly from English-language studies, where listeners judge whether voices originate from AI or humans. Large-scale studies report variable human performance. Warren et al. (2024) found overall accuracies of 63.9–85.8% across three benchmark datasets with over 1200 participants, while Müller et al. (2022) reported that human listeners' accuracy plateaued at about 80% in a gamified ASVspoof2019 detection task (410 participants, 13,229 trials). More recently, Barrington et al. (2025) employed state-of-the-art *ElevenLabs* voice cloning and showed that listeners correctly identified AI-generated voices as synthetic on only about 60% of trials, indicating that high-quality voice clones frequently evade explicit human detection. These behavioral findings suggest that although humans can detect degraded synthetic speech, high-quality AI clones often pass as human.

Importantly, neuroimaging work demonstrates dissociable neural pathways for natural and synthetic identities. Roswandowitz et al. (2024) reported that despite above-chance deepfake identity matching (~69%), natural and deepfake voices engaged distinct neural pathways: natural identities activated the nucleus accumbens (NAcc), a region linked to social reward and bonding, whereas deepfake voices showed weaker NAcc responses but stronger auditory-cortical activity, suggesting synthetic voices lack socially rewarding properties and require greater acoustic-phonetic processing effort. Similarly, Bratan et al. (2025) found that AI-synthesized emotional voices, though judged acoustically similar to natural voices, elicited substantially weaker activation in mirror-neuron and emotion-related regions, instead primarily recruiting memory-related networks. These neural dissociations imply that, despite partial behavioral deception, synthetic voices are not processed as fully natural speaker identities by the human brain.

A second source of evidence comes from research directly exploring identity representation. In identity matching tasks where listeners judge whether two consecutive audio clips are the same speaker, Barrington et al. (2025) found that participants judged AI-cloned voices and their human sources as the same identity in approximately 80% of trials, indicating that listeners often treat cloned and original voices as belonging to a single speaker. However, Chen et al. (2026) used AI-generated voices whose human-likeness had been independently validated as substantially lower than that of natural voices. Under these conditions, listeners judged AI clones and their human sources as the same identity only slightly above chance level, suggesting that awareness of voice source matters for identity perception. Chen et al. (2026) also showed that within AI or human voices separately, identity recognition accuracy reached a 99% ceiling when prosody was consistent and remained above 90% when prosody differed. This indicates that both human and AI voices carry internally coherent identity signatures to which listeners are highly sensitive. We therefore propose that when AI and human voices are examined separately, without requiring cross-category identity recognition, both groups should exhibit broadly similar behavioral patterns of identity discrimination, albeit with lower accuracy under memory-demanding recognition tasks than in short-term matching paradigms.

Beyond cross-category comparisons between AI and human voices, a third line of evidence examines whether AI voices can engage person-specific representation systems when identity familiarity is manipulated within synthetic speech alone. Using functional near-infrared spectroscopy (fNIRS), Zhang et al. (2025) showed that AI-generated maternal voices, cloned from participants' own mothers, elicited significantly stronger activation in listeners' prefrontal and temporal cortices than AI-generated unfamiliar female voices, consistent with engagement of familiarity-, emotion-, and memory-related processes. Although this study did not compare AI against natural speech or test identity recognition per se, it demonstrates that person-specific familiarity signals can be reinstated even in fully synthetic speech, supporting the idea that AI voices may access overlapping person-representation systems.

Taken together, existing evidence shows that AI voices can indeed deceive listeners into perceiving them as human and can engage familiarity-based identity processing to some extent, but their neural signatures could still remain distinct from those of natural voices in certain aspects.

1.4. Theoretical motivation for an EEG investigation of voice identity and prosodic processing in human and AI voices

Existing voice perception models propose a hierarchical architecture in which initial acoustic analyses in temporal voice areas are followed by increasingly abstract representations culminating in person identity nodes (Belin et al., 2004; Young et al., 2020). This hierarchical organization necessarily unfolds over time: voice/nonvoice distinctions emerge as early as 30-150 ms, whereas identity recognition effects emerge from 250 ms onward, representing a subsequent and distinct processing stage (Lamothe et al., 2026). Understanding how these hierarchical stages unfold, therefore, requires methods with high temporal resolution, a dimension that fMRI evidence alone cannot fully address (Young et al., 2020), making EEG uniquely positioned to track the time course of voice identity processing (Lamothe et al., 2026).

Examining AI voices alongside human voices is further motivated by theoretical and empirical considerations. Nussbaum et al. (2025) proposed that naturalness assessments occur at early stages of voice object analysis, while identity authenticity is evaluated at later stages of voice content analysis, suggesting that AI-human distinctions may manifest differently across the processing hierarchy. Providing empirical support, Chen et al. (2025), using the same voice cloning technology and speaker pool as the present study, showed through behavioral ratings that listeners categorically treat AI voices as a perceptual out-group, with this boundary constraining how prosodic cues translate into social impressions. EEG is therefore uniquely positioned to track when and how these categorical distinctions emerge across the processing hierarchy, informing whether human and AI voice identities engage shared or distinct neural mechanisms.

Finally, the interaction between prosodic cues and speaker identity representations further motivates an EEG approach. Pinheiro (2025) highlighted that transient vocal signals, such as emotional prosody and stable characteristics such as speaker identity, interact during voice perception, with EEG uniquely suited to track how these distinct types of information emerge and integrate over time. We extend this framework to prosodic style as a speaker-specific binding cue (Kroczek and Gunter, 2021), motivating the examination of whether prosodic violations elicit speaker-specific expectancy responses and whether such responses differ between human and AI voices.

1.5. The present study

As demonstrated above, existing studies have not simultaneously examined speaker recognition in the face of both speech content and prosodic variations, nor have they directly compared the neural mechanisms of identity processing between AI and human voices. The present study addresses these gaps by comparing identity learning and recognition in human vs. AI voices across manipulations of speech content and prosody. We employ AI clones with prosodic styles cloned separately to minimize acoustic confounds (Roswandowitz et al., 2024). Human and AI voices are examined in separate blocks. Participants learn speaker identities through name-labeling, then perform old/new recognition with different utterances while prosodic consistency is manipulated. We focus on two research questions.

- RQ1: Whether parietal old/new ERP effects emerge across different speech content at the utterance level, and whether such effects differ between human and AI voices.
- RQ2: Whether prosodic violations elicit speaker-specific expectancy responses, and whether such responses differ between human and AI voices.

For RQ1, based on prior evidence (Schweinberger et al., 2011; Zäske et al., 2014, 2017), parietal old/new effects should emerge across changes in speech content when prosody remains consistent. Additionally, we anticipate replicating speech-independent beta band oscillations (16-17 Hz, 290-370 ms) at central and right temporal sites (Zäske et al., 2014). For the AI-human comparison, both voice types should show comparable parietal old/new effects, given that behavioral evidence demonstrates comparable within-category identity discrimination for both human and AI voices in a within-category recognition task (Chen et al., 2026).

For RQ2, prosodic violations should produce late ERP components reflecting speaker-specific expectancy violations analogous to syntactic coupling effects (Kroczek and Gunter, 2021). For human voices, this should manifest as late positivity reflecting prosodic expectancy. For AI voices, the human/AI distinction may engage mechanisms analogous to in-group/out-group categorization in accent perception, given that AI voices may be processed as a categorical out-group (Bratan et al., 2025; Roswandowitz et al., 2024). Drawing on Jiang et al. (2020), who found that listeners elicited late negativity when evaluating vocal expressions of doubt produced by out-group accented speakers, we tentatively predict that AI voices may similarly elicit late negativity for prosodic violations, while acknowledging that this prediction remains exploratory given the paradigm differences.

2. Methods

2.1. Participants

We recruited 43 native Mandarin speakers from universities in Shanghai. Three participants were excluded due to technical issues in the initial experimental setup. The remaining 40 participants consisted of 20 females (age: 21.7 ± 1.7 years; years of education: 17.6 ± 1.9) and 20 males (age: 23.2 ± 1.7 years; years of education: 18.8 ± 1.7). None reported speech or hearing impairments or any psychiatric or neurological disorders. All provided written informed consent, and the study was approved by the Ethics Committee of the Institute of Language Sciences at Shanghai International Studies University. Participants received 50 RMB per hour as compensation. Our sample size of 40 participants is consistent with established EEG studies of speaker recognition, such as Zäske et al. (2014) with 24 participants.

2.2. Stimuli

2.2.1. Stimuli creation

Our audio stimuli were selected from a larger validated corpus of 11,808 audio recordings. The large corpus was created through a multistage procedure: First, 24 native Mandarin speakers (12 females, 12 males) recorded 15 utterances in three prosodic styles (confident, doubtful, neutral; only confident and doubtful were used in the present study), which were used to train speaker-specific AI voice models using Huawei's *Celia* system; part of the acoustic validation for 10 of these speakers is reported in Chen and Jiang (2023). Second, these AI models generated 123 novel utterances in confident and doubtful prosodies (5904 AI recordings: 123 utterances × 2 prosodies × 24 speakers). Critically, each speaker's confident AI clone was trained exclusively on their confident recordings, and their doubtful AI clone exclusively on their doubtful recordings, with no cross-training between prosodic styles. Third, approximately one month later, the same 24 speakers returned to produce human versions of these 123 utterances in both prosodic styles, resulting in 5904 human recordings matched to the AI corpus. Fourth, 48 independent listeners rated 11,808 clips (human and AI, confident and doubtful) on perceived humanlikeness and vocal confidence using 7-point scales. More demographic and procedural information for the speakers and raters (e.g., age, educational background) is reported in Chen et al. (2026).

For the present EEG study, 960 stimuli were selected based on these validation ratings. Utterances were brief Mandarin Chinese statements expressing factual information (e.g., “我的鼠标坏了” [My mouse is broken]), evaluations (e.g., “她很有幽默感” [She has a good sense of humor]), or intentions (e.g., “他们不想上课了” [They don't want to go to class anymore]).

The 24 speakers were divided into four groups of six by selecting every other speaker from height-ranked sequences within each biological sex (e.g., speakers ranked 1st, 3rd, 5th, 7th, 9th, 11th formed one group), creating perceptual similarity gradients while controlling for height-related acoustic cues (VTL and F0). Within each group, stimuli were selected based on validation ratings to ensure that confident prosody received higher confidence ratings than doubtful prosody, and human voices received higher humanlikeness ratings than AI voices (mean differences, not statistically tested).

The audio files were normalized to −30 dBFS and standardized to stereo format (44,100 Hz sampling rate) using *pydub* library (version 0.25.1) in Python 3.11.4 (Robert, 2021) to ensure that all stimuli were presented at the same volume level through headphones during EEG recording.

2.2.2. Stimuli validation

We conducted acoustic and perceptual validation on all 960 voice recordings (480 human-AI speaker pairs). F0 was extracted using *Praat* 6.2.09 (Boersma and Weenink, 2021). Mean F0 was computed for each utterance after using the *extractvowels* plugin from the *Praat Vocal Toolkit* to extract each utterance's vowels (Chen et al., 2026; Corretgé, 2024). Independent raters evaluated all stimuli on two 7-point scales: confidence level (1 = not confident at all, 7 = very confident) and humanlikeness (1 = very machine-like, 7 = very human-like). Statistical analyses used Linear Mixed-Effects Regression (LMER) implemented in R (version 4.3.3) with the *lmerTest* package (Kuznetsova et al., 2017). For F0, the model formula was: f0_mean_hz ~ Source × Confidence_level × Biological_Sex + (1|Speaker) + (1|Item). For perceptual ratings, model formulas were: *perceived_humanlikeness ~ Source × Confidence_level + (1|Speaker) + (1|Item)* and *perceived_confidence ~ Source × Confidence_level + (1|Speaker) + (1|Item)*. Simple effects were examined using the *emmeans* package (Lenth et al., 2021), and effect sizes were quantified using Cohen's *d*.

Our validation analyses confirmed successful experimental manipulations. For F0 (Fig. 1A), human voices differed from AI voices in the doubtful condition ($p < .001$, $d = 0.09$) but not in the confident condition ($p = .146$, $d = -0.03$). Prosody effects were significant for both human ($p < .001$, $d = 0.26$) and AI voices ($p < .001$, $d = 0.14$). For confidence ratings (Fig. 1B; left panel), human voices were rated higher than AI voices in both prosody conditions (confident: $p < .001$, $d = 0.84$; doubtful: $p < .001$, $d = 1.07$). Prosody effects on confidence ratings were significant for both human ($p < .001$, $d = 5.95$) and AI voices ($p < .001$, $d = 2.54$). For humanlikeness ratings (Fig. 1B; right panel), human voices were rated higher than AI voices in both confident ($p < .001$, $d = 4.42$) and doubtful ($p < .001$, $d = 4.56$) conditions. Prosody effects on humanlikeness were non-significant for human voices ($p = .056$, $d = 0.24$) but significant for AI voices ($p < .001$, $d = 0.31$).

To visualize acoustic patterns, we generated mel spectrograms for one representative human-AI matched speaker pair across both prosody conditions using the *librosa* package (version 0.11.0 (McFee et al., 2025)) in Python 3.14. Spectrograms were computed with the following parameters: sampling rate = 22,050 Hz, FFT window size = 2048 samples, hop length = 512 samples, and 128 mel frequency bands. Power spectrograms were converted to a decibel scale relative to peak amplitude. To enable direct visual comparison of acoustic patterns across utterances, we standardized leading and trailing silence to 80 ms for all displayed utterances using the *pydub* library (version 0.25.1) (Robert, 2021). Fig. 1C shows an example mel spectrogram comparison for one male speaker-clone pair.

2.3. Experiment details

Experimental design. Participants completed eight blocks with fixed voice source order (human blocks 1-4, AI blocks 5-8). This fixed ordering was implemented because we could not assume uniform participant responses to alternating between biologically-produced and algorithmically-generated voices across blocks, which could introduce variable task-switching costs and compromise data consistency during identity learning. This design choice is justified by our behavioral results (see Checking phase accuracy and relevant sections in Discussion). Meanwhile, prosody order was counterbalanced across participants: half (n = 20; 10 males, 10 females) experienced confident-then-doubtful prosody (Blocks 1-2, 5-6 confident; Blocks 3-4, 7-8 doubtful), while

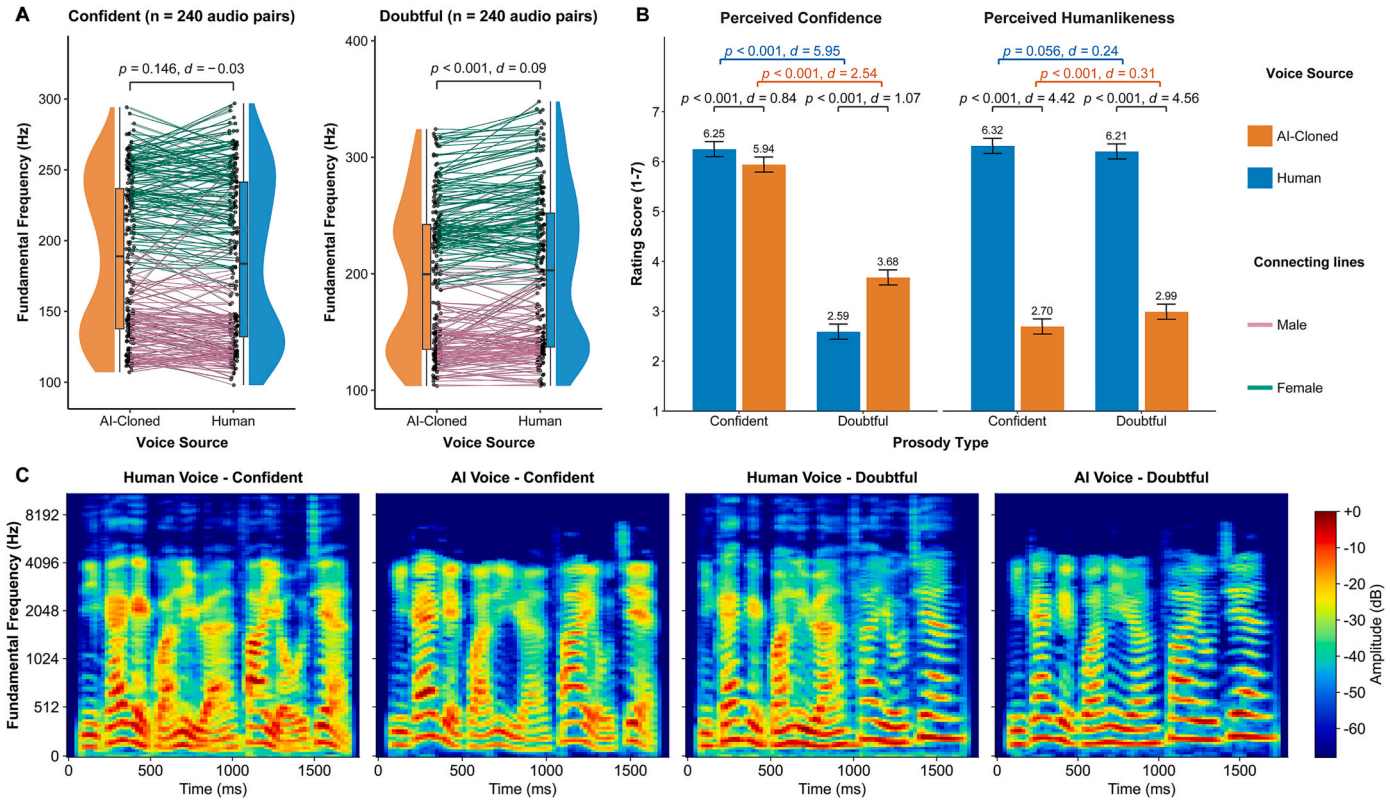

Fig. 1. ***Acoustic and perceptual characteristics of 960 voice stimuli used in the EEG experiment.*** **(A)** F0 distributions for human and AI-cloned voices across prosodic conditions. Connecting lines link matched human-AI speaker pairs, colored by male/female. **(B)** Perceptual ratings showing perceived confidence and humanlikeness across voice sources and prosody types. **(C)** An example showing one male speaker producing "你该给管理员发邮件" [You should send an email to the administrator]. AI voices: generated by independently trained models, where each speaker's confident and doubtful clones were trained on a small subset of their original human recordings (15 utterances per prosody, recorded prior to AI generation). Human voices: recorded approximately one month later, with the speaker instructed to naturally produce the utterance after hearing the AI-generated version. Color intensity represents amplitude (dB).

the other half (n = 20; 10 males, 10 females) experienced doubtful-then-confident prosody. Each block lasted approximately 10 min, with the entire experiment taking 90-100 min, including breaks.

Procedure. The experiment consisted of eight blocks (human and AI voices never mixed within blocks), each comprising three phases: Training, Checking, and Testing (Fig. 2A). During Training (24 trials), participants learned to associate three unfamiliar speakers' voices with Chinese surnames (e.g., "小赵" [Junior ZHAO]). In the Checking phase (12 trials), participants identified speakers using a three-alternative forced-choice task to verify successful identity learning. During Testing (72 trials), participants categorized speakers as OLD (trained) or NEW (untrained). Training and Testing used completely different linguistic materials (utterances). The key manipulation was prosody consistency: participants learned speaker identities through voices in either confident or doubtful prosody during Training, but encountered these speakers in both the same and different prosodic conditions during Testing (Fig. 2B–C). Participants proceeded at their own pace during Training to encode speaker-name associations, but were instructed to respond quickly and accurately during Checking and Testing phases. An accuracy threshold of 10/12 correct responses (≥83%) was required in the Checking phase; accuracy feedback was displayed after each block, and blocks failing to meet this threshold were repeated after a brief break.

EEG recording. EEG data were acquired in a sound-treated, dimly lit, electromagnetically shielded laboratory. Continuous EEG was recorded using a 64-channel elastic cap with passive Ag/AgCl electrodes (ActiCap system, Brain Products, Germany) according to the extended 10-20 system. FCz served as the online reference electrode. Electrode impedances were maintained below 5 kΩ using conductive gel. Signals were digitized at 500 Hz with a bandpass filter of 0.01-100 Hz. Two electrooculography electrodes were placed near the left outer canthus and below the right eye to monitor horizontal and vertical eye movements and blinks. Auditory stimuli were presented binaurally through audiology insert earphones. Participants sat in a comfortable chair approximately 80 cm from a computer monitor in the shielded room. Participants were instructed to minimize head and body movements and to blink only between trials when possible.

2.4. Data analysis

Behavioral data analysis. *Behavioral performance* was analyzed using LMERs implemented in R (version 4.3.3) with the *lmerTest* package (Kuznetsova et al., 2017). The Checking phase aimed to ensure that participants successfully established associations between speaker identities and their voices. Two separate models were fitted for accuracy and reaction time (RT), respectively: *accuracy* ~ *Source* × *Prosody* + *(1|Participant)* and *rt* ~ *Source* × *Prosody* + *(1|Participant)*, where Source (human vs. AI) and Prosody (confident vs. doubtful) were fixed effects. Post-hoc pairwise comparisons were conducted using the *emmeans* package (Lenth et al., 2021) to compare prosody conditions within each voice source. Effect sizes were calculated using Cohen's *d* with pooled standard deviations.

For the Testing phase (Old/New effect), two sets of analyses were conducted. The first analysis examined old/new speaker recognition effects: *accuracy* ~ *Source* × *Old/New* + *(1|Participant)* and *rt* ~ *Source* × *Old/New* + *(1|Participant)*, where Old/New compared old vs. new speakers. For RT analysis here, data were filtered to include only correct trials.

For the prosody expectation effect, only old speaker trials were retained. The analysis used the formula: *accuracy* ~ *Source* × *Prosody* + *(1|Participant)* and *rt* ~ *Source* × *Prosody* + *(1|Participant)*, where Prosody compared same vs. different prosody relative to the Training phase. For RT analysis, data were further filtered to include only trials with correct same/different judgments. Post-hoc comparisons and effect size calculations followed the same procedures as the Checking phase.

RT analyses retained correct trials only. For Old/New recognition, retention rates were 62.2% for Human Old voices (3584 of 5760 trials), 74.4% for Human New voices (4283 of 5760 trials), 60.5% for AI Old voices (3484 of 5760 trials), and 73.5% for AI New voices (4232 of 5760 trials). For the prosody expectation effect analyses on old trials,

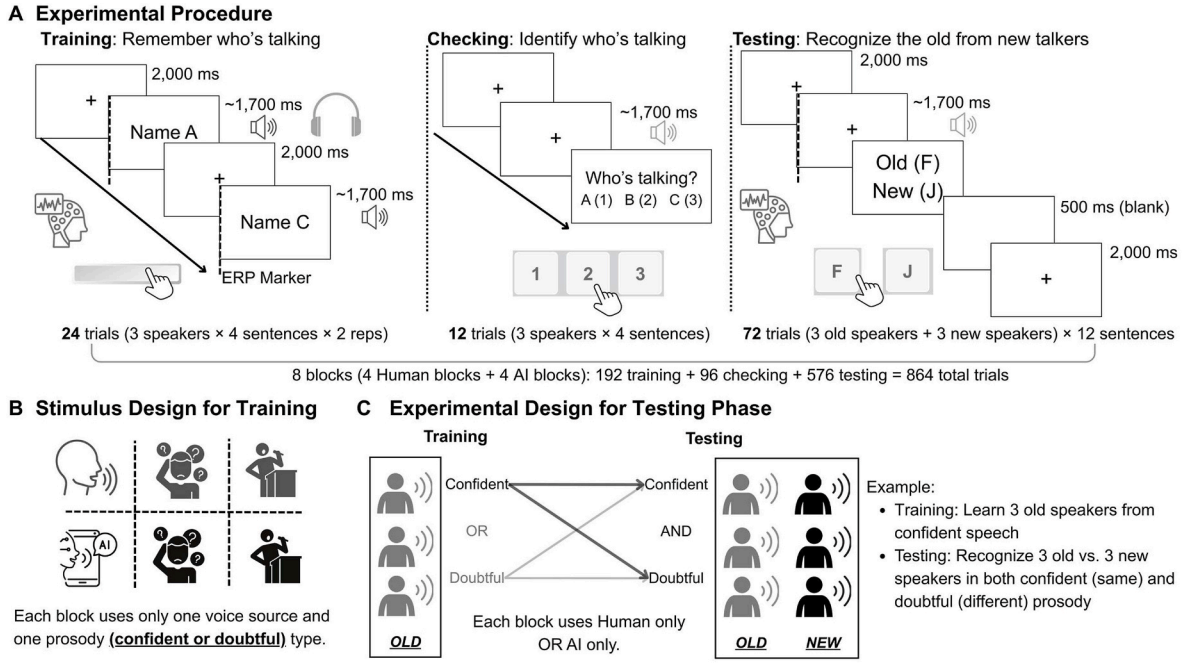

Fig. 2. ***Experimental design and procedure for speaker identity learning and recognition.*** **(A)** Each of 8 blocks included three phases: Training, where participants learned three speakers' names from their voices (24 trials); Checking, where participants identified speakers using the same utterances (12 trials, three-alternative forced-choice); and Testing, where participants categorized speakers as OLD (trained) or NEW (untrained) (72 trials). **(B)** Training stimulus design showing that each block used one voice source (Human or AI) and one prosody (confident or doubtful). **(C)** Testing design where participants distinguished OLD from NEW speakers, with prosody consistency (same vs. different from Training) as the critical manipulation.

retention rates in Same prosody conditions were high for both Human (74.7%, 2152 of 2880 trials) and AI voices (74.5%, 2145 of 2880 trials). However, retention rates dropped in Different prosody conditions for both Human (49.7%, 1432 of 2880 trials) and AI voices (46.5%, 1339 of 2880 trials).

EEG data preprocessing. The preprocessing was conducted using EEGLAB (version 2025.0) (Delorme and Makeig, 2004) in MATLAB (version R2024b). Data were re-referenced offline to the average of bilateral mastoid electrodes. To account for leading silence differences between manually edited human recordings and AI-generated audio, silence duration was quantified for each file using the *pydub* library (version 0.25.1) (Robert, 2021) in Python 3.11.7, identifying silence below −50 dB lasting greater than 5 ms. EEG event markers were then adjusted by these durations to ensure precise time-locking to the acoustic onset. Raw continuous EEG data were band-pass filtered. Bad channels and trials with obvious artifacts were identified through visual inspection; bad channels were interpolated using spherical spline interpolation, and contaminated trials were removed. For ERP analysis, data were filtered at 0.1–40 Hz, epoched from −500 to 1500 ms relative to the adjusted stimulus onset, and baseline-corrected using the −500 to 0 ms pre-stimulus interval.

For time-frequency analysis, data were filtered at 0.1–100 Hz, epoched from −900 to 1500 ms, and baseline-corrected using the −500 to −250 ms interval. For Independent Component Analysis (ICA), data were temporarily high-pass filtered at 1 Hz to improve decomposition quality. ICA decomposition using the extended infomax algorithm was performed on the 1 Hz filtered data, and the resulting unmixing weights were transferred back to the 0.1 Hz filtered data. Components reflecting ocular, muscular, and cardiac artifacts were identified through visual inspection of component topographies and time courses, then removed by back-projection. After artifact removal, data were low-pass filtered at 40 Hz for ERP analysis and 100 Hz for time-frequency analysis.

MVPA analysis. To identify when neural activity patterns discriminated between experimental conditions, we applied MVPA to the time-domain EEG data using linear discriminant analysis (LDA) implemented in the MVPA-Light toolbox for MATLAB (Treder, 2020). ERP epochs (−500 to 1500 ms relative to stimulus onset) were analyzed at each time point, with a classifier trained using the spatial pattern of voltage across all electrodes. Two decoding analyses were performed: (1) distinguishing old vs. new speakers (correct trials only), and (2) distinguishing same vs. different prosody within correctly recognized old speakers. Classification performance was evaluated using 5-fold cross-validation, with decoding accuracy quantified using the area under the receiver operating characteristic curve (AUC) and classification accuracy. This procedure was applied separately for human and AI voices, yielding time-resolved decoding trajectories for each participant. Group-level statistical inference was performed using cluster-based permutation testing (1000 permutations) to control for multiple comparisons across time. Significant clusters were identified using a cluster-forming threshold of $z = 1.96$ (corresponding to $p < .05$), with a family-wise error rate of $\alpha = .05$, as contiguous temporal windows where decoding performance exceeded chance level (AUC = 0.5).

Additionally, time-frequency MVPA was conducted to examine oscillatory power patterns. Given the absence of statistically significant effects, these results are reported in **Supplementary Analysis 1**. Note that ERP waveforms and inset bar plots in the main figures were generated directly from raw EEG data in MATLAB, whereas significance markers (e.g., asterisks) reflect LMER results computed in R.

Old-new effect (MVPA-based clusters). Based on the temporal MVPA results, three significant time windows were identified for AI voices in the old vs. new speaker recognition task. Direct MVPA on human voices did not identify significant clusters, likely because human voices contain inherently greater acoustic variability than AI voices (see Fig. 1C for visual examples), which may have obscured identity-based discrimination signals. We therefore applied these AI-derived time windows to human voice data for comparative analysis. For each cluster, we identified the top 10 electrodes with the highest contribution to decoding accuracy using Haufe-transformed LDA weights (Haufe et al., 2014). Trial-level ERP amplitudes averaged across these 10 electrodes within each time window were analyzed using LMERs (Kuznetsova et al., 2017): *Voltage ~ OldNew + (1|Participant) + (1|Item)*. Type III ANOVA with Satterthwaite's approximation tested the main effect of speaker identity (old vs. new), with effect sizes quantified using partial omega-squared (ω^2).

Old-new effect (Literature-based). To examine condition-related amplitude differences across scalp regions, we conducted LMER analyses on trial-level ERP data extracted from nine electrodes (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4) selected based on prior voice recognition research (Zäske et al., 2014). For each electrode, we averaged voltage values within three predefined time windows: N250 (200-280 ms), P300 (300-380 ms), and LPC (400-800 ms). The same LMER modeling approach was used with statistical inference via Type III ANOVA (Satterthwaite approximation) and effect size quantification via ω^2. Pairwise contrasts between old and new conditions were conducted using estimated marginal means with the *emmeans* package (Lenth et al., 2021). The alpha level was set at 0.05 for all tests.

Speaker-specific cue-binding effect (Literature-based). To investigate whether prosodic consistency between learning and test phases influenced recognition performance, we conducted LMER analyses examining the same vs. different prosody contrast within correctly recognized old speakers. Based on previous research demonstrating late ERP effects for speaker-specific expectancy violations (Kroczek and Gunter, 2021), we analyzed a late time window (500-900 ms). Trial-level ERP amplitudes averaged across these electrodes were analyzed using the model formula *Voltage ~ SameDiff + (1|Participant) + (1|Item)*, testing the main effect of prosodic consistency. All other statistical procedures remained identical to the old vs. new analysis.

3. Results

3.1. Behavioral results

In the Training/Checking phase, participants heard each audio clip twice and then identified the correct name from three options after hearing a single audio presentation. For human voices, we found neither accuracy nor reaction time differed between prosody conditions. In contrast, for AI voices, participants achieved higher accuracy ($z = -2.94$, $p = .003$, $d = 0.61$) and faster responses ($z = 4.40$, $p < .001$, $d = 0.55$) when learning speakers with doubtful prosody (94.7%, 1167 ms) compared to confident prosody (91.6%, 1514 ms). See Fig. 3A–B.

In the Testing phase, participants were expected to classify the three trained familiar speakers as Old and the three untrained novel speakers as New, focusing only on speaker identity. Recognition accuracy was significantly lower for Old compared to New speakers in both human voices (old: 62.2%, new: 74.4%, $z = -14.16$, $p < .001$, $d = 1.04$) and AI voices (old: 60.5%, new: 73.5%, $z = -15.15$, $p < .001$, $d = 1.23$). Reaction times showed no difference between Old and New speakers for either voice type. See Fig. 3C–D.

Listeners learned speaker identities in either confident or doubtful prosody. During testing, they encountered both confident and doubtful prosodies. Among Old speaker trials, those with the same prosody as training/checking yielded significantly higher accuracy than those with different prosody in both human voices (same: 74.7%, different: 49.7%, $z = 20.61$, $p < .001$, $d = 2.11$) and AI voices (same: 74.5%, different: 46.5%, $z = 23.07$, $p < .001$, $d = 2.43$). Similarly, reaction times (analyzed on correct trials only) were significantly faster for same vs. different prosody conditions in both human voices (same: 1046 ms, different: 1251 ms, $z = -3.94$, $p < .001$, $d = 0.44$) and AI voices (same: 1182 ms, different: 1359 ms, $z = -2.45$, $p = .014$, $d = 0.28$). Notably, accuracy in the different prosody conditions dropped to chance level for both voice types. See Fig. 3E–F.

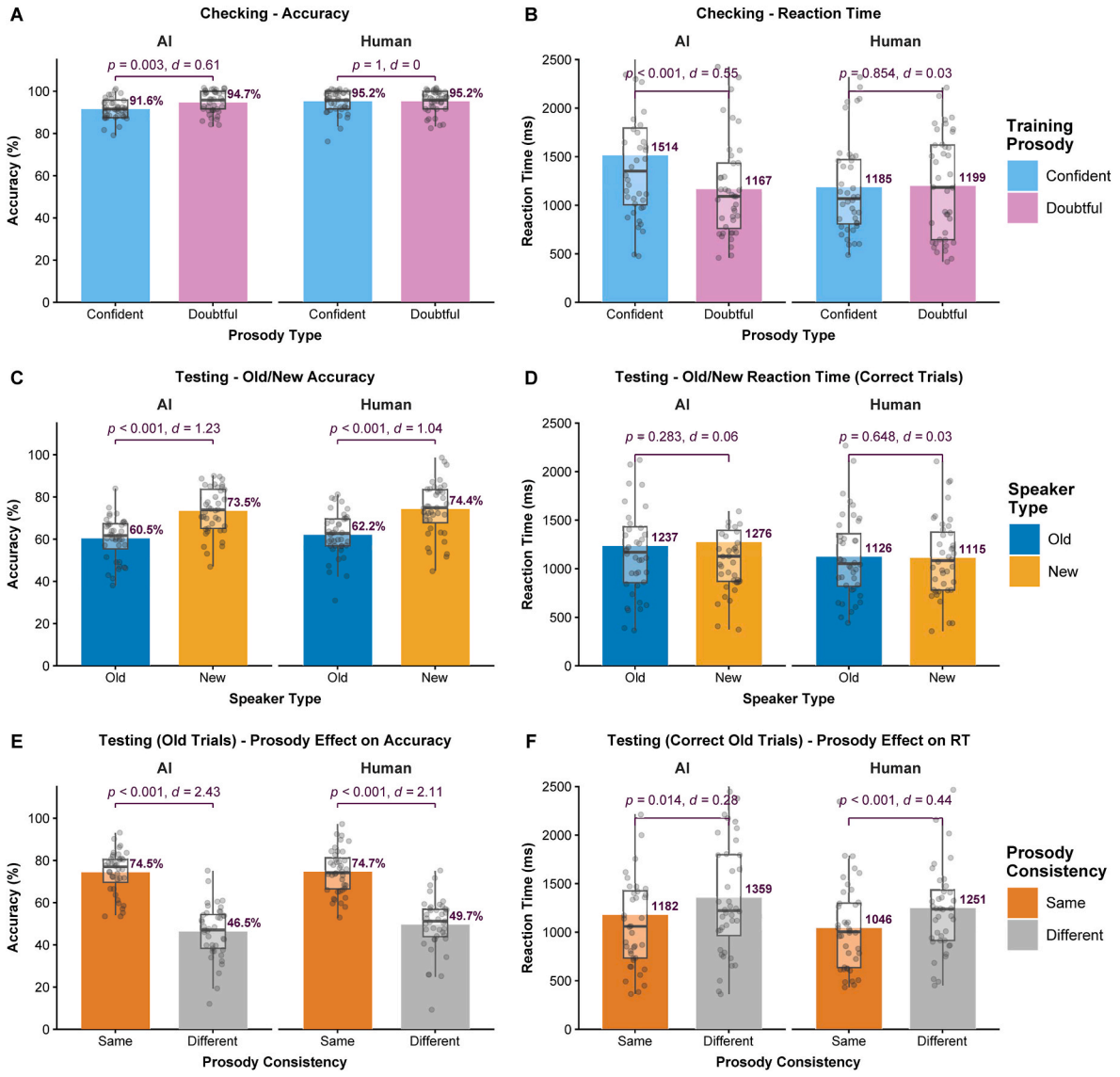

Fig. 3. ***Behavioral performance in voice identity recognition.*** **(A-B)** Accuracy and reaction time during the checking phase, where listeners identified speakers by selecting among three names. **(C-D)** Accuracy and reaction time (correctly old/new trials only) during the testing phase, where listeners judged whether voices belonged to old (trained) or new (untrained) speakers. **(E)** Accuracy for old speaker trials when prosody was the same vs. different between training and testing. **(F)** Reaction time for correctly recognized old speakers when prosody was the same vs. different. Bars represent group means. Boxes indicate the interquartile range with the median line. Individual data points represent each participant's mean.

3.2. Temporal MVPA results

For correctly judged trials, AI voices showed robust neural discrimination of old vs. new speakers in three significant clusters: 662-702 ms (AUC = 0.519 ± 0.052, $p = .029$), 758-844 ms (AUC = 0.525 ± 0.048, $p = .006$), and 866-1498 ms (AUC = 0.520 ± 0.050, $p < .001$), with peak decoding at 682 ms (AUC = 0.537). Human voices showed no significant discrimination (peak AUC = 0.527 at 1392 ms, all $ps > 0.05$; see Fig. 4A–B, E). We also examined prosody consistency effects among correctly recognized old speakers (same vs. different prosody relative to training), but found no significant clusters for either voice type (all $ps > 0.05$), despite substantial behavioral differences (see Figs. 3E and 4C–D, F).

To identify electrodes contributing most to classifier discrimination, we applied the Haufe transformation to the LDA weights and selected the top 10 contributing electrodes within each significant temporal window for AI voices. For the earlier windows (662-702 ms and 758-844 ms), contributing electrodes concentrated over posterior scalp regions, with Pz showing the strongest contribution. For the later window (866-1498 ms), electrodes spanned right frontal and posterior regions, with Pz remaining among the top contributors. These time windows and electrode selections were applied to human voices for parallel visualization and statistical analysis (Fig. 5).

To validate the neural distinctions identified by MVPA, we extracted mean ERP amplitudes within each significant temporal cluster and conducted linear mixed-effects regression comparing old vs. new speakers. For AI voices, old speakers elicited significantly more positive amplitudes than new speakers in the first two windows (cluster 1 [662-702 ms]: F (1, 7525) = 25.07, $p < .001$, $\beta = 1.340$; cluster 2 [758-844 ms]: F (1, 7540) = 16.38, $p < .001$, $\beta = 1.250$). For human voices, only the first window showed a significant old/new difference (cluster 1: F (1, 7589) = 10.05, $p = .002$, $\beta = 0.828$), while the second window showed a marginal trend (F (1, 7592) = 2.80, $p = .094$, $\beta = 0.489$). The later window (866-1498 ms) showed no significant old/new difference for either voice type (AI: F (1, 7537) = 2.40, $p = .122$, $\beta = 0.626$; Human: F (1, 7588) = 1.79, $p = .181$, $\beta = 0.470$). Also see Table S1 and S2 for

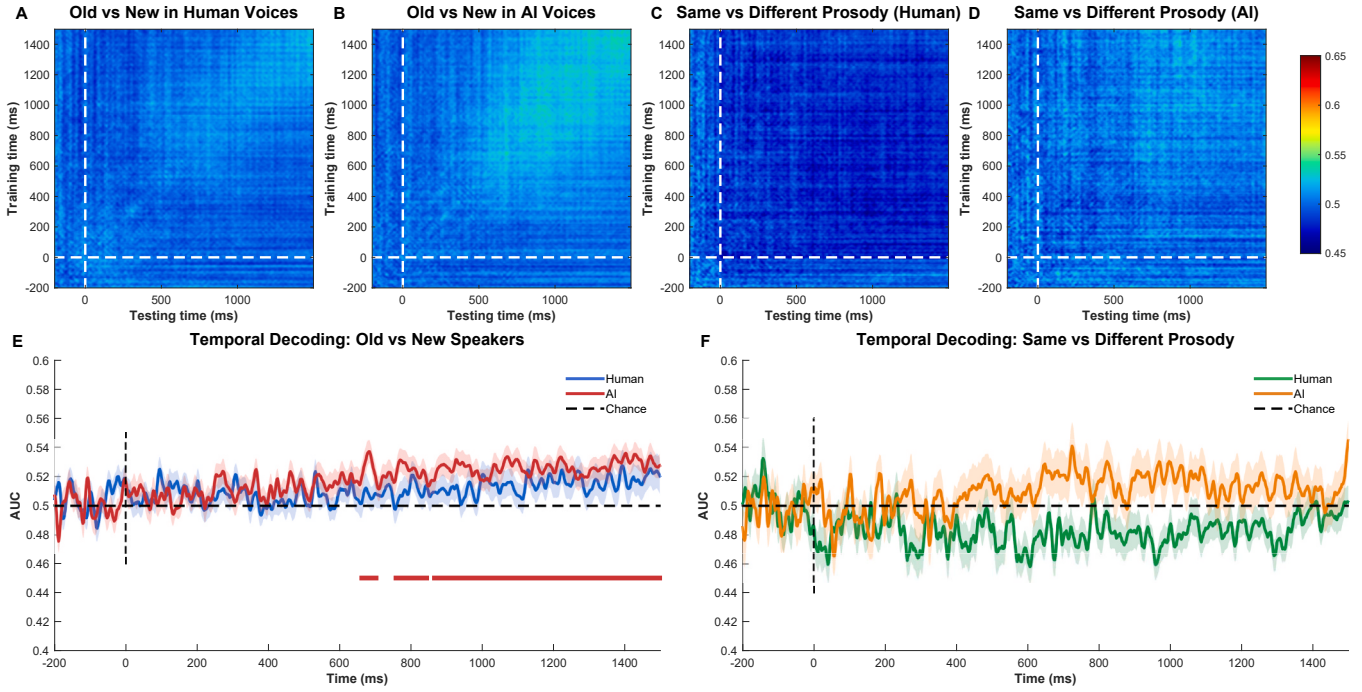

Fig. 4. ***Time-resolved MVPA results for speaker recognition.*** **(A-B)** Time generalization matrices (TGMs) for decoding old vs. new speakers (correct trials) in human and AI voices. **(C-D)** TGMs for decoding same vs. different prosody (relative to training) among correctly recognized old speakers. **(E)** Temporal decoding trajectories for old vs. new speakers; colored bars indicate significant clusters. **(F)** Temporal decoding trajectories for same vs. different prosody; no significant clusters identified. Dashed line = chance level (AUC = 0.50).

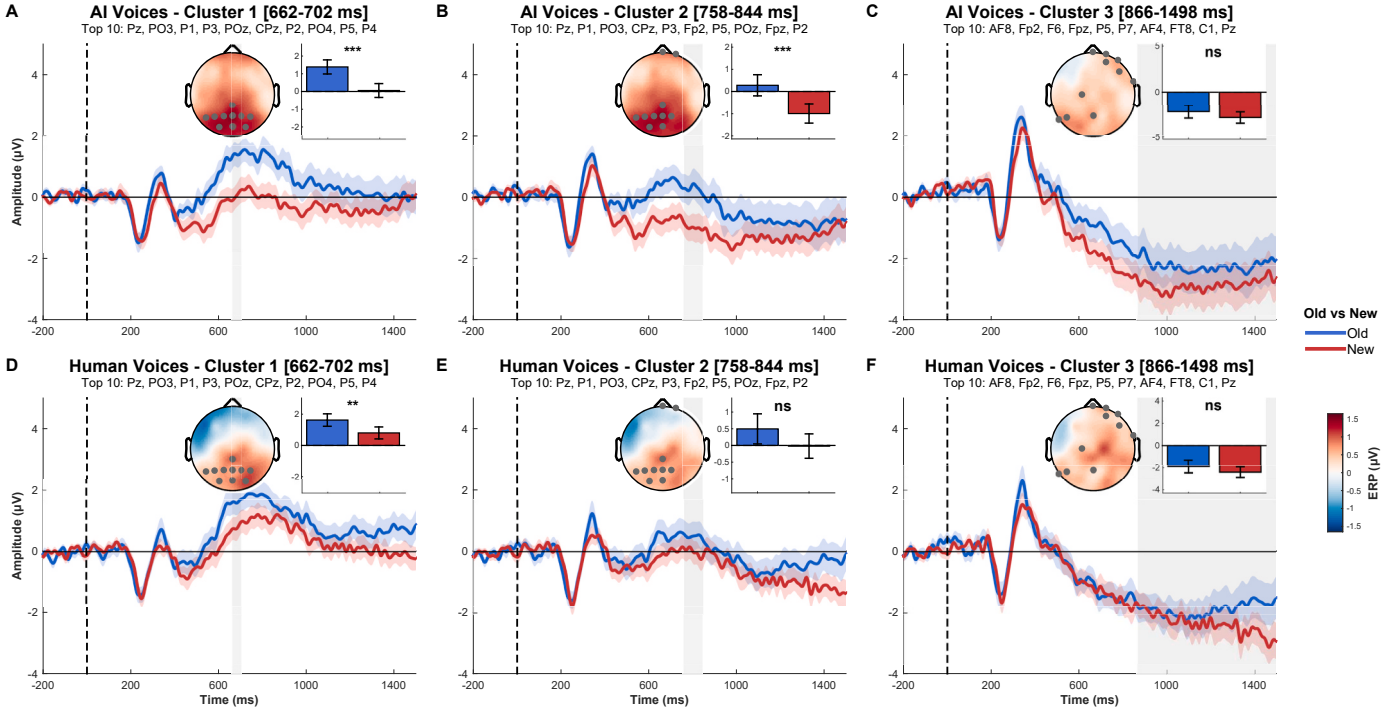

Fig. 5. ***ERP amplitudes for the top contributing electrodes in MVPA-identified windows.*** **(A-C)** ERPs for AI voices in three significant temporal clusters. **(D-F)** ERPs for human voices using the same time windows and electrodes. Each panel shows the ERP time course, topographic map of old-new difference, and mean amplitude comparison. Gray shading = analyzed window. ***$p < .001$, **$p < .01$, *$p < .05$, ns = not significant.

statistical details.

3.3. Speech-content-independent old vs. new effect (N250/P300/LPC)

We analyzed mean amplitudes at nine scalp electrodes (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4) across three time windows (N250: 200-280 ms;

P300: 300-380 ms; LPC: 400-800 ms) using linear mixed-effects regression (see Table S3, Table S4, Figs. 6–8).

In the N250 window, AI voices showed that old speakers elicited significantly greater positivity than new speakers at Pz (F (1, 309,218) = 35.82, $\beta = 0.29$, $p < .001$) and P4 (F (1, 308,515) = 4.52, $\beta = 0.08$, $p = .033$), with no significant differences at other electrodes. For human voices, old speakers elicited significantly greater negativity at six electrodes (ordered by effect size): Pz (F (1, 311,628) = 48.93, $\beta = -0.32$, $p < .001$), P3 (F (1, 311,890) = 40.70, $\beta = -0.24$, $p < .001$), C3 (F (1, 311,860) = 28.83, $\beta = -0.22$, $p < .001$), Fz (F (1, 311,921) = 12.38, $\beta = -0.18$, $p < .001$), C4 (F (1, 311,938) = 8.58, $\beta = -0.12$, $p = .003$), and P4 (F (1, 311,738) = 6.24, $\beta = -0.09$, $p = .013$), with no significant effects at Cz, F3, and F4.

In the P300 window, AI voices showed that old speakers elicited significantly greater positivity than new speakers at five electrodes (ordered by effect size): Pz (F (1, 309,411) = 137.83, $\beta = 0.59$, $p < .001$), P3 (F (1, 308,966) = 47.15, $\beta = 0.28$, $p < .001$), P4 (F (1, 309,072) = 45.46, $\beta = 0.28$, $p < .001$), C3 (F (1, 309,130) = 12.06, $\beta = 0.16$, $p < .001$), and Cz (F (1, 309,257) = 6.37, $\beta = 0.13$, $p = .012$), with no significant differences at frontal electrodes. For human voices, old speakers elicited significantly greater positivity at all nine electrodes (ordered by effect size): Cz (F (1, 311,951) = 294.46, $\beta = 0.83$, $p < .001$), C4 (F (1, 311,904) = 193.24, $\beta = 0.60$, $p < .001$), C3 (F (1, 311,896) = 163.91, $\beta = 0.57$, $p < .001$), F4 (F (1, 311,918) = 160.54, $\beta = 0.66$, $p < .001$), Fz (F (1, 311,930) = 136.27, $\beta = 0.64$, $p < .001$), F3 (F (1, 311,837) = 134.07, $\beta = 0.69$, $p < .001$), P4 (F (1, 311,780) = 32.22, $\beta = 0.22$, $p < .001$), P3 (F (1, 311,890) = 22.68, $\beta = 0.19$, $p < .001$), and Pz (F (1, 311,777) = 4.36, $\beta = 0.10$, $p = .037$).

In the LPC window, AI voices showed that old speakers elicited significantly greater positivity than new speakers at all nine electrodes (ordered by effect size): Pz (F (1, 1,518,133) = 2715.60, $\beta = 1.35$, $p < .001$), P4 (F (1, 1,517,857) = 2364.50, $\beta = 1.02$, $p < .001$), P3 (F (1, 1,517,390) = 1761.10, $\beta = 0.88$, $p < .001$), Cz (F (1, 1,517,675) = 955.48, $\beta = 0.78$, $p < .001$), C4 (F (1, 1,517,723) = 867.51, $\beta = 0.66$, $p < .001$), C3 (F (1, 1,517,294) = 548.98, $\beta = 0.54$, $p < .001$), Fz (F (1, 1,517,963) = 247.42, $\beta = 0.44$, $p < .001$), F4 (F (1, 1,517,178) = 122.73, $\beta = 0.30$, $p < .001$), and F3 (F (1, 1,516,924) = 38.85, $\beta = 0.19$, $p < .001$). For human voices, old speakers elicited significantly greater positivity at five posterior and central electrodes (ordered by effect size): P4 (F (1, 1,529,408) = 1378.20, $\beta = 0.74$, $p < .001$), Pz (F (1, 1,529,319) = 370.31, $\beta = 0.48$, $p < .001$), P3 (F (1, 1,529,579) = 303.10, $\beta = 0.35$, $p < .001$), Cz (F (1, 1,529,560) = 272.36, $\beta = 0.40$, $p < .001$), and C4 (F (1, 1,529,518) = 183.63, $\beta = 0.29$, $p < .001$). Notably, at frontal and left-central electrodes, old speakers elicited significantly greater negativity: F3 (F (1, 1,529,316) = 98.50, $\beta = -0.31$, $p < .001$), Fz (F (1, 1,529,548) = 44.93, $\beta = -0.18$, $p < .001$), F4 (F (1, 1,529,480) = 40.18, $\beta = -0.17$, $p < .001$), and C3 (F (1, 1,529,507) = 14.13, $\beta = -0.09$, $p < .001$).

To examine effect size progression across time windows, we calculated Spearman rank correlations between time window order (N250, P300, LPC) and ω^2 values across all nine electrodes. AI voices showed a significant positive correlation ($\rho = 0.735$, $p < .001$), indicating progressive strengthening from N250 (mean $\omega^2 = 0.000015$) through P300 (0.000088) to LPC (0.000702), with significant electrodes expanding from 2 to 9. In contrast, human voices showed no significant linear trend ($\rho = 0.321$, $p = .103$), with effect sizes peaking at P300 (0.000403) before declining at LPC (0.000196), despite maintaining significance at all nine electrodes from P300 onward.

To summarize, four key patterns emerged. First, both voice types

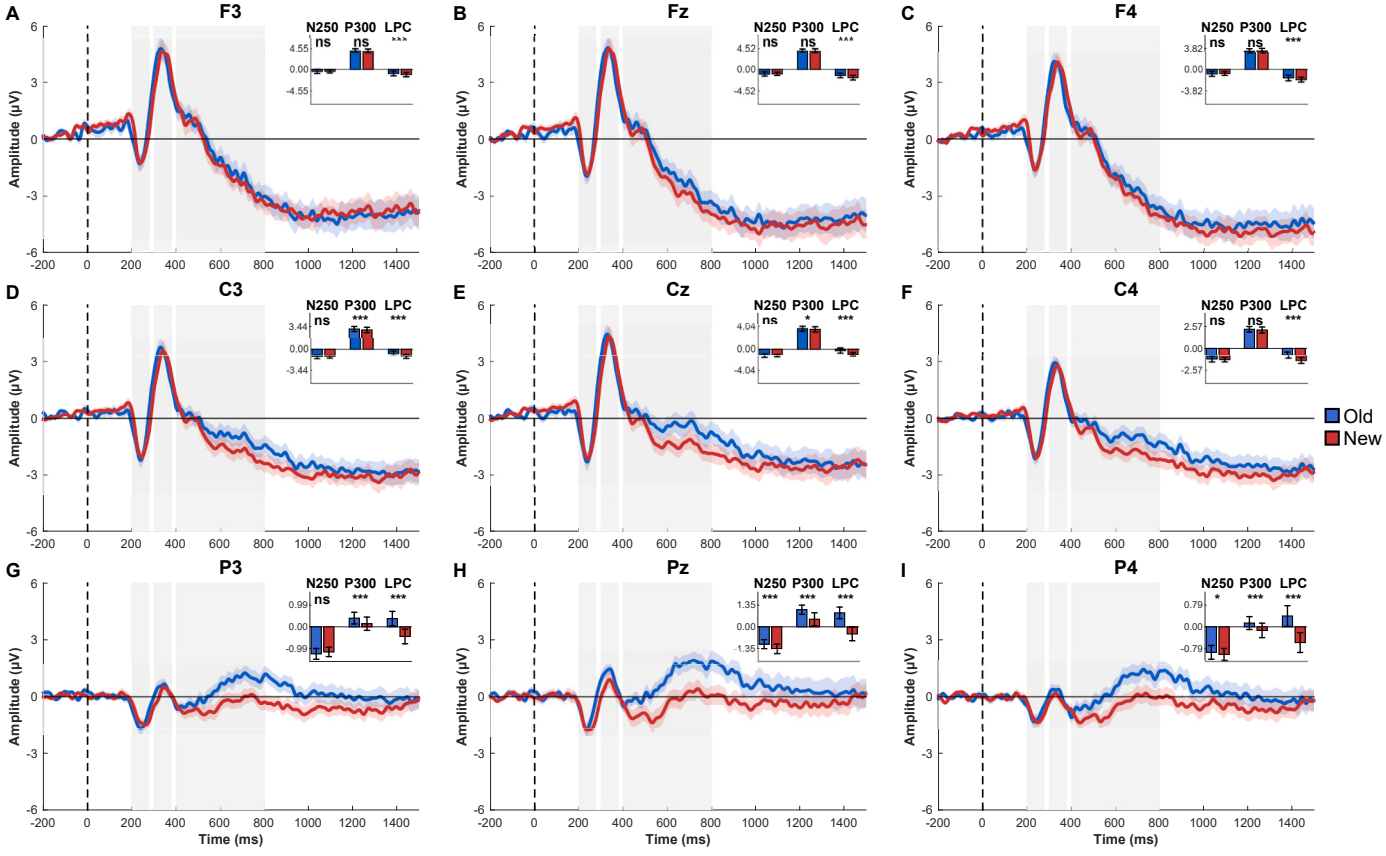

Fig. 6. ***ERP waveforms for old vs. new speaker recognition in AI voices.*** ERPs at nine scalp electrodes. Gray shading indicates analyzed time windows (N250: 200-280 ms; P300: 300-380 ms; LPC: 400-800 ms). Inset bar plots show mean amplitudes for each window. Significance markers: ***$p < .001$, **$p < .01$, *$p < .05$, ns = not significant.

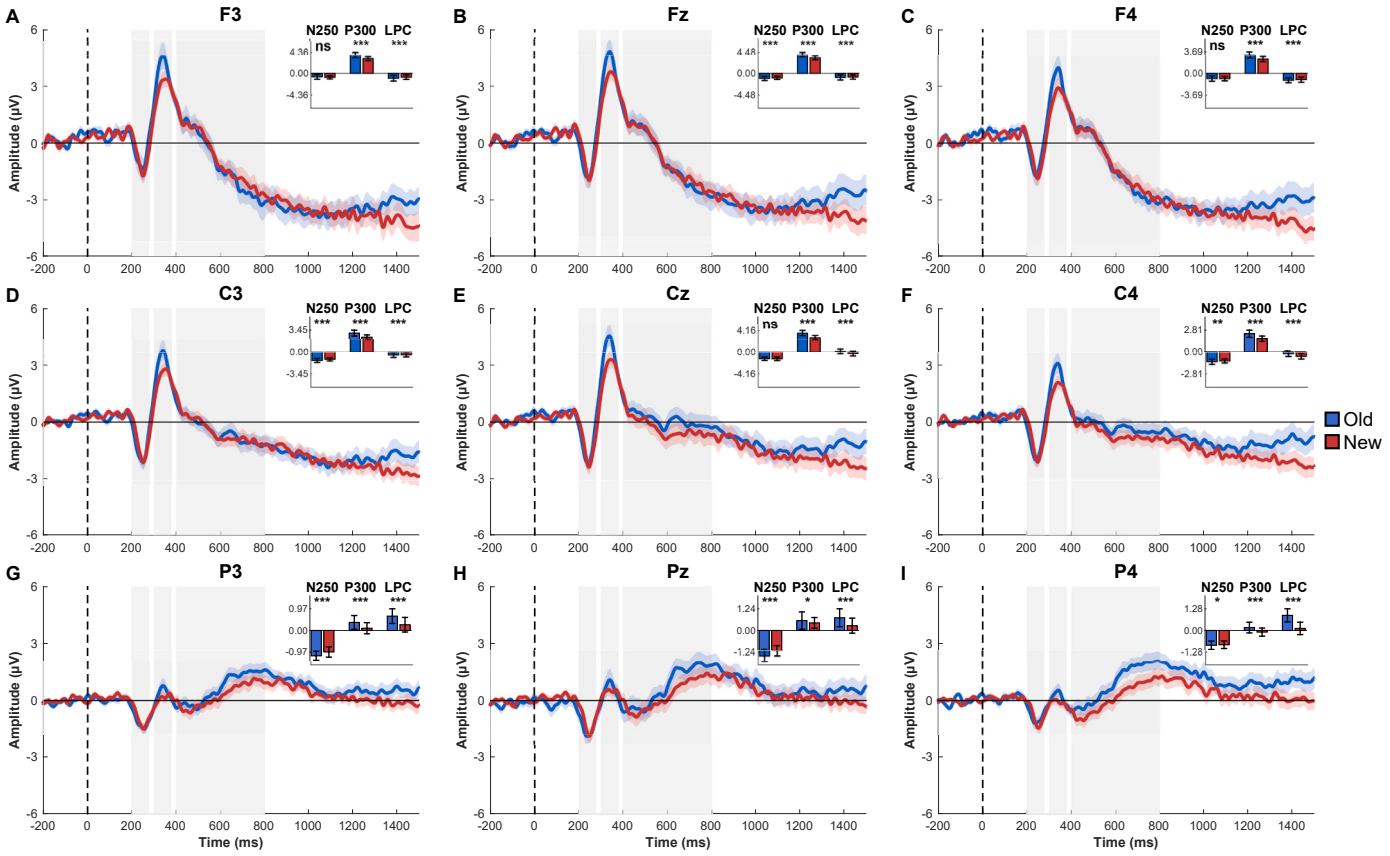

Fig. 7. ***ERP waveforms for old vs. new speaker recognition in human voices.*** ERPs at nine scalp electrodes. Gray shading indicates analyzed time windows (N250: 200-280 ms; P300: 300-380 ms; LPC: 400-800 ms). Inset bar plots show mean amplitudes for each window. Significance markers: ***$p < .001$, **$p < .01$, *$p < .05$, ns = not significant.

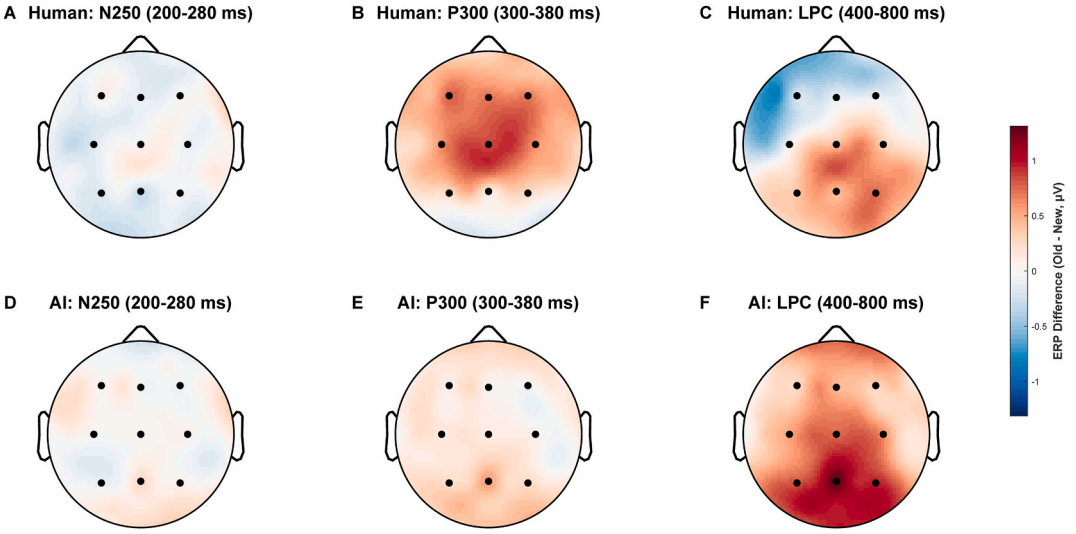

Fig. 8. ***ERP difference maps for old-new speaker recognition.*** **(A-C)** Human voices in three time windows (N250, P300, LPC). **(D-F)** AI voices in the same windows. Black circles mark the nine analyzed electrodes.

demonstrated reliable old/new discrimination across all three time windows, with convergent parietal engagement (Pz) in the LPC window. Second, spatial organization differed markedly: AI voices maintained consistent posterior-centered patterns with uniform positive polarity, while human voices showed more widespread distributions across scalp regions and late frontal polarity reversal. Third, AI voices showed larger neural responses than human voices, with progressive electrode recruitment (N250: 2 electrodes; P300: 5 electrodes; LPC: 9 electrodes), whereas human voices showed broader early engagement that stabilized (N250: 6 electrodes; P300-LPC: 9 electrodes). Fourth, temporal dynamics diverged: AI voices showed significant effect size increases across windows ($\rho = 0.735$, $p < .001$), reflecting progressive strengthening of neural discrimination, while human voices showed no significant linear trend ($\rho = 0.321$, $p = .103$), with early robust engagement.

3.4. Speaker-specific prosody expectation effect (late time window)

We analyzed mean amplitudes at nine scalp electrodes (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4) within the late window (500-900 ms) using linear mixed-effects regression to examine prosody consistency effects (same vs. different prosody relative to training). See Table S5, Table S6, Figs. 9–11.

In the late window, AI voices showed that different prosody (violation) elicited significantly greater negativity than same prosody at all nine electrodes (ordered by effect size): Pz (F (1, 684,902) = 910.16, $\beta = 1.20$, $p < .001$), P3 (F (1, 684,486) = 415.27, $\beta = 0.67$, $p < .001$), F4 (F (1, 682,677) = 196.42, $\beta = 0.59$, $p < .001$), C4 (F (1, 683,909) = 152.71, $\beta = 0.43$, $p < .001$), P4 (F (1, 684,568) = 130.58, $\beta = 0.38$, $p < .001$), Cz (F (1, 684,303) = 114.86, $\beta = 0.42$, $p < .001$), Fz (F (1, 684,344) = 98.06, $\beta = 0.43$, $p < .001$), F3 (F (1, 682,273) = 95.36, $\beta = 0.45$, $p < .001$), and C3 (F (1, 684,377) = 23.33, $\beta = 0.18$, $p < .001$).

For human voices, different prosody elicited significantly greater positivity than same prosody at eight electrodes (ordered by effect size): C4 (F (1, 697,828) = 1252.50, $\beta = -1.17$, $p < .001$), P4 (F (1, 697,779) = 934.66, $\beta = -0.93$, $p < .001$), Cz (F (1, 697,826) = 816.96, $\beta = -1.07$, $p < .001$), P3 (F (1, 697,844) = 527.43, $\beta = -0.70$, $p < .001$), Fz (F (1, 697,747) = 434.09, $\beta = -0.89$, $p < .001$), C3 (F (1, 697,823) = 428.29, $\beta = -0.72$, $p < .001$), Pz (F (1, 697,838) = 385.43, $\beta = -0.74$, $p < .001$), and F4 (F (1, 697,703) = 375.42, $\beta = -0.79$, $p < .001$). Notably, F3 showed an opposite pattern, with different prosody eliciting greater negativity (F (1, 697,694) = 18.45, $\beta = 0.22$, $p < .001$).

Overall, both voice types showed robust prosody consistency effects across all nine electrodes, but with opposite polarities: AI voices showed late negativity for prosody violations, while human voices showed late positivity.

4. Discussion

Our findings provide mixed support for both research questions. For RQ1 (speech-content-independent recognition), the prediction was partially supported: robust parietal old/new effects emerged even when utterances changed completely between training and testing, confirming abstract identity representations beyond episodic memory. Although we anticipated replicating speech-independent beta band oscillations (16-17 Hz), time-frequency MVPA revealed no significant clusters for either voice type. For RQ2 (speaker-specific prosody expectations), the prediction was supported: prosodic violations elicited robust late components, confirming that listeners bind prosodic patterns to speaker identities. Examining human and AI voices in parallel across both RQs, both voice types showed comparable parietal old/new effects, yet MVPA revealed stronger old/new discrimination for AI voices than human voices. For prosodic violations, both voice types elicited late components, though with opposite polarities: late positivity for human voices and late negativity for AI voices. Overall, the presence of shared old/new and prosody binding effects across both voice types supports our central argument that human and AI voice identities engage fundamentally similar neural mechanisms during speaker identity perception, even as group-level differences between voice types introduce systematic nuances in these shared signatures.

4.1. Content-independent speaker recognition and the contribution of name-labeling to identity encoding

Although Schweinberger et al. (2011), as introduced earlier, demonstrated content-independent parietal effects at P3, Pz, and P4 at the syllabic level, to date, no direct ERP evidence had established whether such effects extend to the utterance level. In immediately

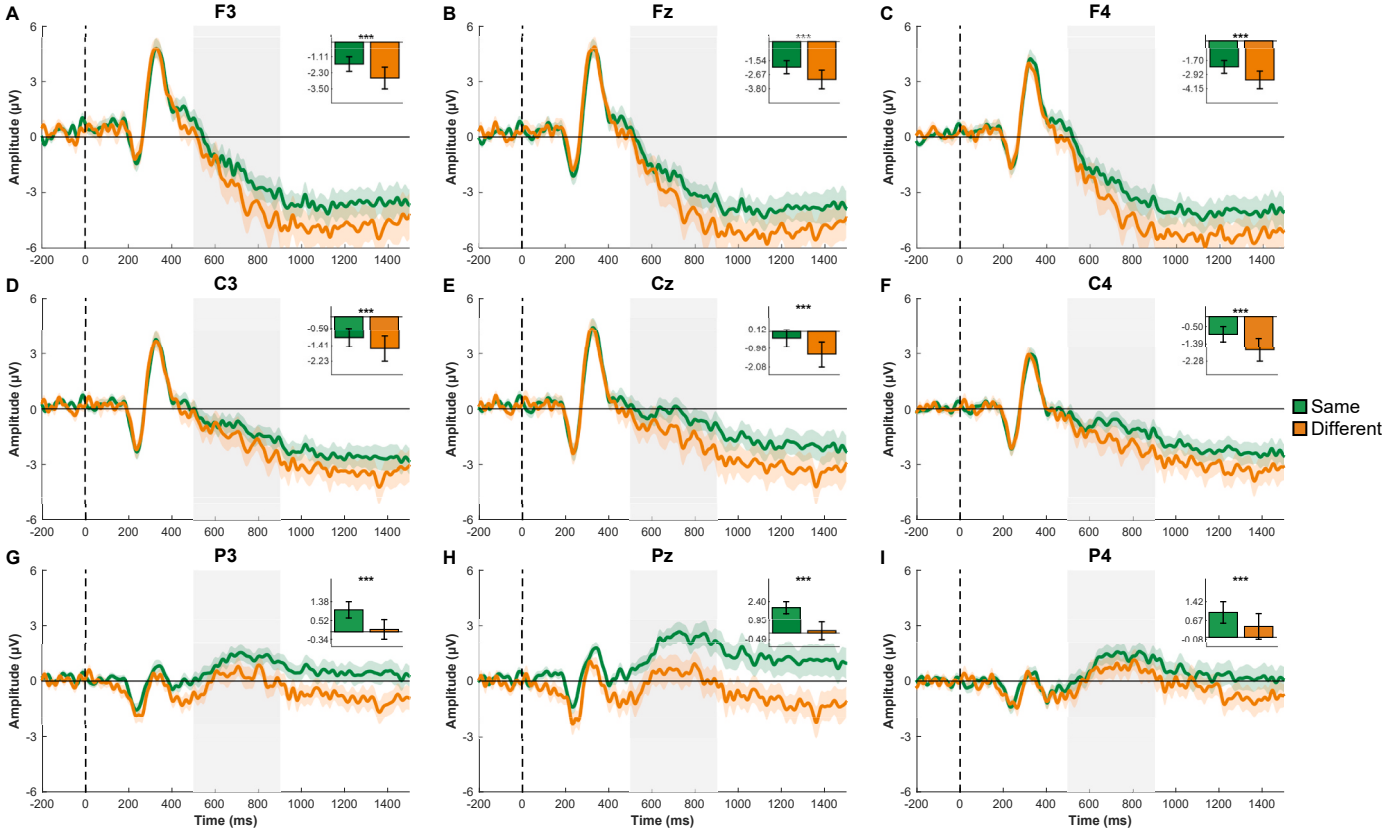

Fig. 9. ***ERP waveforms for speaker-specific prosody effects in AI voices.*** ERPs at nine scalp electrodes comparing same vs. different prosody relative to training. Gray shading indicates the late window (500-900 ms) where prosody violations elicit neural responses. Inset bar plots show mean amplitudes. ***$p < .001$, **$p < .01$, *$p < .05$, ns = not significant.

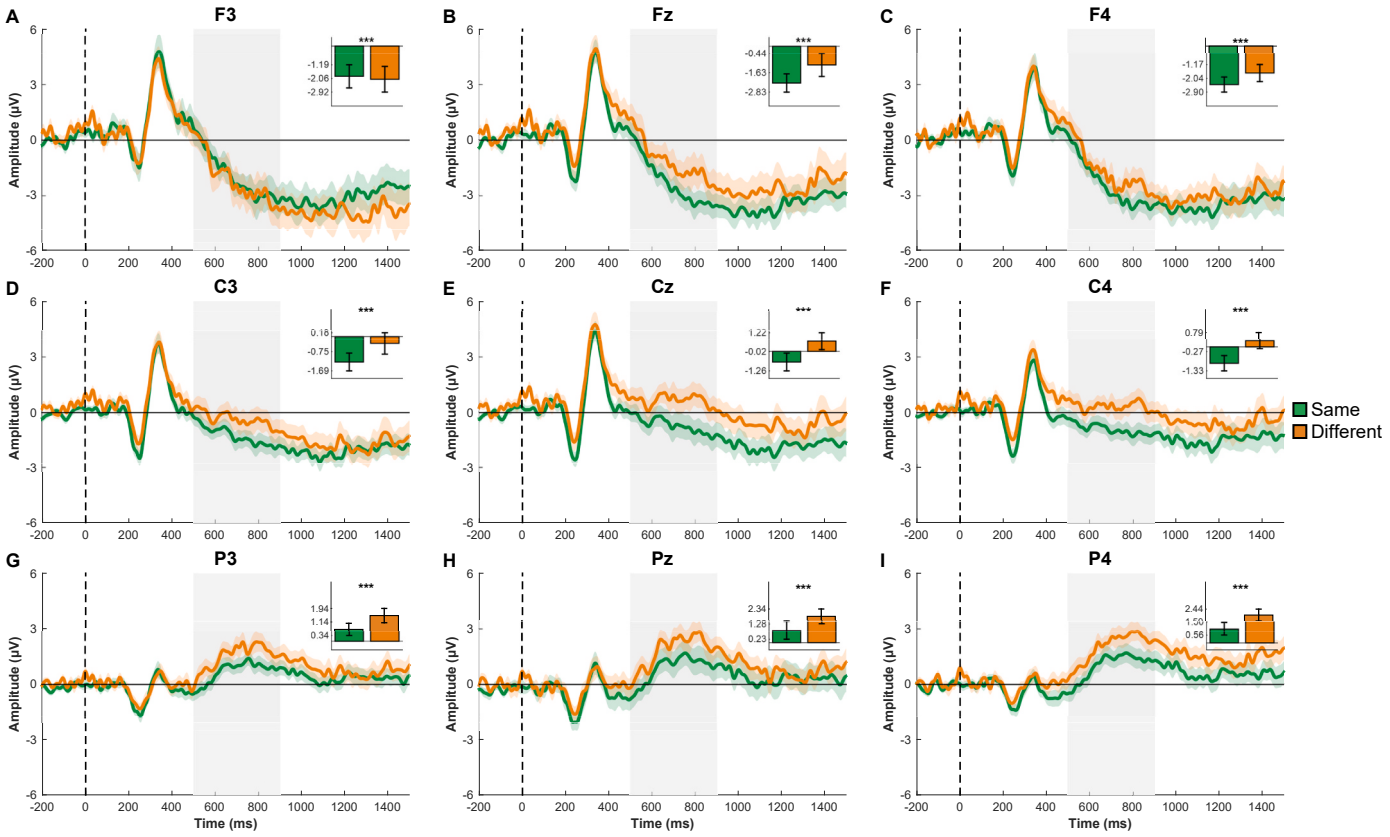

Fig. 10. ***ERP waveforms for speaker-specific prosody effects in human voices.*** ERPs at nine scalp electrodes comparing same vs. different prosody relative to training. Gray shading indicates the late window (500-900 ms) where prosody violations elicit neural responses. Inset bar plots show mean amplitudes. ***$p < .001$, **$p < .01$, *$p < .05$, ns = not significant.

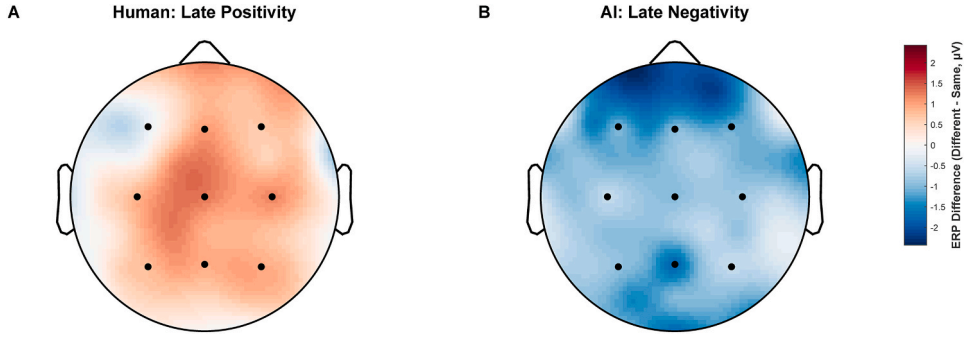

Fig. 11. ***ERP difference maps for speaker-specific prosodic expectancy violations in correctly recognized old speakers.*** **(A)** Human voices show late positivity (500-900 ms). **(B)** AI voices show late negativity (500-900 ms). Black dots indicate the nine analyzed electrodes.

relevant old/new studies at the utterance level, Zäske et al. (2014) reported ~62-70% accuracy (depending on block half) using neutral speech without prosodic variation, while Xu and Armony (2021) found accuracy dropping sharply toward chance (~39-49%) when prosody differed between training and testing (neutral vs. fearful). Our study used naturally varying prosodic styles (confident vs. doubtful) rather than neutral speech, yet replicated the accuracy pattern of Zäske et al. (2014) and showed markedly higher accuracy than Xu and Armony (2021).

Having established above-chance behavioral performance with robust identity encoding, we analyzed EEG data from correctly judged old/new trials to investigate the neural mechanisms underlying content-independent voice recognition. Multiple converging lines of evidence pointed to Pz and parietal regions as central sites for identity processing. First, MVPA identified Pz as a major contributor in the late decoding window for both voice types. Second, univariate analyses revealed significant old/new effects at Pz. Third, effect-size rankings consistently placed Pz among the strongest responders across time windows. Critically, unlike Zäske et al. (2014), who observed parietal LPC only when training and testing used identical utterances, our study demonstrated robust parietal LPC effects even when speech content changed completely between training and testing. This finding indicates that the parietal LPC can reflect speech-content-independent identity retrieval rather than episodic memory for specific utterances.

We speculate that these behavioral and neural differences arise from how identity was encoded in our paradigm. In previous studies using passive exposure (Xu and Armony, 2021; Zäske et al., 2014, 2017), listeners heard voices repeatedly without explicit identity labeling, resulting in near-chance behavioral accuracy and limited neural discrimination when content varied. In contrast, our listeners learned

each voice together with a name, and we ensured this name-voice association was fully established through checking verification (Lavan et al., 2019c). This explicit name-labeling procedure elevated both behavioral recognition accuracy and neural discriminability.

As such, we speculate that names may play a more direct role than other labels in facilitating identity consolidation. Unlike occupations or semantic attributes, names carry no inherent meaning and cannot be retrieved directly from perceptual cues such as faces or voices; instead, they must be accessed through person-identity codes (McWeeny et al., 1987). Because names are cognitively unique identity labels that require stronger encoding than other person information, successfully binding a name to a newly learned voice may strengthen person-level representations. In the framework of the looping mechanism (Maguinness et al., 2018), where repeated recognition episodes progressively strengthen voice prototype representations, name-based learning (McWeeny et al., 1987) may establish more robust identity anchors than arbitrary symbols or numbers, facilitating the consolidation of stable voice identities. However, this speculation lacks direct supporting evidence and requires controlled comparisons in future research.

So far, we cannot definitively attribute these neural signatures solely to name-labeling. On one hand, our trained-familiar voices with name associations still differ qualitatively from genuinely intimate voices. Voices of close family members, long-term friends (Plante-Hébert et al., 2021), or well-known public figures (Rinke et al., 2022) carry autobiographical, affective, and deeply consolidated semantic associations that likely exceed what can be induced through brief laboratory training. On the other hand, our name-labeling procedure may have achieved its effects through mechanisms shared with other explicit identity anchors. Previous studies have linked voices to icons (Cooper et al., 2024; Perrachione et al., 2011) or numbers (Xie and Myers, 2015), and these alternative labels also enhance familiarity relative to passive exposure. Whether names confer unique advantages over other explicit labels, or whether any salient identity anchor would produce similar content-independent representations, remains an empirical question for future research. Controlled comparisons directly contrasting name-labeling with icon-labeling or number-labeling within the same paradigm would be necessary to isolate the specific contribution of names.

Nevertheless, our LPC findings suggest a possible pathway through which name-labeling may elevate newly learned voices toward intimate familiarity. In Plante-Hébert et al. (2021), voice familiarity spans three levels: unfamiliar voices; trained-to-familiar voices acquired through repeated exposure without explicit identity labeling; and intimately familiar voices such as family and friends. Intimate familiarity produces two key neural signatures: strongly parietal spatial distributions extending from Pz across posterior midline electrodes, and robust late positive components reflecting retrieval of rich person-specific representations. Our old/new effects exhibited similar patterns. The spatial distribution was strongly parietal, with Pz as the strongest contributor and additional posterior sites showing reliable discrimination. Moreover, we observed a robust LPC even when speech content varied, suggesting listeners accessed stable voice identity representations rather than episodic traces of specific utterances. These convergences raise the possibility that pairing a voice with a name and ensuring listeners correctly learn this association elevates newly learned voices toward a level of familiarity that approximates key aspects of intimate voice processing. Thus, we speculate that name-voice association learning could represent a foundational process by which new acquaintances transition from strangers to socially familiar individuals (Lavan and McGettigan, 2023; Maguinness et al., 2018; Sidtis and Kreiman, 2012).

We also note that Zäske et al. (2014) additionally reported beta band (16-17 Hz) effects during 290-370 ms for old vs. new voices using univariate power analyses at specific electrode clusters. Our time-frequency MVPA (**Supplementary Analysis 1**) did not replicate this effect. Two methodological differences may account for this discrepancy. First, Zäske et al. (2014) employed univariate analyses testing frequency-specific power differences at individual electrodes, whereas our multivariate decoding classified conditions based on distributed spatial patterns across all electrodes, which may be less sensitive to localized frequency-specific effects. Second, our paradigm introduced prosodic variability (confident vs. doubtful) that was absent in the neutral-prosody-only design of Zäske et al. (2014). Future work could test the replicability of these beta effects.

4.2. Speaker-specific identity representations remain stable across prosodic variation

In predictive processing, the brain continuously generates predictions to minimize mismatches between expected and received sensory inputs (Friston and Kiebel, 2009). We analyzed ERP data from trials where participants successfully identified old speakers and found robust late components, supporting our hypothesis that prosody serves as a speaker-specific predictive cue. Specifically, old speaker-based prosodic violations elicited robust late neural responses in both voice types, but with opposite polarities: human voices showed enhanced late positivity, while AI voices showed enhanced late negativity.

For human voices, this late positivity reflects pragmatic repair mechanisms similar to those observed when listeners reconcile conflicting vocal confidence cues (Jiang and Pell, 2016). When prosodic violations remain communicatively coherent, as in our paradigm where speaker identity stays constant despite prosodic shifts, listeners engage inferential reinterpretation rather than rejecting the speaker representation. This pattern aligns with P600 components elicited by violations of speaker-specific communicative styles, including syntactic structure preferences (), ironic tendencies (Regel et al., 2010), and gesture patterns (Obermeier et al., 2015). Across these domains, late positivities emerge when violations are recoverable, that is, when the mismatch can be reconciled with stored speaker representations without rejecting the speaker's identity or communicative intent. Our findings extend this mechanism to prosodic style, demonstrating that listeners bind prosodic patterns to speaker identities and flexibly update expectations when encountering within-speaker prosodic variation.

In contrast, AI voices elicited late negativity, reflecting effortful reprocessing mechanisms. This pattern parallels findings that atypical vocal signals create processing challenges. Jiang et al. (2020) found that compared to local Canadian English, non-native regional accents (Australian English) and L2-accented speech (French-accented English) combined with doubtful vocal confidence elicited sustained late negativity, reflecting integration difficulties when constructing speaker representations from atypical vocal patterns. Similarly, AI-generated voices may be processed as a distinct type of speech signal analogous to unfamiliar accents (Roswandowitz et al., 2024): although listeners successfully recognize speaker identity, the synthetic origin creates a baseline processing challenge. When prosodic violations occur within this already-atypical vocal framework, listeners must reconcile expectations across prosodic styles while also processing artificially-generated acoustic patterns. This dual challenge pushes processing toward effortful reinterpretation routes involving inhibition and recomputation (Jiang et al., 2013), rather than the flexible updating mechanisms available for natural human voices, where only prosodic expectations require adjustment.

The significance of these findings becomes evident when considering our cloning procedure. Human prosodic variations originated from the same speaker across identical utterances, maintaining consistent identity. AI voices presented a more stringent test: each speaker's confident and doubtful clones were independently trained from separate 15-utterance corpora, with no cross-training between prosodic styles. The AI models then generated 123 novel utterances they had never encountered during training, which human speakers subsequently recorded one month later. Despite this algorithmic separation and the generative nature of AI speech, both voice types elicited robust late-window ERP differences for prosodic violations, demonstrating that listeners formed

unified identity representations. This provides direct evidence that speaker identity remains sufficiently stable across prosodic variations, provided the variations are not overly expressive (Lavan et al., 2019a). Both AI generative models and human listeners successfully extracted shared identity characteristics despite within-speaker prosodic variation. These findings support theories of flexible within-speaker identity processing (Lavan et al., 2019b).

Our study extends research on speaker-specific communicative styles by examining prosody, which differs critically from syntactic or other stylistic cues in that prosody additionally shifts speaker identity representations through VTL and F0 adjustments (Lavan et al., 2019c). This creates a unique processing challenge: listeners must match the current speaker's identity representation with stored representations of familiar talkers despite prosodic differences (Lavan and McGettigan, 2023), which requires greater effort than matching identical prosodic styles (e.g., both neutral). Only after successful identity retrieval can listeners process pragmatic information conveyed by prosodic violations. Thus, prosody introduces an interaction between identity-level adaptation to within-speaker acoustic variation and pragmatic-level processing of communicative style, demonstrating the flexible capacity of the auditory system to accommodate both levels simultaneously.

4.3. Old/new and prosody binding effects: hierarchical processing, parietal networks, and mechanistic considerations

The late parietal old/new effects observed in the present study can be situated within the hierarchical framework of voice processing (Belin et al., 2004), in which initial acoustic analyses in temporal voice areas are followed by increasingly abstract representations culminating in person identity nodes. Under the most stringent test, MVPA identified significant old/new decoding for AI voices in three late clusters (662-1498 ms), with contributing electrodes concentrated at posterior parietal regions, particularly Pz. Although human voices did not independently yield significant MVPA clusters, applying the AI-derived windows revealed significant old/new ERP effects at the same posterior parietal sites. This convergence on posterior parietal regions across both voice types aligns with evidence that speaker identities can be decoded from parietal regions alongside temporal voice areas (Lamothe et al., 2026), suggesting that our late effects reflect processing at later stages of the voice perception hierarchy, corresponding to the retrieval of abstract identity representations rather than early acoustic processing.

The prominence of Pz warrants further consideration in light of Zäske et al. (2014), who observed a parietal LPC at Pz only when training and test utterances were identical, interpreting it as reflecting detailed explicit retrieval of the study episode from episodic memory. Crucially, our findings extend this interpretation: we observed robust Pz effects even when speech content changed completely between learning and testing, suggesting that parietal activity at this site is not limited to episodic retrieval of specific utterances. This aligns with Schweinberger et al. (2011), who demonstrated that parietal effects (at P3, Pz, and P4) reflecting voice identity processing persisted across different syllables, concluding that voice identity processing becomes independent of speech content after ~300 ms. Together, these findings suggest that Pz indexes a more abstract level of identity processing that generalizes beyond speech-content-dependent episodic memory.

Although EEG does not permit strong claims regarding spatial generators, the prominence of Pz invites speculative interpretation in light of converging fMRI evidence. The midline parietal focus may reflect activity within the default mode network, particularly the angular gyrus and precuneus, which most consistently associate with old/new recollection effects and are thought to support the mental re-experiencing of previously encoded events (Kim, 2013; Sestieri et al., 2011). Simultaneous EEG-fMRI recordings have further linked the late parietal old/new effect to activation in the posterior hippocampus and parahippocampal cortex (Hoppstädter et al., 2015). Andics et al. (2010) demonstrated anatomically separable representations in fMRI: a voice-acoustics space in bilateral middle/posterior STS reflecting short-term acoustic similarity processing, and a voice-identity space in regions including the anterior temporal pole and amygdala reflecting long-term stored identity representations. These identity-sensitive regions show functional proximity to the default mode network, suggesting that Pz may index the interface between voice-identity representations in temporal regions and the broader recollection network, though future research combining EEG with methods offering greater spatial resolution is needed to directly test this interpretation.

The late timing of our effects raises the question of whether they necessitate feedback mechanisms from higher-order identity representations back to early acoustic processing stages. Our two effects appear to reflect different levels of top-down processing. The old/new effects are consistent with the activation of voice-identity space (Andics et al., 2010), whereby the brain recruits long-term stored identity representations to match incoming acoustic input. The prosody binding effects go further: once speaker identity is retrieved, stored speaker-specific prosodic representations appear to generate predictions against which incoming prosody is compared, producing expectancy violation signals when mismatched, more consistent with predictive feedback from identity representations to lower-level prosodic processing (Friston and Kiebel, 2009). It should be noted, however, that EEG alone cannot establish the directionality of these neural interactions, and future research will likely require more sophisticated computational approaches to causal inference, such as dynamic causal modeling, to directly address this question.

4.4. AI voices as methodological tools for isolating identity processing?

Despite similar behavioral patterns for the old/new effects, our MVPA analyses revealed a notable pattern: AI voices showed three significant decoding clusters for old/new discrimination, whereas human voices showed no significant clusters independently (Fig. 5). However, human voices showed comparable old/new ERP effects when analyzed using AI-identified time windows and electrodes in linear mixed-effects models. This asymmetry likely stems from reduced prosodic variability in AI voices. Although both voice types conveyed prosodic distinctions (see Chen et al. (2026) and Fig. 1), human voices exhibited substantially larger prosodic differences. Because our old/new analyses collapsed across prosodic conditions, greater prosodic heterogeneity in human voices may have obscured identity patterns in MVPA, whereas AI voices provided cleaner signals for identity extraction.

Beyond prosodic variation, AI voices may exhibit greater internal homogeneity than human voices (Chen et al., 2024). Natural human speech carries rich variation in paralinguistic, emotional, and social-affective dimensions that engage broader neural networks (Roswandowitz et al., 2024). While this richness reflects the complexity of human voice processing, it can introduce confounds when investigating specific cognitive processes. AI-generated voices, by minimizing these extraneous factors, may serve as valuable methodological tools for isolating core identity processing mechanisms, such as investigating neural sensitivity to different levels of voice familiarity (Ma et al., 2026; McGettigan et al., 2025) without confounding effects from emotional, prosodic, or social-affective variation.

Our prosody consistency effect further hints at this methodological advantage. Although neither voice type showed significant MVPA clusters for same vs. different prosody discrimination, AI voices maintained above-chance decoding accuracy while human voices fluctuated around chance levels. Human voices carry richer incidental variation, such as pronunciation inconsistencies or accent features that introduce noise beyond the intended prosodic manipulation (Jiang et al., 2018, 2020; Leemann et al., 2018; Ulbrich and Mennen, 2016). AI-generated voices, by delivering standardized outputs, maximize signal-to-noise ratio for experimentally manipulated dimensions. For example, Di Cesare et al. (2022) introduced a robotic voice lacking prosodic

intonation as a control condition to isolate the contribution of human vocal prosody. AI voice synthesis similarly offers a useful tool for investigating how specific acoustic dimensions influence identity processing by isolating target variables while holding other factors constant.

Meanwhile, we need to highlight that univariate ERP analyses revealed early processing advantages for human voices that MVPA could not detect. In the N250 window (200-280 ms), human voices showed old/new discrimination across six electrodes vs. only two for AI voices, suggesting faster identity extraction. Conversely, MVPA identified significant decoding clusters for AI voices but not human voices. Thus, the two voice types engage temporally distinct processing trajectories: human voices demonstrate early robust discrimination, whereas AI voices show late-stage multivariate decoding advantages.

We also note that human voices demonstrated overall learning advantages in the Checking phase. Within AI voices, confident prosody yielded lower accuracy than doubtful prosody, whereas human voices showed no prosodic differences. This suggests AI voices with confident prosody have greater internal similarity among learned speakers despite our height-matched controls. These findings indicate identity learning operates holistically and is influenced by prosody, with future research needing to attend to differential prosodic effects when using AI voices.

Overall, while AI voices demonstrated MVPA advantages and human voices showed early univariate discrimination, both approaches converged on identifying Pz as a critical contributor. This suggests that despite temporally distinct trajectories, both voice types engage overlapping parietal substrates for identity processing. The choice should depend on research goals: AI voices for minimizing confounding variation, human voices for ecological validity, and early processing dynamics.

4.5. Implications for the perception of AI-generated "humans"

Although both voice types showed old/new and prosody binding effects, systematic differences emerged between human and AI voices: MVPA identified significant decoding clusters only for AI voices, and prosody binding effects showed opposite polarities across voice types. These patterns are consistent with neuroimaging evidence that human and AI voices engage partially distinct neural pathways (Bratan et al., 2025; Roswandowitz et al., 2024), and extend such findings to the temporal dynamics of identity and prosodic processing. It should be noted, however, that the AI voices used in the present study were clearly distinguishable from human voices, selected from a corpus of 11,808 recordings based on perceptual validation ratings, with humanlikeness ratings substantially lower than human voices. Whether the neural distinctions observed here would persist as AI voice quality improves remains an open question.

In the visual domain, AI-generated faces have been shown to exceed human faces in perceived realism, a phenomenon termed hyperrealism (Miller et al., 2023). For voices, however, recent evidence suggests that this threshold has not yet been reached, with voice clones sounding similarly real to human voices but not more so (Lavan et al., 2025). This cross-modal asymmetry underscores the importance of future research examining how the neural mechanisms of AI voice identity processing evolve as synthesis technology continues to advance.

4.6. Limitations and future directions

Our fixed block order (human blocks 1-4, then AI blocks 5-8) requires special attention, as earlier blocks typically show stronger identity encoding advantages than later blocks (Xu and Armony, 2021; Zäske et al., 2014), suggesting alternating human/AI blocks might be preferable. However, our Checking phase results (Fig. 3A) justify this design. Despite appearing in later blocks with potential cumulative learning advantages, AI voices showed systematically lower accuracy than human voices, particularly for AI confident prosody. This persistent disadvantage, despite later positioning, suggests fundamental encoding differences. As discussed above, AI confident voices likely exhibited greater internal perceptual similarity among the three learned speakers. Had we alternated human/AI blocks while counterbalancing prosody order, AI voices might have shown even greater disadvantages, and this would have disrupted our human voice recognition baseline needed for comparison with prior exclusively-human-voice literature.

As we examined only two pragmatically-marked prosodic types (confident and doubtful) rather than more basic emotional prosody, such as happy or sad (Larrouy-Maestri et al., 2025), future research could also extend or replicate our paradigm using emotional prosody (Xu and Armony, 2021), while ensuring within-speaker variation remains within appropriate ranges (Lavan et al., 2019a).

Recall as well that our study analyzed human and AI voices separately rather than including a direct statistical contrast between them, instead inferring shared underlying mechanisms from parallel patterns within each voice type (old/new effects and speaker-specific prosodic expectation violations). As such, our current design cannot address several important questions: can listeners rapidly perceive the distinction between human and AI voices at early processing stages (Jiang and Pell, 2024; Lavan et al., 2024)? Does this human/AI discrimination depend on prosodic cues (Kühne et al., 2020; Rodero and Lucas, 2023)? Future research should investigate these questions to clarify the temporal dynamics and cognitive mechanisms underlying human vs. AI voice discrimination.

A further limitation concerns the use of a single voice cloning system in Mandarin Chinese. Although our stimuli underwent extensive perceptual validation, drawn from 11,808 recordings with 24 speakers represented in both human and AI versions with systematic prosodic variation (Chen et al., 2026), findings may not fully generalize to AI voices produced by other systems or in other languages. Future EEG studies comparing AI voice clones with human voices should attend carefully to stimulus comparability, including perceptual screening and control of within-speaker variation when prosodic manipulation is involved. Alternatively, studies examining human vs. AI voice processing without prosodic manipulation could employ multiple AI TTS systems alongside matched human voices. Such studies could verify perceptual differences between voice types prior to EEG data collection using humanlikeness ratings (Chen et al., 2025), given that highly realistic AI products risk introducing confounds that obscure genuine human vs. AI processing distinctions (Miller et al., 2023).

5. Conclusion

Our study addresses current gaps in voice processing research. First, while parietal old/new ERP effects had previously been demonstrated at the syllabic level across different speech content (Schweinberger et al., 2011), and at the utterance level only when utterances were repeated at test (Zäske et al., 2014), the present study extends these findings by demonstrating robust parietal old/new ERP effects at the utterance level even when speech content changed completely between training and testing, indicating abstract identity representations beyond episodic memory.

Second, listeners construct predictive models of interlocutors' communicative behaviors based on available social cues, forming expectations for speaker-specific patterns (Kroczek et al., 2019; Obermeier et al., 2015; Regel et al., 2010); by manipulating prosodic style under controlled conditions, we extend this principle to the prosodic dimension of voice identity. Within this effect, for human voices, our observed late positivity aligns with speaker-specific communicative style violation effects; for AI voices, the observed late negativity that also signals expectation violations suggests that AI voice perception may resemble accented speech perception (Jiang et al., 2020).

Third and most importantly, the observed comparable old/new effects across voice types indicate that identity perception mechanisms operate at a general level (Giamundo et al., 2025; Schweinberger et al.,

2011; Zäske et al., 2017); our findings extend this principle by demonstrating that such mechanisms transcend the biological vs. algorithmic origins of vocal signals.

Our findings also provide a theoretical basis for understanding AI voice agents in commercial applications: listeners do encode and remember AI speaker identities, suggesting that identity-related attributes may influence user acceptance and long-term engagement with synthetic voices (Brandtzaeg et al., 2022; Jones et al., 2021).

Data and code availability

Behavioral data, EEG data, analysis code, and example stimuli are available at the Open Science Framework: https://osf.io/9b8pq/overview?view_only=360d32c442e4440e89859d379192043a.

Ethics approval

This study was conducted in accordance with the Declaration of Helsinki and approved by the institutional ethics committee. The ethical committee of the Institute of Language Sciences, Shanghai International Studies University, approved the experiment (20230628027).

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

During the preparation of this work, the authors used Claude Sonnet 4.5 (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 utterance clarity. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

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

CRediT authorship contribution statement

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

Acknowledgement

We would like to thank the two anonymous reviewers for their constructive comments and suggestions, which greatly improved the manuscript. We would also like to thank Dr. Xiaolin Zhou for his encouragement, suggestions, and comments on this project throughout WJC's Master's studies at Shanghai International Studies University, from the thesis proposal, through his academic writing course, to serving as thesis committee chair.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2026.109493.

References

Adam-Darque, A., Pittet, M.P., Grouiller, F., Rihs, T.A., Leuchter, R.H.-V., Lazeyras, F., Michel, C.M., Hüppi, P.S., 2020. Neural correlates of voice perception in newborns and the influence of preterm birth. Cerebr. Cortex 30 (11), 5717–5730. https://doi.org/10.1093/cercor/bhaa144.

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

Anikin, A., Pisanski, K., Massenet, M., Reby, D., 2021. Harsh is large: nonlinear vocal phenomena lower voice pitch and exaggerate body size. Proceedings of the Royal Society B 288 (1954), 20210872. https://doi.org/10.1098/rspb.2021.0872.

Apple, 2023. Advancing speech accessibility with personal voice, 2023. https://machinelearning.apple.com/research/personal-voice.

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

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends Cognit. Sci. 8 (3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008.

Belyk, M., Waters, S., Kanber, E., Miquel, M.E., McGettigan, C., 2022. Individual differences in vocal size exaggeration. Sci. Rep. 12 (1), 2611. https://doi.org/10.1038/s41598-022-05170-6.

Boersma, P., Weenink, D., 2021. Praat: Doing Phonetics by Computer, Version 6.2.09. https://www.praat.org/.

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

Bratan, C.A., Marinescu, A., Terecoasa, E., Tebeanu, A.V., Morosanu, B., Franti, E., Dascalu, M., Andrei, A., Tocila-Matasel, C., Ionescu, B., 2025. Mirror neurons cannot be fooled by artificial Voices–a study with implications for education using magnetic resonance imaging (MRI) and convolutional neural network (CNN). International Journal of Education and Information Technologies 19, 120–127. https://doi.org/10.46300/9109.2025.19.12.

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. Proc. Speech Prosody. https://doi.org/10.21437/SpeechProsody.2024-171, 2024, Leiden, The Netherlands.

Chen, W., Pell, M.D., Jiang, X., 2025. Does speech prosody shape social perception equally for AI and human voices? A 16-Dimension rating study. Preprints. https://doi.org/10.20944/preprints202510.1492.v1.

Chen, W., Pell, M.D., Jiang, X., 2026. Prosodic cues strengthen human-AI voice boundaries: listeners do not easily perceive human speakers and AI clones as the same person. Comput. Hum. Behav.: Artificial Humans 7, 100261. https://doi.org/10.1016/j.chbah.2026.100261.

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.

Cooper, A., Eitel, M., Fecher, N., Johnson, E., Cirelli, L.K., 2024. Who is singing? Voice recognition from spoken versus sung speech. JASA Express Letters 4 (6). https://doi.org/10.1121/10.0026385.

Corretgé, R., 2024. Praat Vocal Toolkit. https://www.praatvocaltoolkit.com.

Darwin, C.J., Brungart, D.S., Simpson, B.D., 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. J. Acoust. Soc. Am. 114 (5), 2913–2922. https://doi.org/10.1121/1.1616924.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134 (1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009.

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

Fish, K., Rothermich, K., Pell, M.D., 2017. The sound of (in)sincerity. J. Pragmat. 121, 147–161. https://doi.org/10.1016/j.pragma.2017.10.008.

Friston, K., Kiebel, S., 2009. Predictive coding under the free-energy principle. Phil. Trans. Biol. Sci. 364 (1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300.

Giamundo, M., Trapeau, R., Thoret, E., Renaud, L., Brochier, T., Belin, P., 2025. Voice identity invariance by anterior temporal lobe neurons. Sci. Adv. 11 (35). https://doi.org/10.1126/sciadv.adv7033 eadv7033.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067.

Hoppstädter, M., Baeuchl, C., Diener, C., Flor, H., Meyer, P., 2015. Simultaneous EEG–fMRI reveals brain networks underlying recognition memory ERP old/new effects. Neuroimage 116, 112–122. https://doi.org/10.1016/j.neuroimage.2015.05.026.

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. Q. J. Exp. Psychol. 73 (1), 55–79. https://doi.org/10.1177/1747021819865833.

Jiang, X., Li, Y., Zhou, X., 2013. Even a rich man can afford that expensive house: ERP responses to construction-based pragmatic constraints during sentence comprehension. Neuropsychologia 51 (10), 1857–1866. https://doi.org/10.1016/j.neuropsychologia.2013.06.009.

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.

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., 2016. The feeling of another's knowing: how “mixed messages” in speech are reconciled. J. Exp. Psychol. Hum. Percept. Perform. 42 (9), 1412–1428. https://doi.org/10.1037/xhp0000240.
Jiang, X., Pell, M.D., 2017. The sound of confidence and doubt. Speech Commun. 88, 106–126. https://doi.org/10.1016/j.specom.2017.01.011.
Jiang, X., Pell, M.D., 2024. Tracking dynamic social impressions from multidimensional voice representation. Trends Cognit. Sci. https://doi.org/10.1016/j.tics.2024.08.005.
Jones, V.K., Hanus, M., Yan, C., Shade, M.Y., Blaskewicz Boron, J., Maschieri Bicudo, R., 2021. Reducing loneliness among aging adults: the roles of personal voice assistants and anthropomorphic interactions. Front. Public Health 9, 2021. https://doi.org/10.3389/fpubh.2021.750736 [Brief Research Report].
Khanjani, Z., Watson, G., Janeja, V.P., 2023. Audio deepfakes: a survey. Front. Big Data 5, 2022. https://doi.org/10.3389/fdata.2022.1001063 [Review].
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. Front. Neurorob. 14, 593732. https://doi.org/10.3389/fnbot.2020.593732.
Kim, H., 2013. Differential neural activity in the recognition of old versus new events: an activation likelihood estimation meta-analysis. Hum. Brain Mapp. 34 (4), 814–836. https://doi.org/10.1002/hbm.21474.
Kim, J., Toutios, A., Lee, S., Narayanan, S.S., 2020. Vocal tract shaping of emotional speech. Comput. Speech Lang 64, 101100. https://doi.org/10.1016/j.csl.2020.101100.
Kroczek, L.O.H., Gunter, T.C., Rysop, A.U., Friederici, A.D., Hartwigsen, G., 2019. Contributions of left frontal and temporal cortex to sentence comprehension: evidence from simultaneous TMS-EEG. Cortex 115, 86–98. https://doi.org/10.1016/j.cortex.2019.01.010.
Kroczek, L.O.H., Gunter, T.C., 2021. The time course of speaker-specific language processing. Cortex 141, 311–321. https://doi.org/10.1016/j.cortex.2021.04.017.
Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Software 82 (13), 1–26. https://doi.org/10.18637/jss.v082.i13.
Lamothe, C., Giamundo, M., Belin, P., 2026. Voice information processing by the primate brain. Trends Cognit. Sci. https://doi.org/10.1016/j.tics.2026.02.005.
Larrouy-Maestri, P., Poeppel, D., Pell, M.D., 2025. The sound of emotional prosody: nearly 3 decades of research and future directions. Perspect. Psychol. Sci. 20 (4), 623–638. https://doi.org/10.1177/17456916231217722.
Lavan, N., Burston, L.F., Ladwa, P., Merriman, S.E., Knight, S., McGettigan, C., 2019a. Breaking voice identity perception: expressive voices are more confusable for listeners. Q. J. Exp. Psychol. 72 (9), 2240–2248. https://doi.org/10.1177/1747021819836890.
Lavan, N., Burton, A.M., Scott, S.K., McGettigan, C., 2019b. Flexible voices: identity perception from variable vocal signals. Psychon. Bull. Rev. 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), e0332692. https://doi.org/10.1371/journal.pone.0332692.
Lavan, N., Knight, S., McGettigan, C., 2019c. Listeners form average-based representations of individual voice identities. Nat. Commun. 10 (1), 1–9. https://doi.org/10.1038/s41467-019-10295-w.
Lavan, N., Rinke, P., Scharinger, M., 2024. The time course of person perception from voices in the brain. Proc. Natl. Acad. Sci. 121 (26), e2318361121. https://doi.org/10.1073/pnas.2318361121.
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.
Leemann, A., Kolly, M.-J., Nolan, F., Li, Y., 2018. The role of segments and prosody in the identification of a speaker's dialect. J. Phonetics 68, 69–84. https://doi.org/10.1016/j.wocn.2018.02.001.
Lenth, R., Singmann, H., Love, J., Buerkner, P., Herve, M., 2021. Emmeans: estimated marginal means, Aka least-squares means. The Comprehensive R Archive Network. R package version 1.5. 1.)[Computer software]. https://cran.r-project.org/web/packages/emmeans/index.html.
Ma, Y., Niu, Z., Zhang, X., Zhang, Y., Liu, Z., Yu, K., Wang, R., 2026. Differentiation of physically and psychologically familiar voices and their roles in spoken word processing: evidence from ERPs and neural oscillation. Neuropsychologia 221, 109309. https://doi.org/10.1016/j.neuropsychologia.2025.109309.
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.
Mathias, S.R., von Kriegstein, K., 2019. Voice processing and voice-identity recognition. In: Siedenburg, K., Saitis, C., McAdams, S., Popper, A.N., Fay, R.R. (Eds.), Timbre: Acoustics, Perception, and Cognition. Springer International Publishing, pp. 175–209. https://doi.org/10.1007/978-3-030-14832-4_7.
Mauchand, M., Vergis, N., Pell, M.D., 2020. Irony, prosody, and social impressions of affective stance. Discourse Process. 57 (2), 141–157. https://doi.org/10.1080/0163853X.2019.1581588.
McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., Malek, A., Raffel, C., Lostanlen, V., van Niekirk, B., Lee, D., Cwitkowitz, F., Zalkow, F., Nieto, O., Ellis, D., Mason, J., Lee, K., Steers, B., Südholt, D., 2025. Librosa/librosa: 0.11.0. Zenodo. https://doi.org/10.5281/zenodo.15006942.
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.
McWeeny, K.H., Young, A.W., Hay, D.C., Ellis, A.W., 1987. Putting names to faces. Br. J. Psychol. 78 (2), 143–149. https://doi.org/10.1111/j.2044-8295.1987.tb02235.x.
Miller, E.J., Steward, B.A., Witkower, Z., Sutherland, C.A.M., Krumhuber, E.G., Dawel, A., 2023. AI hyperrealism: why AI faces are perceived as more real than human ones. Psychol. Sci. 34 (12), 1390–1403. https://doi.org/10.1177/09567976231207095.
Müller, N.M., Pizzi, K., Williams, J., 2022. Human perception of audio deepfakes. https://doi.org/10.1145/3552466.3556531.
Neuhaus, T.J., Scherer, R.C., Whitfield, J.A., 2024. Gender perception of speech: dependence on fundamental frequency, implied vocal tract length, and source spectral tilt. J. Voice. https://doi.org/10.1016/j.jvoice.2024.01.014.
Nussbaum, C., Frühholz, S., Schweinberger, S.R., 2025. Understanding voice naturalness. Trends Cognit. Sci. 29 (5), 467–480. https://doi.org/10.1016/j.tics.2025.01.010.
Obermeier, C., Kelly, S.D., Gunter, T.C., 2015. A speaker's gesture style can affect language comprehension: ERP evidence from gesture-speech integration. Soc. Cognit. Affect Neurosci. 10 (9), 1236–1243. https://doi.org/10.1093/scan/nsv011.
Pell, M.D., Skorup, V., 2008. Implicit processing of emotional prosody in a foreign versus native language. Speech Commun. 50 (6), 519–530. https://doi.org/10.1016/j.specom.2008.03.006.
Perrachione, T.K., Del Tufo, S.N., Gabrieli, J.D., 2011. Human voice recognition depends on language ability. Science 333 (6042). https://doi.org/10.1126/science.1207327, 595-595.
Pinheiro, A.P., 2025. Behind a voice there is a speaker: why vocal emotion research needs to become ‘personal’. Affective Science 6 (3), 562–574. https://doi.org/10.1007/s42761-025-00317-w.
Pisanski, K., Anikin, A., Reby, D., 2022. Vocal size exaggeration May have contributed to the origins of vocalic complexity. Philos. Trans. R. Soc. B 377 (1841), 20200401. https://doi.org/10.1098/rstb.2020.0401.
Plante-Hébert, J., Boucher, V.J., Jemel, B., 2021. The processing of intimately familiar and unfamiliar voices: specific neural responses of speaker recognition and identification. PLoS One 16 (4), e0250214. https://doi.org/10.1371/journal.pone.0250214.
Regel, S., Coulson, S., Gunter, T.C., 2010. The communicative style of a speaker can affect language comprehension? ERP evidence from the comprehension of irony. Brain Res. 1311, 121–135. https://doi.org/10.1016/j.brainres.2009.10.077.
Rinke, P., Schmidt, T., Beier, K., Kaul, R., Scharinger, M., 2022. Rapid pre-attentive processing of a famous speaker: electrophysiological effects of angela Merkel's voice. Neuropsychologia 173, 108312. https://doi.org/10.1016/j.neuropsychologia.2022.108312.
Robert, J., 2021. Pydub. In: Python Package Index. https://pypi.org/project/pydub/.
Rodero, E., Lucas, I., 2023. Synthetic versus human voices in audiobooks: the human emotional intimacy effect. New Media Soc. 25 (7), 1746–1764. https://doi.org/10.1177/14614448211024142.
Roswandowitz, C., Kathiresan, T., Pellegrino, E., Dellwo, V., Frühholz, S., 2024. Cortical-striatal brain network distinguishes deepfake from real speaker identity. Commun. Biol. 7 (1), 711. https://doi.org/10.1038/s42003-024-06372-6.
Schweinberger, S.R., Walther, C., Zäske, R., Kovács, G., 2011. Neural correlates of adaptation to voice identity. Br. J. Psychol. 102 (4), 748–764. https://doi.org/10.1111/j.2044-8295.2011.02048.x.
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.
Sestieri, C., Corbetta, M., Romani, G.L., Shulman, G.L., 2011. Episodic memory retrieval, parietal cortex, and the default mode network: functional and topographic analyses. J. Neurosci. 31 (12), 4407–4420. https://doi.org/10.1523/jneurosci.3335-10.2011.
Sidtis, D., Kreiman, J., 2012. In the beginning was the familiar voice: personally familiar voices in the evolutionary and contemporary biology of communication. Integr. Psychol. Behav. Sci. 46 (2), 146–159. https://doi.org/10.1007/s12124-011-9177-4.
Sidtis, D.V.L., Zäske, R., 2021. Who we are. In: The Handbook of Speech Perception, pp. 365–397. https://doi.org/10.1002/9781119184096.ch14.
Sun, Y., Ming, L., Sun, J., Guo, F., Li, Q., Hu, X., 2023. Brain mechanism of unfamiliar and familiar voice processing: an activation likelihood estimation meta-analysis. PeerJ 11, e14976. https://doi.org/10.7717/peerj.14976.
Treder, M.S., 2020. MVPA-light: a classification and regression toolbox for multi-dimensional data. Front. Neurosci. 14, 289. https://doi.org/10.3389/fnins.2020.00289.
Ulbrich, C., Mennen, I., 2016. When prosody kicks in: the intricate interplay between segments and prosody in perceptions of foreign accent. Int. J. BiLing. 20 (5), 522–549. https://doi.org/10.1177/1367006915572383.
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: In: Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24). Association for Computing Machinery, New York, NY, USA, pp. 2696–2710. https://doi.org/10.1145/3658644.3670325.
Xie, X., Myers, E., 2015. The impact of musical training and tone language experience on talker identification. J. Acoust. Soc. Am. 137 (1), 419–432. https://doi.org/10.1121/1.4904699.
Xu, Y., 2019. Prosody, tone, and intonation. In: The Routledge Handbook of Phonetics. Routledge, pp. 314–356. 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. Q. J. Exp. Psychol. 74 (7), 1185–1201. https://doi.org/10.1177/1747021821998557.

Yamagishi, J., Veaux, C., King, S., Renals, S., 2012. Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction. Acoust Sci. Technol. 33 (1), 1–5. https://doi.org/10.1250/ast.33.1.

Young, A.W., Frühholz, S., Schweinberger, S.R., 2020. Face and voice perception: understanding commonalities and differences. Trends Cognit. Sci. 24 (5), 398–410. https://doi.org/10.1016/j.tics.2020.02.001.

Zhang, W., Li, J., Ji, L., Cheng, X., Sun, D., Jiang, Y., Chen, F., Zhou, Y., Choi, C., Cheng, H., Cai, S., 2025. fNIRS experimental study on the impact of AI-synthesized familiar voices on brain neural responses. Sci. Rep. 15 (1), 16872. https://doi.org/10.1038/s41598-025-92702-5.

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.

Zäske, R., Volberg, G., Kovács, G., Schweinberger, S.R., 2014. Electrophysiological correlates of voice learning and recognition. J. Neurosci. 34 (33), 10821–10831. https://doi.org/10.1523/JNEUROSCI.0581-14.2014.