Voice-cloning artificial-intelligence speakers can also mimic human-specific vocal expression

Wenjun Chen[1,3] and Xiaoming Jiang*[1,2]

[1]Institute of Language Sciences, Shanghai International Studies University, Shanghai, 201620, China

[2]Key Laboratory of Language Science and Multilingual Artificial Intelligence, Shanghai International Studies University, Shanghai, 201620, China

[3]School of Communication Sciences and Disorders, McGill University, Montréal, H3G1A8, Canada

*Corresponding author address: Xiaoming Jiang, PhD, 1550 Wenxiang Road, Shanghai 201620, China. Email: xiaoming.jiang@shisu.edu.cn

Wenjun Chen: wenjun.chen@mail.mcgill.ca | https://orcid.org/0000-0003-3870-5041

Xiaoming Jiang: xiaoming.jiang@shisu.edu.cn | https://orcid.org/0000-0002-5171-9774

Statements and Declarations

Funding: This research was supported by the National Natural Science Foundation of China (Grant No. 32471109) awarded to X.J. W.C. is supported by a McGill-China Scholarship Council (CSC) Joint Scholarship.

Competing Interests: We have no competing interests.

Acknowledgments: We would like to thank Dr. Marc Pell for providing resources for data analysis. We also thank Dr. Maria Cutumisu for offering an insightful machine learning course (taken by W.C.), which informed our analytical approach.

Author Contributions: Wenjun Chen: Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft. Xiaoming Jiang: Conceptualization, Methodology, Supervision, Funding acquisition, Resources, Writing – review & editing.

Ethics approval: This study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was granted by the Ethics Committee of the Institute of Linguistics, Shanghai International Studies University (Approval No. 20230628027). Written informed consent was obtained from all participants prior to their participation in the study.

Consent to Participate and Publish:: Informed consent was obtained from all individual participants included in the study. Informed consent for publication of data derived from this study was obtained from all participants prior to their participation.

Data availability: All data and analysis code supporting the findings of this study are publicly available via the Open Science Framework (OSF) at https://osf.io/3c7rd/overview?view_only=a194fd9070ad4ca18b2eb7d5c4d303f4.

Abstract

As AI voice assistants become increasingly prevalent, psychological research on human-AI voice interaction requires stimuli that control for both speaker identity and prosodic characteristics. We tested whether voice cloning systems trained on prosodically homogeneous recordings can reproduce target prosody when generating novel utterances. Ten Mandarin speakers produced 30 sentences (15 Geography, 15 Trivia statements) in confident, doubtful, and neutral intonation. For each speaker and prosody, we trained separate AI voice clones on either the Geography or the Trivia recordings; these clones then generated all 30 sentences, yielding 2,700 utterances (900 Human, 900 AI-Geography, 900 AI-Trivia) for analysis. Machine learning classification using 88 eGeMAPS acoustic features achieved high within-source accuracy (mean 0.85) and substantial cross-source generalization (mean 0.65), with spectral flux emerging as a top contributor across all three sources. Linear mixed-effects models confirmed that confident prosody exhibited higher spectral flux, longer estimated vocal tract length, and lower fundamental frequency (F0) than doubtful prosody across human and AI speakers; notably, spectral flux showed no sex effect, suggesting it indexes prosodic differences rather than speaker-level anatomical variation. Time-resolved F0 decoding discriminated confident from doubtful prosody most reliably in late utterance windows across all sources, although human speakers showed prosodic markers more broadly distributed across the utterance than AI speakers. We demonstrate that voice cloning preserves not only speaker identity but also prosodic style, while AI-generated speech still forms a distinguishable acoustic population relative to human speech. Voice cloning thus provides a viable method for preparing more controlled human-AI stimuli for psychological research.

Keywords: Voice cloning, voice style transfer, affective computing, vocal confidence, machine learning

1. Introduction

According to the American Psychological Association's 2026 Trends Report[1], AI chatbots are now common in digital life, with many platforms (e.g., ChatGPT) offering voice-enabled modes that approximate human cadence and tone, raising concerns about synthetic relationships and their psychological implications (Cross & Kappas, 2025; McGettigan et al., 2025). However, rigorous psychological research on human-AI voice interaction requires careful control of potential confounds: When comparing human and AI speech, researchers should better match not only the content of utterances but also speaker identity (Lavan et al., 2025; Roswandowitz et al., 2024) and prosodic (Fan & Liu, 2025) characteristics. Against this background, we aim to provide a stimulus preparation approach for future psychological research on human-AI voice interaction and digital companion effects.

In human speech, speakers modulate prosody, namely the intonational, rhythmic, and voice quality properties (Cutler et al., 1997), to achieve communicative intentions and signal internal states, attitudes, and emotions (Hammerschmidt & Jürgens, 2007; Hirschberg, 2002; Trott et al., 2023). For instance, the word "yes" can convey certainty (with falling pitch) or doubt (with rising pitch) (Jiang & Pell, 2017). However, prosody in text-to-speech (TTS) systems lacks a biological foundation and is instead generated computationally. At the implementation level, two paradigms are commonly employed: (1) explicit control, where prosodic parameters (e.g., pitch, duration, energy) are directly manipulated through linguistic rules or user-defined specifications (Pamisetty & Sri Rama Murty, 2023; Raitio et al., 2020), and (2) learned representations, where models extract prosody embeddings from reference speech or training data (Klimkov et al., 2019; Skerry-Ryan et al., 2018). Namely, AI prosody can either be manually programmed by adjusting acoustic parameters according to predefined

[1] https://www.apa.org/monitor/2026/01-02/trends-digital-ai-relationships-emotional-connection

rules, or learned from data by training models to replicate prosodic patterns observed in human speech recordings.

The present study tests this latter paradigm in practice, applying it to voice style transfer across three prosodic conditions: neutral, doubtful, and confident. If successful, this approach could be extended into a pipeline for preparing comparable AI/human speech stimuli for other psychologically relevant prosodic contrasts, while controlling speaker identity as a confound.

1.1. Voice Cloning as a Potential Method for Preparing Human-AI Prosody-Comparable Stimuli

Building on the above technical considerations, we now examine how voice cloning may serve as a method for stimulus preparation. Early human-AI voice comparison research relied on off-the-shelf synthetic voices or voice-transformation-manipulated human voices, comparing them with unrelated human speakers (Mullennix et al., 2003; Rodero, 2017). While this approach revealed perceptual differences between human and synthetic voices, it confounded prosodic differences with speaker identity differences, since synthetic and human voices did not share the same vocal identity. A more recent trend has been to control for speaker identity by using voice cloning technology (Lavan et al., 2025; Roswandowitz et al., 2024), which enables both human and AI speech to share the same vocal identity representation (operationalized as the anatomical-acoustic vocal signature), thereby allowing researchers to better attribute observed behavioral and neural responses to the human vs. AI source rather than to speaker identity confounds. For instance, such confounds can include perceived body size, as signaled by vocal tract length (VTL), which influences judgments of dominance (Pisanski et al., 2022; Pisanski & Reby, 2021). If human and AI voices do not share the same speaker identity, listeners' responses to confounding factors may be

mistakenly attributed to the human-AI distinction rather than to speaker-specific characteristics.

Given that prosody in TTS systems can be generated through learned representations (Klimkov et al., 2019; Skerry-Ryan et al., 2018), we speculate that while voice cloning is primarily designed to replicate speaker identity, the underlying technical implementation may also capture prosodic characteristics, thereby inadvertently cloning prosody as a byproduct. For instance, Apple's Personal Voice enables users to record 150 audio samples to create a synthetic voice that replicates their vocal identity for TTS-based communication in applications such as FaceTime[2]. A reason why voice cloning can replicate speaker identity from audio recordings lies in mel-spectrograms, which encode speaker-specific timbral characteristics (formant frequencies, spectral envelope) that enable voice identification (Jia et al., 2018). At the same time, however, mel-spectrograms also capture temporal-frequency dynamics (pitch contours, energy, rhythm) critical for prosody recognition (Meng et al., 2019; Ong et al., 2023).

Given this dual encoding capacity, we hypothesize that voice cloning models trained on recordings of a specific prosodic style may capture and reproduce that prosodic pattern alongside speaker identity, even though the technology is not explicitly designed for this purpose.

1.2. Biological Foundations of Prosody: Modulation of Speech Production

Human speech production operates through neuro-muscular control of vocal organs; for example, real-time magnetic resonance imaging (MRI) studies have directly observed speakers modulating vocal tract length during emotional expression, with happy speech

[2] https://machinelearning.apple.com/research/personal-voice

displaying shorter VTL than angry or sad speech (Kim et al., 2020). Considering that both VTL and fundamental frequency (F0) can be directly estimated through acoustic analysis (Anikin et al., 2024) and exhibit an inverse physiological relationship (Neuhaus et al., 2024) , we speculate that pragmatically marked prosody, which extends beyond basic emotional expression, may also involve systematic VTL modulation. For example, in Canadian English, research on vocal confidence has revealed that doubtful expressions display elevated mean F0 compared to confident expressions (Jiang & Pell, 2017).

The present study investigates this in Mandarin Chinese by testing whether (1) doubtful speech exhibits elevated mean F0 as observed in English (Jiang & Pell, 2017), and (2) acoustically estimated VTL is longer for confident vs. doubtful prosody. If our results confirm that confident speech acoustically exhibits longer estimated VTL than doubtful speech, this would link our findings to literature on acoustic size exaggeration (Belyk et al., 2022; Pisanski & Reby, 2021) and origins of vocalic complexity (Pisanski et al., 2022). Specifically, this would support the "fast-and-frugal" principle (Pisanski et al., 2022), which posits that speakers can achieve significant communicative effects through minimal and metabolically efficient articulatory adjustments; in our case, speakers modulate VTL (lengthening or shortening) to express their feeling of knowing through pragmatically marked prosody (Swerts & Krahmer, 2005).

A more intriguing question is whether AI-generated speech also exhibits the acoustic effects that, in humans, arise from vocal apparatus modulation. After all, parameters such as F0 and VTL are typically estimated from the audio output of biological speech production. Although AI voice cloning lacks the underlying biological substrate, we apply the same acoustic estimation procedures to AI-generated speech as we do to human speech. Our rationale is straightforward: if modulation patterns observed in human speech production also

emerge in AI-cloned speech, this would suggest that AI voice cloning replicates human-specific vocal expression at the acoustic level.

1.3. AI-Generated Speech as a "Dialect": Extending Dialect Theory to Synthetic Speech

Vocal expression contains both universal regularities and culturally-specific features (Scherer, 1986). The dialect theory of emotional expression (Elfenbein, 2013) holds that speakers from shared linguistic backgrounds form systematic in-group acoustic patterns, yielding an in-group advantage in classification: machine learning classifiers trained and tested within the same speaker group outperform those trained and tested across groups, although cross-group performance still exceeds chance (Laukka et al., 2014).

Beyond basic emotion, this dialect framework has been extended to pragmatically marked prosody. Using machine learning classifiers to decode confident vs. doubtful expressions across native Canadian English, Quebecois-French-accented English, and Australian-accented English, classification accuracy was higher when training and testing within the same accent than across accents, while cross-accent classification still exceeded chance (Jiang & Pell, 2018). This established that vocal confidence, like basic emotion, exhibits dialect-like in-group acoustic patterns.

The present study extends this paradigm in a new direction. Rather than treating different human accents as the in-group/out-group contrast, we treat human and AI-generated speech as the contrasting populations. Treating AI voices as a distinct social group has precedent in the social-psychological literature: Edwards et al. (2019) applied Social Identity Theory and the Computers-Are-Social-Actors paradigm to listener perception of AI voices, demonstrating that listeners categorize AI-generated speech using the same in-group/out-group frameworks they apply to human speech. Existing research on synthetic speech detection further shows that AI-generated speech systematically differs from human speech

in detectable acoustic properties, with characteristic artifacts arising from generative models (Malik, 2019; Sahidullah et al., 2015), suggesting that AI speech may form its own population-level acoustic dialect.

Accordingly, we hypothesize that machine learning classifiers will successfully decode confident vs. doubtful prosody across human and AI speech, demonstrating that core prosodic patterns transfer between these populations; however, cross-source classification accuracy will be lower than within-source accuracy, reflecting AI-internal acoustic regularities that distinguish synthetic from human speech.

1.4. The Current Study

Building on the framework outlined above, the present study tests whether voice cloning systems trained on prosodically homogeneous recordings reproduce target prosody when generating novel utterances, and whether AI-generated speech as a population exhibits distinctive acoustic patterns relative to human speech. We employed three complementary analyses on a parallel human–AI dataset.

First, we used machine learning classification on 88 eGeMAPS acoustic features (Eyben et al., 2015) to test whether prosodies could be distinguished within and across sources (human, AI-Geography, AI-Trivia), and to identify the most influential features (Jiang & Pell, 2018). Second, we examined these features using linear mixed-effects models, comparing prosody on both anatomical features (F0, VTL) and anatomy-independent dynamic features (spectral flux). Third, we used time-resolved decoding of normalized F0 contours to test whether prosodic markers of confidence are distributed throughout the utterance or concentrated in specific temporal windows (Goupil & Aucouturier, 2021). The neutral condition served as a baseline where applicable. We operationalize this investigation through three research questions:

- RQ1: Can machine learning classifiers using 88 eGeMAPS features distinguish prosodies within and across human and AI speech, and do the most influential features reflect anatomical properties (e.g., F0) or anatomy-independent prosodic dynamics?
- RQ2: How do confident and doubtful prosodies differ at the level of individual acoustic features?
- RQ3: Do F0-based prosodic markers distinguishing confident from doubtful prosody emerge across the entire utterance, or are they concentrated near specific temporal regions (e.g., utterance endings)?

Together, these three lines of evidence assess whether voice cloning, while designed to replicate speaker identity, can simultaneously perform prosodic style transfer.

2. Materials and Methods

2.1. Stimulus Preparation

2.1.1. Participants

Five males (Age = 22.8 ± 3.0 years; Years of education = 19.4 ± 3.0 years; Height = 182.2 ± 5.8 cm) and five females (Age = 22.0 ± 1.7 years; Years of education = 19.0 ± 1.2 years; Height = 167.4 ± 4.3 cm) standard Mandarin speakers from *** University were recruited with financial compensation (55 RMB per hour). All participants had considerable experience in acting performance, speech, or music training and demonstrated high proficiency in Mandarin Chinese, as evidenced by the Putonghua Proficiency Test (scores: 87–91 out of 100). None reported any history of speech-hearing impairment or neurological or psychiatric disorders. All participants provided written informed consent prior to participation.

2.1.2. Recording

Recordings were conducted in a sound-attenuated laboratory. An Audio-Technica AT2035 cardioid condenser microphone was connected to a Dell G3-3579 laptop via a Komplete Audio 6 Mk2 sound card, with *Praat* 6.2.09 (Boersma & Weenink, 2021) serving as the recording software.

The human recording followed an established procedure (Jiang & Pell, 2017). Participants completed three blocks in fixed order: neutral, doubtful, and confident. Within each block, 30 sentences (15 Trivia, 15 Geography; see **Supplementary Table 1**) were presented in randomized order. For the neutral block, participants read target statements directly without prompts. For doubtful and confident blocks, each trial began with a background prompt (e.g., "*You are playing a knowledge testing game, and you are asked, Frogs only nod their heads and do not shake them, aren't they?*"), followed by a probability phrase (e.g., "I am certain" or "I'm not sure") and the target statement. After each production, participants rated their subjective confidence on a 7-point Likert scale (1 = not confident at all, 7 = very confident). Each sentence was produced twice, and the better production was selected based on the speaker's self-ratings and the acoustic impression judged by the first author. The recording session lasted approximately 2 hours, including self-paced breaks.

2.1.3. AI Voice Cloning

To generate AI speakers capable of producing different prosodies, a Huawei Nova 9 smartphone was connected to the laptop via a portable (Changba Live No. 1) sound card converter, allowing human recordings to be input directly into Huawei's *Xiaoyi* (Celia) voice cloning service (Version 11.0.44.306) at 30% system volume. For each speaker at each prosody level, two AI models were created: one trained on the 15 Geography sentences (AI-Geography) and one on the 15 Trivia sentences (AI-Trivia), resulting in 60 AI models total

(10 speakers × 3 prosodies × 2 statement types). Each AI model then generated all 30 sentences (both Geography and Trivia statements) while screen recording captured the output. The 60 videos were converted to WAV audio, and individual utterances were extracted. See **Figure 1**.

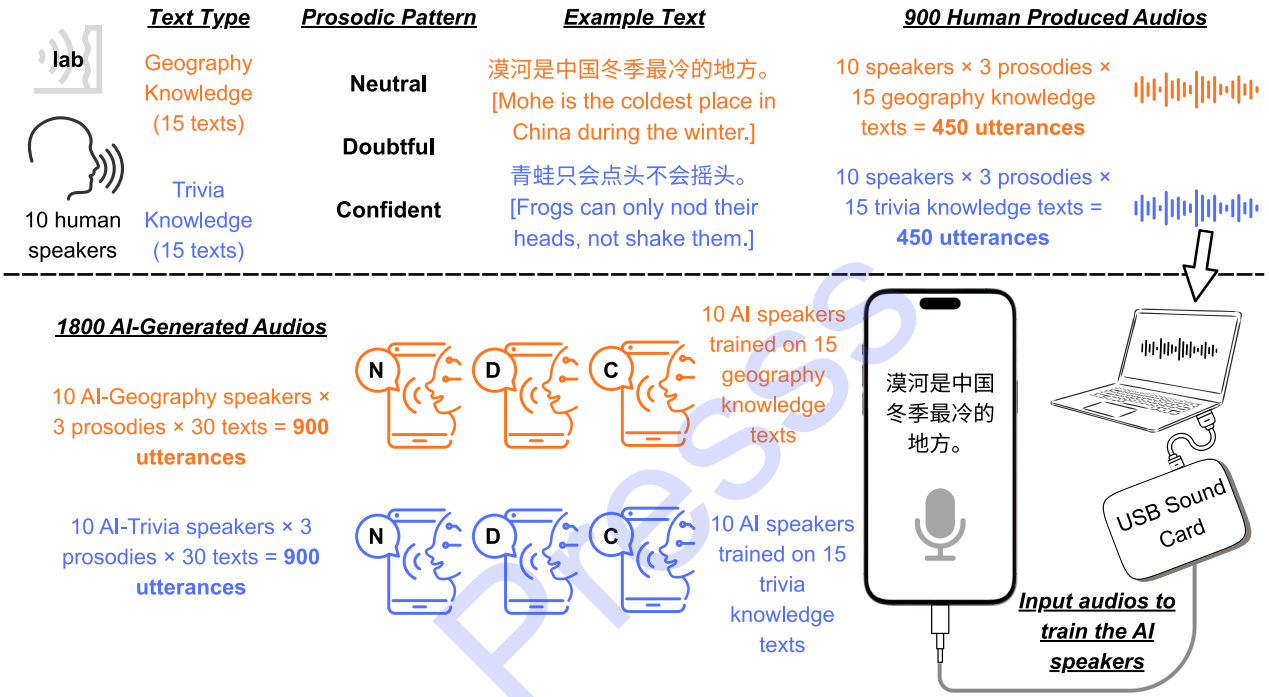


Figure 1. ***Workflow Of The 2,700-Audio Dataset Preparation.*** Ten human speakers produced 900 utterances across three prosodic conditions (neutral, doubtful, confident), using 30 unique texts: 15 Geography knowledge statements and 15 Trivia knowledge statements. To create AI voice cloning models, each speaker's recordings were used as training inputs. For example, Speaker 1's 15 Geography sentences produced in neutral prosody were fed into a smartphone (via a computer and sound card) to train an AI-Geography speaker with neutral prosody. This process was repeated for doubtful and confident prosodies, yielding three AI-Geography speakers for Speaker 1. Similarly, Speaker 1's 15 Trivia sentences across three prosodies were used to create three AI-Trivia speakers. This procedure was applied to all 10 speakers, generating 60 AI models total (10 speakers × 3 prosodies × 2 statement types). Each AI model then generated all 30 texts, producing 900 AI-Geography utterances and 900 AI-Trivia utterances. The final dataset comprises 2,700 audio files (900 Human + 900 AI-Geography + 900 AI-Trivia) for subsequent acoustic analysis. N = Neutral, D = Doubtful, C = Confident.

2.1.4. Audio Preprocessing

All audio files, including both Human recordings and AI-generated speech, were normalized together using Root Mean Square (RMS) normalization to a target level of -30

dBFS. Mel spectrograms were generated using *Librosa* 0.11.0 (McFee et al., 2015) with the following parameters: `sampling rate = 22,050 Hz, FFT window size = 2,048, hop length = 512, and 128 mel-frequency bins`. These spectrograms provide a visual example of acoustic patterns across confidence levels and audio sources (see **Figure S1**).

2.2. Stimulus Validation

To verify that speakers produced the intended prosodic manipulations and to test whether AI voice clones preserved these prosodic patterns, we conducted perceptual validation followed by acoustic analyses linking acoustic features to confidence ratings.

Perceptual Validation Procedure. Approximately one month after the initial recording session, participants returned to the laboratory to rate only their own human recordings. Sentences were presented using OpenSesame 3.3.14 (Mathôt et al., 2012), and participants rated how confident each audio sample sounded on a 7-point Likert scale (1 = not at all confident, 7 = very confident). Stimuli were played through Hewlett-Packard (HP) GH10 headphones at a comfortable volume level and presented randomly across three blocks with self-paced rest periods between blocks. The validation task took approximately 10 minutes to complete.

Acoustic-Rating Correlation Analysis. To extend the validation to AI voice clones, we tested whether acoustic features (mean F0 and estimated VTL) correlated with confidence ratings across all three sources. For Original Human Speakers, we used the self-ratings obtained from perceptual validation. For AI voice clones, each AI-generated utterance was assigned the rating from its corresponding source human recording, allowing us to test whether AI clones preserved the acoustic-rating relationships present in their source speakers. Both F0 and VTL were standardized within sex using z-scores to remove sex-related differences. Pearson and Spearman correlations quantified the strength of acoustic-rating

relationships within each source. Ratios of AI-to-Human correlation coefficients indexed the degree of prosody preservation in voice cloning.

Results: Human Self-Ratings Confirmed Successful Prosodic Manipulation. Self-rated confidence scores showed clear differentiation among conditions: confident (M = 6.57, SD = 0.75), neutral (M = 4.08, SD = 0.44), and doubtful (M = 1.39, SD = 0.56). Paired t-tests confirmed that confident ratings were significantly higher than neutral, $t(9) = 18.13$, $p < .001$, $d = 5.73$, and neutral ratings were significantly higher than doubtful, $t(9) = 23.02$, $p < .001$, $d = 7.28$. These results confirm that speakers successfully produced vocal expressions corresponding to the three intended confidence levels.

Results: AI Voice Clones Preserved Acoustic-Rating Relationships. F0 and VTL showed expected directional relationships with confidence ratings across all three sources (**Figure S2**). For F0, all sources showed negative correlations (Original Human: r = -0.16, ρ = -0.11; AI-Geography: r = -0.28, ρ = -0.21; AI-Trivia: r = -0.32, ρ = -0.28; all $p < .001$), indicating that higher F0 was associated with lower confidence. For VTL, all sources showed positive correlations (Original Human: r = 0.18, ρ = 0.21; AI-Geography: r = 0.27, ρ = 0.24; AI-Trivia: r = 0.23, ρ = 0.21; all $p < .001$), indicating that longer estimated VTL was associated with higher confidence. Notably, AI voice clones showed amplified correlation strengths compared to human speakers: for F0, AI-Geography showed 1.70× and AI-Trivia showed 1.97× the human correlation; for VTL, AI-Geography showed 1.44× and AI-Trivia showed 1.25× the human correlation. These results indicate that AI voice clones not only preserved but amplified the acoustic-rating relationships present in their source recordings.

2.3. Data Analysis

2.3.1. eGeMAPS-Based Prosody Classification

We employed XGBoost classifiers to evaluate prosodic pattern learning across three data sources: Original Human Speakers, AI Speakers Trained on Geography Texts (AI-

Geography), and AI Speakers Trained on Trivia Texts (AI-Trivia). All models used 88 eGeMAPS acoustic features (Eyben et al., 2015) extracted from audio recordings using openSMILE.

Data were partitioned into training, validation, and test sets using a 60/20/20 speaker-wise split (random seed = 42), ensuring that no speaker appeared in multiple partitions to prevent data leakage. Hyperparameters were optimized via grid search with 5-fold stratified cross-validation on the Human training data, evaluating 3,072 parameter combinations. Optimal parameters (**Table 1A**) were applied identically to all subsequent models to ensure fair comparison.

Table 1. Hyperparameter Grid Search Results

A. XGBoost Classification (eGeMAPS Features)

Hyperparameter	Search Range	Optimal Value	Best CV F1
learning_rate	[0.01, 0.05, 0.1, 0.2]	0.2	0.8641
max_depth	[3, 5, 7, 10]	3	
n_estimators	[50, 100, 200, 300]	100	
colsample_bytree	[0.3, 0.5, 0.7, 1.0]	0.7	
alpha	[0, 5, 10, 20]	0	
min_child_weight	[1, 3, 5]	3	

B. Random Forest Temporal Decoding (F0 Contours)

Hyperparameter	Search Range	Optimal Value	Best CV Accuracy
n_estimators	[50, 100, 200]	200	0.7833
max_depth	[3, 5, 7, 10]	7	
min_samples_split	[2, 5, 10]	5	
min_samples_leaf	[1, 2, 4]	2	

Note. Hyperparameters were optimized using grid search with 5-fold stratified cross-validation on Human training data for both analyses. Best CV F1 = Best cross-validation F1 score (macro-averaged); Best CV Acc = Best cross-validation accuracy. Performance metrics shown in the first row of each section represent model performance using the optimal parameter combination.

Nine train/test combinations were constructed: three within-domain models (trained and tested on the same source) and six cross-domain models (trained on one source, tested on another). Performance was evaluated using overall accuracy, macro-averaged F1 score,

precision, recall, and RMSE. For three-class classification (confident, doubtful, neutral), ROC curves and AUC values were computed using the one-vs-rest approach. Feature importance values were extracted from within-domain models using XGBoost's built-in `feature_importances_` attribute. Spearman rank-order correlations were computed to compare feature utilization strategies across models. To compare within-domain vs. cross-domain performance, we used Welch's t-test, Mann-Whitney U test, and permutation test (10,000 permutations), with effect sizes quantified using Hedges' g.

2.3.2. Within-Source Analysis of Spectral Flux, F0, and VTL

For each acoustic measure, we fitted linear mixed-effects models separately for each source. Spectral Flux was extracted using openSMILE based on the eGeMAPS feature set, while estimated F0 and VTL were extracted from vowel-only voice streams using *Praat* with *Vocal Toolkit*. Models were fitted using the formula: `Measure ~ Confidence_Level + Sex + (1|Speaker) + (1|Item)`. Main effects were assessed via Type III ANOVA with Satterthwaite's degrees of freedom. Estimated marginal means and pairwise comparisons (Neutral vs. Confident, Neutral vs. Doubtful, Confident vs. Doubtful) were computed using the emmeans package. Effect sizes were quantified using Cohen's d for paired samples.

2.3.3. Time-Resolved F0 Decoding

To examine when prosodic information distinguishing confident from doubtful prosody emerges over time, we extracted F0 contours from entire audio recordings using the YIN algorithm (`fmin = 75 Hz, fmax = 500 Hz`) (McFee et al., 2015). Each F0 contour was interpolated to 100 time points, z-score normalized, and smoothed using a Savitzky-Golay filter (`window length = 15, polynomial order = 3`). Neutral prosody was excluded from this analysis as it may overlap with confident productions (Jiang & Pell, 2017).

We used a sliding window approach with 10% window width and 5% step size, yielding 19 time windows spanning 0-100% of normalized utterance duration. Random Forest classifiers were used for classification, with hyperparameters optimized via grid search on the Human training data using the full F0 contour (**Table 1B**). The same parameters were applied to all sources and time windows.

Data were split using a 60/20/20 speaker-wise split (random seed = 42). For each window, classifiers were trained on combined train + validation data and evaluated on the held-out test set. Standard errors were computed via bootstrap resampling (1,000 iterations). Statistical significance against chance level (0.50) was assessed using bootstrap-based t-tests, with false discovery rate (FDR) correction across windows using the Benjamini-Hochberg method ($\alpha = 0.05$).

2.3.4. Software and Reproducibility

Python analyses (eGeMAPS classification, F0 temporal decoding) were conducted using scikit-learn 1.6.1, XGBoost 3.1.2 (Chen, 2016), librosa 0.11.0 (McFee et al., 2015), scipy 1.15.3, and statsmodels 0.14.5 for FDR correction. Acoustic feature extraction used Praat Vocal Toolkit [3] for F0 and VTL estimation, and openSMILE (version 2.6.0) (Eyben et al., 2010) for eGeMAPS features (Eyben et al., 2010). R analyses (linear mixed-effects models) were conducted in R 4.3.3, using lmerTest 3.1.3 (Kuznetsova et al., 2017) and emmeans 1.11.0 (Lenth et al., 2021). All random seeds were set to 42 for reproducibility. Statistical significance was set at $\alpha = .05$.

[3] https://www.praatvocaltoolkit.com

3. Results

3.1. eGeMAPS-Features-Based Classification Demonstrates Comparable Prosodic Patterns Across Human and AI Speakers

Within-source classification far exceeded chance across all three sources, indicating that AI-cloned voices contain rich acoustic information for distinguishing prosodic confidence. When models were trained and tested on the same source, classification was highly accurate across all three sources: Human/Human achieved 88% accuracy, AI-Geography/AI-Geography 87%, and AI-Trivia/AI-Trivia 81% (**Table 2**; **Figure 2A** to **2C**). The mean within-source accuracy of 0.85 (*SD* = 0.04) far exceeded the 3-class chance level (0.33).

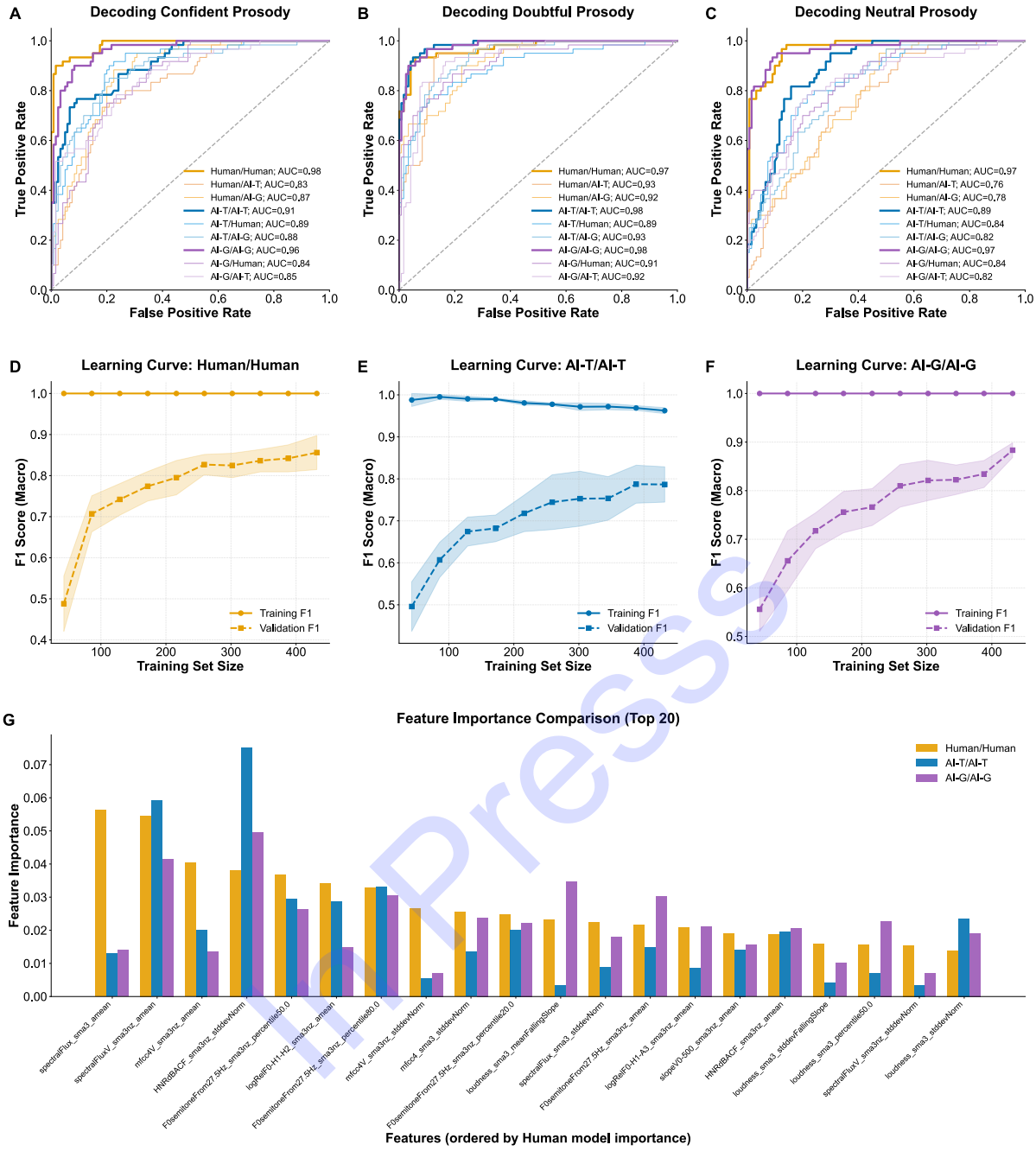
Cross-source models retained meaningful generalization despite a significant performance drop, suggesting that core prosodic patterns transfer between human and AI-cloned speakers. When models trained on one source were applied to test data from a different source, accuracy dropped but remained well above chance. Cross-source accuracies ranged from 0.56 (Human/AI-T) to 0.71 (AI-G/AI-T), with a mean of 0.65 (*SD* = 0.05). The performance gap between within and cross-source models was substantial and statistically significant (gap = 0.21; Welch's $t(7) = 6.63$, $p = .001$, Hedges' $g = 3.94$; permutation test $p = .011$). All six cross-source models exceeded chance level.

Doubtful prosody was the most discriminable category across all conditions, with AI-Trivia showing particularly distinctive doubtful patterns and asymmetric separability among prosodic categories. Per-class AUC values revealed that doubtful prosody achieved the highest discriminability across all conditions, ranging from 0.89 to 0.98 (**Figure 2B**). This pattern was especially striking for AI-Trivia. Confusion matrices (**Figure S3**) showed that AI-Trivia's doubtful prosody was correctly identified by 91.7% to 100% of cross-source

models. In contrast, AI-Trivia's confident prosody was sometimes misclassified as doubtful when tested by Human-trained models (47% misclassification rate).

Learning curves confirmed healthy model behavior without overfitting, demonstrating progressive generalization with increased training data. For the Human/Human model, validation F1 scores improved steadily as training data increased, rising from 0.56 (smallest training set) to 0.88 (full training set; **Table S2**). The gap between training and validation performance narrowed from 0.44 to 0.12 (**Figure 2D** to **2F**). AI-Geography showed a similar trajectory. AI-Trivia exhibited a slightly larger train-validation gap (0.17), but learning remained stable.

AI-Trivia showed significant convergence with human feature weighting strategies, while AI-Geography achieved high accuracy through alternative feature combinations. All three within-source models relied on similar acoustic dimensions for classification, including F0 percentiles, spectral flux, and MFCC4 (**Table 3**; **Figure 2G**). Spearman rank-order correlations of feature importance values showed that AI-Trivia exhibited moderate but significant convergence with the Human model ($\rho = 0.51$, $p = .023$), whereas AI-Geography did not show significant convergence with the Human model ($\rho = 0.26$, $p = .268$). This pattern indicates that AI-Trivia partially replicated human-like feature weighting strategies, while AI-Geography achieved comparable classification through a different combination of acoustic features.



406

407 **Figure 2.** ***XGBoost Classification Performance, Learning Curves, and Feature***
408 ***Importance.*** **(A-C)** ROC curves for three-class prosodic confidence classification using
409 eGeMAPS features. Panel **A**: Confident prosody (one-vs-rest); Panel **B**: Doubtful prosody;
410 Panel **C**: Neutral prosody. Thick lines indicate within-domain models (same training and test
411 source), thin lines indicate cross-domain models. Colors represent training sources: orange
412 (Human), blue (AI-Trivia), purple (AI-Geography). **(D-F)** Learning curves showing training
413 vs. validation F1 scores across training set sizes. Panel **D**: Human/Human; Panel **E**: AI-
414 Trivia/AI-Trivia; Panel **F**: AI-Geography/AI-Geography. Shaded regions indicate ±1 SD
415 across 5-fold cross-validation. **(G)** Feature importance comparison for top 20 acoustic
416 features, ordered by Human model importance. Orange bars: Human/Human; Blue bars: AI-
417 Trivia/AI-Trivia; Purple bars: AI-Geography/AI-Geography. All models used identical

418 hyperparameters optimized on Human training data. Classification employed 60/20/20
419 speaker-wise splits (N=2,700 utterances).

420

Table 2. XGBoost Classification Performance Across Nine Train/Test Combinations

Combination	Acc_Conf ↑	AUC_Conf ↑	Acc_Doub ↑	AUC_Doub ↑	Acc_Neut ↑	AUC_Neut ↑	Test_Acc ↑	F1_Macro ↑	Precision ↑	Recall ↑	RMSE ↓	Time(s) ↓
Human/Human	0.93	0.98	0.92	0.97	0.8	0.97	0.88	0.88	0.89	0.88	0.52	0.26
Human/AI-T	0.32	0.83	1	0.93	0.37	0.76	0.56	0.53	0.63	0.56	0.87	0.24
Human/AI-G	0.58	0.87	0.92	0.92	0.38	0.78	0.63	0.61	0.64	0.63	0.88	0.32
AI-T/AI-T	0.75	0.91	0.9	0.98	0.77	0.89	0.81	0.81	0.81	0.81	0.77	0.25
AI-T/Human	0.95	0.89	0.38	0.89	0.6	0.84	0.64	0.63	0.72	0.64	0.88	0.31
AI-T/AI-G	0.7	0.88	0.55	0.93	0.75	0.82	0.67	0.67	0.7	0.67	0.89	0.31
AI-G/AI-G	0.88	0.96	0.88	0.98	0.85	0.97	0.87	0.87	0.87	0.87	0.54	0.42
AI-G/Human	0.83	0.84	0.58	0.91	0.58	0.84	0.67	0.67	0.72	0.67	0.93	0.31
AI-G/AI-T	0.58	0.85	0.92	0.92	0.62	0.82	0.71	0.7	0.71	0.71	0.86	0.29

Note. Arrows indicate desirable direction (↑ higher, ↓ lower). Combination format: Train Source/Test Source. All accuracy metrics (Acc_Conf, Acc_Doub, Acc_Neut, Test_Acc, F1_Macro, Precision, Recall) reflect three-class classification (Confident/Doubtful/Neutral), where chance level = 0.33. AUC values (AUC_Conf, AUC_Doub, AUC_Neut) reflect one-vs-rest binary discrimination (e.g., Confident vs. all other classes), where chance level = 0.50. All models used identical hyperparameters optimized on Human training data (Table 1). Within-domain models (same train/test source) achieved mean test accuracy 0.85±0.04, significantly outperforming cross-domain models (0.65±0.05; gap = 0.21; Welch's $t(7) = 6.63$, $p = .001$, Hedges' $g = 3.94$; permutation test $p = .011$). Doubtful prosody showed highest discriminability across all conditions (AUC: 0.89-0.98). Classification employed 60/20/20 speaker-wise train/validation/test splits on 88 eGeMAPS acoustic features (N = 2,700 utterances). AI-T = AI-Trivia; AI-G = AI-Geography.

Table 3. Feature Importance Values for Top 20 Acoustic Features Across Within-Domain Models

Feature	Human/Human	AI-T/AI-T	AI-G/AI-G
spectralFlux_sma3_amean	0.06	0.01	0.01
spectralFluxV_sma3nz_amean	0.05	0.06	0.04
mfcc4V_sma3nz_amean	0.04	0.02	0.01
HNRdBACF_sma3nz_stddevNorm	0.04	0.08	0.05
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	0.04	0.03	0.03
logRelF0-H1-H2_sma3nz_amean	0.03	0.03	0.01
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	0.03	0.03	0.03
mfcc4V_sma3nz_stddevNorm	0.03	0.01	0.01
mfcc4_sma3_stddevNorm	0.03	0.01	0.02
F0semitoneFrom27.5Hz_sma3nz_percentile20.0	0.02	0.02	0.02
loudness_sma3_meanFallingSlope	0.02	0.00	0.03
spectralFlux_sma3_stddevNorm	0.02	0.01	0.02
F0semitoneFrom27.5Hz_sma3nz_amean	0.02	0.01	0.03
logRelF0-H1-A3_sma3nz_amean	0.02	0.01	0.02
slopeV0-500_sma3nz_amean	0.02	0.01	0.02
HNRdBACF_sma3nz_amean	0.02	0.02	0.02
loudness_sma3_stddevFallingSlope	0.02	0.00	0.01
loudness_sma3_percentile50.0	0.02	0.01	0.02
spectralFluxV_sma3nz_stddevNorm	0.02	0.00	0.01
loudness_sma3_stddevNorm	0.01	0.02	0.02

Note. Features ordered by Human/Human model importance (descending). Importance values represent each feature's average contribution to classification accuracy across XGBoost decision trees (range: 0-1). All three within-domain models used identical hyperparameters optimized on Human training data. Spearman correlations: Human-AI-T $\rho = 0.507$ ($p = .023$); Human-AI-G $\rho = 0.260$ ($p = .268$). The significant correlation between Human and AI-Trivia models indicates moderate convergence in feature utilization strategies, whereas AI-Geography models employed divergent feature weighting approaches. AI-T = AI-Trivia; AI-G = AI-Geography.

3.2. Estimated Mean Acoustic Values Are Largely Consistent Across Three Sources

Spectral Flux distinguished all three confidence levels across every source, with confident producing the highest values and doubtful the lowest. All three sources showed strong main effects of confidence level (Human: $F(2, 859) = 515.02$, $p < .001$; AI-Geography: $F(2, 859) = 624.33$, $p < .001$; AI-Trivia: $F(2, 859) = 387.39$, $p < .001$), with all pairwise comparisons reaching significance (all $p < .001$) and large effect sizes for confident vs. doubtful (Human $d = 1.74$; AI-Geography $d = 1.54$; AI-Trivia $d = 1.32$; **Figure 2A** to **2C**). Notably, Spectral Flux showed no main effect of sex in any source (all $p > .349$), indicating that it reflects sex-independent prosodic dynamics.

Estimated F0 showed identical doubtful > confident > neutral hierarchies across all three sources. Estimated marginal means revealed consistent patterns (Human: 166, 196, 209 Hz; AI-Geography: 167, 191, 209 Hz; AI-Trivia: 166, 184, 208 Hz; **Figure 2D** to **2F**). All main effects of confidence level were highly significant (Human: $F(2, 859) = 530.09$, $p < .001$; AI-Geography: $F(2, 859) = 952.73$, $p < .001$; AI-Trivia: $F(2, 859) = 648.73$, $p < .001$), and all pairwise comparisons reached significance with comparable effect sizes for neutral vs. doubtful (Human $d = -0.69$; AI-Geography $d = -0.73$; AI-Trivia $d = -0.70$).

Estimated VTL was largely consistent across sources, though AI-Trivia did not differentiate confident from neutral. All three sources showed significant main effects of confidence level on estimated VTL (Human: $F(2, 859) = 32.88$, $p < .001$; AI-Geography: $F(2, 859) = 119.07$, $p < .001$; AI-Trivia: $F(2, 859) = 93.86$, $p < .001$). Human and AI-Geography showed the expected confident > neutral > doubtful pattern with all pairwise contrasts reaching significance (Human Neutral vs. Confident: $p = .020$, $d = -0.08$; AI-Geography Neutral vs. Confident: $p < .001$, $d = -0.24$). AI-Trivia preserved the doubtful contrast but showed no difference between confident and neutral estimated VTL ($\beta = -0.02$, $t(859) = -0.37$, $p = .926$, $d = -0.01$; **Figure 2G** to **2I**).

Sex effects on F0 and VTL reflected expected biological differences across all sources. Female speakers produced higher F0 (all $p < .001$) and shorter VTL than male speakers (Human: $F(1, 8) = 41.79$, $p < .001$; AI-Geography: $F(1, 8) = 15.03$, $p = .005$; AI-Trivia: $F(1, 8) = 26.16$, $p < .001$). The preservation of these biological sex effects in AI-cloned voices indicates that voice cloning systems captured anatomical characteristics in addition to prosodic patterns.

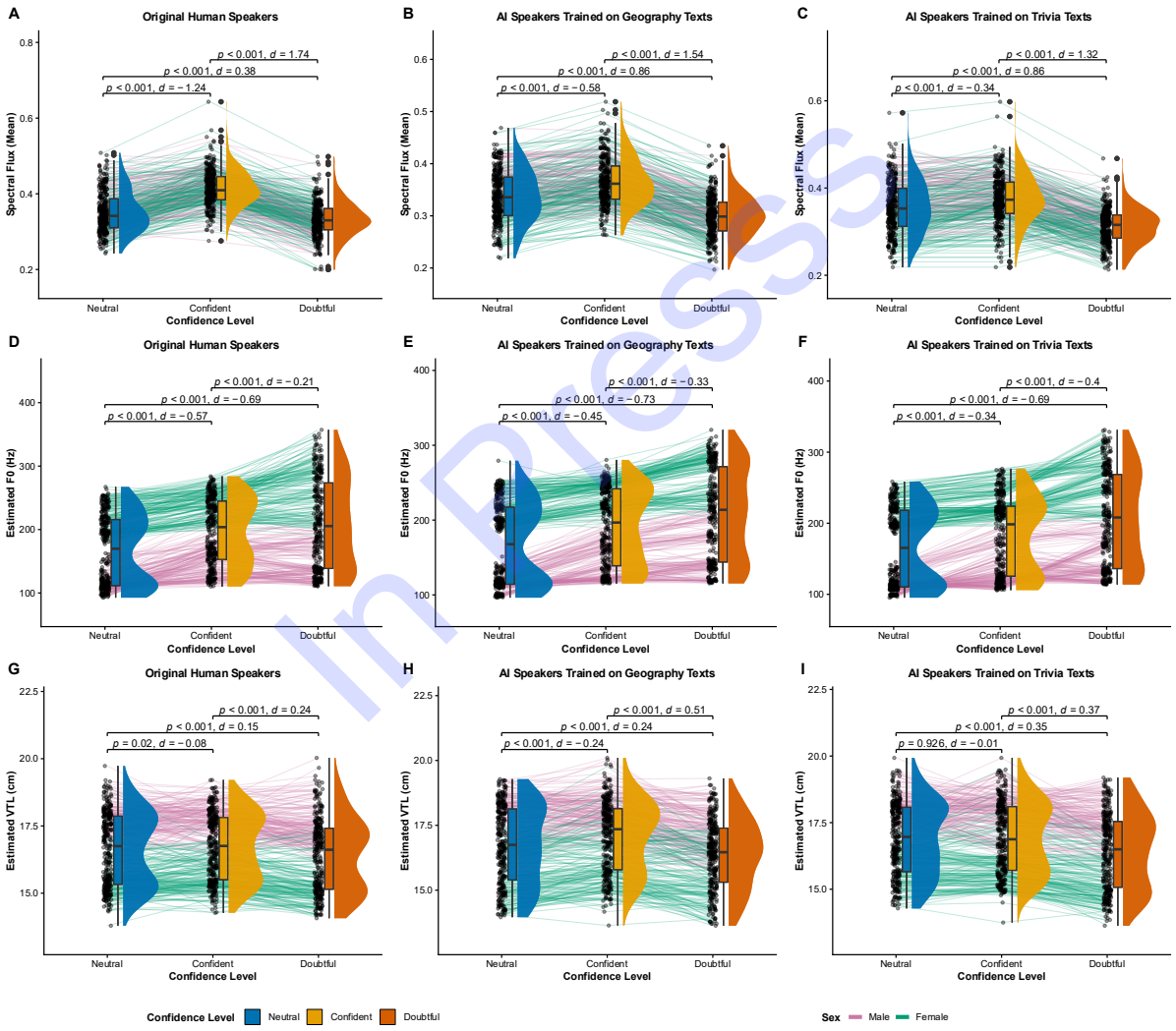


Figure 3. ***Spectral Flux, F0, and VTL Across Prosodic Confidence Levels.*** Within-source comparisons of three acoustic measures across confidence levels. Columns: Human (left), AI-Geography (center), AI-Trivia (right). Rows: **(A-C)** Spectral Flux (mean); **(D-F)** Estimated F0 (Hz); **(G-I)** Estimated VTL (cm). Raincloud plots combine violin distributions, boxplots, and individual data points. Connecting lines represent repeated measures within triads (same speaker, same item, same statement source), colored by sex (pink = male, green = female).

Brackets indicate pairwise comparisons from linear mixed-effects models with *p*-values and Cohen's *d* effect sizes annotated. N = Neutral, C = Confident, D = Doubtful.

3.3. Time-Resolved F0 Decoding Reveals Late-Utterance Markers of Confident vs. Doubtful Prosody Across Three Sources

All three sources peaked in late utterance windows, indicating that confidence-related F0 markers emerge most clearly near utterance endings. Human and AI-Trivia both peaked in the 85-95% window (both *Acc* = 0.64), and AI-Geography peaked in the 90-100% window (*Acc* = 0.68; **Figure 4**). All three peaks reached significance after FDR correction (Human: $p = .010$; AI-Trivia: $p = .019$; AI-Geography: $p = .019$).

Human speakers showed the most distributed prosodic markers across the utterance. For Human speakers, 7 of 19 windows reached significance (36.8%), covering early (0-20%), mid (45-65%), and late (75-100%) regions, with accuracies ranging from 0.60 to 0.64 (**Figure 4A**).

AI-Trivia showed prosodic markers concentrated only at utterance endings. AI-Trivia produced significant decoding in just 2 of 19 windows (10.5%), both at the very end (85-95%: *Acc* = 0.64; 90-100%: *Acc* = 0.62; **Figure 4B**).

AI-Geography showed two separate peaks and the highest overall accuracy. AI-Geography produced significant decoding in 2 of 19 windows (10.5%): one in the early-to-mid utterance (20-30%: *Acc* = 0.62) and one at the very end (90-100%: *Acc* = 0.68; **Figure 4C**). Notably, the early-to-mid window where AI-Geography reached significance was not significant in the original Human speakers.

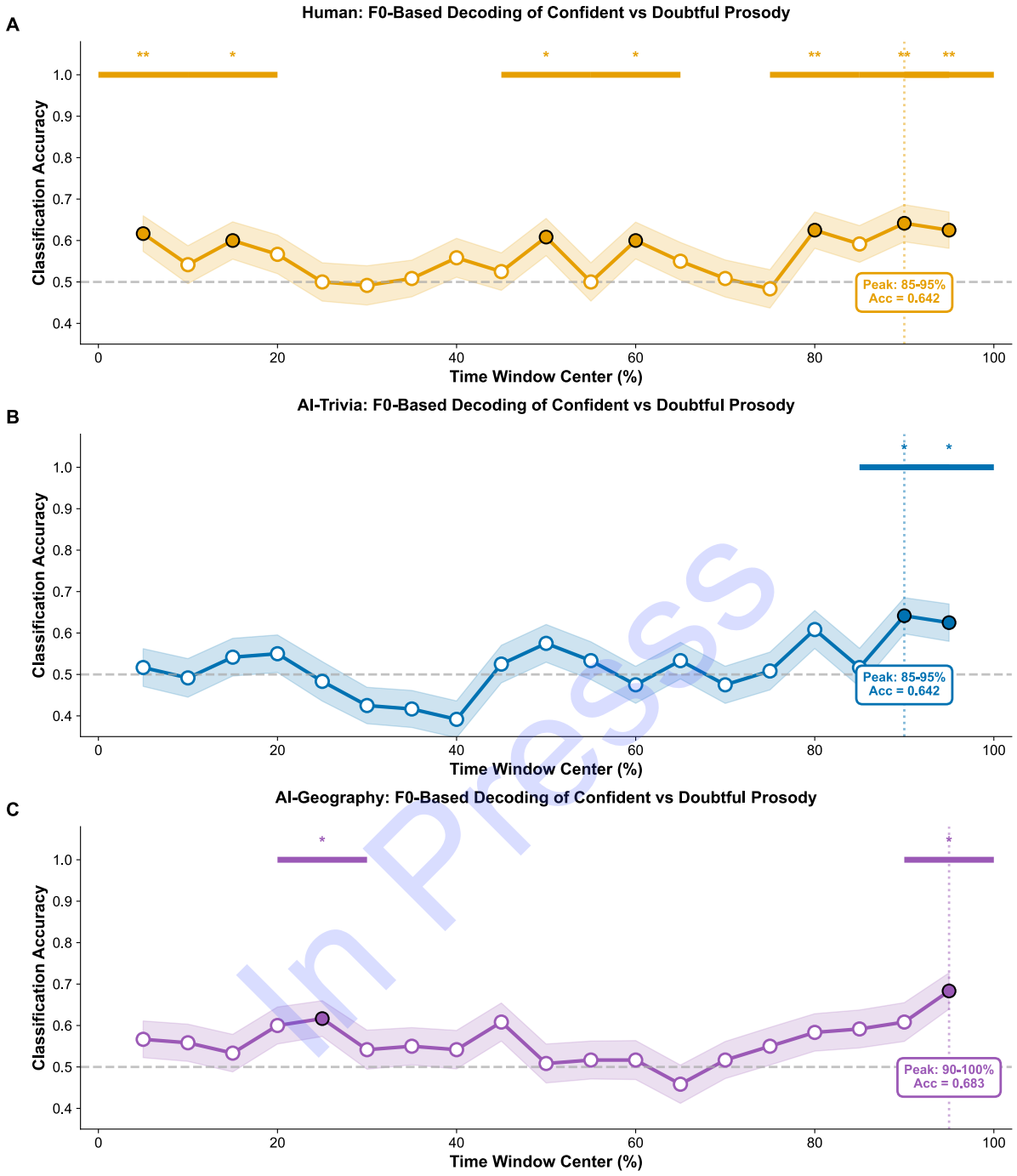


Figure 4. ***F0-Based Temporal Decoding of Confident vs. Doubtful Prosody.*** Time-resolved classification accuracy for distinguishing confident from doubtful prosody across normalized utterance duration. **(A)** Human speakers; **(B)** AI-Trivia speakers; **(C)** AI-Geography speakers. Classification used Random Forest models trained on F0 contours extracted from 10% sliding windows (5% step). Filled circles indicate windows significantly exceeding chance (0.5) after FDR correction; hollow circles indicate non-significant windows. Horizontal bars show significant windows: * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded regions represent ±1 SD from bootstrap resampling. Peak classification windows are annotated.

4. Discussion

We provide converging evidence that voice cloning services, originally designed to replicate speaker identity, can additionally reproduce prosodic patterns when trained on prosodically homogeneous recordings, although the reproduction is not a perfect match to human speech. For RQ1, machine learning classification using 88 eGeMAPS features successfully distinguished prosodies within and across human and AI sources, with spectral flux emerging as a top contributor in all three sources, suggesting that anatomy-independent prosodic dynamics play a central role. For RQ2, linear mixed-effects models revealed that confident prosody exhibited higher spectral flux, longer estimated VTL, and lower F0 than doubtful prosody across all three sources. For RQ3, time-resolved F0 decoding showed confidence-related markers peaking in late utterance windows across all sources, although Human speakers showed more broadly distributed markers than AI speakers.

4.1. Voice Cloning Captures Prosody Alongside Identity

As hypothesized, AI voice cloning captured prosodic patterns alongside speaker identity (Klimkov et al., 2019; Skerry-Ryan et al., 2018), even though such systems are primarily designed to clone vocal identity for users. Three converging lines of evidence support this conclusion. First, machine learning classification using 88 eGeMAPS features achieved high within-source accuracy (mean 0.85) across human, AI-Geography, and AI-Trivia speech. Second, linear mixed-effects models revealed parallel patterns across all three sources: confident prosody exhibited higher spectral flux, longer estimated VTL, and lower F0 than doubtful prosody. Notably, spectral flux emerged as a top contributor in all three sources and showed no biological sex effect, distinguishing it from F0 and VTL, which carry both prosodic information and speaker-anatomical variation. The fact that voice cloning replicated not only identity-related features (F0, VTL) but also this identity-independent

dynamic feature strengthens the claim that prosodic style transfer operates at multiple acoustic levels. Third, time-resolved F0 decoding identified late-utterance windows (85-100%) as peak regions for prosodic discrimination across all sources, indicating that AI-generated speech preserved the temporal structure of confidence-related markers.

These findings empirically support the dual encoding hypothesis articulated in our Introduction: mel-spectrograms capture both speaker-specific timbral characteristics enabling identity replication and temporal-frequency dynamics critical for prosody (Jia et al., 2018; Meng et al., 2019; Ong et al., 2023). When voice cloning models are trained on prosodically homogeneous recordings, they appear to inadvertently learn the prosodic style alongside the speaker's vocal identity. This byproduct effect transforms voice cloning from a tool for identity replication into a viable method for prosodic style transfer.

These findings have methodological implications for psychological research on human-AI voice interaction(Cross & Kappas, 2025), which requires controlled comparisons between human and AI voices with distinct prosodic profiles. Our paradigm addresses a key challenge: maintaining identical speaker identity while systematically varying prosody across human and AI sources. For example, Gampe et al. (2023) presented children with human-produced speech in monotone vs. lively styles, labeling monotone recordings as "AI" and lively recordings as "human." While this design controlled for voice identity (Roswandowitz et al., 2024), it relied on an operational definition equating prosodic impoverishment with AI speech. Current voice synthesis technology increasingly produces prosodically rich AI speech (e.g., Sora 2 by OpenAI), and the monotone-as-AI operationalization may not generalize beyond children, particularly to younger adults with extensive AI voice exposure (Herrmann, 2023), for whom genuine AI-generated speech may be more ecologically appropriate. Our voice cloning approach provides an alternative: researchers can train AI models on prosodically homogeneous human recordings and generate stimuli preserving both speaker

identity and intended prosodic characteristics (Li et al., 2025), enabling more rigorously controlled investigations of human-AI voice perception.

4.2. AI-Generated Speech as an Acoustic Population: Evidence for an "AI Dialect"

Although voice cloning successfully reproduced prosodic patterns at the acoustic level, AI-generated speech as a population still exhibited distinctive acoustic patterns relative to human speech, supporting the existence of an acoustic dialect. Several findings converge on this conclusion: (1) within-source classification accuracy substantially exceeded cross-source accuracy (mean 0.85 vs. 0.65; gap = 0.21, p = .001); (2) AI-Trivia showed significant feature-weighting convergence with Human (ρ = 0.51), whereas AI-Geography did not (ρ = 0.26); and (3) AI-Trivia's doubtful prosody was correctly identified by 91.7-100% of cross-source models. This pattern parallels the in-group advantage observed in human accent-based classification (Jiang & Pell, 2018; Laukka et al., 2014) and aligns with synthetic speech detection literature (Malik, 2019; Sahidullah et al., 2015), suggesting that different training-content domains produce different "sub-dialects" within the broader AI acoustic population, much like different linguistic backgrounds yield different human dialects (Scherer, 1986).

Temporal dynamics provide further insight into how voice cloning learns prosodic patterns. Human speakers showed prosodic markers across 7 of 19 distributed temporal windows, suggesting continuous dynamic modulation of the vocal apparatus throughout the utterance. Both AI sources, by contrast, concentrated their markers in only 2 of 19 windows and peaked near the utterance ending, consistent with the late-utterance pattern observed in other languages (Goupil et al., 2021; Jiang & Pell, 2017) and reinforcing the cross-language robustness of utterance-final prosodic markers. Voice cloning thus successfully captured the most discriminative late-utterance window but did not replicate the broader temporal distribution of human prosodic modulation. More strikingly, AI-Geography produced a

significant decoding window (20-30%) that was not significant in the original human recordings on which it was trained, indicating that voice cloning can also introduce acoustic markers absent in the source human speech. Together, these temporal patterns may reflect one signature of the AI "dialect" (used as an analogy only, see below): voice cloning learns the most prominent prosodic landmarks but does not faithfully reproduce the continuous dynamic variability of human speech production.

Beyond the temporal dynamics discussed above, another candidate signature of the AI "dialect" lies in detectable acoustic differences between human and AI speech as a whole. Large-scale studies on synthetic speech detection suggest a partial dissociation between algorithmic and perceptual detection of these differences. On the one hand, ML algorithms can reliably distinguish AI-generated speech from human speech (Sahidullah et al., 2015), based on systematic acoustic differences arising from the inherently more linear generation process of synthetic voices compared to the nonlinear physiological processes of human speech production (Malik, 2019). On the other hand, human listeners perform less reliably than algorithms (~70-80% accuracy under realistic conditions), and are particularly susceptible to misclassification when AI speech contains human-like features such as natural-sounding prosody or breathing, or when listeners attribute audio artifacts to recording quality rather than synthesis (Müller et al., 2022; Warren et al., 2024).

We therefore stress that our use of "AI dialect" is strictly an analogy, adopted because no established terminology yet captures this phenomenon. Unlike within-language accent dialects, which emerge from shared linguistic-cultural backgrounds and remain perceptually salient to listeners, the patterns observed here arise from algorithmic generation processes and are structurally more complex than any human dialect contrast. In essence, what we describe as the AI "dialect" refers to systematic acoustic regularities shared across AI-generated speech that distinguish it from human speech as a population, manifested both in

the temporal dynamics (Jiang & Pell, 2017) of prosodic markers and in detectable acoustic signatures (Malik, 2019; Sahidullah et al., 2015) at the broader level of human-AI speech contrast.

4.3. Articulatory Effort and the "Fast-and-Frugal" Logic of Confidence-Doubt Prosody

In human speech, speakers modulate their vocalizations to express various prosodies according to social context, including basic emotions (Kim et al., 2020), social dominance, masculinity/femininity (Pisanski et al., 2018), authority (Sorokowski et al., 2019), and attractiveness (Pisanski & Rendall, 2011; Xu et al., 2013). Underlying these phenomena is a fast-and-frugal principle (Pisanski et al., 2022):: rather than requiring years of growth and extensive metabolic investment to increase actual body size, speakers can instantaneously modulate vocal tract length to achieve competitive advantages at negligible energetic cost.

Our findings extend this fast-and-frugal principle to the cognitive-pragmatic domain of epistemic certainty (Goupil & Aucouturier, 2021). Although prompts included epistemic phrases ("I am certain"/"I'm not sure") to elicit prosodic styles (Vromans et al., 2024), target utterances remained lexically identical (Jiang & Pell, 2017). Speakers conveyed confidence vs. doubt through prosodic modulation alone, without altering propositional content. The core logic is that signaling epistemic commitment requires articulatory effort: speakers exert greater vocal effort when projecting confidence to maximize social gains (e.g., persuasion, perceived authority), and reduce such effort when conveying doubt. Our acoustic findings reflect this principle in two complementary ways. First, modulation of estimated VTL and F0 directly indexes vocal apparatus adjustment for epistemic signaling (Section 4.1). Second, human speakers distributed prosodic markers across multiple temporal windows throughout the utterance, with the strongest contrast at the utterance ending; voice cloning, by contrast, captured the strong utterance-final F0 contrast but did not replicate the broader, distributed

636 temporal modulation observed in human speech (Section 4.2). This dissociation suggests that
637 the fast-and-frugal principle, grounded in continuous neuro-muscular control of the vocal
638 apparatus, cannot be fully replicated by data-driven models lacking such a physiological
639 substrate. Human speakers remain distinctive in their capacity to dynamically modulate the
640 vocal tract throughout an utterance, while AI captures the most discriminative acoustic
641 landmarks but not the full continuous variability.

642 The spectral flux finding offers an additional perceptual perspective on this principle.
643 The auditory system is particularly sensitive to spectral change over time (Chen et al., 2012;
644 Kluender et al., 2003), and spectral flux specifically captures the acoustic dynamics that the
645 brain relies on to parse continuous speech into syllabic and phonemic units (Giroud et al.,
646 2024). That confident prosody exhibited higher spectral flux than doubtful prosody, with no
647 biological sex effect, indicates that confident speech contains richer dynamic spectral
648 variation than doubtful speech, regardless of speaker anatomy. This aligns with the fast-and-
649 frugal logic: confident speakers invest greater articulatory effort, producing more dynamic
650 spectral variation; doubtful speakers reduce this effort, yielding flatter spectral content. Voice
651 cloning reproduced this contrast across all sources, indicating that data-driven models can
652 capture acoustic signatures of effortful vs. reduced vocal modulation even without an
653 articulatory substrate.

654

655 **4.4. Limitations and Implications**

656 Several limitations of the present study warrant consideration. First, our sample
657 comprised 10 Mandarin speakers with experience in acting, speech, or music training, which
658 may limit generalizability to broader populations and to spontaneous speech. Second, our
659 findings were obtained using a single voice cloning system; other architectures may differ in
660 how they encode identity and prosody, potentially producing different "AI dialects." Third,

we examined a single language and a single prosodic contrast (confident, doubtful, neutral); whether voice cloning can reproduce other prosodic contrasts across other languages remains an open question.

Although we observed correlations between acoustic parameters (F0, VTL) and confidence ratings in human speech, and similar (indeed amplified) patterns when assigning AI-generated utterances the ratings from their source recordings (**Figure S2**), we did not treat these as primary evidence for prosodic cloning. Two considerations motivated this decision. First, collecting perceptual ratings for all 2,700 audio files would have imposed an excessive burden. Second, speech prosody simultaneously encodes subjective confidence and objective accuracy through distinct acoustic features, with accuracy encoded beyond speakers' metacognitive awareness (Goupil & Aucouturier, 2021); assuming AI-generated speech receives identical ratings to its source recordings overlooks this complexity. We therefore relied on more direct evidence (cross-source classification, temporal decoding, feature importance convergence) to demonstrate prosodic preservation.

Although our cross-domain design generated both training and novel sentences (e.g., AI-Geography trained on 15 Geography sentences then generated all 30 sentences including 15 novel Trivia sentences), we recommend that future research use entirely novel sentences not present in the AI training corpus (Lavan et al., 2025). While neural TTS systems generate new acoustic waveforms for each synthesis rather than retrieving recordings (Li et al., 2025; Tan et al., 2021), entirely novel sentences better capture real-world AI voice assistant scenarios. We therefore propose the following workflow: (1) train AI voice clones on prosodically homogeneous human recordings using one sentence set; (2) generate entirely different novel sentences using these trained AI speakers; (3) if resources permit, conduct third-party perceptual validation to assess both the AI-versus-human distinction (Nussbaum

et al., 2025) and the intended prosodic manipulation, applying rigorous filtering before deploying stimuli.

Finally, we acknowledge that voice in human-AI interaction is more than an acoustic signal: it functions as a medium of communicative practice through which speakers give an account of themselves, and critical scholarship has analyzed how voice technologies can embed identitarian and ideological assumptions, reducing voice to a measurable sound object (Napolitano, 2022a, 2022b). The present study deliberately focuses on the acoustic level, treating "speaker identity" in a narrow operational sense as anatomical-acoustic vocal signature rather than social or biographical identity. Our acoustic-level analysis provides methodological foundations for psychological research on human-AI voice interaction, but does not address the broader social, ethical, and critical dimensions of synthetic voice technology, which warrant continued scholarly attention.

5. Conclusion

As AI voice assistants become increasingly prevalent in daily life, the psychological implications of synthetic relationships emerge as pressing concerns (Cross & Kappas, 2025; McGettigan et al., 2025). Rigorous psychological experiments comparing human vs. AI voices require carefully prepared stimuli that control for speaker identity (Roswandowitz et al., 2024) while systematically manipulating prosodic characteristics (Gampe et al., 2023). The current study provides such a paradigm and offers a template for future psychological research to prepare more controlled voice stimuli that vary in prosody, particularly given that AI voices are no longer monotone and can now sound human-like.

References

Anikin, A., Barreda, S., & Reby, D. (2024). A practical guide to calculating vocal tract length and scale-invariant formant patterns. *Behavior Research Methods*, *56*(6), 5588-5604. https://doi.org/10.3758/s13428-023-02288-x

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

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

Chen, J., Baer, T., & Moore, B. C. J. (2012). Effect of enhancement of spectral changes on speech intelligibility and clarity preferences for the hearing impaired. *The Journal of the Acoustical Society of America*, *131*(4), 2987-2998. https://doi.org/10.1121/1.3689556

Chen, T. (2016). XGBoost: A Scalable Tree Boosting System. *Cornell University*.

Cross, E. S., & Kappas, A. (2025). Social Robotics Is Not (Just) About Machines, It Is About People: Psychology's Role in Developing Social Machines. *Annual review of psychology*. https://doi.org/10.1146/annurev-psych-040325-025951

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, *40*(2), 141-201. https://doi.org/10.1177/002383099704000203

Edwards, C., Edwards, A., Stoll, B., Lin, X., & Massey, N. (2019). Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions. *Computers in Human Behavior*, *90*, 357-362. https://doi.org/10.1016/j.chb.2018.08.027

Elfenbein, H. A. (2013). Nonverbal dialects and accents in facial expressions of emotion. *Emotion Review*, *5*(1), 90-96. https://doi.org/10.1177/1754073912451332

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., & Narayanan, S. S. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190-202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., & Schuller, B. (2010). *Opensmile: the munich versatile and fast open-source audio feature extractor* Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy. https://doi.org/10.1145/1873951.1874246

Fan, G., & Liu, D. (2025). When Machines Speak with Feeling: Investigating Emotional Prosody, Authenticity, and Trust in AI vs. Human Voices. *Proceedings of the Annual Meeting of the Cognitive Science Society*, https://escholarship.org/uc/item/8vr8s6h8

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

Giroud, J., Trébuchon, A., Mercier, M., Davis, M. H., & Morillon, B. (2024). The human auditory cortex concurrently tracks syllabic and phonemic timescales via acoustic spectral flux. *Science Advances*, *10*(51), eado8915. https://doi.org/10.1126/sciadv.ado8915

Goupil, L., & Aucouturier, J.-J. (2021). Distinct signatures of subjective confidence and objective accuracy in speech prosody. *Cognition 212*, 104661. https://doi.org/10.1016/j.cognition.2021.104661

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., & Aucouturier, J.-J. (2021). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common

prosodic signature. *Nature Communications*, *12*(1), 861. https://doi.org/10.1038/s41467-020-20649-4

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical Correlates of Affective Prosody. *Journal of Voice*, *21*(5), 531-540. https://doi.org/10.1016/j.jvoice.2006.03.002

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

Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication*, *36*(1), 31-43. https://doi.org/10.1016/S0167-6393(01)00024-3

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Lopez Moreno, I., & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, *31*.

Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, *88*, 106-126. https://doi.org/10.1016/j.specom.2017.01.011

Jiang, X., & Pell, M. D. (2018). Predicting confidence and doubt in accented speakers: Human perception and machine learning experiments. *Proceedings of Speech Prosody*, https://doi.org/10.21437/SpeechProsody.2018-55

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

Klimkov, V., Ronanki, S., Rohnke, J., & Drugman, T. (2019). Fine-grained robust prosody transfer for single-speaker neural text-to-speech. *arXiv preprint arXiv:1907.02479*. https://doi.org/10.48550/arXiv.1907.02479

Kluender, K. R., Coady, J. A., & Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, *41*(1), 59-69. https://doi.org/https://doi.org/10.1016/S0167-6393(02)00093-6

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1 - 26. https://doi.org/10.18637/jss.v082.i13

Laukka, P., Neiberg, D., & Elfenbein, H. A. (2014). Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations. *Emotion* *14*(3), 445. https://doi.org/10.1037/a0036048

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

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

Li, Y. A., Han, C., & Mesgarani, N. (2025). StyleTTS: A Style-Based Generative Model for Natural and Diverse Text-to-Speech Synthesis. *IEEE Journal of Selected Topics in Signal Processing*, *19*(1), 283-296. https://doi.org/10.1109/JSTSP.2025.3530171

Malik, H. (2019, 2019). Fighting AI with AI: fake speech detection using deep learning. *2019 AES International Conference on Audio Forensics*,

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314-324. https://doi.org/10.3758/s13428-011-0168-7

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. *Proceedings of the 14th Python in Science Conference*,

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

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network. *IEEE Access 7*, 125868-125881. https://doi.org/10.1109/ACCESS.2019.2938007

Mullennix, J. W., Stern, S. E., Wilson, S. J., & Dyson, C.-l. (2003). Social perception of male and female computer synthesized speech. *Computers in Human Behavior*, *19*(4), 407-424. https://doi.org/10.1016/S0747-5632(02)00081-X

Müller, N. M., Pizzi, K., & Williams, J. (2022). *Human Perception of Audio Deepfakes* Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, Lisboa, Portugal. https://doi.org/10.1145/3552466.3556531

Napolitano, D. (2022a). AI voice between anthropocentrism and posthumanism: Alexa and voice cloning. *Journal of Interdisciplinary Voice Studies*, *7*(1), 35-49. https://doi.org/10.1386/jivs_00053_1

Napolitano, D. (2022b). Reuniting speech-impaired people with their voices: sound technologies for disability and why they matter for organisation studies. *PuntOorg International Journal*, *7*(1), 6-21. https://doi.org/10.19245/25.05.pij.7.1.2

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. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2024.01.014

Nussbaum, C., Frühholz, S., & Schweinberger, S. R. (2025). Understanding voice naturalness. *Trends in Cognitive Sciences*, *29*(5), 467-480. https://doi.org/10.1016/j.tics.2025.01.010

Ong, K. L., Lee, C. P., Lim, H. S., Lim, K. M., & Alqahtani, A. (2023). Mel-MViTv2: Enhanced Speech Emotion Recognition With Mel Spectrogram and Improved Multiscale Vision Transformers. *IEEE Access 11*, 108571-108579. https://doi.org/10.1109/ACCESS.2023.3321122

Pamisetty, G., & Sri Rama Murty, K. (2023). Prosody-TTS: An End-to-End Speech Synthesis System with Prosody Control. *Circuits, Systems, and Signal Processing*, *42*(1), 361-384. https://doi.org/10.1007/s00034-022-02126-z

Pisanski, K., Anikin, A., & Reby, D. (2022). Vocal size exaggeration may have contributed to the origins of vocalic complexity. *Philosophical Transactions of the Royal Society B*, *377*(1841), 20200401. https://doi.org/10.1098/rstb.2020.0401

Pisanski, K., Oleszkiewicz, A., Plachetka, J., Gmiterek, M., & Reby, D. (2018). Voice pitch modulation in human mate choice. *Proceedings of the Royal Society B*, *285*(1893), 20181634. https://doi.org/10.1098/rspb.2018.1634

Pisanski, K., & Reby, D. (2021). Efficacy in deceptive vocal exaggeration of human body size. *Nature Communications*, *12*(1), 968. https://doi.org/10.1038/s41467-021-21008-7

Pisanski, K., & Rendall, D. (2011). The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *The Journal of the Acoustical Society of America*, *129*(4), 2201-2212. https://doi.org/10.1121/1.3552866

Raitio, T., Rasipuram, R., & Castellani, D. (2020). Controllable neural text-to-speech synthesis using intuitive prosodic features. *arXiv preprint arXiv:2009.06775*. https://doi.org/10.48550/arXiv.2009.06775

Rodero, E. (2017). Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices. *Computers in Human Behavior*, *77*, 336-346. https://doi.org/10.1016/j.chb.2017.08.044

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

Sahidullah, M., Kinnunen, T., & Hanilçi, C. (2015). A comparison of features for synthetic speech detection. *Proc. Interspeech 2015*. https://doi.org/10.21437/Interspeech.2015-472

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin*, *99*(2), 143. https://doi.org/10.1037/0033-2909.99.2.143

Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R., Clark, R., & Saurous, R. A. (2018). Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. *Proceedings of the 35th International Conference on Machine Learning*, https://proceedings.mlr.press/v80/skerry-ryan18a.html

Sorokowski, P., Puts, D., Johnson, J., Żółkiewicz, O., Oleszkiewicz, A., Sorokowska, A., Kowal, M., Borkowska, B., & Pisanski, K. (2019). Voice of Authority: Professionals Lower Their Vocal Frequencies When Giving Expert Advice. *Journal of Nonverbal Behavior*, *43*(2), 257-269. https://doi.org/10.1007/s10919-019-00307-0

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of memory and language*, *53*(1), 81-94. https://doi.org/10.1016/j.jml.2005.02.003

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

Trott, S., Reed, S., Kaliblotzky, D., Ferreira, V., & Bergen, B. (2023). The Role of Prosody in Disambiguating English Indirect Requests. *Language and Speech*, *66*(1), 118-142. https://doi.org/10.1177/00238309221087715

Vromans, R. D., van de Ven, C. C. M., Willems, S. J. W., Krahmer, E. J., & Swerts, M. G. J. (2024). It is, uh, very likely? The impact of prosodic uncertainty cues on the perception and interpretation of spoken verbal probability phrases. *Risk Analysis*, *44*(10), 2496-2515. https://doi.org/10.1111/risa.14319

Warren, K., Tucker, T., Crowder, A., Olszewski, D., Lu, A., Fedele, C., Pasternak, M., Layton, S., Butler, K., & Gates, C. (2024, 2024). " Better Be Computer or I'm Dumb": A Large-Scale Evaluation of Humans as Audio Deepfake Detectors. *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*, Salt Lake City, UT, USA. https://doi.org/10.1145/3658644.3670325

Xu, Y., Lee, A., Wu, W.-L., Liu, X., & Birkholz, P. (2013). Human vocal attractiveness as signaled by body size projection. *PLoS One*, *8*(4), e62397. https://doi.org/10.1371/journal.pone.0062397

Ethics approval: This study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was granted by the Ethics Committee of the *** University (Approval No. ***). Written informed consent was obtained from all participants prior to their participation in the study.

Data availability: All data and analysis code supporting the findings of this study are publicly available via the Open Science Framework (OSF) at https://osf.io/3c7rd/overview?view_only=a194fd9070ad4ca18b2eb7d5c4d303f4.