Title

Voice-Cloning Artificial-Intelligence Speakers Can Also Mimic Human-Specific Vocal Expression

Author Names:

Wenjun Chen
Institute of Linguistics, Shanghai International Studies University, China, 201620
Email addresses: 0213100578@shisu.edu.cn

Xiaoming Jiang
Institute of Linguistics, Shanghai International Studies University, China, 201620
Key Laboratory of Language Science and Multilingual Artificial Intelligence, Shanghai International Studies University, Shanghai 201620, China
Email addresses: xiaoming.jiang@shisu.edu.cn

Highlights:

- Voice cloning tools replicate vocal confidence, independent of linguistic content.
- Confidence by AI and humans relies differently on spectral and acoustic features.
- Classifications between two AI sources have an in-group advantage over AI-to-human.
- AI fails to replicate MFCC features through speech from human sources.
- Human and AI talkers "lengthen" vocal tracts in more confident than doubtful states.

Abstract

Purpose: This study explored how well vocal-identity-cloning AI can replicate human-specific vocal expressions by mimicking the acoustic and spectral markers humans use.

Methods: Ten Mandarin-speaking participants produced 900 audio clips across three confidence conditions (confident, doubtful, neutral), which were used to train AI models, generating 1,800 AI-cloned clips. In total, 2,700 clips were analysed. Acoustic features and spectral markers were extracted. Machine learning classifications using eXtreme Gradient Boosting (XGBoost) with and without cross-validation were employed to classify confidence levels, and linear mixed-effects models (LMEMs) examined the effects of confidence on key acoustic features in varied sources.

Main Findings: Both human and AI speech used Chroma constant-Q transform (Chroma_cqt) to differentiate confidence levels, with confident speech consistently displaying higher Chroma_cqt and longer vocal tract length (VTL). AI speech mirrored human speech by using Chroma_cqt, but it relied more on spectral features like spectral bandwidth, compensating for its inability to replicate the natural dynamic variability in human speech, particularly in features like Amplitude and VTL. AI models classified AI-generated speech more accurately than human speech, though cross-source classifications still performed above chance. AI struggled to replicate human-like

variability in Mel Frequency Cepstral Coefficients (MFCC), failing to differentiate clearly between neutral and confident states.

Conclusions: The study highlights that voice cloning tools can effectively replicate vocal confidence, independent of linguistic content, using different acoustic markers in both human and AI speech. While AI systems can mirror some features of human vocal confidence, they still struggle with capturing the nuanced variability and dynamic range inherent in human emotional expression. Future work should focus on enhancing AI's ability to model these complex acoustic variations to improve realism in emotional speech synthesis.

Keywords: Voice cloning; Affective computing; Vocal confidence; AI speakers; Speech communication

1. Introduction

Imagine asking a non-native English speaker in London to pronounce the name of a less-known village like Happisburgh (pronounced "Hayes-bruh"). The speaker may sound hesitant due to unfamiliarity. This hesitation is conveyed through their tone of voice. In contrast, a Text-to-Speech (TTS) system, like Apple's Siri, would likely pronounce it as "Hap-pis-burgh", without regard for correctness or familiarity. While TTS systems generate human-like speech through models such as WaveNet, Tacotron, or FastSpeech [1], it is unclear if AI can replicate the human ability to encode confidence through tone of voice. In human communication, paralinguistic cues play a crucial role in conveying both stable traits (like biological sex and age) and short-term states (such as emotion and confidence) [2]. These cues help listeners decode the speaker's identity and state. Both human listeners and computational models can effectively recognise states like sleepiness, emotion, and confidence – a speaker's internal "feeling of knowing" [2, 3]. This study investigates the extent to which AI can replicate human vocal expressions of confidence, focusing on the acoustic mechanisms underlying these expressions in both human and AI-generated speech. By utilising voice cloning TTS technology, the research aims to examine how confidence is encoded and assess how effectively AI-generated speech mimics these paralinguistic vocal patterns.

Previous research on human-computer interaction with AI voice has studied topics such as Amazon's Alexa having a female voice persona [4], AI voices with familiar or unfamiliar accents [5], and the preference for and trustworthiness of human-generated voices over AI-generated voices [6, 7]. However, these studies lacked efforts to match the identity of human and AI speakers. In contrast, Rodero [8] conducted acoustic analyses to ensure that human and AI speakers' fundamental frequency (F0) ranges were similar. Despite this, human raters perceived both synthetic voices (e.g., Siri and Loquendo) and human-manipulated voices (morphed using KaleiVoiceCope software) as less effective for advertisements compared to original, non-manipulated human voices [8]. In another study, the same woman acted as both the "AI" speaker and spoke with her natural voice. Children altered their interaction style, becoming less active when they believed they were interacting with an "AI" speaker, even though it was the same woman speaking in either a monotone or lively tone [9]. These studies attempted to compare human and AI speakers [8, 9], but variations in paralinguistic features, such as speaker identity or speech prosody, still existed and require more rigorous experimental control.

To address this gap, Huawei's Xiaoyi, a conversational agent using voice cloning technology, is utilised to replicate both speaker identity and speech prosody. While past human-computer

interaction (HCI) studies have used various TTS services, such as DECtalk, KaleiVoiceCope, Siri, Loquendo, and Microsoft Mary, which also involved voice-cloning technology based on specific human speaker embeddings, this study differs. This study is a 3 confidence level (confident vs. doubtful vs. neutral) × 2 sources (human vs. AI) design that contrasts how vocal confidence is encoded in speech produced by source human and AI-cloned speech.

The significance of exploring whether AI encodes vocal expressions in a manner similar to humans lies in existing studies that revealed an "in-group bias" when listeners decoded accented English sentences. Jiang and Pell [10] used machine learning to classify prosody groups and found that the algorithm had an advantage in classifying confident versus doubtful prosody in Canadian native, Quebecois-French, and Australian-English ("regional accent") speakers. However, while all classifications were above the chance level, accuracy decreased when training and testing were conducted across different accents. This computational approach motivates the current study to treat AI versus human speech similarly to "accented" vs. "non-accented" speech and investigate the commonalities in vocal confidence across these categories.

This study investigates key phonetic cues such as F0 (fundamental frequency), mean Amplitude, and Harmonics-to-Noise Ratio (HNR), which have been shown to be essential for distinguishing vocal confidence [10]. Additionally, perceptually salient vocal characteristics like Vocal Tract Length (VTL) are examined. VTL, which varies with the shape and length of the vocal tract, correlates with F0 and can indicate different mental states, such as confidence or doubt [11]. Evidence from both human and animal studies suggests that VTL modulation may signal dominance or size, implying that this mechanism may also play a role in conveying confidence in human speech [12, 13]. The ability to manipulate VTL and F0 may have contributed to the development of complex vocal communication in humans, particularly in expressing paralinguistic information like confidence [20]. Other features explored include Mel Frequency Cepstral Coefficients (MFCC), which simulate human auditory perception and are widely used in emotion and speaker recognition tasks [14]. Although the specific contribution of MFCC to vocal confidence is not fully understood, it remains an important tool in computational paralinguistics. Chroma-based features, such as Chroma STFT and Chroma CQT, are used to represent pitch and timbre, while Root Mean Square (RMS) energy captures the loudness of speech, both contributing to emotion and confidence classification [10, 15]. Other features like spectral centroid, spectral contrast, and spectral bandwidth are also considered, as they reflect aspects of speech brightness, frequency distribution, and timbre changes, which may encode confidence levels in speech [16]. A comprehensive list of features explored in this study is summarised in **Table 1**, including acoustic reflects basic physical properties of sound like pitch, loudness, and vocal energy (VTL, F0, Amplitude, HNR, RMS, and Utempo) and spectral features analyse the frequency content of sound, capturing tonal harmony and complexity (Chroma_stft, Chroma_cqt, Chroma_cens, MFCC, Spectral Centroid, Bandwidth, Contrast, Flatness, Rolloff, Tonnetz, and ZCR).

----------Insert **Table 1** about here----------

Table 1
17 Features for Machine Learning Classification

Acoustic Parameter	Auditory Perception	Measurement
Vocal Tract Length (VTL) [17]	A longer VTL is associated with a deeper, more mature voice.	Measured by the length of the vocal tract from glottis to lips.
Fundamental Frequency (F0) [10]	Lower F0 is linked to a deeper pitch.	Calculated by determining the base frequency of the voice signal.

Amplitude [10]	Increased Amplitude is perceived as a louder, more forceful voice.	Average Amplitude over the duration of the speech signal.
Harmonic-to-Noise Ratio (HNR)[10]	Lower HNR suggests more noise elements relative to harmonics, leading to a rougher voice.	Ratio of harmonic sound to noise in the speech.
Chroma_stft [18]		Uses Fourier transform to extract pitch-related features.
Chroma_cqt [16, 18]	Signifies tonal harmony and stability, contributing to confidence.	A chroma representation based on constant-Q transform.
Chroma_cens [18]		Chroma features normalised over time.
Mel Frequency Cepstral Coefficients (MFCC) [16]	Richer voice texture and timbre, clearer speech.	Calculated using Mel scale filter banks to model auditory perception.
Root Mean Square (RMS) [16]	Higher RMS indicates a louder and more authoritative voice.	Calculated as the square root of the mean squared Amplitude.
Spectral Centroid [16]	A higher spectral centroid leads to a brighter, clearer sound.	Determined by calculating the "center of mass" of the spectrum's energy.
Spectral Bandwidth [16]	Higher spectral bandwidth is associated with a more complex and dynamic voice.	The range between low and high frequencies contains most of the energy.
Spectral Contrast [18]	Higher spectral contrast adds texture and richness to the voice.	Difference in energy between spectral peaks and valleys.
Spectral Flatness [18]	Higher spectral flatness leads to a noisier sound with less harmonic content.	Measure how much the signal resembles white noise.
Spectral Rolloff [16]	Higher values indicate a brighter sound.	The rate at which spectral energy decreases across frequencies.
Tonnetz [19]	Higher values relate to more harmonious and structured tonal patterns.	Measures relationships between tones or pitch classes.
Zero Crossing Rate (ZCR) [16]	Higher ZCR values are associated with noisier, more abrupt sounds.	Count the number of times the signal waveform crosses zero.
Utempo (Speech Tempo) [20]	Faster tempo is associated with confidence and decisiveness.	Tempo is measured in beats per minute.

This study seeks to characterise vocal confidence in humans through acoustic/spectral features and assess how AI can replicate this human-specific emotion, addressing three research questions. 1) Can AI-cloned speakers effectively replicate the vocal confidence encoding mechanism observed in humans? If yes, 2) Is it feasible to consistently predict confidence levels across human and AI-generated speech? 3) How do those features, particularly vocal characteristics such as VTL, contribute to the portrayal of human vocal confidence?

In this study, ten human speakers were invited to produce 30 statements related to trivia and geography, using neutral, doubtful, and confident tones. One month later, they returned to the lab to rate their own produced sentences; the results suggest a clear distinction among the three confidence levels. The audio recordings of these statements were then used separately to train AI models designed to replicate speaker identity and confidence-related prosodies. These trained models were then applied to generate new statements via TTS based on novel text inputs. As a result, a total of 2,700 Chinese audio samples were collected from three sources: human speakers, AI models trained on trivia statements, and AI models trained on geography statements, each

reading the same 30 sentences. Two machine learning classification studies and one set of linear mixed-effects models (LMEMs) were conducted: 1) A ten-fold cross-validation XGBoost classification method was used to evaluate the importance of these features; 2) 1,000 iterations of classifications without cross-validation were performed to test the "in-group" advantage; and 3) LMEMs were fitted to visualise how important features signifies the vocal confidence.

2. Materials and Methods

This study recruited ten volunteers to express 30 statements in three confidence conditions. The obtained recordings were split into two halves according to the linguistic information per speaker – with half about trivia knowledge and the other half conveying highly-known knowledge about geography. Audios in each half were imported to Huawei's *Xiaoyi* service[1] one after another and thus made 10 speakers * 3 confidence levels * 2 sources (the trivia half and geography half) = 60 AI models, henceforth AI Trivia and AI-Geography. The feature extraction was followed by data analysis. The overall workflow is shown in **Fig. 1**.

----------Insert **Fig. 1** about here----------

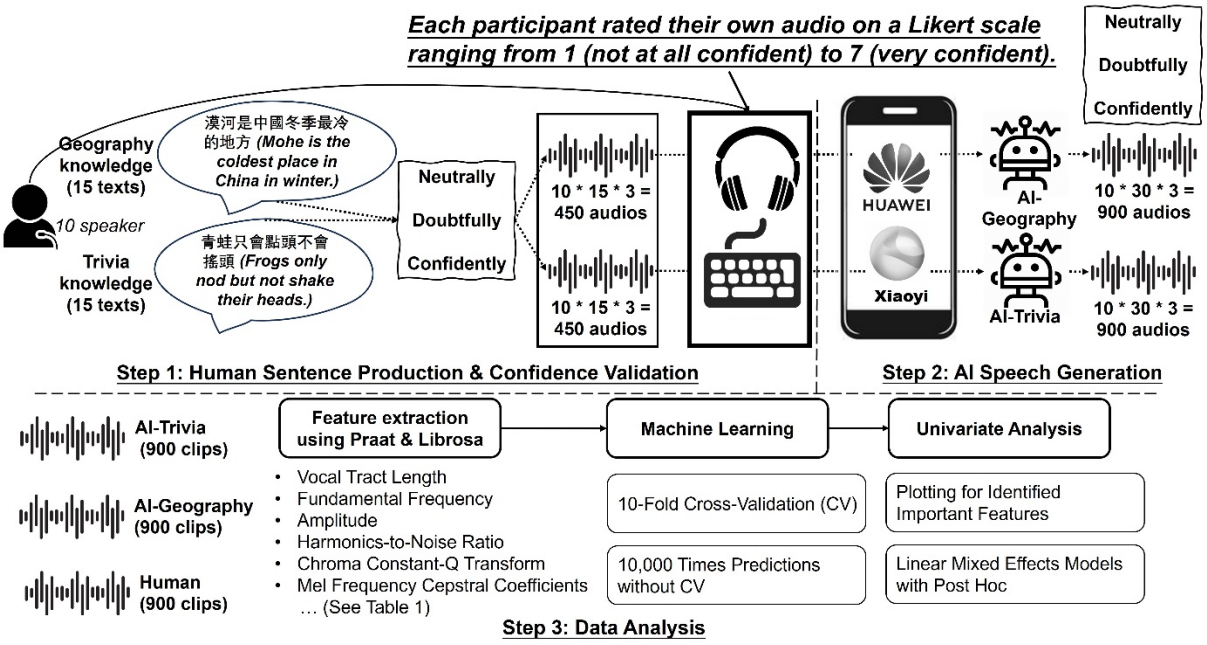

Fig. 1. Illustration of the three-step process: Step 1: Human speakers produced 900 sentences and self-assessed their own audio using a 7-point Likert scale (1 = not confident at all, 7 = very confident). Step 2: Using 15 sentences related to Geography knowledge, an AI-Geography talker was generated. Similarly, an AI-Trivia talker was generated using 15 sentences related to Trivia knowledge. Both the AI-Geography and AI-Trivia talkers then produced 30 sentences (covering both Geography and Trivia topics), matching the content of the 30 sentences initially produced by

[1] Huawei's - *Xiaoyi* (https://devicepartner.huawei.com/cn/solutions/product/hey-celia/) is a voice assistant like Apple's Siri. It provides a personalised voice clone service that allows the AI *Xiaoyi* to speak in the cloned vocal identity.

the human speakers. Step 3: Acoustic parameters were extracted, followed by multivariate and univariate data analysis.

2.1. Audios samples preparations

2.1.1. Human participants

Five males (Age=22.8±2.71 years; Years of education=19.4±2.73 years; Height=182.2±5.15 cm) and five females (Age=22±1.54 years; Years of education=19±1.1 years; Height=167.4±3.88 cm) standard Mandarin speakers from Shanghai International Studies University were recruited (with reimbursement). All had considerable experience in acting performance, speech or music training. All were reported to have high proficiency in Mandarin Chinese, evidenced by the Putonghua Proficiency Test (scored 87~91 out of 100). None of them reported any history of speech-hearing impairment or neurological or psychiatric disorders. All participants were given informed consent before entering the study. The study was approved by the Research Ethics Committee of the Institute of Linguistics, Shanghai International Studies University. The study was performed in accordance with the Declaration of Helsinki.

2.1.2. Audio recording

The recording took place in a sound-attenuated laboratory, where Audio-Technica AT2035 Cardioid Condenser Microphone was powered by Komplete Audio 6 Mk2 Sound Card, connecting to *Praat* 6.2.09, the sound recorder running on Dell G3-3579 (PC). The participants sat comfortably 20 centimetres away from the microphone. They read 30 prescribed sentences, consisting of 15 Trivia knowledge (Length=17±5.1), such as *"Frogs only nod their heads and do not shake them"* and 15 China highly-known geographical knowledge (Length=14.47±2.48), such as *"Mohe is the coldest place in China in winter"* (see **Supplementary Table 1**) required by Xiaoyi Smart Assistance, a conversational agent. All participants went through three independent blocks (in neutral, doubtful, and confident tone of voice), where each sentence was consecutively produced two times. In confident and doubtful blocks, participants first saw a screen showing texts such as *"You are playing a knowledge testing game, and you are asked, Frogs only nod their heads and do not shake them, aren't they?"*. Sentences were fully randomised per participant. Their vocal expression was elicited with a preceding lexical phrase of probability, such as *"I am certain"* or *"I'm not sure"*, randomly assigned to each text item per confidence condition [10]. To encourage the speakers' self-awareness during the recording, they were asked to rate their subjective confidence level after each sentence expression on a 7-Likert Scale, where 1 stood for "not at all confident" and 7 denoted "very confident". In total, 1,800 sentences (10 speakers * 3 confidence levels * 30 texts * 2 repetitions) were recorded. The better-expressed one of two clips was selected based on the speakers' explicit rating and the acoustic impression of how the sound represented the intended confidence level judged by the first author. All sentences were recorded at a single Channel, with a sampling rate of 44,100 Hz and saved as *wav* files. Recordings were edited to include only the target statements but not the preceding phrases.

2.1.3. Two sets of audios generated by AI models

The purpose of this process was to create AI-cloned models that replicate each individual speaker's vocal expressions for confident, doubtful, and neutral tones. A *Huawei Nova 9* cell phone was connected to the PC via *Changba Live No. 1 Sound Card Converter (2021-1)*, which allowed the simulation of the phone's microphone input with prepared *wav* files without signal loss. Sixty AI models (10 speakers * 3 confidence levels * 2 sources) were constructed by inputting (at a volume of 30% in the PC) each audio in Human set into Huawei's *Xiaoyi* service (Version: 11.0.44.306).

Two types of AI sources were models separately built upon Trivia sentences or highly-known Geography sentences, which were split into 30 AI-Trivia models and 30 AI-Geography models. For each AI model, Huawei's *Xiaoyi was* summoned to read 30 lines of text that had been expressed by human speakers while the screen recording was going on, thus forming 60 videos. All 60 videos were then converted into *wav* audio with *GoldWave* for Windows (Version: 6.65). The study further separated each long audio into audio clips through Python script based on the silence between each articulation. All 1,800 sentences (60 models * 30 sentences (15 Geography sentences + 15 highly-known Trivia sentences)) were hereby generated, followed by a normalisation manipulation at 70 dB SPL (henceforth Set AI-Trivia and Set AI-Geography).

2.2. Data analysis

2.2.1. Perception Study: Vocal confidence validation

All 900 human sounds (henceforth Set Human) were normalised at 70dB SPL with *Praat* for perceptual validation. To verify the robustness of vocal confidence, the same participants (n=10) were invited back to the laboratory to rate only their own recordings one month later. Sentences of their own voice were presented with *OpenSesame* [21], and they were asked to rate how confident the audio sounded on a 7-Likert scale, with 1 denoting "not at all confident" to 7 for "very confident". All recordings were played through Hewlett-Packard (HP) GH10 headphones at a comfortable volume level. The stimuli were presented randomly in three blocks.

LMEM was performed with the formula of *"Subjective Confidence Rating ~ Intended Confidence Level * Biological Sex + Text from Geography or Trivia + (1|Speaker)"* using lme4-package [22], followed by a subsequent post hoc comparison with emmeans when necessary [23]. An estimation of the effect size of the effect of interest - ηp^2 was provided using the test-statistic approximation method (https://easystats.github.io/effectsize/articles/anovaES.html). The small, medium, and large effect size is generally referred to as $\eta p^2 = .01$, $\eta p^2 = .06$, and $\eta p^2 = .14$ [24]. Results are reported in **Table 2**.

2.2.2. Feature extraction

Considering that VTL estimation was typically conducted on vowels or voiced segments [25], this study extracted the voiced parts of all 2,700 audios from the human and AI voice-cloning sources through the *Extract Vowels* functions of *Praat Vocal Toolkit* (www.Praatvocaltoolkit.com/), generating sets of Text-Grid annotated voiced parts for each audio. The mean VTL for each voiced part was estimated with the *Calculate Vocal Tract Length* function of the *Vocal Toolkit* and then averaged for each audio. Similarly, given that F0 calculation performs better in vowels than consonants [26] and in order to minimise the impact of unvoiced parts' influence over F0 estimation, this study extracted Mean F0 from each voiced part before averaging them for all 2,700 voiced-part-only audios. An LMEM was performed to confirm the equivalence of the extracted vowel number in each audio (about the same text) between the speaker's biological sex and sources (Set Human, AI-Trivia, or AI-Geography), using the formula: *"Number of extracted vowels in each audio ~ Source * Biological Sex + (1|Item) + (1| speaker)"*. No main effect was found for either Source (p=.936) or Biological Sex (p=.090). The interaction of Source and Biological sex was identified but with a small effect size ($F(2, 2657)=4.63$, p=.009, ηp^2=.003). These results suggested that humans and AI speakers both followed the phonemic rules in Chinese when articulating each text in spoken language.

In addition to Mean VTL and Mean F0 from voiced parts, 13 other spectral or beat-related acoustic parameters were extracted from the complete audios through *Librosa* [27]. The 13 features

included Chroma_stft, Chroma_cqt, Chroma_cens, MFCC, Root Mean Square, Spectral Centroid, Spectral Bandwidth, Spectral Contrast, Spectral Flatness, Spectral Rolloff, Tonnetz, ZCR, and Utempo. Among them, Utempo was already one-dimensional data. The other 12 spectral features were reduced to numeric numbers without time course and phase information through *numpy.mean()* function of Python for further analysis [28]. Despite 1-dimensional features could lose a certain amount of information than 2-dimensional spectral features and thus influence vocal states classification performance of machine learning studies [29], this study ensured all features were 1-dimensional so as to make them comparable to VTL and F0 for machine learning classification studies. Both Mean Amplitude and Mean HNR were calculated from the complete audios as global prosodic measures using *Praat* [3]. Altogether, this study obtained 17 features.

2.2.3. Statistical analysis

Analysis 1: Cross-validation-based machine learning classifications

This analysis investigates each feature's importance in signalling vocal confidence in each source. A machine learning 10-fold cross-validation (CV) classification study was conducted using the XGBoost package in python (version 3.9). XGBoost is a machine learning framework with proven high-performance scalability that implements gradient boosting to combine multiple weak learners into strong learners in the decision trees [29]. Seventeen features were inputted into machine learning to classify vocal confidence by humans and AI. Δ VTL and Δ F0 were excluded from the machine learning model because they have been reported to provide less benefit in perceptual studies and heavily depend on VTL and F0 themselves [30]. Seventeen other features were fed into the algorithm.

The XGBoost model was built using the *train()* function, and 10-fold cross-validation with the *cv()* function was applied to tune the model and prevent overfitting. In this setup, the dataset was split into 10 equal parts, with each fold used once for testing and the rest for training. Shapley values, grounded in cooperative game theory, were calculated to quantify the contribution of each feature to model predictions [31]. For each fold, Shapley values were computed for all 17 features by evaluating how the model's predictions changed with or without each feature. These values were averaged across all instances in each fold and then averaged again across the ten folds for robustness. The final importance scores were visualised through SHAP bar plots, illustrating how each feature contributed to classifying vocal confidence across datasets. The results are shown in **Figure 2 (A to C)** for Set Human, Set AI-Trivia, and Set AI-Geography.

Analysis 2: Machine learning training/testing studies to evaluate the predictability

This part of the study constructed 1,000 "non-reinforced listeners" in addition to the previously discussed "reinforced listeners" in the 10 fold-CV machine learning. Unlike the reinforced models that underwent K-fold CV, these non-reinforced algorithms were each exposed to a random combination of subsets once before performing classification tasks. The data for each listener was split into a Training Set (80%) and a Testing Set (20%). Performance metrics for each of the 1,000 classifications included overall accuracy, accuracy for each confidence condition, root-mean-squared errors (RMSE), and F1-score (macro). RMSE measures the spread of prediction errors from true values by calculating the square root of the average squared differences between predicted and actual values [32]. The F1-score (macro) assesses the accuracy of binary classifications by averaging the harmonic mean of precision and recall for each class, thus providing equal weight to each class [32]. The performance outcomes were aggregated to derive average accuracies and model performances, depicted in **Table 3**. A representative "listener" was

chosen based on a matching RMSE and F1-score (macro) to the averages from the 1,000 tests, with its ROC curve displayed in **Figure 2D** to illustrate in-group advantages visually.

Additionally, an ANOVA was conducted to validate the presence of an in-group advantage, defined as the model's improved predictive accuracy when trained and tested on the same source set compared to different sets. The analysis followed the formula *"Overall Accuracy of each iteration in the 1,000 runs ~ Training * Testing"*, incorporating three levels of data sources in both the Training and Testing variables. A pairwise comparison was performed using the argument *"pairwise ~ Training * Testing"* to evaluate all possible combinations between these levels. The detailed results of the ANOVA and pairwise comparisons are presented in **Supplementary Table 2**.

Analysis 3: Univariate analyses of feature's role in portraying vocal confidence

Six acoustic features – Chroma_cqt, Mean Vocal Tract Length (VTL), Amplitude, Mean Fundamental Frequency (F0), Spectral Bandwidth, and MFCC – were selected based on their importance scores from the 10-fold cross-validation classification. These features are compared across AI-Geography, AI-Trivia, and human sources, with groupings by biological sex and confidence levels.

In recognising the known impact of Biological Sex on the perception of vocal confidence and the existing knowledge of women's and men's biological differences in certain laryngeal features such as VTL [33], a set of LMEMs was formulated through *"VARIABLE ~ Confidence Levels * Biological Sex + (1| Height) + (1|Item)"* while separating the data according to three sources. *VARIABLE* are six selected features according to importance scores from 10-fold CV classification. Plots are shown in **Fig. 3**, with p-values of pairwise contrasts using "Confidence Levels | Biological Sex" and "Biological Sex | Confidence Levels" annotated. The corresponding emmeans values comparison were shown in **Table 4**. Meanwhile, the inferential statistics were shown in **Supplementary Table 3** (between biological sexes in one particular confidence level) and **Supplementary Table 4** (between confidence levels in male or female).

To verify the correlation between VTL and F0 [11], as well as to explore the possible relationship between VTL and Chroma_cqt the most important parameter to classify vocal confidence and VTL, the *"lm(Mean_vtl ~ Mean_F0)"* and *"lm(Mean_vtl ~ Chroma_cqt)"* was employed (See **Fig. 4**).

3. Results

3.1. Perception study: Vocal confidence validation

The LMEMs revealed the main effect of intended confidence ($F(2,880)= 5972.24$, $p<<2e\text{-}16$, $\eta p^2 =.93$) but not that of Biological Sex ($F(1,8)= .25$, $p= .63$) or text ($F(1,8885)= 3.59$, $p= .06$). The interaction between intended confidence and Biological Sex ($F(2,885)=7.42$, $p= .0006379$, $\eta p^2 =.02$) was found. Post hoc results showed that, using a 95% confidence interval (CI), rating scores were ranked as confident (6.57±.05) > neutral (4.08±.05) > doubtful (1.39±.05) from high to low (not shown in **Table 2**). Also, the ratings for the intended confidence level were consistent across biological sexes, although females rated their doubtful speech lower than males (β=-.24, t=-2.37, p=.0307; **Table 2**). These findings validated the perceptual differences between the three intended confidence levels.

----------Insert **Table 2** about here----------

Table 2
Post-hoc Analysis of Subjective Confidence Ratings by Intended Confidence Levels and Biological Sex

Interaction		*β*	t	*p*
Female	Confident-Doubtful	5.36	80.13	<.0001
	Confident-Neutral	2.55	38.17	<.0001
	Doubtful-Neutral	-2.81	-41.96	<.0001
Male	Confident-Doubtful	5	74.74	<.0001
	Confident-Neutral	2.43	36.28	<.0001
	Doubtful-Neutral	-2.57	-38.47	<.0001
Confident	Female-Male	.12	1.19	.2533
Doubtful	Female-Male	-.24	-2.37	.0307
Neutral	Female-Male	-.01	-.07	.9483

3.2. Analysis 1: Identifying key features of vocal confidence using 10-Fold cross-validation

For the Human context (**Fig. 2A**), three levels of vocal confidence classification are predominantly influenced by Chroma_cqt (0.15), Chroma_cens (0.08), and Amplitude (.05). RMS and MFCC also contribute significantly, each adding 0.03, while VTL and Spectral Contrast each have a lesser influence at 0.01. In the AI-Trivia context **(Fig. 2B)**, Chroma_cqt stands out at 0.17, marking a noticeable increase from the human model. Spectral Bandwidth (0.04) and F0 at 0.03 each emphasise a shift toward spectral characteristics and pitch. Chroma_cens at 0.07 and Chroma_stft at 0.02 also play roles alongside RMS, Spectral Rolloff, Spectral Flatness, and Spectral Contrast, each at 0.01, and VTL, also contributing at 0.01. For AI-Geography (**Fig. 2C**), similar trends are observed with Chroma_cqt (0.15) and Chroma_cens (0.09) leading the influences. An increased emphasis on VTL at 0.03 aligns with the model's focus. Spectral Bandwidth again stands at 0.04, with adjustments in F0 and RMS each marked at 0.02, and contributions from Chroma_stft and Spectral Rolloff at 0.02 and 0.01, respectively.

Across three contexts, Chroma_cqt proves key for portraiting vocal confidence in both human and AI contexts. Similarly, VTL was found to be important in three sources, too. While other spectral features like Chroma_cens, F0, and Spectral Bandwidth show similar patterns of relevance in AI models, a stark contrast emerges with the use of Amplitude. Unlike in the human context where Amplitude is crucial, it is notably absent in AI models, highlighting a shift from amplitude-dependent dynamics in humans to a spectral-focused approach in AI.

----------Insert **Fig. 2** about here----------

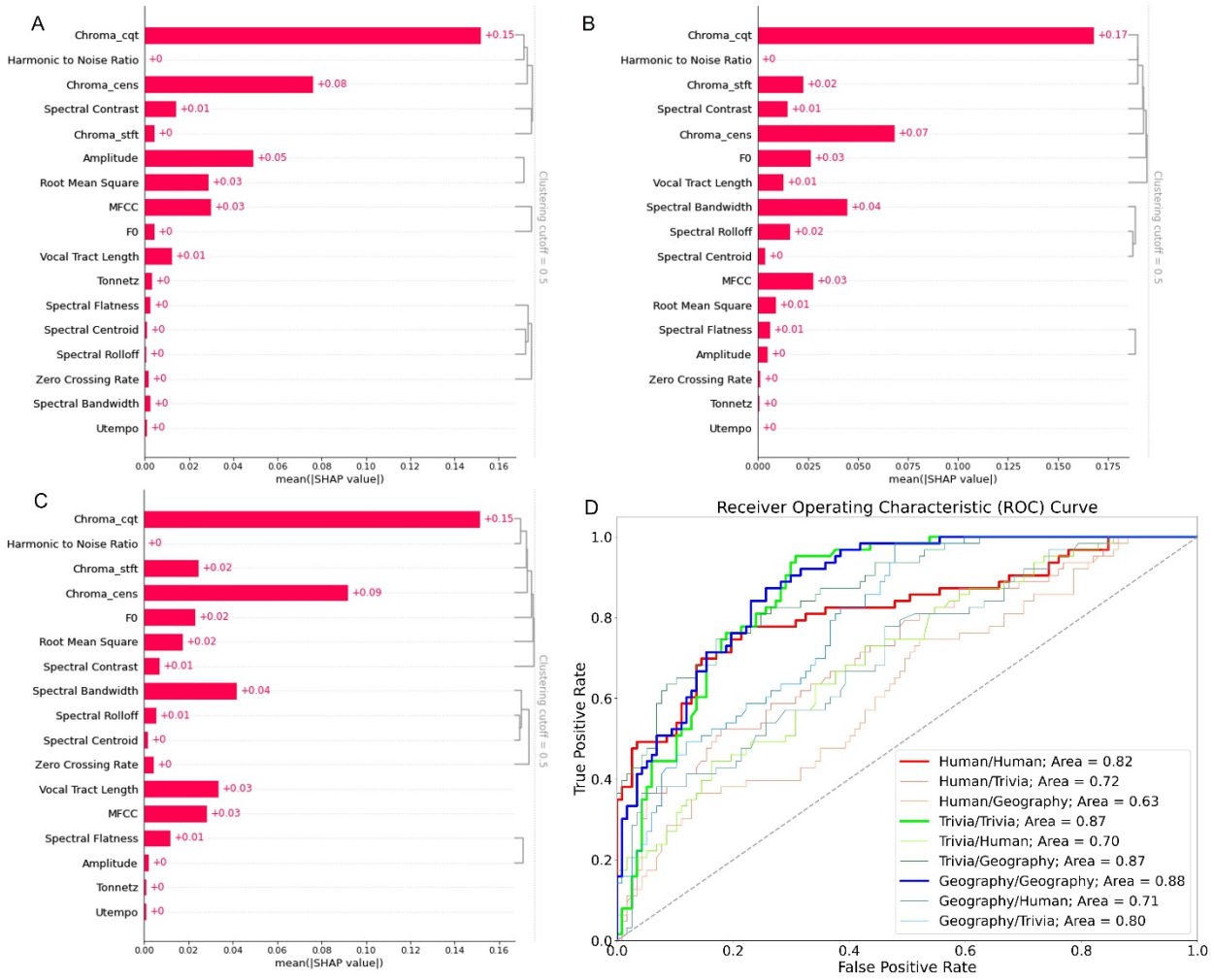

Fig. 2. Importance scores for 17 features reported by vocal confidence differentiating algorithms based on 10-fold cross-validation. The datasets used were (A) Human, (B) AI-Trivia, and (C) AI-Geography. As features (e.g., RMS energy and Amplitude) could be correlated with each other, a clustering cutoff analysis was conducted using the SHAP (SHapley Additive exPlanations, https://shap.readthedocs.io/) function to reduce acoustic dimensions. The clustering cut-off of 0.5 indicates factors that shared more than 50% of their explanatory power. The ROC curve for the "listener" of random_seed=44,523 was illustrated in (D) as it shared similarity in overall accuracy, F1 Score (macro), and RMSE with the 1,000 averaged results.

3.3. Analysis 2: Evaluating vocal confidence prediction with 1,000 non-cross-validated Iterations

The results from the 1,000 iterations, shown in **Table 3**, reveal two clear in-group advantages. First, models performed best when trained and tested on data from the same source. For example, the Human/Human model (H/H) achieved the highest overall accuracy of 0.72, while AI-Geography/Human (AIg/H) and AI-Trivia/Human (AIt/H) models showed lower accuracies of 0.51 and 0.38, respectively. Similarly, when the Human model was tested on AI data, accuracies dropped, with H/AIg at 0.54 and H/AIt at 0.53. Secondly, AI models exhibited better performance when tested on other AI data compared to human data. For instance, AIg/AIt achieved an accuracy

of 0.67, and AIt/AIg reached 0.69, both higher than AIg/H (0.51) and AIt/H (0.38). All accuracy levels exceeded the chance level (1/3), underscoring meaningful classification of vocal confidence.

The ROC curve in **Fig. 2D** also demonstrated such in-group advantage. The ANOVA analysis of *"Overall Accuracy of each iteration in the 1,000 runs ~ Training * Testing"* revealed significant main effects of both training (F=2175, p<2e-16, ηp^2=.33) and testing (F=8335, p<2e-16, ηp^2=.65), as well as their interaction effect (F=16123, p<2e-16, ηp^2=.88). The pairwise contrast in **Supplementary Table 2** yielded several findings. H/H showed better performance than AIt/H (β=.34, p<.0001) and AIt/AIt (β=.04, p<.0001). AIg/H demonstrated superior performance than H/H (β=-.21, p<.0001) and AIt/H (β=.13, p<.0001). AIg/AIg consistently outperformed other conditions, particularly when compared to AIt/H (β=.38, p<.0001) and H/AIg (β=.22, p<.0001). AIt/AIg underperformed when compared to H/AIg (β=-.15, p<.0001) and H/AIt (β=.16, p<.0001). Finally, the AI-Trivia model performed equally well when trained on AI-Geography and AI-Trivia datasets for testing on AI-Trivia data; see AIt/AIg - AIt/AIt (β=0, t=2.09, p=1).

Altogether, the 1,000 training and testing iterations confirmed that models performed best when trained and tested on data from the same source. AI models generally achieved higher accuracy when tested on other AI data compared to human data, demonstrating the in-group advantage. Despite this, the above-chance-level accuracies when training and testing across human and AI sources suggest that AI models are still capable of replicating human-specific vocal confidence to a significant degree.

----------Insert **Table 3** about here----------

Table 3

Machine Learning Classification Accuracies from 1,000 Iterations, a Representative Trial, and Model Reliability Indicators

Training/Testing[a]	Overall Accuracy [b]	Accuracy[c]			RMSE[d]	f1-score (macro)[e]
		Confident	*Neutral*	*Doubtful*		
The accuracy results from 1,000 classifications						
H/H	.72	.84	.78	.82	.67	.72
AIg/H	.51	.63	.69	.7	1.01	.45
AIt/H	.38	.4	.68	.68	1.24	.26
AIt/AIt	.69	.78	.74	.85	.69	.68
AIg/AIt	.67	.78	.74	.83	.72	.66
H/AIt	.53	.74	.7	.63	.98	.51
AIg/AIg	.75	.83	.81	.87	.63	.74
AIt/AIg	.69	.8	.76	.82	.71	.67
H/AIg	.54	.75	.71	.62	.98	.51

[a] AIg for AI-Geography; AIt for AI-Trivia; H for Human.

[b] The accuracy was calculated as (TP + TN) / (TP + FP + TN + FN).

c Accuracy of class i = (TPi + TNi) / (TPi + FPi + TNi + FNi), where TPi is the number of true positives for class i, FPi is the number of false positives for class i, TNi is the number of true negatives for class i, and FNi is the number of false negatives for class i.

[d] The root-mean-squared error (RMSE) was used to indicate the model fit.

[e] This study tackled a multi-classification problem where three confidence levels were classified. The F1 score (macro) was used to represent the averaged accuracy across the three confidence levels and indicate the model fit. It was calculated by averaging each category's F1 score (F1 score = 2 * (precision * recall) / (precision + recall)), where precision is TP/(TP+FP) and recall is TP/(TP+FN).

3.4. Analysis 3: The effects of confidence levels on acoustic cues in Human and AI sources

Group-level differences across the three confidence levels (Confident, Neutral, Doubtful) were observed, as shown in **Fig. 3**. Firstly, for Chroma_cqt, in confident speech, it consistently displayed higher values than neutral, which in turn was higher than doubtful across both biological sexes and the three sources (Human, AI-Geography, AI-Trivia). Secondly, VTL followed a similar trend to Chroma_cqt, except in two cases: among females in Human speech, there were no significant differences between confident and neutral; similarly, in AI-Trivia's male group, confident and neutral conditions did not differ significantly. Additionally, in AI-Geography, no significant difference was observed between doubtful and neutral for either sex. Thirdly, Amplitude, identified as an important feature for machine learning classification (SHAP: 0.05 for Human, 0 for AI-Geography and AI-Trivia), exhibited an inverse trend in the confident condition, where it had the lowest values compared to neutral and doubtful. Among males, doubtful speech showed the highest Amplitude, followed by neutral, with confident speech being significantly lower. However, no significant difference was found between doubtful and neutral in females, although both were larger than confident. Fourth, for F0, doubtful speech consistently exhibited the highest values, followed by confident, with neutral speech having the lowest values across all three sources. Fifth, Spectral Bandwidth, important in both AI models (0.04 for AI-Geography and AI-Trivia, 0 for Human), displayed a consistent ranking: confident speech had the highest values, followed by neutral and then doubtful. Notably, in AI-Geography, no significant differences were found between males and females, and in females, neutral and doubtful were not significantly different. Sixth, for MFCC, doubtful speech consistently exhibited higher values than neutral, which was followed by confident speech across all three sources. An exception was found in AI-Geography's female group, where no significant difference was observed between confident and doubtful conditions.

----------Insert **Fig. 3** about here----------

Fig. 3. Comparisons in (A) Chroma_cqt, (B) Mean VTL in centimetres (cm), (C) Amplitude in decibels (dB), (D) Mean F0 in Hertz (Hz), (E) Spectral Bandwidth, (F) MFCC, in AI-Geography and AI-Trivia TTS models, and human, grouped by biological sex and confidence levels.

----------Insert **Table 4** about here----------

Table 4

Post Hoc Contrast Results of Confidence Level and Biological Sex on Six Features per Source

Feature	Contrast on CL[a]	Contrast on BS[b]	AI-Geography		AI-Trivia		Human	
			emmean	lower.CL, upper.CL	emmean	lower.CL, upper.CL	emmean	lower.CL, upper.CL
Chroma_cqt	C	F	.37	[.35,.39]	.36	[.34,.39]	.43	[.41,.45]
	C	M	.43	[.41,.45]	.44	[.42,.46]	.49	[.47,.51]
	D	F	.31	[.29,.33]	.31	[.29,.33]	.37	[.35,.39]
	D	M	.33	[.32,.35]	.33	[.31,.35]	.39	[.37,.41]

	N	F	.34	[.32,.36]	.35	[.32,.37]	.4	[.38,.42]
	N	M	.42	[.41,.44]	.42	[.39,.44]	.45	[.43,.47]
	C	F	15.76	[14.85,16.67]	15.62	[14.91,16.33]	15.34	[14.84,15.84]
	C	M	17.83	[17.02,18.64]	17.65	[17.02,18.29]	17.43	[16.99,17.88]
Mean VTL	D	F	15.18	[14.28,16.09]	14.95	[14.24,15.66]	14.99	[14.49,15.49]
	D	M	17.02	[16.21,17.83]	17.29	[16.65,17.92]	17.21	[16.77,17.66]
	N	F	15.24	[14.34,16.15]	15.43	[14.72,16.14]	15.19	[14.69,15.69]
	N	M	17.63	[16.82,18.44]	17.74	[17.1,18.37]	17.38	[16.93,17.83]
	C	F	31.78	[28.35,35.2]	31.04	[26.75,35.33]	63.52	[61.67,65.37]
	C	M	32.08	[28.71,35.44]	30.71	[26.47,34.94]	61.42	[59.75,63.08]
Amplitude	D	F	29.22	[25.79,32.64]	26.73	[22.44,31.01]	67.21	[65.36,69.07]
	D	M	26.55	[23.18,29.91]	25.93	[21.7,30.17]	66.63	[64.97,68.29]
	N	F	31.82	[28.39,35.24]	28.44	[24.15,32.73]	67.05	[65.2,68.9]
	N	M	30.96	[27.59,34.33]	28.14	[23.91,32.38]	65.83	[64.17,67.49]
	C	F	236.51	[208.79,264.22]	231.11	[204.46,257.76]	239.06	[209.45,268.68]
	C	M	148.69	[123.89,173.48]	137.96	[114.11,161.8]	153.32	[126.82,179.83]
Mean F0	D	F	261.08	[233.36,288.8]	264.82	[238.17,291.48]	266.35	[236.73,295.97]
	D	M	157.54	[132.75,182.34]	150.25	[126.4,174.09]	149.58	[123.08,176.08]
	N	F	218.04	[190.32,245.76]	217.66	[191.01,244.31]	213.67	[184.06,243.29]
	N	M	117.12	[92.33,141.92]	114.7	[90.86,138.55]	115.27	[88.77,141.78]
	C	F	1182.04	[1106.42, 1257.67]	1306.9	[1225.73, 1388.07]	2129.21	[2018.53, 2239.90]
	C	M	1162.02	[1093.18, 1230.85]	1277.95	[1204.18, 1351.72]	2143.18	[2043.53, 2242.83]
Spectral Bandwidth	D	F	1069.58	[993.96, 1145.20]	1232.4	[1151.23, 1313.57]	2164.17	[2053.49, 2274.85]
	D	M	1023.38	[954.55, 1092.22]	1124.92	[1051.15, 1198.68]	2104.66	[2005.00, 2204.31]
	N	F	1083.31	[1007.69, 1158.93]	1224.89	[1143.72, 1306.06]	2145.1	[2034.41, 2255.78]
	N	M	1065.66	[996.82, 1134.49]	1140.88	[1067.11, 1214.65]	2085.37	[1985.72, 2185.02]
	C	F	-29.79	[-32.58,-27.01]	-27.84	[-30.69,-24.99]	-26.08	[-30.01,-22.14]
	C	M	-24.11	[-26.6,-21.62]	-22.21	[-24.76,-19.66]	-20.13	[-23.65,-16.6]
MFCC	D	F	-30.08	[-32.86,-27.3]	-28.79	[-31.64,-25.94]	-23.83	[-27.76,-19.89]
	D	M	-23.16	[-25.65,-20.67]	-21.63	[-24.18,-19.08]	-16.36	[-19.88,-12.84]
	N	F	-27.68	[-30.46,-24.9]	-26.37	[-29.22,-23.52]	-21.03	[-24.97,-17.09]
	N	M	-20.7	[-23.19,-18.21]	-19.76	[-22.31,-17.2]	-13.36	[-16.88,-9.84]

[a] C for Confident; D for Doubtful; N for Neutral. CL for Confidence Levels.

[b] F for Female; M for Male. BS for Biological Sex.

Correlation analyses revealed a negative correlation between Mean VTL and Mean F0 and a positive correlation between Mean VTL and Chroma_cqt across AI-Geography, AI-Trivial, and Human conditions (**Fig. 4**).

----------Insert **Fig. 4** about here----------

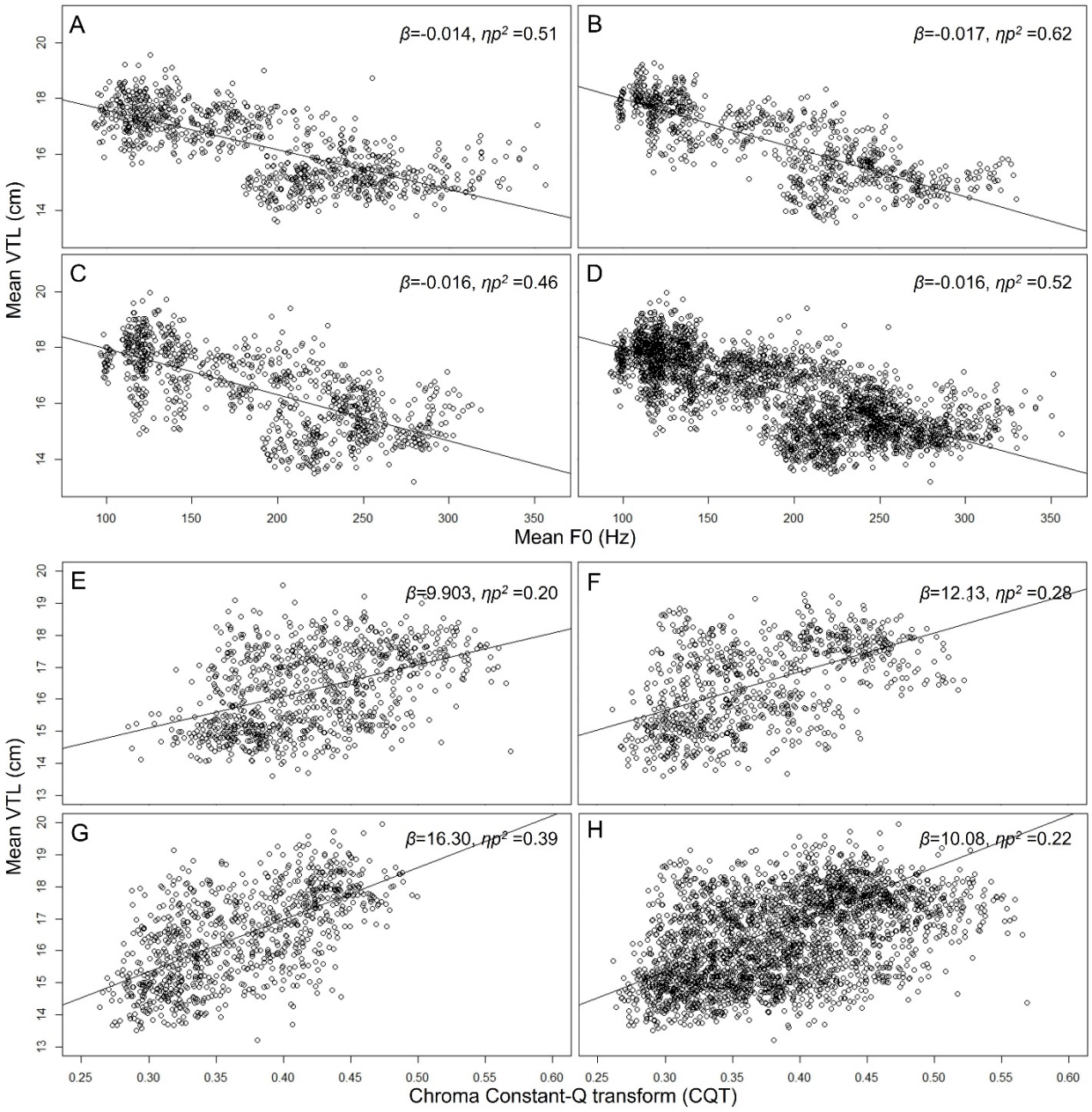

Fig. 4. Correlation plots between Mean VTL (cm) and Mean F0 are shown in (A-(E), with each point representing an individual speaker's average values across three confidence levels. The correlation between VTL and Chroma_cqt is shown in (F)-(H), with each point representing a single sentence from the confidence levels. The x- and y-axes in each plot correspond to Mean VTL and Mean F0 (Hz) or VTL and Chroma_cqt, as indicated.

4. Discussion

4.1. Spectral vs. acoustic features: Contrasts in human and AI speech

One key finding of this study is the central role of Chroma_cqt in distinguishing emotional states in both human and AI-generated speech. Chroma_cqt, which reflects pitch and tonal harmony, emerged as the most important feature for conveying vocal confidence across both sources. In confident speech, Chroma_cqt values are consistently higher, while doubtful speech is associated with lower values, indicating that both human and AI speakers adjust tonal qualities similarly to express emotions. Furthermore, Chroma_cqt is positively correlated with VTL and negatively correlated with F0, reinforcing its connection to confident speech. Voices with higher Chroma_cqt and longer VTL tend to sound "brighter" and more resonant, much like higher musical notes [16], which often project a more engaging and confident vocal quality. This correlation reflects the tonal adjustments that both humans and AI models use to signify emotional states, aligning with musical theory, where higher notes are often linked to brightness and assertiveness.

In AI-generated speech, Chroma_cens and spectral bandwidth play pivotal roles in classifying vocal confidence. Chroma_cens, which measures the stability of tonal content over time, helps AI models track how consistently pitch is maintained, making it effective for differentiating between emotional states like confidence and doubt. Meanwhile, spectral bandwidth, which reflects the range and richness of frequencies, adds emotional depth to the voice by enhancing its harmonic complexity. In contrast, these features are less crucial for human speech, where Chroma_cqt, Amplitude, RMS, MFCC, and VTL take precedence. Humans naturally rely on a broader range of cues, such as loudness and vocal tract length, to convey emotions, while AI compensates for this lack of innate emotional expressiveness by focusing on more structured spectral patterns. This distinction highlights that AI prioritises tonal stability and spectral richness, whereas human speech employs more varied, dynamic cues for expressing emotions. AI-generated speech, thus, leans heavily on tonal consistency and frequency complexity, which may result from the AI's need to simulate human-like expressiveness through more measurable and structured acoustic elements.

This study reveals differing emphases on acoustic and spectral features in human and AI-generated speech when conveying vocal confidence, doubt, and neutrality. Human speech tends to prioritise acoustic features like Amplitude, VTL, RMS, and F0, which are naturally modulated to reflect emotions, making them more dynamic and varied. In contrast, AI-generated speech relies more heavily on spectral features such as Chroma_cqt, Chroma_cens, and spectral bandwidth, using structured tonal stability and frequency complexity to compensate for the absence of natural emotional modulation. This indicates that while humans instinctively adjust physical properties like loudness and pitch, AI systems depend on more formulaic spectral patterns to replicate emotional expression. It is hypothesised that this difference arises from AI's reliance on precise, consistent patterns in the sound spectrum, favouring spectral elements for more controlled emotion simulation across various contexts.

4.2. In-group advantage and the "dialect" of AI speech: Challenges in replicating human emotional expression

This study highlights an in-group advantage, where machine learning models trained and tested within the same data source—whether AI or human—achieved higher accuracy than when applied across sources. Models performed significantly better when classifying data from their own group, suggesting that they are more attuned to familiar vocal patterns. However, accuracy dropped in AI-to-human or human-to-AI classifications, reflecting the difficulty in bridging the differences

between AI-generated and human speech. Interestingly, AI models trained on different datasets (e.g., AIg/AIt at 0.67) demonstrated cross-dataset generalisation, pointing to shared patterns within AI-generated speech. This mirrors Dialect Theory [34-36], which posits that individuals from similar linguistic backgrounds exhibit consistent emotional expression patterns. Similarly, AI-generated speech develops "dialects," making intra-AI classification more reliable.

The lower accuracy in cross-group classifications emphasises AI's challenge in replicating the nuanced variability of human emotional expression, such as pitch modulation and subtle emotional shifts, as noted. While AI captures features like spectral bandwidth and tonal stability, it struggles with the fluid complexity of human speech. For AI to fully replicate human emotional expression, future improvements must integrate the natural variability found in human communication, much like how human dialects encode subtle emotional and cultural cues.

4.3. Interplay of VTL, F0, and Chroma_cqt in conveying vocal confidence

In this study, Vocal Tract Length (VTL), F0 (fundamental frequency), and Chroma_cqt emerge as key features in expressing vocal confidence, reflecting both the physical modulation of speech and its emotional undertones. VTL was positively correlated with confidence, with longer VTL indicating a deeper, more authoritative voice. This elongation of the vocal tract, achieved by adjusting the larynx and resonance chambers, is linked to evolutionary mechanisms where greater body size, or the impression of it, signals dominance [14]. As VTL increases, F0 – the vibration frequency of the vocal cords – decreases, producing a lower-pitched voice often associated with calmness and control, while higher F0 is typically linked to doubt or anxiety. Moreover, VTL showed a positive correlation with Chroma_cqt, a spectral feature that captures tonal brightness. Higher Chroma_cqt values, indicative of higher notes on the musical scale, were linked to confidence, suggesting that confident voices are both deeper and tonally "brighter," combining resonance and tonal stability. This interaction between VTL, F0, and Chroma_cqt reveals the intricate vocal adjustments humans make to express confidence, a pattern that AI models can effectively replicate, though through algorithmic approximations of these physical traits.

4.4. MFCC patterns in confidence levels: Variability and AI limitations

MFCC values were consistently higher in doubtful speech compared to neutral and confident speech across both human and AI-generated data. This reflects the greater vocal variability and complexity associated with doubt, including hesitations and unpredictable tonal shifts. Doubtful speech, being less controlled, leads to more dynamic timbral fluctuations, which MFCC captures effectively. In contrast, lower MFCC values in confident speech suggest a more stable and controlled vocal profile, aligning with the perception of confidence as clear, steady, and deliberate. This difference likely arises because speakers expressing doubt rely on more varied and complex vocal strategies due to emotional instability or uncertainty. Confident speakers, by comparison, maintain a consistent vocal delivery, producing a simpler acoustic profile with fewer fluctuations.

While human data consistently showed a clear pattern of doubtful speech exhibiting the highest MFCC values, followed by neutral and then confident speech, AI-generated speech displayed less variation in certain cases. In some AI models, differences between confidence levels were not as pronounced, suggesting that AI struggles to fully replicate the natural variability present in human speech. This lack of variation indicates that AI models may not be capturing the subtle fluctuations in timbre and speech texture that distinguish emotional states in human speakers. The more uniform MFCC values in AI-generated speech suggest that AI has yet to fully learn the intricate

vocal patterns that convey doubt and confidence in human expression, highlighting an area for further improvement in AI speech synthesis.

4.5. Limitations

While this study provides valuable insights into how human and AI speech convey vocal confidence, several limitations should be addressed. First, most acoustic features analysed, aside from F0, do not directly reflect laryngeal mechanisms crucial for voice production. Future research should incorporate measures of laryngeal function to better understand how vocal adjustments contribute to emotional expression [9, 13, 37].

Second, the study lacks an exploration of how listeners perceive AI versus human speech, particularly how knowledge of the speaker's identity (AI or human) affects social judgments [38]. The impact of phenomena like the uncanny valley, where imperfect human likeness creates discomfort, and the Eliza effect, where human-like behaviours are attributed to AI [39, 40], also remains unexplored.

Additionally, the use of a predefined, non-interactive AI model limits the generalizability of the findings. Future work should investigate how AI can learn to convey emotional cues in real-time, interactive settings. Broader datasets, including various languages and cultural contexts, are also needed to ensure these results apply across different populations.

5. Conclusion

This study examined how vocal confidence is encoded in human and AI-generated speech, highlighting key similarities and differences in acoustic and spectral features. Both human and AI speakers modulated VTL and Chroma_cqt to signal confidence. However, AI systems rely more on spectral features like spectral bandwidth, while human speech emphasised acoustic cues such as Amplitude and RMS, reflecting AI's difficulty in replicating the natural variability of human speech. AI struggled particularly with MFCC, showing less differentiation between confident and neutral speech, unlike the clear pattern of greater vocal variability in human speech. Despite an observed in-group advantage, where machine learning models performed better on data from the same source, AI systems still faced challenges capturing the nuanced emotional shifts in human speech. While AI voice-cloning technology has advanced, it remains limited in fully replicating human vocal variability. Future improvements should focus on enhancing AI's ability to mirror the dynamic complexity of human emotional expression, incorporating more natural acoustic fluctuations and broader datasets across different languages and cultural contexts.

Declarations

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

Data availability statement

Data will be made available on upon request to the authors.

Acknowledgements

We wanted to express our gratitude to Yanbing Hu for his insightful suggestions during our earlier data analysis and visualisation.

Funding

This work was supported by the Natural Science Foundation of China (Grant No. 31971037); the 'Shuguang Programme' supported by the Shanghai Education Development Foundation and Shanghai Municipal Education Committee (Grant No. 20SG31); the Natural Science Foundation of Shanghai (22ZR1460200); the Supervisor Guidance Programme of Shanghai International Studies University (2022113001); and the Major Programme of the National Social Science Foundation of China (Grant No. 18ZDA293).

References

[1] N. Kaur, P. Singh, Conventional and contemporary approaches used in text to speech synthesis: A review, Artif. Intell. Rev. 56(7) (2023) 5837-5880.
[2] B. Schuller, A. Batliner, Computational paralinguistics: emotion, affect and personality in speech and language processing, John Wiley & Sons2013.
[3] X. Jiang, M.D. Pell, Predicting confidence and doubt in accented speakers: Human perception and machine learning experiments, In *Proceedings of Speech Prosody* (2018) *pp. 269-273.*
[4] L. Fortunati, A. Edwards, C. Edwards, A.M. Manganelli, F. de Luca, Is Alexa female, male, or neutral? A cross-national and cross-gender comparison of perceptions of Alexa's gender and status as a communicator, Comput. Hum. Behav. 137 (2022) 107426.
[5] C. Edwards, A. Edwards, B. Stoll, X. Lin, N. Massey, Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions, Comput. Hum. Behav. 90 (2019) 357-362.
[6] W.J. Mitchell, C.-C. Ho, H. Patel, K.F. MacDorman, Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management, Comput. Hum. Behav. 27(1) (2011) 402-412.
[7] J. Kim, K. Merrill Jr, K. Xu, S. Kelly, Perceived credibility of an AI instructor in online education: The role of social presence and voice features, Comput. Hum. Behav. 136 (2022) 107383.
[8] E. Rodero, Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices, Comput. Hum. Behav. 77 (2017) 336-346.
[9] A. Gampe, K. Zahner-Ritter, J.J. Müller, S.R. Schmid, How children speak with their voice assistant Sila depends on what they think about her, Comput. Hum. Behav. 143 (2023) 107693.
[10] X. Jiang, M.D. Pell, The sound of confidence and doubt, Speech Commun. 88 (2017) 106-126.
[11] L. Nagels, E. Gaudrain, D. Vickers, P. Hendriks, D. Başkent, Development of voice perception is dissociated across gender cues in school-age children, Sci. Rep. 10(1) (2020) 1-11.

[12] K. Pisanski, D. Reby, Efficacy in deceptive vocal exaggeration of human body size, Nat. Commun. 12(1) (2021) 1-9.
[13] M. Belyk, S. Waters, E. Kanber, M.E. Miquel, C. McGettigan, Individual differences in vocal size exaggeration, Sci. Rep. 12(1) (2022) 1-12.
[14] K. Pisanski, A. Anikin, D. Reby, Vocal size exaggeration may have contributed to the origins of vocalic complexity, Philos. Trans. R. Soc. B 377(1841) (2022) 20200401.
[15] P.A. Abhang, B.W. Gawali, Correlation of EEG images and speech signals for emotion analysis, Br. J. Appl. Sci. Technol. 10(5) (2015) 1-13.
[16] Y.-P. Huang, R. Mushi, Classification of Cough Sounds Using Spectrogram Methods and a Parallel-Stream One-Dimensional Deep Convolutional Neural Network, IEEE Access 10 (2022) 97089-97100.
[17] T. Koelewijn, E. Gaudrain, T. Tamati, D. Başkent, The effects of lexical content, acoustic and linguistic variability, and vocoding on voice cue perception, J. Acoust. Soc. Am. 150(3) (2021) 1620-1634.
[18] B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, 2015, pp. 18-24.
[19] J.L. Besada, E. Bisesi, C. Guichaoua, M. Andreatta, The Tonnetz at First Sight: Cognitive Issues of Human–Computer Interaction with Pitch Spaces, Music & Science 7 (2024) 20592043241246515.
[20] V. Dellwo, A. Leemann, M.-J. Kolly, Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors, J. Acoust. Soc. Am. 137(3) (2015) 1513-1528.
[21] S. Mathôt, D. Schreij, J. Theeuwes, OpenSesame: An open-source, graphical experiment builder for the social sciences, Behav. Res. Methods 44(2) (2012) 314-324.
[22] A. Kuznetsova, P.B. Brockhoff, R.H.B. Christensen, Package 'lmertest', R Packag. Vers. 2(0) (2015) 734.
[23] R. Lenth, H. Singmann, J. Love, P. Buerkner, M. Herve, Emmeans: Estimated marginal means, aka least-squares means, R Packag. Vers. 1(1) (2018) 3.
[24] S. Olejnik, J. Algina, Measures of effect size for comparative studies: Applications, interpretations, and limitations, Contemp. Educ. Psychol. 25(3) (2000) 241-286.
[25] T. Sakata, N. Ikeda, Y. Ueda, A. Watanabe, Vocal Tract Length Estimation Using Accumulated Means of Formants and Its Effects on Speaker-Normalization, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 1049-1064.
[26] D. Fogerty, L.E. Humes, The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences, J. Acoust. Soc. Am. 131(2) (2012) 1490-1501.
[27] B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, In Proceedings of the 14th python in science conference (2015) pp. 18-24.
[28] X. Chen, Z. Li, S. Setlur, W. Xu, Exploring racial and gender disparities in voice biometrics, Sci. Rep. 12(1) (2022) 1-12.
[29] F. Javanmardi, S.R. Kadiri, P. Alku, A comparison of data augmentation methods in voice pathology detection, Comput. Speech Lang. 83 (2023) 101552.
[30] N. El Boghdady, E. Gaudrain, D. Başkent, Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users?, J. Acoust. Soc. Am. 145(1) (2019) 417-439.
[31] K. Aas, M. Jullum, A. Løland, Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, Artificial Intelligence 298 (2021) 103502.
[32] D.M. Belete, M.D. Huchaiah, Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results, Int. J. Comput. Appl. 44(9) (2022) 875-886.
[33] D.R.R. Smith, R.D. Patterson, The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age, J. Acoust. Soc. Am. 118(5) (2005) 3177-3186.
[34] K.R. Scherer, Vocal affect expression: a review and a model for future research, Psychol. Bull. 99(2) (1986) 143.
[35] P. Laukka, D. Neiberg, H.A. Elfenbein, Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations, Emotion 14(3) (2014) 445.

[36] X. Jiang, S. Paulmann, J. Robin, M.D. Pell, More than accuracy: Nonverbal dialects modulate the time course of vocal emotion recognition across cultures, J. Exp. Psychol. Hum. Percept. Perform. 41(3) (2015) 597.
[37] N. Lavan, P. Rinke, M. Scharinger, The time course of person perception from voices in the brain, PNAS 121(26) (2024) e2318361121.
[38] Y. Mou, K. Xu, The media inequality: Comparing the initial human-human and human-AI social interactions, Comput. Hum. Behav. 72 (2017) 432-440.
[39] M. Mori, K.F. MacDorman, N. Kageki, The Uncanny Valley [From the Field], IEEE Robot. Autom. Mag. 19(2) (2012) 98-100.
[40] S.Y. Kim, B.H. Schmitt, N.M. Thalmann, Eliza in the uncanny valley: Anthropomorphizing consumer robots increases their perceived warmth but decreases liking, Mark. Lett. 30 (2019) 1-12.