Article

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

Wenjun Chen and Xiaoming Jiang *

Posted Date: 13 December 2023

doi: 10.20944/preprints202312.0807.v1

Keywords: Voice cloning; Affective computing; Vocal confidence; Social AI; Human-computer interaction

Article

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

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

[1] Institute of Linguistics, Shanghai International Studies University, Shanghai 201620, China
[2] Key Laboratory of Language Science and Multilingual Intelligence Applications, Shanghai International Studies University, Shanghai 201620, China
* Correspondence: xiaoming.jiang@shisu.edu.cn (X.J.)

Abstract: This study investigated the capability of vocal-identity-cloning Artificial Intelligence (AI) to encode human-specific confident, doubtful, and neutral-intending emotive states. Linear mixed-effects models and machine learning classification with eXtreme Gradient Boosting were employed to examine the underlying acoustic signatures from 2,700 audio clips, comprising of sentences spoken by human speakers and two sets of equivalences (AI-Geography/AI-Trivia, based on the trained text) generated by voice-cloning models designed to clone human speakers' identities. Compared with neutral-intending voice, human speakers lengthened their vocal tract, raised the fundamental frequency, and increased Chroma constant-Q transform when they intended to be confident; An opposite pattern was shown when they intended to be doubtful. The two sets of AI sounds displayed a similar pattern to human speech, suggesting a shared mechanism for encoding vocal expression across sources. The 1,000 times training-testing classification models reported an in-group advantage for AI sources. The algorithms, trained on AI-Geography/AI-Trivia, resulted in higher accuracies when tested within these AI sources than when tested on human audio. All between-source classifications reported above-chance-level (1/3) accuracies. These findings highlighted that voice-cloning AI, the widely used conversational agent, can learn and generate human-specific vocally-expressed confidence.

Keywords: voice cloning; affective computing; vocal confidence; social ai; human-computer interaction

1. Introduction If one invites a native English speaker on a random street in London to pronounce a less-frequently known small village in Norfolk, England, with a confusing pronunciation - *Happisburgh* (*Hayes-bruh*), he/she may find the speaker sounds hesitant due to their lack of knowledge or familiarity. Here, the human speaker can encode his or her intended confidence towards what they say through their tone of voice. However, in most cases, an Artificial Intelligence (AI) speaker (voice assistant/Text-to-Speech (TTS) service such as Apple's Siri that synthesises human-like speeches) simply sounds the word out regardless of their knowledge or familiarity, even if they pronounced it wrong as *Hap-pis-burgh*. TTS composes of a text analysis component and a speech synthesis component, where the speech synthesis component can be built through different models such as WaveNet, Tacotron, or FastSpeech to produce natural-sounding speech that sounds like a human speaker (Arik et al., 2018; Williams & King, 2019; Yang et al., 2022). It is unknown whether (and if so, how) such a human-specific confidence encoding process can be expressed by AI speakers.

In human voice communication, paralinguistic vocal information is essential in encoding a variety of speaker-related information, including both stable traits or identity (e.g. biological sex, age, and personality) and dynamic or short-term states (e.g., emotion) for further decoding by listeners (Schuller & Batliner, 2013). Both human listeners and computational models based on algorithms for paralinguistic features are capable of identifying speakers' short-term states, such as sleepiness (Egas-López & Gosztolya, 2021), intentions and attitudes (Ishi et al., 2008), as well as vocal expression of emotions (Anagnostopoulos et al., 2015; Kaya et al., 2017), and confidence that signals the speaker's feeling of knowing (Jiang & Pell, 2018).

To what extent can AI also encode vocal expression as a human does? How can AI's capability of cloning human beings' acoustic-articulatory mechanisms underlying vocal expressions be quantified? One of the mostly studied synthetic approaches was the AI voice cloning technology that creates personalised voice assistants. To illustrate, the algorithm utilises a speaker encoder network to extract a target speaker's utterance-level embeddings (Shi et al., 2020), which are passed down to a Tacotron 2 network to synthesise speech conditioned on the speaker embeddings that process speaker identity (Arik et al., 2018). Notably, such a cloning process can also copy the original speech prosody; for example, Lux et al. (2022) extracted and normalised fine-grained prosodic features from reference audio and applied them to new voices using a FastSpeech 2 model and an utterance-level speaker embedding; a perceptual study that involved 32 human listeners and a speaker's embedding test (a technique to display the similarity of speaker models before and after the application of the speaker embedding in clusters in a low-dimensional 2D plot), confirmed the possibility of the algorithm to mimic not only speaker identity but also replicate the speech prosody in AI voice cloning.

This study aims to address the above questions by examining both the pattern of encoding vocally-expressed confidence with articulatory and acoustic cues in human target speeches and in AI-generated speeches via voice cloning TTS technology.

1.1. Research gaps that warrant the adoption of voice cloning TTS for studying the encoding mechanisms underlying vocal expressions

Existing human-computer interaction studies on voice cloning AI mostly focus on the human perceptual outcome and their performance during the interaction with the speeches produced by voice cloning TTS, although few examined expressive voices or compared AI speeches with human speeches used to synthesise AI speeches with the TTS. Behavioural studies demonstrated that human listeners displayed high consensus in reporting the biological sex of Amazon's Alexa to be female (Fortunati et al., 2022). Moreover, listeners prefer AI speeches which were synthesised in their native accents as compared with those of different accents. Listeners who judged the voices generated from TTS to be older in age also rated the same voices as more credible (Edwards et al., 2019), and they also believed the AI voice sounding human-like as more credible than those that sounded machine-like (Kim et al., 2022). Neurophysiological studies also showed differential event-related potentials and event-related spectral perturbations towards the onset of the most relative to the least preferred AI voices (Li et al., 2023).

However, these studies did not investigate how human reacts differently to AI and human voices, which another group of studies have attempted. Listeners preferred the human female voice over a speech by AI female speakers (Mullennix et al., 2003). Human participants would exaggerate their preference for human voices when they explicitly knew they were rating between AI and human voices, as compared to their performance in the Implicit Association Task, the response latency test that measures the strength of human's association of human voices with positive and synthesised voices with negative, or vice versa (Mitchell et al., 2011). In these studies, no efforts have been made to match the identity of the human and the AI speakers. Rodero (2017) performed acoustic analyses that ensured the similarity of F0 ranges of human and AI speakers. Human raters perceived both synthetic voices (by Siri and Loquendo) and human-manipulated voices that morphed the original voice using KaleiVoiceCope software as less effective advertisement tool when compared to original human non-manipulated voices (Rodero, 2017). In another study, the same human acted as an 'AI' speaker or spoke with her own voice. Naive children shifted their interactive style to be less active when they were convinced they were playing a game with an 'AI' speaker as compared to the human speaker, even though the 'AI' and the human speaker were the same woman, except she expressed information in a monotone or lively tone (Gampe et al., 2023). These studies attempted to construct comparable human versus AI speaker groups (Gampe et al., 2023; Rodero, 2017); however, variations can still exist in paralinguistic features between speaker groups, including speaker identity or speech prosody, which need more effective experimental control.

To address this gap, this study posits Huawei's *Xiaoyi*, a conversational agent service, an accessible AI voice clone technology that serves the purpose of cloning speaker identity and speech prosody (namely, confidence levels in this study) and aims to provide empirical evidence to attest so. The SV2TTS is constructed by three key components – Encoder, Synthesiser and Vocoder, largely draw upon the Mel spectrogram that can be converted to Mel Frequency Cepstral Coefficients (MFCC) for acoustic analysis (Jia et al., 2018), and such a spectrogram carries not only speaker's vocal identity embeddings but also paralinguistic information (Hossain & Muhammad, 2019; Zhao et al., 2019). It is worth noting that past HCI studies relied upon accessible TTS services that were also using voice-cloning TTS since their used products, such as DECtalk, KaleiVoiceCope, Siri, Loquendo, and Microsoft Mary, were all built upon a specific original human speaker's embeddings. The present study is similar to the past studies in terms of using the produced audio by voice-cloning TTS as research target, but different from the aforementioned HCI studies given that the present study included the audios presented by the target human speakers – where human speakers expressed their vocal confidence.

The perceptual difference between human and AI speakers raised the issue of whether human and AI speakers encode vocal expressions similarly or differently (Gampe et al., 2023; Rodero, 2017). If they do share a similarity, how can acoustic measurements and machine learning prediction provide supporting evidence (Jiang & Pell, 2017)? To address this gap, data-driven computational studies on acoustic features should be introduced to investigate the affective factors in audio by AI and humans (Rodero, 2017; Schuller & Batliner, 2013). In human-human interaction (HHI), Jiang and Pell (2017) prepared prosody-varied human recordings stating the same text in different accent groups (Canadian-English, Quebecois-French, and Australian-English) and performed a computational study using supervised machine learning models that classified the doubtful and confident prosodies through the acoustic measures inputted - mean pitch (F0), variation of pitch, and vocal quality measures (HNR). The computational paralinguistic experiment reported the consensus importance of F0 in categorising audio clips by three groups of accent users and reported an in-group bias of predicting novel vocal expression – training with Canadian-English to testing on Canadian-English generated a higher classification accuracy than training with Canadian-English to testing on Australian-English. Despite the study by Jiang and Pell (2017) adopting few acoustic features and only performing machine learning on confident and doubtful conditions, it still provided a methodological paradigm that encouraged the present study to view AI speakers as accent users in the computational research, thus validating the generalisability of vocal confidence across human and AI speakers.

1.2. Linguistic phonetic cues to differentiate human vocally-expressed confidence

Linguistic phonetic studies demonstrated a profile of suprasegmental and segmental cues for speakers to extend their feeling of knowing in speech, thereby forming speaker confidence. Speaker confidence is a type of emotive communication that signals the transient mental state of the talker's subjective certainty towards the statement they are making based on the concept and/or words they retrieved from their metacognitive judgement; and often, such a level of confidence to convey pragmatic intention is a stable mental state that is independent of speech content (Boduroglu et al., 2014; Mori & Pell, 2019; Nelson & Russell, 2011).

When speakers encode confidence in short sentences in English, the unconfident voice is signalled by a higher F0, mean amplitude and Harmonics-to-Noise Ratio (HNR), a slower speech rate, and more pauses, whereas the confident voice was observed to have a higher amplitude, lower HNR, and greater variations in F0 and amplitude (Jiang & Pell, 2017). Similarly, the acoustic analysis of vowels of the Chinese Wuxi dialect reported a similar F0 modulation subjected to the intended speaker's confidence in the findings in English, where the dialect speakers raised their F0 to express unconfident feelings (Ji et al., 2022). Spectral information, such as the formant peak values, was also relevant. A following study using machine learning classifications with XGBoost (eXtreme Gradient Boosting) further confirmed that the mean F0, F0 variation, and HNR were crucial features for distinguishing perceived confident and doubtful voices across accented and native English speakers (Jiang & Pell, 2018).

In addition to the findings from acoustic analyses on produced vocal sounds, perceptual studies further confirmed that listeners reliably use F0 to represent speaker levels of confidence (Jiang & Pell, 2015; Jiang & Pell, 2017; Monetta et al., 2008). Studies manipulated F0 from auditory sentences to higher or lower with *Praat* showed according to changes in the perceived level of confidence (Goupil et al., 2021; Guyer et al., 2019); see (Guyer et al., 2021). Consistent with the dialect theory of vocal communication (Elfenbein & Ambady, 2002; Jiang & Pell, 2018), these findings suggest that different speaker groups could encode confidence in different vocal dialects but also follow a universal encoding mechanism in human speeches during human-human interaction (Ji et al., 2022; Scherer, 1997). Despite the fact that F0 is considered to reflect a biological modulator of encoding speaker identity (Lavan, Knight, et al., 2019), neither acoustic analyses nor human perceptual experiments have sufficiently addressed the biological significance of F0 modulation in different confidence levels.

1.3. Laryngeal and acoustic features of human vocal expressions

While few studies have directly reported how speakers vary their laryngeal structures when speaking confidently or doubtfully, many showed that speaker identity is reliably represented and distinguished, given how the vocal tract varies in its shape and length (Belin et al., 2004). Studies on speaker identity have consistently reported the role of Vocal Tract Length (VTL) in signalling speakers' biological sex and age (Lavan, Knight, et al., 2019; Meister et al., 2016; Rachman et al., 2022; Smith & Patterson, 2005). The VTL measures the curvilinear distance along the midline of the tract, from the glottis to the intersection, with a line drawn tangentially to the lips, growing from 6 to 8 cm in infancy to 15 cm for females to 18 cm for males in adulthood (Vorperian et al., 2005).

The VTL has been reported to correlate with F0 across speakers (Nagels et al., 2020), and therefore this study suspected that VTL and F0 could also show a reliable range of variation as the speakers' mental state is confident/doubtful/neutral (Boduroglu et al., 2014; Mori & Pell, 2019; Nelson & Russell, 2011). The bridge between the apparent acoustic dimension of VTL and F0 could be the vocal size that signs listeners' perceived quantity of the speakers' anatomical property (Fuller et al., 2014). Specifically, the VTL modulation was widely associated with vocal size deception (for example, vocalisers lengthen their VTL to sound larger to intimidate enemies) in the animal world (Charlton et al., 2008; Pfefferle & Fischer, 2006; Reby & McComb, 2003) as well as in human beings (Belyk et al., 2022; Pisanski & Reby, 2021; Waters et al., 2021). Such evidence leads to an association between laryngeal mechanisms and paralinguistic information in speech communication (Belin et al., 2004). It is thus deduced that humans' vocal size exaggeration ability dates back before human language and could contribute to the origins of the vocalic complexity of language (Pisanski et al., 2022; Pisanski et al., 2016). As such, VTL, this anatomically-related vocal cue that was found to be positively correlated to F0 (Nagels et al., 2020), is suspected of signalling vocal expression as F0 could do, as evidenced in the traditional vocal emotion research (Jiang & Pell, 2017). This study predicts the VTL modulation to be found when human speakers convey paralinguistic signals, such as confidence levels in language communication.

Mel Frequency Cepstral Coefficients (MFCC), which simulate human hearing perception, are measured with the shape of a spectral envelope and represent the short-term power spectrum of voice (X. Chen et al., 2022). MFCC is a frequently-cited parameter for computational paralinguistics, including vocal emotion recognition (Koduru et al., 2020; S. Liu et al., 2021) and musical instrument or genre classification (Bhalke et al., 2016; Friberg et al., 2014). MFCC is effective for speaker identification tasks (Hansen et al., 2017; Tirumala et al., 2017); for example, males have a higher value of one-dimensional MFCC than females (X. Chen et al., 2022). Still, how MFCC specifically contributes to the characterisation of vocal confidence remained unanswered, with limited research having noted that MFCC is an effective tool to depict verbal expressions but not showing how exactly (Hossain & Muhammad, 2019; Zhao et al., 2019).

Chroma-based features, or pitch-class profiles (PCP), are typical representations of the musical octave with 12 varied halftones and represent the relationship between the degree of changes in timbre and the musical aspect of harmony. Chroma Short-time Fourier Transform (STFT), Chroma Constant-Q transform (CQT), and Chroma Energy Normalized (CENS) were reported to be efficient parameters to classify basic vocal emotions, including neutral, calm, happy, sad, angry, fearful, disgusted, and surprised (Alnuaim et al., 2022). With such, this study aims to provide a further understanding of how Chroma-based features contribute to characterising speaker confidence levels.

Root Mean Square energy (RMS) which measures the loudness of the speech signal and is calculated by adding the audio's mean squares of the amplitudes up, was proven to be salient in classifying emotions and thus also considered to be useful (Abhang & Gawali, 2015) as amplitude alone does (Jiang & Pell, 2018). Still, the relative importance of RMS and Amplitude compared with others to signal paralinguistic information remained to be revealed for further studies. Spectrum centroid, which measures the mass centre in the spectrum of a voice and signals speech brightness, is also suspected of playing a role in signalling different levels of confidence (Huang et al., 2019), despite its limited ability to support consistent predictions of music (Schubert, 2004). Considering the homology of music and speech and *Librosa* (https://*Librosa*.org/; Version: 0.9.2) has been reported as a reliable tool to visualise various acoustic features, including the aforementioned Chroma-based features, the present study also extracted a list of extra accessible features through *Librosa* (Er, 2020; McFee et al., 2015). Specifically, the additional features were Spectral Bandwidth that measures the width of a band of frequencies at half the maximum intensity (Abel & Fingscheidt, 2017; Cramer & Huggins, 1958), Spectral Contrast that measures the difference between the peaks and valleys of the spectrum of a speech signal (Leek & Summers, 1996; Nogueira et al., 2016), Spectral Flatness that reflects how much the speech signal resembles white noise (Kim & Stern, 2011; Madhu, 2009), Spectral Rolloff that measures how fast the spectrum of a speech signal decreases with frequency (Chandwadkar & Sutaone; Stolar et al., 2018), Tonnetz (German for 'tone network') that shows the triadic relationships of the perfect fifth and the major third among the 12 pitch classes of the chromatic scale (Milne & Holland, 2016), Zero Crossing Rate (ZCR) that measures how many times the signal changes from positive to zero to negative or from negative to zero to positive (Song et al., 2021), and Utempo, named the static Tempo with a uniform prior, that measures the speed of a musical piece or speech signal, usually expressed in beats per minute (BPM) (Kong et al., 2004).

1.4. The present study

This study aims to characterize human vocal confidence through acoustic features and assess how AI can mimic human-specific vocal confidence, thus constructing three research questions. Firstly, how do acoustic features contribute to depicting human-specific vocal confidence, especially with the laryngeal-related cues? Secondly, can AI-cloned speakers mimic human beings' observed vocal confidence encoding mechanism? If yes, then thirdly, is predicting confidence levels in human and AI speeches across sources viable?

In this study, ten human speakers were invited to produce 30 statements of trivia/geography knowledge neutrally, doubtfully, and confidently. Audios for trivial and geography knowledge statements were further separately utilised for training AI models that replicate speaker identity and confidence prosodies. Hereby, along with human speech, 2,700 Chinese audios from three sources, produced by Humans, AI-Trivia text-based algorithms, and AI-Geography text-based algorithms that read 30 same sentences, were obtained. After extracting a set of 19 acoustic cues, linear mixed-effects models (LMEM) were performed on each of these features per sources of humans and AI. Ten-fold cross-validation XGBoost Classification methods were applied to produce importance scores based on these features (Jiang & Pell, 2018). Further model comparisons were performed to compare accuracies between models trained and tested on human speeches and different sources of AI with 1000 times simulations.

2. Materials and methods This study recruited ten human 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 (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.) 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 univariate analysis with LMEM and multivariate studies with XGBoost to address relevant questions.

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. The study was approved by the Research Ethics Committee of the Institute of Linguistics, Shanghai International Studies University. Participants provided written informed consent prior to the commencement of the experiments.

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 S1) required by *Xiaoyi* Smart Assistance, a conversational agent.

All participants sequentially went through three independent blocks (instructing the portrayal of sentences in neutral, doubtful, and confident tone of voice, respectively), 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 (Jiang & Pell, 2017). 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-portrayed repetition 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 such that they only included the main statements but not the preceding phrases.

2.1.3. 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 (Mathôt et al., 2012), 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 (Kuznetsova et al., 2015), followed by a subsequent post hoc comparison with emmeans when necessary (Lenth et al., 2018). 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 ηp^2 = .01, ηp^2 = .06, and ηp^2 = .14 (Olejnik & Algina, 2000). The LMEMS revealed the main effect of intended confidence (F(2,880)= 5972.24, p<<2e-16, ηp^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, ηp^2 =.02) was found. Post hoc results showed that: (1) under the level of confidence of 95%, rating scores were ranked as confident (6.57±.05) > neutral (4.08±.05) > doubtful (1.39±.05) from high to low; (2) the ratings on the intended confidence level rating were consistency across biological sex; and (3) females rated their doubtful speech lower that males would do (Table 1). These findings validated the perceptual differences between the three intended confidence levels.

Table 1. 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

2.1.4. Two sets of audio generated by AI models

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 * (15 Geography sentences + 15 highly-known Trivia sentences)) were hereby generated, followed by a normalisation manipulation at 70 dB SPL (henceforth Set AI-Geography and Set AI-Geography).

2.2. Data analysis

2.2.1. Acoustic feature extraction

Considering that VTL estimation was typically conducted on vowels or voiced segments (Sakata et al., 2021), this study extracted the voiced parts of all 2,700 audios from the human and AI voice-cloning sources through *Extract Vowels* functions of *Praat Vocal Toolkit* (www.*Praat*vocaltoolkit.com/), generating sets of Text-Grid annotated voiced parts for each audio. The mean VTL for each voiced part was estimated with *Calculate Vocal Tract Length* function of the *Vocal Toolkit* and then averaged for each audio. ΔVTL was calculated by first estimating the Mean VTL of the small voiced parts of each audio (before re-joining) and then subtracting the minimum Mean VTL from the maximum Mean VTL of each audio. Similarly, given that F0 calculation performs better in vowels than consonants (Fogerty & Humes, 2012) 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. ΔF0 was calculated by directly subtracting the minimum from the maximum of the mean F0 of the voiced parts from the audio.

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 Δ/Mean VTL and Δ/Mean F0 from voiced parts, 13 other spectral or beat-related acoustic parameters were extracted from the complete audios through *Librosa* (McFee et al., 2015). 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 one-dimensioned data already. 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 (X. Chen et al., 2022; Oliphant, 2006). 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 (Javanmardi et al., 2022), this study ensured all acoustic 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 also calculated from the complete audios as global prosodic measures using *Praat* (Jiang & Pell, 2018).

Altogether, this study obtained 19 features: VTL and F0, and Δ VTL and Δ F0 from voiced parts, Amplitude and HNR from the complete sentences through *Praat*, and 13 other features from the whole audios through *Librosa*.

2.2.2. Statistical analysis

Experiment 1: Univariate analyses of each acoustic feature's role in predicting vocal confidence

Firstly, an LMEM was built on 'VARIABLE ~ Source * Confidence Levels + Biological Sex + (1| Height) + (1|Item)' to evaluate the main effect and interaction influence of Source and Confidence Levels using R (see Table 2). The 'VARIABLE' was looped through all 19 features. The pairwise contrast (Source and Confidence Levels) results per source and per confidence level were shown in Supplementary Table S2.

In recognising the known impact of Biological Sex in 2.1.3 on the perception of vocal confidence and the existing knowledge of women's and men's biological difference in certain laryngeal features such as VTL (Smith & Patterson, 2005), another set of LMEMs were formulated through 'VARIABLE ~ Confidence Levels * Biological Sex + (1| Height) + (1|Item)' while separating the data according to three sources. The main effect and interaction findings of all features were reported in Table 3. The

doi:10.20944/preprints202312.0807.v1

contrast results per biological sex and per confidence level was shown in Supplementary Table S3 and S4. Further pairwise contrasts using 'Confidence Levels | Biological Sex' and 'Biological Sex | Confidence Levels' were conducted, and the generated *p*-values were annotated in descriptive plotting of each acoustic feature as grouped by confidence levels and biological sex (see Figures 1 and 2 and Supplementary Figure S1, S2, and S3). The corresponding emmeans values comparison were shown in Table 4, and the inferential statistics were shown in Supplementary Table S5 and S6.

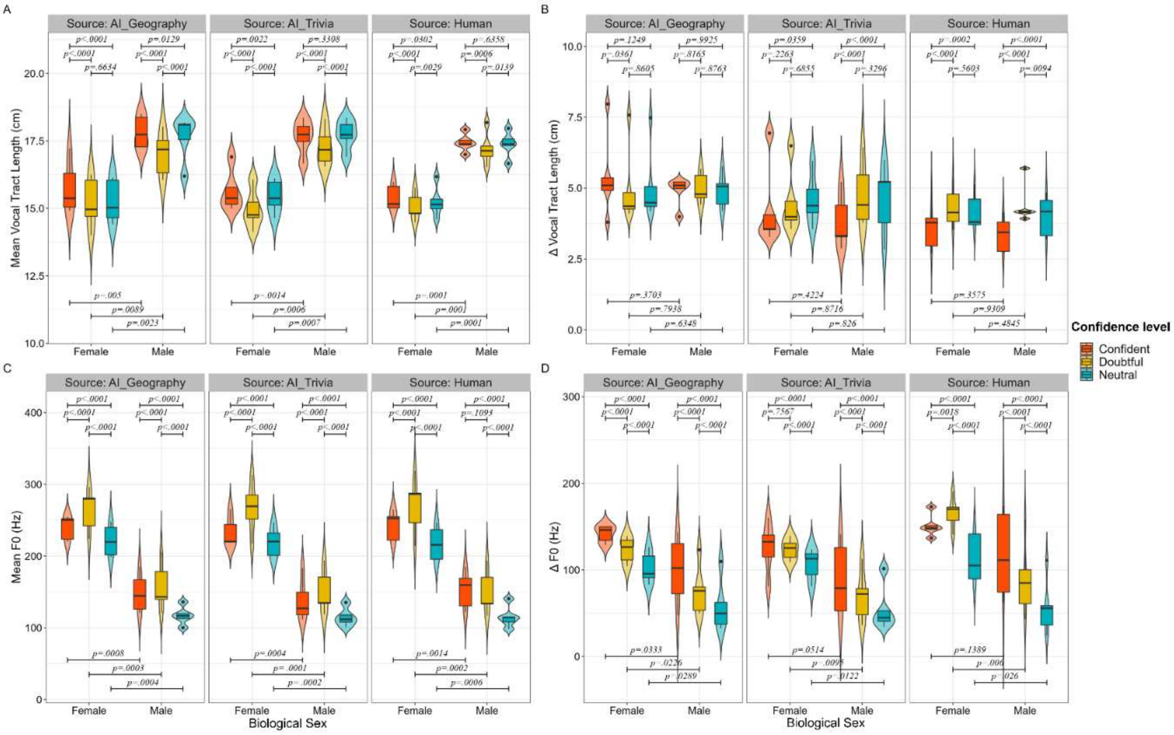

Figure 1. Comparisons in (a) Mean VTL in centimetres (cm), (b) Δ VTL in cm, (c) Mean F0 in Hertz (Hz), and (d) Δ F0 in AI-Geography and AI-Trivia TTS models, and human, differentiated by biological sex and confidence levels.

To verify the known correlation between VTL and F0 (Nagels et al., 2020), as well as to explore the possible relationship between Chroma_cqt (which was reported as the most important parameter to classify vocal confidence in later analysis (Figure 4)) and VTL, the lm(Mean_vtl ~ Mean_F0) was employed (See Figure 3 for the plots).

Demonstrations of two-dimensional MFCC spectrograms were also shown in Supplementary Figure S4. The visualisation of each speaker's height and VTL in three confidence levels were additionally shown in Supplementary Figure S5 (Smith & Patterson, 2005).

Experiment 2: Optimised machine learning classification studies to explore the relative contribution of each acoustic feature

This experiment is to further address the first two questions by informing each acoustic feature's importance in signalling vocal confidence in each source. A machine learning 10-fold cross-validation 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 (Chen & Guestrin, 2016). Seventeen acoustic features were inputted into machine learning to classify vocal confidence by humans and AI. Δ VTL and Δ F0 were not put into machine learning due to reported less importance compared with other features (El Boghdady et al., 2019). Seventeen features were fed into the algorithms, including two extracted from the voiced parts of the audios - VTL and F0 - and the other 15 extracted from the complete original sentences. The other 15 parameters included in the analysis were Mean Amplitude, Mean HNR, Chroma_stft, Chroma_cqt, Chroma_cens, MFCC, Root Mean Square, Spectral Centroid, Spectral Bandwidth, Spectral Contrast, Spectral Flatness, Spectral Rolloff, Tonnetz, ZCR, and Utempo.

The XGBoost model was constructed through the *train()* function. The 10-fold cross-validation was employed to tune the model into a 'reinforced listener' with the *cv()* function. Cross-validation was deployed to prevent overfitting, which occurs when a model is trained too well on the training data and performs poorly on new, unseen data. In this 10-fold cross-validation study, the dataset was split into 10 parts/folds of equal size. Each fold was chosen in turn for testing and the remaining parts for training. This study repeated the training and testing ten times, with each time using a different fold as the testing fold. For each fold, the algorithm utilises the trained model to make predictions on the test set while also calculating the Shapley values for each acoustic feature. Here, the Shapley values calculate the importance of a feature by comparing what a model predicts with and without the feature, thus serving as a function to explain how each feature inputted into the algorithm contributes to the prediction of a certain model. In this study, each fold generated a Shapley value for each of the 17 acoustic features by averaging the Shapley values of that feature over all instances in the test set of that fold. The study then averaged the Shapley values of that feature over all ten folds. Hence, the study gained a more robust estimate of how important a certain feature is for the predictions made by the models trained on different subsets of data. Hereby, the importance score for each acoustic feature was obtained. By plotting the SHAP values using *shap.plots.bar()*, the importance scores for each acoustic parameter were visually presented in Figure 4 (A to C) to report how each feature contributes to classifying vocal confidence by Set Human, Set AI-Trivia, and Set AI-Geography.

Experiment 3: Machine learning training/testing studies to evaluate the predictability

The above-mentioned machine learning constructed 'reinforced listeners' to investigate how each acoustic features contribute to signifying vocal confidence in the speech by human and AI speakers. This study further constructed 1,000 naïve 'non-reinforced listeners' who did not go through any K-fold cross-validation. Each 'non-reinforced listener' had one chance to study a random combination of subsets and then directly perform the classification task. For each classification among the 1,000, the data was split into Training Set (80%) and Testing Set (20%). The study recorded the classification performance, specifically the overall accuracy, accuracy for classifying each confidence condition, root-mean-squared errors (RMSE), and F1-score (macro), which signify the models' performance. RMSE was calculated by taking the square root of the mean of the squared differences between the predicted and actual values, measuring how spreading out the prediction errors are from the true values (Belete & Huchaiah, 2022). F1-score (macro) was calculated by taking the harmonic mean of precision and recall (obtaining F1-score) and then averaging the obtained F1-score of each class, giving equal weights to each class, thus measuring the accuracy of a binary classification model by combining both precision and recall metrics (Belete & Huchaiah, 2022; Q. Chen et al., 2022). Hence, the performance for one 'listener' was available. The study assigned 1,000 randomly generated numbers (within 1 to 100,000) to the argument 'random_seed' of the algorithms, thus generating 1,000 'listeners' who were fed with different Training and Testing sets. The averaged accuracies and model performance of the 1,000 times were demonstrated in Table 5. A representative seed was selected based on the similarity of RMSE and F1-score (macro) of the designed seed and the averaged 1,000 tests, and its corresponding representative receiver operating characteristic (ROC) curve was shown in Figure 4D to demonstrate the visualised in-group advantage.

Additionally, an ANOVA was performed to further attest to the in-group advantage, following 'Overall Accuracy of each iteration in the 1,000 ones ~ Training * Testing'. The in-group advantage can be defined as the increased predictive accuracy of the model trained and tested on the same source set relative to that on different sets. Three levels of data sources were included in Training and Testing variables. The pairwise argument was set as 'pairwise ~ Training * Testing' to test all possible comparisons between the levels of the two variables. The resulting output is reported in Supplementary Table S5 using the *summary()* function.

3. Results*3.1. The effects of confidence levels on acoustic cues in Human and AI sources*

While not separating the sources, the analysis for 'VARIABLE ~ Source * Confidence Levels + Biological Sex + (1| Height) + (1|Item)' revealed significant main effects across all 19 acoustic features at both the source and confidence level dimensions (p<.05), as illustrated in Table 2. For the main effect related to sources, only modest effect sizes were observed (ηp^2 ranging from .01 to .06) for the following features: ΔVTL, ΔF0, Mean VTL, Mean F0, Tonnetz, and Utempo. For the main effect of confidence levels, Tonnetz was manifested at a negligible level (ηp^2<.01), whereas Mean VTL, Spectral Flatness, Δ VTL, Amplitude, and Utempo demonstrated small effect sizes (ηp^2 ranging between .01 and .06) (Note that the small, medium, and large effect size is generally referred to as ηp^2 = .01, ηp^2 = .06, and ηp^2 = .14 (Olejnik & Algina, 2000). The ηp^2 in this study was calculated through the test-statistic approximation method (https://easystats.github.io/effectsize/articles/anovaES.html).).

Table 2. Main Effects and Interaction Results of Confidence Levels and Human/AI Sources on 19 Acoustic Features.

Feature[a]	Parameter[b]	F	*p*	*ηp2*
Mean VTL	S	23.72	<.0001	.02
	CL	169.09	<.0001	.11
	S:CL	12.91	<.0001	.02
Δ VTL	S	152.94	<.0001	.1
	CL	27.85	<.0001	.02
	S:CL	13.99	<.0001	.02
Mean F0	S	23.54	<.0001	.02
	CL	1676.71	<.0001	.56
	S:CL	16.07	<.0001	.02
Δ F0	S	106.39	<.0001	.07
	CL	357.47	<.0001	.21
	S:CL	18.96	<.0001	.03
Chroma_cqt	S	890.34	<.0001	.4
	CL	1563.93	<.0001	.54
	S:CL	12	<.0001	.02
Chroma_cens	S	593.21	<.0001	.31
	CL	906.45	<.0001	.41
	S:CL	11.45	<.0001	.02
Chroma_stft	S	4944.14	<.0001	.79
	CL	1320.71	<.0001	.5
	S:CL	21.59	<.0001	.03
MFCC	S	1900.18	<.0001	.59
	CL	738.66	<.0001	.36
	S:CL	91.88	<.0001	.12
Spectral Contrast	S	2289.89	<.0001	.63
	CL	645.33	<.0001	.33
	S:CL	8.06	<.0001	.01
Spectral Bandwidth	S	39003.19	<.0001	.97
	CL	255.01	<.0001	.16

	S:CL	52.24	<.0001	.07
Root Mean Square	S	465.29	<.0001	.26
	CL	720.99	<.0001	.35
	S:CL	8.49	<.0001	.01
Amplitude	S	6614.12	<.0001	.83
	CL	12.2	<.0001	.01
	S:CL	35.08	<.0001	.05
Tonnetz	S	21.7	<.0001	.02
	CL	5.07	.01	0
	S:CL	3.71	.01	.01
Spectral Flatness	S	8068.02	<.0001	.86
	CL	76.46	<.0001	.05
	S:CL	23.7	<.0001	.03
Spectral Centroid	S	13001.55	<.0001	.91
	CL	485.93	<.0001	.27
	S:CL	9.79	<.0001	.01
Spectral Rolloff	S	15655.64	<.0001	.92
	CL	507.01	<.0001	.28
	S:CL	16.67	<.0001	.02
ZCR	S	353.08	<.0001	.21
	CL	312.98	<.0001	.19
	S:CL	16.78	<.0001	.02
Utempo	S	25.22	<.0001	.02
	CL	7.36	<.0001	.01
	S:CL	2.93	.02	0
HNR	S	1242.46	<.0001	.48
	CL	1919.14	<.0001	.59
	S:CL	11.51	<.0001	.02

[a] All statements are mean values unless otherwise indicated. [b] CL for Confidence Levels; BS for Biological Sex; CL*BS for the interaction between CL and BS.

Table 3. Main Effects and Interaction Results of Confidence Levels and Human/AI Sources on 19 Acoustic Features.

Feature[a]	Parameter[b]	AI-Geography			AI-Trivia			Human		
		F value	*p*	ηp^2	F value	*p*	ηp^2	F value	*p*	ηp^2
Mean VTL	CL	97.45	<.0001	.19	93.06	<.0001	.18	23.45	<.0001	.05
	BS	16.64	.0047	.7	30.75	.0009	.81	59.09	.0001	.89
	CL:BS	15.34	<.0001	.03	8.34	.0003	.02	1.24	.2909	.0
Δ VTL	CL	1.14	.3194	.0	25.84	<.0001	.06	41.81	<.0001	.09
	BS	.34	.5784	.05	.09	.768	.01	.39	.5531	.05
	CL:BS	2.47	.0855	.01	8.53	.0002	.02	1.86	.1566	.0
Mean F0	CL	940.44	<.0001	.69	573.07	<.0001	.57	586.33	<.0001	.58
	BS	38.42	.0004	.85	46.98	.0002	.87	35.7	.0006	.84

	CL:BS	37.8	<.0001	.08	38.39	<.0001	.08	70.52	<.0001	.14
Δ F0	CL	187.32	<.0001	.3	92.45	<.0001	.18	146.48	<.0001	.25
	BS	7.65	.0279	.52	9.44	.018	.57	7.76	.0271	.53
	CL:BS	.52	.5961	.0	12.18	<.0001	.03	23.89	<.0001	.05
Chroma_cqt	CL	799.11	<.0001	.65	699.05	<.0001	.62	485.46	<.0001	.53
	BS	28.35	.0011	.8	17.23	.0043	.71	13.67	.0077	.66
	CL:BS	107.62	<.0001	.2	88.07	<.0001	.17	37.17	<.0001	.08
Chroma_cens	CL	471.02	<.0001	.52	333.36	<.0001	.44	297.6	<.0001	.41
	BS	4.22	.0793	.38	2.49	.1585	.26	1.68	.2359	.19
	CL:BS	57.4	<.0001	.12	65.7	<.0001	.13	33.71	<.0001	.07
Chroma_stft	CL	793.22	<.0001	.65	860.15	<.0001	.67	269.23	<.0001	.39
	BS	30.04	.0009	.81	48.89	.0002	.87	27.4	.0012	.8
	CL:BS	103.92	<.0001	.2	116.46	<.0001	.21	28.17	<.0001	.06
MFCC	CL	317.99	<.0001	.43	178.32	<.0001	.29	485.33	<.0001	.53
	BS	17.22	.0043	.71	16.08	.0051	.7	9.93	.0161	.59
	CL:BS	18.8	<.0001	.04	18.99	<.0001	.04	12.3	<.0001	.03
Spectral Contrast	CL	372.68	<.0001	.46	299.31	<.0001	.41	201.5	<.0001	.32
	BS	19.07	.0033	.73	43.37	.0003	.86	10.89	.0131	.61
	CL:BS	134.05	<.0001	.24	84.07	<.0001	.16	46.32	<.0001	.1
Spectral Bandwidth	CL	482.9	<.0001	.53	317.25	<.0001	.43	8.95	.0001	.02
	BS	.47	.5137	.06	2.81	.1379	.29	.32	.5866	.04
	CL:BS	6.97	.001	.02	30.98	<.0001	.07	29.89	<.0001	.07
Root Mean Square	CL	318.07	<.0001	.43	232.09	<.0001	.35	299.36	<.0001	.41
	BS	16.55	.0047	.7	10.15	.0154	.59	2.49	.1584	.26
	CL:BS	15.43	<.0001	.03	22.48	<.0001	.05	9.67	<.0001	.02
Amplitude	CL	43.66	<.0001	.09	25.67	<.0001	.06	337.11	<.0001	.44
	BS	1.18	.3138	.15	.19	.6737	.03	1.56	.2512	.18
	CL:BS	5.07	.0064	.01	.09	.9105	.0	8.26	.0003	.02
Tonnetz	CL	5.92	.0028	.01	5.45	.0044	.01	.27	.7642	.0
	BS	.08	.7914	.01	1.88	.2121	.21	.06	.818	.01
	CL:BS	2.05	.129	.0	4.4	.0125	.01	1.18	.3084	.0
Spectral Flatness	CL	112.77	<.0001	.21	66.99	<.0001	.14	43.03	<.0001	.09
	BS	.51	.4984	.07	.99	.3527	.12	6.96	.0337	.5
	CL:BS	13.8	<.0001	.03	1.93	.1454	.0	17.33	<.0001	.04
Spectral Centroid	CL	348.6	<.0001	.45	238.86	<.0001	.36	169.73	<.0001	.28
	BS	3.21	.1164	.31	2.84	.1357	.29	10.01	.0159	.59
	CL:BS	11.66	<.0001	.03	28.17	<.0001	.06	12.93	<.0001	.03
Spectral Rolloff	CL	411.26	<.0001	.49	267.66	<.0001	.38	113.0	<.0001	.21
	BS	1.14	.322	.14	1.51	.2586	.18	4.37	.0749	.39
	CL:BS	14.41	<.0001	.03	21.84	<.0001	.05	18.66	<.0001	.04
ZCR	CL	207.19	<.0001	.33	136.84	<.0001	.24	173.61	<.0001	.29
	BS	3.87	.0898	.36	.19	.6798	.03	6.08	.0431	.46
	CL:BS	18.38	<.0001	.04	74.3	<.0001	.15	35.97	<.0001	.08
Utempo	CL	5.34	.005	.01	7.35	.0007	.02	2.38	.0928	.01
	BS	2.32	.1752	.26	4.41	.0754	.4	.31	.5962	.04
	CL:BS	1.06	.3484	.0	2.4	.0912	.01	3.7	.0252	.01
HNR	CL	996.63	<.0001	.7	749.12	<.0001	.64	696.65	<.0001	.62
	BS	24.93	.0016	.78	22.75	.002	.77	9.61	.0174	.58
	CL:BS	94.25	<.0001	.18	72.06	<.0001	.14	24.49	<.0001	.05

doi:10.20944/preprints202312.0807.v1

[a] All statements are mean values unless otherwise indicated. [b] CL for Confidence Levels; BS for Biological Sex; CL*BS for the interaction between CL and BS.

Regarding interactions, no large effect sizes (ηp^2>.14) were detected for any parameter. A majority of the acoustic features, including Amplitude, Δ F0, Chroma_stft, Spectral Flatness, Mean VTL, Δ VTL, Mean F0, Chroma_cqt, Chroma_cens, Spectral Rolloff, ZCR, and HNR, exhibited modest effect sizes (ηp^2 ranging between .01 and .06). Meanwhile, Spectral Contrast, Root Mean Square, Tonnetz, Spectral Centroid, and Utempo were found to only negligible effect sizes (ηp^2<.01).

The left side of Supplementary Table S2 presented the results of the pairwise analysis concerning the sources. In a comparative analysis between two subsets of AI speech sources, no significant differences (p>.10) were observed in Mean VTL, Chroma_cqt, Spectral Contrast, and Root Mean Square. Moreover, Chroma_stft and HNR revealed only marginally significant differences (.05<p<.10). In contrast, significant differences (p<.05) across all acoustic parameters were observed when contrasting clips generated by AI-Geography with human speech or audio produced by AI-Trivia with human speech.

The right side of Supplementary Table S2 outlines the results of the pairwise analysis related to confidence levels. When contrasting the confident with the doubtful speech, significant differences (p<.05) were observed across all features. However, when confident and neutral speech were contrasted, only Amplitude and Utempo failed to show significant differences (p>.05), whereas the remaining 17 features did. When comparing doubtful and neutral speech, no significant differences (p>.05) were found for ΔVTL, Spectral Bandwidth, Tonnetz, Spectral Centroid, Spectral Rolloff, and Utempo, while significant differences were observed for the other features.

Supplementary Table S3 delineates a comparative analysis across three confidence levels (C for Confident, D for Doubtful, and N for Neutral) derived from three sources. Within the context of AI-Geography, non-significance (p>.05) was observed in eight scenarios: Δ VTL (C – D, C – N, D – N), Amplitude (C – N), Tonnetz (C – D, C – N), and ZCR (D-N). In the AI-Trivia subset, non-significance (p > .05) was found in nine conditions: Mean VTL (C - N), Δ VTL (D - N), MFCC (C - D), Spectral Bandwidth (D - N), Tonnetz (C – N, D - N), Spectral Centroid (D - N), Spectral Rolloff (D - N), and Utempo (C - N). In the human subset, eight instances did not meet the threshold of significance (p>.05): Spectral Bandwidth (C - D), Tonnetz (C – D, C – N, D - N), and Utempo (C – D, C – N, D - N). All remaining conditions across the three sources demonstrated significant variations.

Both sources and confidence levels showed significant main and interaction effects across all 19 acoustic features. However, relatively small main effects were observed in Mean F0 and Mean VTL that signal biological modulation. And, still, nuances of non-significances were observed in the pairwise results.

3.2. Similar effect of biological sex and its interaction with confidence levels between human and AI sources

This analysis engaged three different datasets separately from AI-Geography, AI-Trivia, and Human sets. The aim was twofold: first, to describe the mechanisms of vocal confidence, and second, to probe the potential of AI in mimicking such mechanisms. Key acoustic features were Mean VTL, Mean F0, Chroma_cqt, Chroma_cens, Amplitude, and MFCC. These were chosen based on their significant scores in the predictive model trained to characterise human vocal confidence (Figure 4A) and their biological relevance to encoding the vocal expression.

Table 3 presents both the main effect and the interaction of biological sex and confidence levels. For the main effect of biological sex, Chroma_cens and Amplitude did not exert a significant effect over the acoustic features. However, the other four did present a significant main effect. The above patterns for the six features were consistently observed in AI-Geography, AI-Trivia, and Human. Regarding the main effect of confidence levels, all parameters demonstrated main effects across all three sources. In the case of interaction between biological sex and confidence levels, non-significant interaction effects in Amplitude of AI-Trivia (p=.91) and Mean VTL of Human (p=.29) were noted. Aside from these two conditions for the six parameters, all other conditions demonstrated interaction effects.

A subsequent pairwise analysis was performed on confidence levels by subtracting the values of males from that of females (Supplementary Table S4). The results revealed significant differences in Mean VTL (AIg: -2.1; AIt: -2.23; Human: -2.17), Mean F0 (97.43; 103.56; 100.3), Chroma_cqt (-.05; -.06; -.04), and MFCC (-6.53; -6.47; -7.03) across AI-Geography, AI-Trivia, and Human sources, respectively. Inferential statistical analyses comparing the F – M values across three confidence levels and the C – D/C – N/D – N values across two biological sexes are shown in Supplementary Tables S5 and 6, respectively. The associated significance is annotated in Figures 1 and 2.

To distinguish between the three confidence levels, estimated marginal means (emmeans), which account for all factors in the model, and the confidence intervals were shown in Table 4. Firstly, when considering Mean VTL, for males, the ranking order for the vocal expression was as follows: Confident > Neutral > Doubtful in both the AI-Geography and Human datasets, while in the AI-Trivial dataset, the ranking was Neutral > Confident > Doubtful. For females, however, the ranking remained consistent as Confident > Neutral > Doubtful across all three data sources. Secondly, as for Mean F0, males exhibited a ranking of Doubtful > Confident > Neutral in the AI-Geography and AI-Trivia datasets, whereas the Human dataset showed a different trend with Confident > Doubtful > Neutral. For females, the ranking of Doubtful > Confident > Neutral remained the same across all three sources. Thirdly, when observing Chroma_cqt, the pattern was uniform for both sexes across all three data sources, with a ranking of Confident > Neutral > Doubtful. Fourth, for MFCC, the male ranking across all sources was consistently Neutral > Doubtful > Confident. The female ranking, on the other hand, followed a pattern of Neutral > Confident > Doubtful in the AI-Geography and AI-Trivia datasets but showed a different order of Neutral > Doubtful > Confident in the Human dataset.

Due to the similar rankings and biological importance of Mean VTL, Mean F0, and Chroma_cqt, correlation studies were performed. The findings 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 (Figure 3).

Humans lengthened their vocal tract when they were confident, resulting in higher Chroma_cqt and lower F0, and shortened it when they were doubtful, causing the opposite effects; this pattern was validated through human/AI speaker sources and male/female biological sexes. Likewise, other important features (see 3.3) showed agreement across biological sexes and sources.

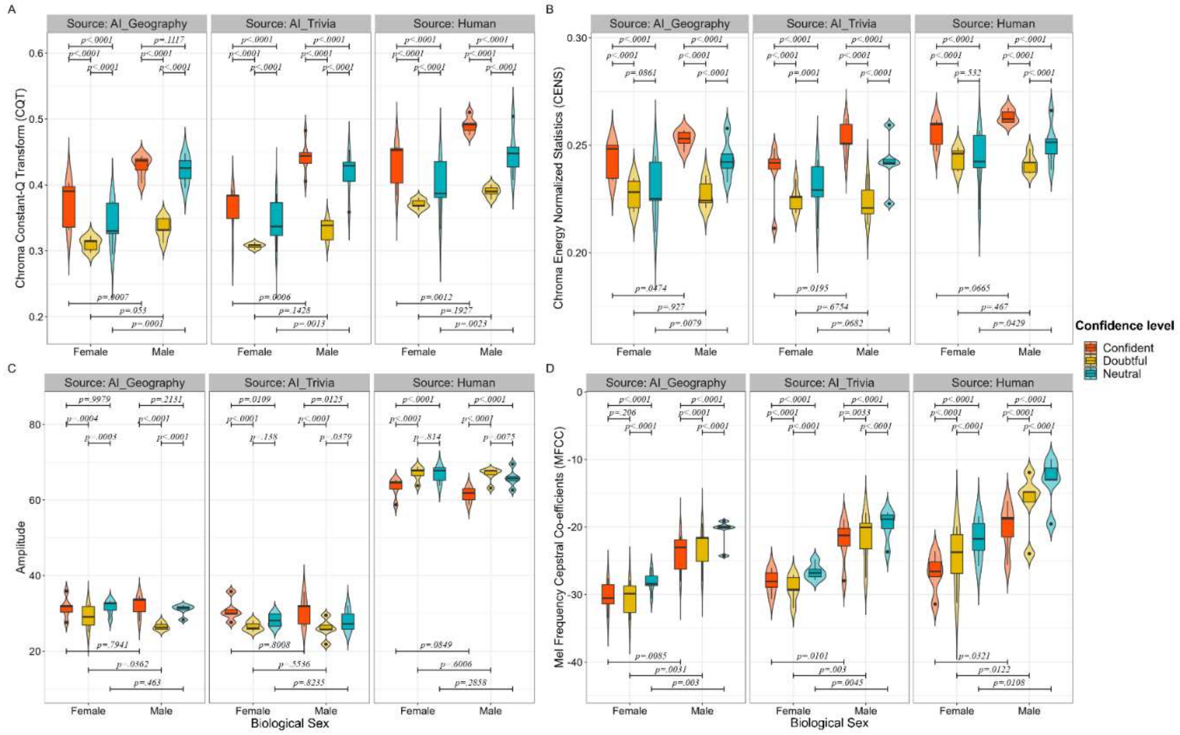

Figure 2. Comparisons in (a) Chroma_cqt, (b) Chroma_cens, (c) Amplitude in decibels (dB), and (d) MFCC features in AI-Geography and AI-Trivia TTS models, and human, differentiated by biological

sex and confidence levels.

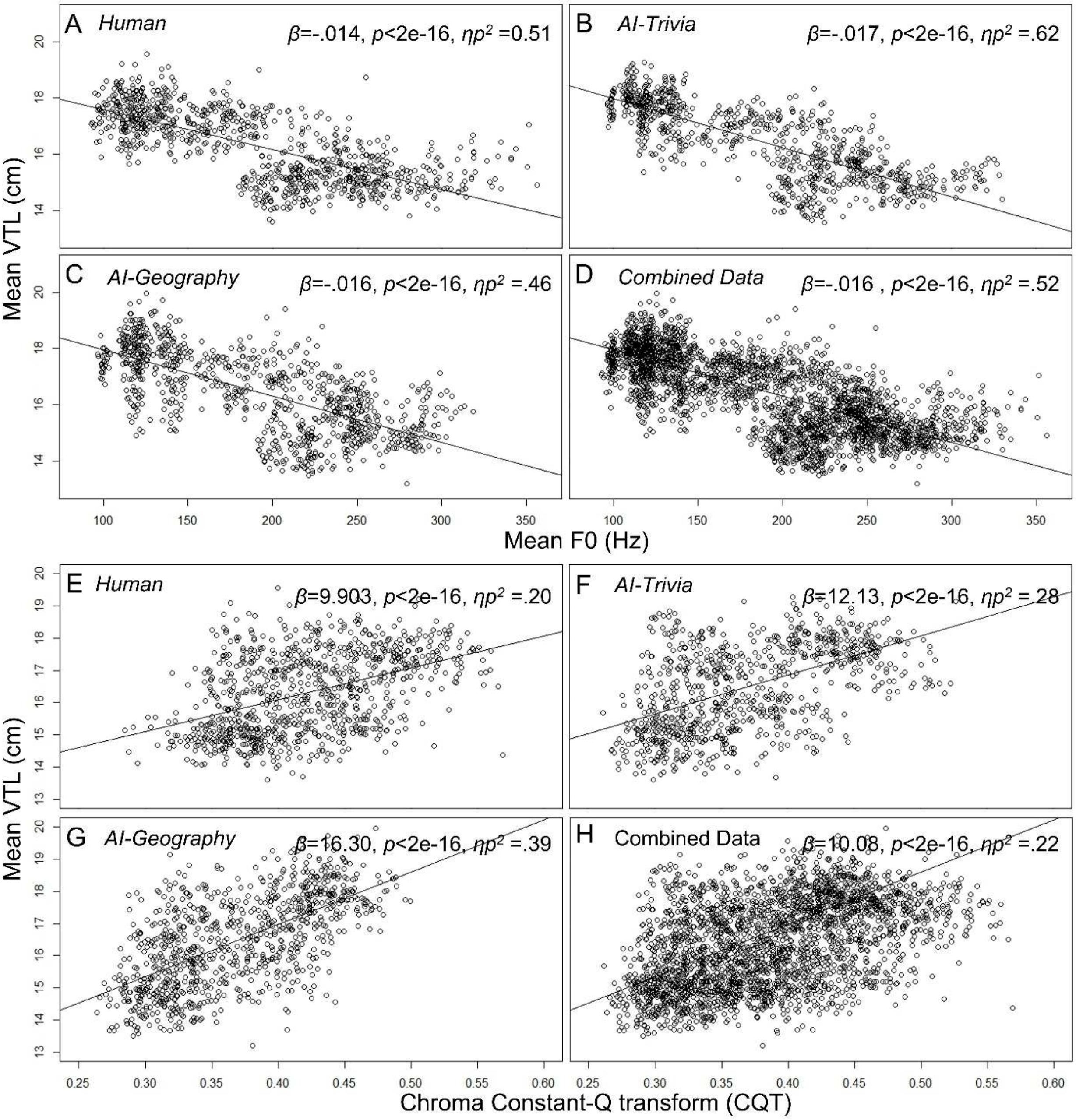

Figure 3. Correlation plots between Mean VTL (cm) and Mean F0 are shown in (a)-(e), with each point representing an individual speaker's average values across three confidence levels. The correlation between VTL and Chroma_cqt is shown in (f)-(h), with each point representing a single sentence from the 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.

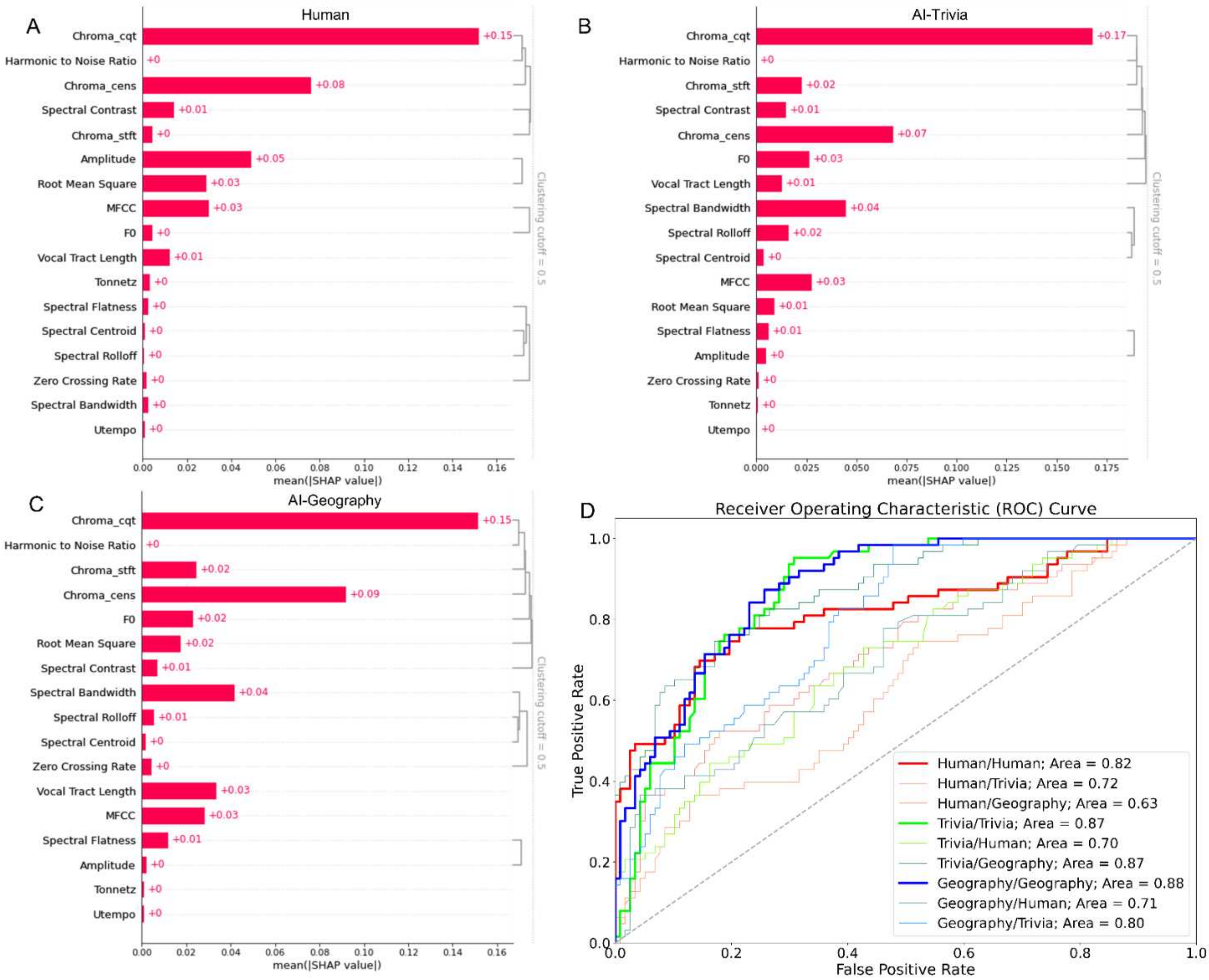

Figure 4. Importance scores for 17 features reported by vocal confidence differentiating algorithms based on 10-fold cross-validation. The datasets used were (a) Human, (b) AI-Trivia, and (c) AI-Geography. As features (e.g., RMS energy and Amplitude) could be correlated with each other, a clustering 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.

Table 4. Main Effects and Interaction Results of Confidence Levels and Human/AI Sources on 19 Acoustic Features.

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
Mean VTL	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]
	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]
Δ VTL	C	F	5.6	[4.33,6.87]	4.46	[3.14,5.77]	3.77	[3.07,4.48]
	C	M	4.9	[3.76,6.04]	3.82	[2.64,5]	3.38	[2.74,4.01]
	D	F	5.2	[3.93,6.47]	4.67	[3.36,5.99]	4.46	[3.75,5.16]

	D	M	5.0	[3.86,6.14]	4.8	[3.62,5.98]	4.42	[3.79,5.06]
	N	F	5.28	[4.01,6.55]	4.78	[3.46,6.1]	4.32	[3.61,5.02]
	N	M	4.92	[3.78,6.06]	4.61	[3.43,5.79]	4.02	[3.38,4.66]
Mean F0	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]
	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]
Δ F0	C	F	140.62	[112.93,168.31]	127.69	[99.56,155.82]	150.37	[115.5,185.24]
	C	M	99.24	[74.44,124.04]	90.27	[65.08,115.45]	117.22	[85.98,148.45]
	D	F	122.03	[94.34,149.72]	125.65	[97.52,153.78]	165.47	[130.61,20.34]
	D	M	76.48	[51.68,101.28]	69.59	[44.41,94.78]	89.43	[58.19,120.66]
	N	F	101.28	[73.59,128.97]	108.08	[79.95,136.21]	112.67	[77.8,147.54]
	N	M	58.37	[33.57,83.17]	54.84	[29.65,80.03]	57.24	[26,88.47]
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]
Chroma_cens	C	F	.24	[.24,.25]	.24	[.23,.25]	.26	[.25,.26]
	C	M	.25	[.25,.26]	.25	[.24,.26]	.26	[.26,.27]
	D	F	.23	[.22,.23]	.23	[.22,.23]	.24	[.24,.25]
	D	M	.23	[.22,.23]	.22	[.22,.23]	.24	[.24,.25]
	N	F	.23	[.22,.24]	.23	[.22,.24]	.24	[.24,.25]
	N	M	.24	[.24,.25]	.24	[.23,.25]	.25	[.25,.26]
MFCC	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]
	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]
Amplitude	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]
	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]

[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.

3.3. The important features signalling vocal confidence of humans and AI

Seven key audio features showed high importance scores for accurately classifying confidence levels in human audio: Chroma_cqt, Chroma_cens, Root Mean Square, MFCC, Spectral Contrast, Vocal Tract Length, and Amplitude (Figure 4A).

However, in the case of AI-Trivia, only six of these features were found to be important, with Amplitude not contributing significantly to classification accuracy. In addition, Spectral Bandwidth, F0, Chroma_stft, Spectral Rolloff, and Spectral Flatness were also identified as important features for AI-Trivia classification (Figure 4B).

For AI-Geography, Amplitude was found to be of no importance, while Spectral Bandwidth, Chroma_stft, F0, Spectral Rolloff, and Spectral Flatness showed greater contribution than in human audios (Figure 4C).

Therefore, the importance scores for Amplitude and additional acoustic features were similar in both AI models, despite some scattered differences in the values.

3.4. Training and predicting vocal confidence across sources

The 1,000 times averaged results in Table 5 witnessed two in-group advantages. Firstly, as expected, the models that were trained and tested on their respective data demonstrated the highest overall accuracy (Jiang & Pell, 2018). For instance, H/H achieved an accuracy of .72, while the accuracy of AIg/H and AIt/H was reduced to .51 and .38, respectively. Likewise, H/H achieved an accuracy of .72, while H/AIg and H/AIt achieved accuracies of .54 and .53, respectively. Secondly, yet most importantly, when AI models were tested on one another's data, their overall accuracies were higher than when tested on human data. For example, AIg/AIt had an accuracy of .67, AIt/AIg had an accuracy of .69, while that of AIg/H and AIt/H was .51 and .38. All accuracy levels were above the chance level (1/3). The ROC curve analysis in Figure 4D also demonstrated such in-group advantage.

The ANOVA analysis of 'Overall Accuracy ~ 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 S7 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 study confirmed that models trained and tested on their respective data showed higher accuracy levels, and AI models generally performed better when tested on each other's data than when tested on human data, hence the in-group advantage. Still, the above-chance-level accuracies when training and testing across humans and AI suggested AI's robust capacity to replicate human-specific vocal confidence.

Table 5. Machine Learning Classification Accuracies from 1,000 Iterations, 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	.40	.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	.80	.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).

4. Discussion*4.1. Characterising human vocal confidence through VTL*

The mean VTL was reported to encode speaker confidence, with the confident voice showing the longest VTL, followed by neutrally-intending and the doubtful voice the shortest. Meanwhile, longer and shorter Mean VTL was associated with vocal modulation of vocal tract length, with lengthening the VTL with the aim of displaying a larger body size (Anikin et al., 2022). Hence, the confident sound is described as a state where human speakers extend their vocal tract causing a lower Mean F0 to sound more dominant (Puts et al., 2007). Studies have noted the importance of Mean F0 and Mean VTL that characterise 'who is talking' or speaker identity (Lavan, Knight, et al., 2019). The current study contributed to the argument that speaker identity and long-term traits, and short-term states, such as speaker emotions, are intertwined (Belin et al., 2004; Lavan, Burton, et al., 2019; Mileva & Lavan, 2023; Schuller & Batliner, 2013; Sorokowski et al., 2019).

Moreover, the mean VTL was positively correlated with Chroma_cqt and negatively correlated with Mean F0, which followed robust values ranking of Confident > Neutral > Doubtful. Chroma_cqt represents the twelve different pitch classes in the speech signal, which correspond to the notes C, C#, D, D#, E, F, F#, G, G#, A, A#, and B from lowest to highest in the Western music scale (Huang & Mushi, 2022). Higher Chroma_cqt suggests the speech sample was closer to the higher note in the music scale. To produce a confident speech, the speaker could increase Chroma_cqt to sound brighter (Collier & Hubbard, 2004).

These findings based on the vocal expression portrayed by Chinese speakers expanded findings in the English context (Jiang & Pell, 2017, 2018). Despite not taking VTL into account, the previous studies did suggest that the unconfident voice showed a higher Mean F0, consistent with the lower Mean VTL reported in the current study. Vocal size exaggeration has been associated with the evolution of the human speech-oral-motor system and cited as a common ability across species, and the human speaker has incorporated this intuitive capacity of vocal anatomy-sound coordination into the vocal communication system (Anikin et al., 2021; Pisanski et al., 2022; Pisanski & Reby, 2021). The current study, therefore, strengthened empirical evidence from a cross-linguistic perspective and displayed a more informative depiction of the vocal modulation mechanism underlying speaker expression through the analysis of acoustic cues signalling the anatomical structure.

4.2. AI speakers can imitate human-specific vocal confidence

The present study demonstrated that an AI algorithm designed to clone speaker identity could also mimic human vocal confidence levels. The importance scores showed that two sets of AI data, akin to human audio, relied on the same range of acoustic features, from Chroma_cqt to VTL in a list of seven cues, to encode vocal confidence. Such a mutual utilisation of important acoustic features could lead to the above-chance level classification accuracies observed when training and testing across different data sources, as well as similar value rankings in Chroma_cqt, VTL, and MFCC.

As compared with human speakers, the AI models exhibited a greater reliance on additional features, such as Spectral Bandwidth and F0, which were not deemed important when encoding human confidence. It is possible that, due to these additional features involved in creating a multivariate pattern of representation, AI performed even higher accuracies when training and testing within AI sources than when training or testing on human data.

AI's strong already cloning ability in HCI is in line with the dialect theory in HHI. The Dialect theory about human communication highlighted that individuals from different cultural backgrounds or group identities share similar patterns of expressing emotions (Scherer, 1997), despite variations caused by culture- or group-specific norms (e.g., mother tongue). Such encoding rules within and between speaker cultures and groups have been supported by perceptual studies which showed common and differential neural responses of decoding vocally-expressed confidence and doubt in native and accented speakers (Jiang et al., 2015; Laukka et al., 2014; Pell et al., 2009; Scherer, 1997). Systematically modulating the anatomical structure of the larynx allowed humans to produce different types of speech that conveyed their varied levels of 'feelings of knowing'. Here, AI is proven to be capable of replicating such peculiar speech differences while even cloning the speaker-specific identity signalling 'who is talking' by simulating a speech-motor/laryngeal control of the vocal box in humans to serve the purpose of communicating certain pragmatic meanings that can be 'perceived'. The AI's capability to learn and replicate human-specific speaker identity and emotive vocal states could pose an empirical threat to the modern daily activities that heavily demand HCI, where the internet connects computers that play sounds out unceasingly. These sounds may contain speech signals that are either authentic or faked and may vary in tone and emotion, thus making it difficult for human listeners to discern the speaker's group identity being human or AI. This could raise a realistic concern that should motivate the advancement of institutional regulations on speech synthesis. For instance, to counter the voice deepfake, some legislation such as *Defending Each and Every Person from False Appearances by Keeping Exploitation Subject (DEEP FAKES) to Accountability Act* has been approved, which requires deep fake creators to mark their content with an indelible digital watermark (Langa, 2021).

4.3. Implications and limitations

This study demonstrated how voice-cloning service meant for human speaker identity cloning also captures and replicates vocal confidence across affective levels. Such capacity is important because a wide range of research has attested that the human brain distinctively responds to individuals with different individual speaker identities (Kroczek & Gunter, 2021; Puhacheuskaya & Järvikivi, 2022). The *human-machine co-behaviour* theories have identified an emerging trend to examine the long-term dynamics of hybrid systems and the ways that human social interactions could be modified by the introduction of intelligent machines (Rahwan et al., 2019). By demonstrating AI's capacity to clone the acoustic encoding of vocally-expressed confidence, the current study serves as an interface that updates studies on the synthesis of affective speech from an engineering perspective and affective computing with decoding approaches (Gunes et al., 2019; Gunes et al., 2011; Habib et al., 2019). The current findings also provide a basis for recognising communicative meanings from alternative modalities through psychological and neurophysiological data analysis (Cross & Ramsey, 2021; Di Cesare et al., 2022; Kuriki et al., 2016; Li et al., 2023; Nummenmaa et al., 2023; Saarimäki et al., 2022; Tamura et al., 2015). In existing HCI studies, AI as a social agent, does not have equivalent emotional intelligence (EQ) as humans, and human counterparts can immediately detect its group identity (Mou et al., 2019). However, future researches on HCI with AI, powered by large language models such as ChatGPT and affective speech synthesis technology, can be challenged due to the higher human likeness marked by emotive states like vocal confidence and perceived personalities (2021); and such challenges are expected to influence human's performance in HCI scenarios in various settings such as cooperation, competition, coordination, learning and communication, e.g., self-driving cars should adopt assertive and dominant synthesis voices to ensure responsible driving from drivers (Wong et al., 2019; Yoo et al., 2022).

Still, this study could suffer some limitations, which can be addressed in further studies. On the one hand, how speakers sharing individual identities but with distinct group identities (here determined by the speaker source of AI or human) could pose an impact on the perceptual responses of human listeners remains unanswered – how does 'knowing the speaker to be AI/human' influence social judgment (Chen et al., 2023; Gampe et al., 2023; Mou & Xu, 2017). Possible scenarios include customer service, education, health care and entertainment, where the tone, emotion, feedback, encouragement, empathy, trustworthiness, humour and personality of the agent could influence the social judgment of the human listeners in a variety of dimensions such as satisfaction, loyalty, motivation, learning, well-being, adherence, enjoyment and engagement (Baird et al., 2018; Canbek & Mutlu, 2016; Chattaraman et al., 2019; Hu et al., 2022; McLean et al., 2021; Reicherts et al., 2022; Rodero & Lucas, 2021).

Furthermore, the current study serves as an interface for clinical trials involving special populations. For example, behavioural data reported children with autism spectrum disorder (ASD) to have a weak ability to perceive psychological attributes of humanness in singing (Kuriki et al., 2016); and such have raised proposals to construct an emotional speech database in Mexican Spanish for future studies in ASD (Duville et al., 2021). However, it is noteworthy that Duville et al. (2021) indeed attempted to construct human versus AI comparable emotive speech sets, yet, their AI voices were obtained through modifying human speech, which could distort speaker identity encoded in the speech; the current study employed voice-cloning techniques to ensure the similarity of both emotive states and speaker identity and have provided validation evidence accordingly. With such foundations, future studies can investigate the behavioural and neural mechanisms of decoding emotions and the speaker's individual and group identity in vocal speeches in abnormal populations (Cha et al., 2021; Chevallier et al., 2011; Frijia et al., 2021; Golan et al., 2006; Järvinen-Pasley et al., 2008; Jones et al., 2009), thus informing fields such as automatic diagnosis through possibly interaction tasks (Baki et al., 2022; Cummins et al., 2020; Siddi et al., 2023), or neural mechanism investigation (Gallagher et al., 2000; M. Liu et al., 2021).

On the other hand, how vocal expression conveying pragmatic intentions in AI speech affects the perceived group identity of human-like AI – how does paralinguistic and linguistic information influence the judgment of a speaker's group identity, e.g., human-like (Jiang & Pell, 2015; Ko et al., 2023; Lee, 2010; Melo et al., 2023; Pelachaud, 2017). The current study showcased AI's strong ability to learn and mimic human vocal confidence when providing text-to-speech services, and such AI-generated speeches were found to be comparable to the original human-specific vocal confidence. These features could make emotive AI-produced audios potentially confuse human listeners regarding the human-ness of the speech, thus leading to questions such as the uncanny valley and Eliza effect. As perceptual effects, the uncanny valley is a phenomenon where people feel uneasy or repulsed by human-like robots or animations that are not quite realistic enough (Mori et al., 2012), while the Eliza effect is a tendency to attribute the anthropomorphism in emotions and intentions to artificial intelligence systems that mimic human speeches or behaviours (Kim et al., 2019).

Importantly, it should be acknowledged that the AI speech in this study acquired vocal confidence expression in a non-interactive task using a pre-defined model. Future research could examine how AI can learn to encode human emotional cues directly from interactive tasks and how this would influence the perception and cognition of human participants (Gampe et al., 2023; Nasir et al., 2022; Salam et al., 2023).

5. Conclusions This study illuminates the capacity of the AI voice cloning tool to also replicate human vocal confidence. The use of a similar set of important acoustic features characterising vocal confidence, especially through the parameters related to laryngeal control, highlights the noteworthy mimicking ability of AI. Collectively, this work underscores the emerging complexities and opportunities at the intersection of voice-cloning AI and computer communication, advocating for a more nuanced examination of how humans and conversational agents interact with and influence each other from psychological, neurological, and engineering perspectives.

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).

Data Availability Statement: The data supporting the findings of this research can be made available by the corresponding authors upon a reasonable request.

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

Conflicts of Interest: The authors declare no conflict of interest.

ReferencesAbel, J., & Fingscheidt, T. (2017). Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(1), 71-83.

2. Abhang, P. A., & Gawali, B. W. (2015). Correlation of EEG images and speech signals for emotion analysis. *British Journal of Applied Science & Technology*, *10*(5), 1-13.
3. Alnuaim, A. A., Zakariah, M., Shukla, P. K., Alhadlaq, A., Hatamleh, W. A., Tarazi, H., Sureshbabu, R., & Ratna, R. (2022). Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier. *Journal of Healthcare Engineering*, *2022*.
4. Anagnostopoulos, C.-N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, *43*(2), 155-177.
5. Anikin, A., Pisanski, K., Massenet, M., & Reby, D. (2021). Harsh is large: nonlinear vocal phenomena lower voice pitch and exaggerate body size. *Proceedings of the Royal Society B*, *288*(1954), 20210872.
6. Anikin, A., Pisanski, K., & Reby, D. (2022). Static and dynamic formant scaling conveys body size and aggression. *Royal Society Open Science*, *9*(1), 211496.
7. Arik, S., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural voice cloning with a few samples. *Advances in neural information processing systems*, *31*.
8. Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummins, N., & Schuller, B. (2018). The perception and analysis of the likeability and human likeness of synthesized speech. In Interspeech (pp. 2863-2867). .
9. Baki, P., Kaya, H., Çiftçi, E., Güleç, H., & Salah, A. A. (2022). Speech analysis for automatic mania assessment in bipolar disorder. *arXiv preprint arXiv:2202.06766*.
10. Belete, D. M., & Huchaiah, M. D. (2022). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, *44*(9), 875-886.
11. Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, *8*(3), 129-135.
12. Belyk, M., Waters, S., Kanber, E., Miquel, M. E., & McGettigan, C. (2022). Individual differences in vocal size exaggeration. *Scientific reports*, *12*(1), 1-12.
13. Bhalke, D. G., Rao, C. B., & Bormane, D. S. (2016). Automatic musical instrument classification using fractional fourier transform based-MFCC features and counter propagation neural network. *Journal of Intelligent Information Systems*, *46*(3), 425-446.
14. Boduroglu, A., Tekcan, A. İ., & Kapucu, A. (2014). The relationship between executive functions, episodic feeling-of-knowing and confidence judgements. *Journal of Cognitive Psychology*, *26*(3), 333-345.
15. Canbek, N. G., & Mutlu, M. E. (2016). On the track of artificial intelligence: Learning with intelligent personal assistants. *Journal of Human Sciences*, *13*(1), 592-601.
16. Cha, I., Kim, S.-I., Hong, H., Yoo, H., & Lim, Y.-k. (2021). Exploring the use of a voice-based conversational agent to empower adolescents with autism spectrum disorder. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-15).
17. Chandwadkar, D. M., & Sutaone, M. S. (2012). Role of features and classifiers on accuracy of identification of musical instruments. In *2012 2nd National Conference on Computational Intelligence and Signal Processing (CISP)* (pp. 66-70). IEEE.
18. Charlton, B. D., Reby, D., & McComb, K. (2008). Effect of combined source (F 0) and filter (formant) variation on red deer hind responses to male roars. *The Journal of the Acoustical Society of America*, *123*(5), 2936-2943.
19. Chattaraman, V., Kwon, W.-S., Gilbert, J. E., & Ross, K. (2019). Should AI-Based, conversational digital assistants employ social-or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior*, *90*, 315-330.
20. Chen, Q., Allot, A., Leaman, R., Islamaj, R., Du, J., Fang, L., Wang, K., Xu, S., Zhang, Y., & Bagherzadeh, P. (2022). Multi-label classification for biomedical literature: an overview of the BioCreative VII LitCovid Track for COVID-19 literature topic annotations. *Database*, *2022*.
21. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
22. Chen, W., Hu, Y., & Jiang, X. (2023). A Social Psycholinguistics Perspective: Encoding and Decoding Mechanisms for Speakers' Individual and Group Identities. *Journal of Foreign Languages*.
23. Chen, X., Li, Z., Setlur, S., & Xu, W. (2022). Exploring racial and gender disparities in voice biometrics. *Scientific reports*, *12*(1), 1-12.
24. Chevallier, C., Noveck, I., Happé, F., & Wilson, D. (2011). What's in a voice? Prosody as a test case for the Theory of Mind account of autism. *Neuropsychologia*, *49*(3), 507-517.
25. Collier, W. G., & Hubbard, T. L. (2004). Musical scales and brightness evaluations: Effects of pitch, direction, and scale mode. *Musicae Scientiae*, *8*(2), 151-173.
26. Cramer, E. M., & Huggins, W. H. (1958). Creation of pitch through binaural interaction. *The Journal of the Acoustical Society of America*, *30*(5), 413-417.
27. Cross, E. S., & Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human–machine interactions. *Trends in cognitive sciences*, *25*(3), 200-212.

28. Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V. S., Christensen, H., Blackburn, D., Schuller, B. W., Magimai-Doss, M., & Strik, H. (2020, 2020). A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition. In *Interspeech 2020* (pp. 2182-2186).
29. Di Cesare, G., Cuccio, V., Marchi, M., Sciutti, A., & Rizzolatti, G. (2022). Communicative and affective components in processing auditory vitality forms: An fMRI study. *Cerebral Cortex*, *32*(5), 909-918.
30. Duville, M. M., Alonso-Valerdi, L. M., & Ibarra-Zarate, D. I. (2021). Electroencephalographic Correlate of Mexican Spanish Emotional Speech Processing in Autism Spectrum Disorder: To a Social Story and Robot-Based Intervention. *Frontiers in Human Neuroscience*, *15*, 626146.
31. 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.
32. Egas-López, J. V., & Gosztolya, G. (2021, 6-11 June 2021). Deep Neural Network Embeddings for the Estimation of the Degree of Sleepiness. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7288-7292). IEEE.
33. El Boghdady, N., Gaudrain, E., & Başkent, D. (2019). Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users? *The Journal of the Acoustical Society of America*, *145*(1), 417-439.
34. Elfenbein, H. A., & Ambady, N. (2002). Is there an in-group advantage in emotion recognition? *Psychological bulletin*, *128*, 243-249.
35. Fogerty, D., & Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, *131*(2), 1490-1501.
36. Fortunati, L., Edwards, A., Edwards, C., Manganelli, A. M., & de Luca, F. (2022). Is Alexa female, male, or neutral? A cross-national and cross-gender comparison of perceptions of Alexa's gender and status as a communicator. *Computers in Human Behavior*, *137*, 107426.
37. Friberg, A., Schoonderwaldt, E., Hedblad, A., Fabiani, M., & Elowsson, A. (2014). Using listener-based perceptual features as intermediate representations in music information retrieval. *The Journal of the Acoustical Society of America*, *136*(4), 1951-1963.
38. Frijia, E. M., Billing, A., Lloyd-Fox, S., Rosas, E. V., Collins-Jones, L., Crespo-Llado, M. M., Amadó, M. P., Austin, T., Edwards, A., & Dunne, L. (2021). Functional imaging of the developing brain with wearable high-density diffuse optical tomography: a new benchmark for infant neuroimaging outside the scanner environment. *Neuroimage*, *225*, 117490.
39. Fuller, C. D., Gaudrain, E., Clarke, J. N., Galvin, J. J., Fu, Q.-J., Free, R. H., & Başkent, D. (2014). Gender categorization is abnormal in cochlear implant users. *Journal of the Association for Research in Otolaryngology*, *15*(6), 1037-1048.
40. Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind'in verbal and nonverbal tasks. *Neuropsychologia*, *38*(1), 11-21.
41. 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.
42. Golan, O., Baron-Cohen, S., & Hill, J. (2006). The Cambridge mindreading (CAM) face-voice battery: Testing complex emotion recognition in adults with and without Asperger syndrome. *Journal of autism and developmental disorders*, *36*, 169-183.
43. 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), 1-17.
44. Gunes, H., Celiktutan, O., & Sariyanidi, E. (2019). Live human–robot interactive public demonstrations with automatic emotion and personality prediction. *Philosophical Transactions of the Royal Society B*, *374*(1771), 20180026.
45. Gunes, H., Nicolaou, M. A., & Pantic, M. (2011). Continuous analysis of affect from voice and face. In A. A. Salah & T. Gevers (Eds.), *Computer Analysis of Human Behavior* (pp. 255-291). Springer.
46. Guyer, J. J., Briñol, P., Vaughan-Johnston, T. I., Fabrigar, L. R., Moreno, L., & Petty, R. E. (2021). Paralinguistic features communicated through voice can affect appraisals of confidence and evaluative judgments. *Journal of Nonverbal Behavior*, *45*(4), 479-504.
47. Guyer, J. J., Fabrigar, L. R., & Vaughan-Johnston, T. I. (2019). Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion. *Personality and Social Psychology Bulletin*, *45*(3), 389-405.
48. Habib, R., Mariooryad, S., Shannon, M., Battenberg, E., Skerry-Ryan, R. J., Stanton, D., Kao, D., & Bagby, T. (2019). Semi-supervised generative modeling for controllable speech synthesis. *arXiv preprint arXiv:1910.01709*.

49. Hansen, J. H. L., Nandwana, M. K., & Shokouhi, N. (2017). Analysis of human scream and its impact on text-independent speaker verification. *The Journal of the Acoustical Society of America*, *141*(4), 2957-2967.
50. Hossain, M. S., & Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, *49*, 69-78.
51. Hu, P., Lu, Y., & Wang, B. (2022). Experiencing power over AI: The fit effect of perceived power and desire for power on consumers' choice for voice shopping. *Computers in Human Behavior*, *128*, 107091.
52. Huang, Y.-P., & Mushi, R. (2022). Classification of Cough Sounds Using Spectrogram Methods and a Parallel-Stream One-Dimensional Deep Convolutional Neural Network. *IEEE Access*, *10*, 97089-97100.
53. Huang, Y., Tian, K., Wu, A., & Zhang, G. (2019). Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. *Journal of Ambient Intelligence and Humanized Computing*, *10*(5), 1787-1798.
54. Ishi, C. T., Ishiguro, H., & Hagita, N. (2008). Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication*, *50*(6), 531-543.
55. Järvinen-Pasley, A., Wallace, G. L., Ramus, F., Happé, F., & Heaton, P. (2008). Enhanced perceptual processing of speech in autism. *Developmental Science*, *11*(1), 109-121.
56. Javanmardi, F., Kadiri, S. R., Kodali, M., & Alku, P. (2022). Comparing 1-dimensional and 2-dimensional spectral feature representations in voice pathology detection using machine learning and deep learning classifiers. In *Interspeech*. International Speech Communication Association.
57. Ji, Y., Hu, Y., & Jiang, X. (2022). Segmental and suprasegmental encoding of speaker confidence in Wuxi dialect vowels. *Frontiers in psychology*, *13*.
58. 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*.
59. Jiang, X. (2021). Perceptual Attributes of Human-Like Animal Stickers as Nonverbal Cues Encoding Social Expressions in Virtual Communication. In J. Xiaoming (Ed.), *Types of Nonverbal Communication* (pp. Ch. 5). IntechOpen.
60. Jiang, X., Paulmann, S., Robin, J., & Pell, M. D. (2015). More than accuracy: Nonverbal dialects modulate the time course of vocal emotion recognition across cultures. *Journal of experimental psychology: human perception and performance*, *41*(3), 597.
61. Jiang, X., & Pell, M. D. (2015). On how the brain decodes vocal cues about speaker confidence. *Cortex*, *66*, 9-34.
62. Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, *88*, 106-126.
63. Jiang, X., & Pell, M. D. (2018). Predicting confidence and doubt in accented speakers: Human perception and machine learning experiments. In *Proceedings of Speech Prosody* (pp. 269-273).
64. Jones, C. R. G., Happé, F., Baird, G., Simonoff, E., Marsden, A. J. S., Tregay, J., Phillips, R. J., Goswami, U., Thomson, J. M., & Charman, T. (2009). Auditory discrimination and auditory sensory behaviours in autism spectrum disorders. *Neuropsychologia*, *47*(13), 2850-2858.
65. Kaya, H., Salah, A. A., Karpov, A., Frolova, O., Grigorev, A., & Lyakso, E. (2017). Emotion, age, and gender classification in children’s speech by humans and machines. *Computer Speech & Language*, *46*, 268-283.
66. Kim, J., Merrill Jr, K., Xu, K., & Kelly, S. (2022). Perceived credibility of an AI instructor in online education: The role of social presence and voice features. *Computers in Human Behavior*, *136*, 107383.
67. Kim, S. Y., Schmitt, B. H., & Thalmann, N. M. (2019). Eliza in the uncanny valley: Anthropomorphizing consumer robots increases their perceived warmth but decreases liking. *Marketing letters*, *30*, 1-12.
68. Kim, W., & Stern, R. M. (2011). Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise. *Speech Communication*, *53*(1), 1-11.
69. Ko, S., Barnes, J., Dong, J., Park, C., Howard, A., & Jeon, M. (2023). The effects of robot voices and appearances on users emotion recognition and subjective perception. *International Journal of Humanoid Robotics*.
70. Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology*, *23*(1), 45-55.
71. Kong, Y.-Y., Cruz, R., Jones, J. A., & Zeng, F.-G. (2004). Music perception with temporal cues in acoustic and electric hearing. *Ear and hearing*, *25*(2), 173-185.
72. Kroczek, L. O. H., & Gunter, T. C. (2021). The time course of speaker-specific language processing. *Cortex*, *141*, 311-321.
73. Kuriki, S., Tamura, Y., Igarashi, M., Kato, N., & Nakano, T. (2016). Similar impressions of humanness for human and artificial singing voices in autism spectrum disorders. *Cognition*, *153*, 1-5.
74. Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package ‘lmertest’. *R package version*, *2*(0), 734.
75. Langa, J. (2021). Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *BUL Rev.*, *101*, 761.

76. 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.
77. Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic bulletin & review*, *26*(1), 90-102.
78. Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications*, *10*(1), 2404.
79. Lee, E.-J. (2010). The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. *Computers in Human Behavior*, *26*(4), 665-672.
80. Leek, M. R., & Summers, V. (1996). Reduced frequency selectivity and the preservation of spectral contrast in noise. *The Journal of the Acoustical Society of America*, *100*(3), 1796-1806.
81. Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R package version*, *1*(1), 3.
82. Li, M., Guo, F., Wang, X., Chen, J., & Ham, J. (2023). Effects of robot gaze and voice human-likeness on users' subjective perception, visual attention, and cerebral activity in voice conversations. *Computers in Human Behavior*, *141*, 107645.
83. Liu, M., Li, B., & Hu, D. (2021). Autism spectrum disorder studies using fMRI data and machine learning: a review. *Frontiers in Neuroscience*, *15*, 697870.
84. Liu, S., Zhang, M., Fang, M., Zhao, J., Hou, K., & Hung, C.-C. (2021). Speech emotion recognition based on transfer learning from the FaceNet framework. *The Journal of the Acoustical Society of America*, *149*(2), 1338-1345.
85. Lux, F., Koch, J., & Vu, N. T. (2022). Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)* (pp. 962-969). IEEE.
86. Madhu, N. (2009). Note on measures for spectral flatness. *Electronics letters*, *45*(23), 1195-1196.
87. 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.
88. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, 2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (Vol. 8, pp. 18-25).
89. McLean, G., Osei-Frimpong, K., & Barhorst, J. (2021). Alexa, do voice assistants influence consumer brand engagement?–Examining the role of AI powered voice assistants in influencing consumer brand engagement. *Journal of Business Research*, *124*, 312-328.
90. Meister, H., Fürsen, K., Streicher, B., Lang-Roth, R., & Walger, M. (2016). The use of voice cues for speaker gender recognition in cochlear implant recipients. *Journal of Speech, Language, and Hearing Research*, *59*(3), 546-556.
91. Melo, C. M. d., Gratch, J., Marsella, S., & Pelachaud, C. (2023). Social Functions of Machine Emotional Expressions. *Proceedings of the IEEE*, 1-16.
92. Mileva, M., & Lavan, N. (2023). Trait impressions from voices are formed rapidly within 400 ms of exposure. *Journal of Experimental Psychology: General*.
93. Milne, A. J., & Holland, S. (2016). Empirically testing Tonnetz, voice-leading, and spectral models of perceived triadic distance. *Journal of Mathematics and Music*, *10*(1), 59-85.
94. Mitchell, W. J., Ho, C.-C., Patel, H., & MacDorman, K. F. (2011). Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. *Computers in Human Behavior*, *27*(1), 402-412.
95. Monetta, L., Cheang, H. S., & Pell, M. D. (2008). Understanding speaker attitudes from prosody by adults with Parkinson's disease. *Journal of neuropsychology*, *2*(2), 415-430.
96. Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, *19*(2), 98-100.
97. Mori, Y., & Pell, M. D. (2019). The look of (un) confidence: visual markers for inferring speaker confidence in speech. *Frontiers in Communication*, *4*, 63.
98. Mou, W., Gunes, H., & Patras, I. (2019). Alone versus in-a-group: A multi-modal framework for automatic affect recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *15*(2), 1-23.
99. Mou, Y., & Xu, K. (2017). The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior*, *72*, 432-440.
100. 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.
101. Nagels, L., Gaudrain, E., Vickers, D., Hendriks, P., & Başkent, D. (2020). Development of voice perception is dissociated across gender cues in school-age children. *Scientific reports*, *10*(1), 1-11.
102. Nasir, J., Bruno, B., Chetouani, M., & Dillenbourg, P. (2022). A Speech-based Productive Engagement Metric for Real-time Human-Robot Interaction in Collaborative Educational Contexts. *IEEE transactions on affective computing*.

103. Nelson, N. L., & Russell, J. A. (2011). When dynamic, the head and face alone can express pride. *Emotion*, *11*(4), 990.
104. Nogueira, W., Rode, T., & Büchner, A. (2016). Spectral contrast enhancement improves speech intelligibility in noise for cochlear implants. *The Journal of the Acoustical Society of America*, *139*(2), 728-739.
105. Nummenmaa, L., Malèn, T., Nazari-Farsani, S., Seppälä, K., Sun, L., Santavirta, S., Karlsson, H. K., Hudson, M., Hirvonen, J., & Sams, M. (2023). Decoding brain basis of laughter and crying in natural scenes. *Neuroimage*, *273*, 120082.
106. Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary educational psychology*, *25*(3), 241-286.
107. Oliphant, T. E. (2006). *A guide to NumPy* (Vol. 1). Trelgol Publishing USA.
108. Pelachaud, C. (2017). Conversing with Social Agents That Smile and Laugh. In *INTERSPEECH* (p. 2052).
109. Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, *37*(4), 417-435.
110. Pfefferle, D., & Fischer, J. (2006). Sounds and size: identification of acoustic variables that reflect body size in hamadryas baboons, Papio hamadryas. *Animal behaviour*, *72*(1), 43-51.
111. 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.
112. Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice modulation: a window into the origins of human vocal control? *Trends in cognitive sciences*, *20*(4), 304-318.
113. Pisanski, K., & Reby, D. (2021). Efficacy in deceptive vocal exaggeration of human body size. *Nature Communications*, *12*(1), 1-9.
114. Puhacheuskaya, V., & Järvikivi, J. (2022). I was being sarcastic!: The effect of foreign accent and political ideology on irony (mis) understanding. *Acta Psychologica*, *222*, 103479.
115. Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. C. (2007). Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), 340-344.
116. Rachman, L., Jebens, A., & Baskent, D. (2022). Phonological but not lexical processing alters the perceptual weighting of mean fundamental frequency and vocal-tract length cues for voice gender categorisation. *The Journal of the Acoustical Society of America*, *151*(4), A262-A262.
117. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., Shariff, A., Tenenbaum, J. B., & Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477-486.
118. Reby, D., & McComb, K. (2003). Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. *Animal behaviour*, *65*(3), 519-530.
119. Reicherts, L., Rogers, Y., Capra, L., Wood, E., Duong, T. D., & Sebire, N. (2022). It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks. *ACM Transactions on Computer-Human Interaction*, *29*(3), 1-41.
120. 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.
121. Rodero, E., & Lucas, I. (2021). Synthetic versus human voices in audiobooks: The human emotional intimacy effect. *new media & society*, 14614448211024142.
122. Saarimäki, H., Glerean, E., Smirnov, D., Mynttinen, H., Jääskeläinen, I. P., Sams, M., & Nummenmaa, L. (2022). Classification of emotion categories based on functional connectivity patterns of the human brain. *Neuroimage*, *247*, 118800.
123. Sakata, T., Ikeda, N., Ueda, Y., & Watanabe, A. (2021). Vocal Tract Length Estimation Using Accumulated Means of Formants and Its Effects on Speaker-Normalization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1049-1064.
124. Salam, H., Celiktutan, O., Gunes, H., & Chetouani, M. (2023). Automatic Context-Aware Inference of Engagement in HMI: A Survey. *IEEE transactions on affective computing*, 1-20.
125. Scherer, K. R. (1997). The role of culture in emotion-antecedent appraisal. *Journal of personality and social psychology*, *73*(5), 902.
126. Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music perception*, *21*(4), 561-585.
127. Schuller, B., & Batliner, A. (2013). Computational paralinguistics: emotion, affect and personality in speech and language processing. John Wiley & Sons.
128. Shi, Y., Huang, Q., & Hain, T. (2020). H-vectors: Utterance-level speaker embedding using a hierarchical attention model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7579-7583). IEEE.

129. Siddi, S., Bailon, R., Giné-Vázquez, I., Matcham, F., Lamers, F., Kontaxis, S., Laporta, E., Garcia, E., Lombardini, F., & Annas, P. (2023). The usability of daytime and night-time heart rate dynamics as digital biomarkers of depression severity. *Psychological medicine*, 1-12.

130. Smith, D. R. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, *118*(5), 3177-3186.

131. Song, X., Qiao, X., Hao, D., Yang, L., Zhou, X., Xu, Y., & Zheng, D. (2021). Automatic recognition of uterine contractions with electrohysterogram signals based on the zero-crossing rate. *Scientific reports*, *11*(1), 1-10.

132. 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.

133. Stolar, M. N., Lech, M., Stolar, S. J., & Allen, N. B. (2018). Detection of adolescent depression from speech using optimised spectral roll-off parameters. *Biomedical Journal*, *2*, 10.

134. Tamura, Y., Kuriki, S., & Nakano, T. (2015). Involvement of the left insula in the ecological validity of the human voice. *Scientific reports*, *5*(1), 8799.

135. Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, *90*, 250-271.

136. Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, *117*(1), 338-350.

137. Waters, S., Kanber, E., Lavan, N., Belyk, M., Carey, D., Cartei, V., Lally, C., Miquel, M., & McGettigan, C. (2021). Singers show enhanced performance and neural representation of vocal imitation. *Philosophical Transactions of the Royal Society B*, *376*(1840), 20200399.

138. Williams, J., & King, S. (2019, 2019). Disentangling Style Factors from Speaker Representations. In *Interspeech* (pp. 3945-3949).

139. Wong, P. N. Y., Brumby, D. P., Babu, H. V. R., & Kobayashi, K. (2019). Voices in Self-Driving Cars Should be Assertive to More Quickly Grab a Distracted Driver's Attention. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 165-176).

140. Yang, Z., Wu, Z., & Jia, J. (2022). Speaker Characteristics Guided Speech Synthesis. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

141. Yoo, Y., Yang, M.-y., Lee, S., Baek, H., & Kim, J. (2022). The effect of the dominance of an in-vehicle agent's voice on driver situation awareness, emotion regulation, and trust: A simulated lab study of manual and automated driving. *Transportation research part F: traffic psychology and behaviour*, *86*, 33-47.

142. Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control*, *47*, 312-323.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.