



Perception and social evaluation of cloned and recorded voices: Effects of familiarity and self-relevance

Victor Rosi^{*} , Emma Soopramanien[✉] , Carolyn McGettigan

Department of Speech, Hearing and Phonetic Sciences, University College London, UK

ARTICLE INFO

Keywords:

Voice cloning
Voice identity
Social evaluation
First impressions

ABSTRACT

Modern speech technologies enable the artificial replication, or cloning, of the human voice. In the present study, we investigated whether listeners' perception and social evaluation of state-of-the-art voice clones depend on whether the clone being heard is a replica of the self, a friend, or a total stranger. We recorded and cloned the voices of familiar pairs of adult participants. Forty-seven of these experimental participants (and 47 unfamiliar controls) rated the Trustworthiness, Attractiveness, Competence, and Dominance of cloned and recorded samples of their own voice and their friend's voice. We observed that while familiar listeners found clones to sound less (or similarly) trustworthy, attractive, and competent than recordings, unfamiliar listeners showed an opposing profile in which clones tended to be rated higher than recordings. Within this, familiar listeners tended to prefer their friend's voice to their own, although perceived similarity of both self- and friend-voice clones to the original speaker identity predicted higher ratings on all trait scales. Overall, we find that familiar listeners' impressions are sensitive to the perceived accuracy and authenticity of cloning for voices they know well, while unfamiliar listeners tend to prefer the synthetic versions of those same voice identities. The latter observation may relate to the tendency of generative voice synthesis models to homogenise speaking accents and styles, such that they more closely approximate (preferred) norms.

1. Introduction

Synthetic voices have been part of everyday human life for some time, from the iconic sound of Stephen Hawking, to satnav devices and public service announcers. Recent rapid development of voice synthesis and voice conversions technologies, often employing artificial intelligence (AI), now allows users to choose and/or design bespoke voices for use in applications from advert voiceovers to conversational agents and vocal avatars for online gaming. In some scenarios, voice synthesis aims to closely replicate specific real human identities. For example, an actor may agree for their voice identity to be resynthesised for use in widespread marketing campaigns or multilingual dubbing, or a person living with motor neuron disease (plwMND) may choose to replicate their voice identity for use in augmentative and alternative communication (AAC). A commonly-used term for synthesising a specific voice identity, in particular using AI, is "voice cloning". Voice cloning, in comparison to voice synthesis in general, implicates additional factors to the study of human perception of synthetic voices, because cloned voices are designed to invoke perceptual representations of familiar human

identities of varying self-relevance (i.e., one's own voice clone vs. the clone of a friend; McGettigan et al., 2024). Therefore, it is crucial to explore how human listeners socially perceive these voices, whether they replicate the voice of a stranger, a loved one, or their own.

1.1. Literature review

The human voice is a rich and dynamic source of information about a person, conveying not only the speaker's linguistic messages but also cues to their psychological state, emotions, and identity (Lavan, 2023a; Lavan et al., 2019; McGettigan, 2015; Scott & McGettigan, 2016). Human adult listeners extract percepts of physical characteristics (e.g. sex, age) and social traits (e.g. Trustworthiness, Dominance) from other human voices rapidly (Lavan, 2023b; Lavan, Rinke, & Scharinger, 2024; Mileva & Lavan, 2023) and with high agreement (Lavan, 2023b; McAleer et al., 2014; Mileva & Lavan, 2023). The evaluation of social traits – either from faces or voices – can be grouped into two key dimensions: warmth and competence, forming a universal social space across sensory modalities (Cuddy et al., 2008; McAleer et al., 2014;

^{*} Corresponding author. full address: Chandler House – 2 Wakefield Street, WC1N 1PF, London, UK.

E-mail address: v.rosi@ucl.ac.uk (V. Rosi).

<https://doi.org/10.1016/j.chbah.2025.100143>

Received 18 December 2024; Received in revised form 13 March 2025; Accepted 19 March 2025

Available online 25 March 2025

2949-8821/© 2025 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Todorov et al., 2008). Warmth reflects traits like likeability, trustworthiness, and attractiveness, while competence also signals dominance and hierarchy. It has furthermore been shown that these social evaluations can influence onward decision making, for example voting choices in elections (Klofstad et al., 2015; Mileva et al., 2020; Tigue et al., 2012). Thus, the perceived sound of a voice – regardless of the accuracy of that percept (Lavan, 2024) – has implications for how its owner might be judged and treated by other humans.

Modern synthesized voices may convey the same types of perceptual information as natural voices. Past work has shown that listeners tend to prefer listening to recordings of naturally-produced human voices compared with synthetic stimuli (Balas & Pacella, 2017; Bruder et al., 2023, pp. 170–171; Cabral et al., 2017; Kühne et al., 2020). The current reality is that in many cases synthetic voices no longer sound robotic and monotone across the board, instead sounding very like natural human voices (especially when enhanced with humanlike conversational markers like dysfluencies and filled pauses; Dinkar et al., 2023). Indeed, recent work examining human listeners' perception of 17 traits from 46 synthetic voices found evidence that the social evaluation of these artificial voice identities was underpinned by dimensions of Valence and Dominance (Shiramizu et al., 2022) as had been previously reported in studies of human voice perception (Baus et al., 2019; Guldner et al., 2024; Mahrholz et al., 2018; McAleer et al., 2014) – although we note that Shiramizu and colleagues did not explicitly link their findings to subjective or objective naturalness measures.

Similarly to human voice perception, there are implications for how the sounds of synthetic voices might affect how human listeners react to artificial agents and/or human users of voice synthesis. For example, there is an active discussion in fields such as computer science and human computer interaction about the impacts of increasing naturalness or “humanlikeness” in artificially-generated voices (Abercrombie, Cercas-Curry, Dinkar et al., 2023). Authors have considered how giving a humanlike vocal persona to a non-human agent, and thus inducing anthropomorphism of that agent by human users, increases the risk of negative outcomes such as inappropriate human-computer trusting behaviours, or even verbal sexual harassment of agents (e.g. agents with ostensibly feminine personae; Cercas Curry et al., 2021; Cercas Curry & Rieser, 2018). Some authors suggest that, while computer agents still do not behave in fully human-like ways, it may be advisable to provide them with voices that clearly signal the non-human status of the agent (Abercrombie, Cercas-Curry, Dinkar et al., 2023; Wilson & Moore, 2017). It is therefore crucial to improve our understanding of synthetic voice perception as the state-of-the-art evolves.

Voice cloning technologies allow for the artificial replication of any voice, including those that are familiar. Previous research on natural human voices suggests that listeners are likely to have sophisticated mental representations of personally-familiar other voices, allowing for highly accurate identity recognition (Kanber et al., 2022) and robust speech recognition (Holmes & Johnsrude, 2020; Johnsrude et al., 2013). There is also a body of work indicating that the self-voice has special status amongst familiar voices, in terms of its perceptual prioritisation (Kirk & Cunningham, 2024; Payne et al., 2021a, 2024; Rosi et al., 2024), the brain's response to hearing it (Graux et al., 2015; Pinheiro et al., 2023), and its social evaluation (e.g., perceived Attractiveness; Hughes & Harrison, 2013; Peng et al., 2019, 2020). Thus, it is important to understand not only how synthetic familiar voices are perceived (and how this differs from unfamiliar perception), but also whether perception differs depending on the level of self-relevance (i.e., self vs other).

1.2. Research gap and motivations

Most research on the perception of artificial voices today focuses on their realism and naturalness (Balas & Pacella, 2017; Bruder et al., 2023, pp. 170–171; Cabral et al., 2017; Kühne et al., 2020). Moreover, while some recent studies have explored the perception of cloned voices (Barrington & Farid, 2024; Lavan, Irvine, et al., 2024), they have not

examined the social trait evaluation of voice clones from familiar voices or the self-voice. This is of theoretical interest as research suggests differences in how familiar and unfamiliar voices are perceived (Holmes & Johnsrude, 2020; Johnsrude et al., 2013; Kanber et al., 2022), with one's own voice representing a unique case that may elicit distinct responses (Hughes & Harrison, 2013; Kirk & Cunningham, 2024; Payne et al., 2021b; Rosi et al., 2024). In this study, we extend these theoretical questions about human voice perception to synthesized voices for comparison.

1.3. Study design and predictions

In this study, we investigated the social trait evaluations of human and synthetic voices. We recruited human participants in familiar pairs of friends, where each member of the pair was recorded, and their voice resynthesised using state-of-the-art voice cloning. In a perceptual experiment, we collected ratings of four social traits – Attractiveness, Trustworthiness, Dominance, and Competence – as participants listened to original recordings and AI-generated speech in their own voice and their friend's voice (Task A). We selected these traits because they align with the two primary dimensions of the social space representative of the social evaluation of voices (Guldner et al., 2024; McAleer et al., 2014; Oleszkiewicz et al., 2017). Specifically, Attractiveness and Trustworthiness derive from the first dimension, while Competence and Dominance derive from the second dimension. To examine how overall familiarity affected ratings, we additionally tested a matched control group of unfamiliar listeners who each rated one of the voice pairs. Further tasks measured the perceived similarity of the stimuli to the original speaker (Task B - familiar listeners only), the perceived similarity between recordings and cloned stimuli from the same speaker (Task C), as well as the listeners' ability to distinguish cloned speech from recordings (Task D).

We predicted that being familiar with the voice identities would be associated with higher sensitivity to the cloning technology, in the form of higher discriminability of clones from natural voice recordings and lower ratings of perceived similarity between the clones and the original identities (self, friend), compared with the responses of unfamiliar listeners. Within familiar listeners, we further predicted higher sensitivity to, and lower perceived similarity of, cloned stimuli of the self-voice compared with the friend's voice. Indeed, the self-voice is usually experienced multimodally via both air- and bone-conduction sensory information (Maurer & Landis, 2009; Orepic et al., 2023) – likely leading listeners to be more inclined to reject a self-voice clone as a good likeness of the self. However, a different profile of results might be expected if the self-voice already sounds somewhat unfamiliar in recordings compared with during speaking – here, we might expect that listeners could be relatively less sensitive to distinguishing clones from recordings for their own voice, compared with their friend's voice.

Given previous research showing greater accuracy in perception of familiar versus unfamiliar identities, as well as prioritisation of the self-voice, we predicted that manipulating familiarity and self-relevance would also impact the social evaluations of voices via trait ratings. For overall effects of familiarity, we predicted that familiar listeners might enjoy hearing the voices more than unfamiliar listeners (McGettigan, 2015) and thus afford them more positive evaluations, but might disfavour cloned voices if these are perceived as inauthentic. For the effect of self-relevance (self vs. friend), some evidence suggests that listeners explicitly dislike the sound of their own (recorded) voice (Holzman et al., 1967; Lee et al., 2005; Naunheim et al., 2023), while other research has suggested at least some implicit preference of the self-voice relative to other (i.e. friend and unfamiliar) voices, specifically for ratings of Attractiveness (Hughes & Harrison, 2013; Peng et al., 2019, 2020). Thus, it is unclear which of the two identities might be afforded more favourable ratings by familiar listeners.

2. Methods

2.1. Participants

100 adult participants took part in the study. The total number of participants was determined by the study's budget and timeframe. 50 experimental participants (Mean age = 22, Age range = 19–31, 36 female, 12 male, 2 preferred not to say) – comprising 25 pairs of close friends, relatives, or partners – were recruited through the UCL Psychology subject pool, as well as local networks such as UCL clubs and societies through emails, social media, and personal networks. Pairs of participants indicated how long they have known each other (ranging from two months to beyond five years), and how often they speak to each other (ranging from daily to monthly). They were first-language speakers of English with a variety of self-reported accents (e.g., British, American, South Asian, Hong Kong). These participants were aware that they were signing up for a study on voice cloning, in which their own and a friend's voice would be cloned. 50 control participants (Mean age = 37, Age range = 18–65, 23 female, 27 male) were recruited online through Prolific (prolific.co). They were first-language speakers of English. Experimental participants completed an in-lab voice recording session plus four additional online testing sessions – results from the first of these online sessions are reported here. In total, 47 experimental participants completed the testing session and are included in the final dataset, along with 47 control participants who heard the same stimuli but were unfamiliar with the speakers.

Ethical Approval was obtained (SHaPS-2023-CM-038), and all participants gave their informed consent prior to the testing. Participants were rewarded at a rate of £9/hour for the recording session, plus £9/hour for the remaining online sessions – a further bonus of £9 was offered to participants who completed all online sessions, resulting in a maximum payment of £36. Control participants completed one online session only and were rewarded at a rate of £9/hour.

2.2. Materials

2.2.1. Words and sentences

For each speaker and sound type (i.e., clone, recording), the voice samples consisted of 12 words (i.e., colours “blue”, “green”, “pink”, and “red”; digits “one”, “two”, “seven”, “nine”; one syllable words “had”, “hard”, “heed”, “who'd”) and four sentences from the IEEE Harvard Sentences corpus ('IEEE Recommended Practice for Speech Quality Measurements', 1969). All items were included in the online listening tasks in both recorded and cloned versions.

2.2.2. Voice clones

Prior to the listening test, a recording session was conducted with experimental participants (see *Design & Procedure*). A subset of these recordings was used as stimuli in the online listening tasks, while the remaining recordings were used to clone the participants' voice identities. The cloning procedure was as follows: First, for each experimental participant we created a concatenated voice sample from the recording session, including read stories and spontaneous speech (~7–8 min). The concatenated audio file was fed to the *Instant Voice Cloning* tool, a generative speech synthesis model from ElevenLabs (<https://elevenlabs.io>) to generate a voice clone for that participant. ElevenLabs' text-to-speech functionality was then used to generate the required word and sentence stimuli for use in listening tasks. The generated stimuli were inspected aurally for audio artefacts and regenerated or edited where possible to generate artefact-free versions. Stimuli were also inspected to verify that the cloned voice bore a reasonable accent match to the original talker, as the *Instant Voice Cloning* tool does not guarantee accent accuracy. In cases where a mismatch was detected (e.g. UK speaker cloned with US accent), stimuli were iteratively regenerated until a suitable accent match was achieved. This was not possible in all cases, thus the final stimulus set retains some of the expected variability in the

outcomes of this cloning tool.

For both voice clones and voice recordings, we trimmed silences and removed breaths, audio artefacts, and other noises using *Librosa* (McFee et al., 2015). All stimuli were resampled to 22050 Hz and RMS-normalised using the Python library *ffmpeg-normalise* (<https://github.com/slhck/ffmpeg-normalize>).

2.3. Design & Procedure

Fig. 1 gives an overview of the experimental procedure for experimental participants and control participants.

2.3.1. Recording session

Recording sessions were conducted in a soundproof booth at UCL's Department of Psychology. The voices of experimental participants were collected using a RODE NT1-A microphone, with recording instructions provided on-screen via the Gorilla Experiment Builder testing platform (Anwyl-Irvine et al., 2020). The session comprised two main parts. In the first part, participants read aloud 12 words (repeated twice) and 35 Harvard sentences. A single recording sample was selected for each of the 12 words and four sentences (see Words and Sentences) to compile the set of voice samples used in the listening test. The second part of the recording session involved collecting various speech styles (i.e., read speech, spontaneous speech) from the participants, which served as references for creating cloned voice identities (see *Voice clones*). Read speech samples were obtained by having participants read Aesop's fable *The North Wind & the Sun* story and *Arthur the rat* (Sweet, 1890), while spontaneous speech samples involved participants describing two Diapix images (Baker & Hazan, 2011) for 1 min 30 s and providing a 2-min summary of their morning routine.

2.3.2. Online testing

Participants were invited to do the listening test online via the Gorilla Experiment Builder (gorilla.sc). They were asked to use a computer to complete the session and ensure that they were in a quiet environment with minimal background noise. They were also asked to wear headphones and to do a sound check prior to the task. During the test, they used a mouse or trackpad to submit responses.

For experimental participants, the listening test consisted of four tasks, Task A, Task B, Task C, and Task D. Control participants only completed Task A, Task C, and Task D.

Before completing the listening tasks, all participants were made aware that they would hear a mixture of real recordings and voice clones. This was to control for the fact that the experimental participants already expected to hear voice clones, based on the information they were given when recruited to the study.

2.3.3. Task A - trait ratings

In this task, participants rated recorded and cloned stimuli from two voice identities on four trait scales: Attractiveness, Competence, Dominance, and Trustworthiness. During each trial, participants listened to a voice sample and rated the intensity of a specific trait using a slider ranging from 0 (not at all <trait>) to 100 (very <trait>). For example, when assessing Attractiveness, participants responded to the question, “How attractive is the voice you just heard?” from *not at all attractive* to *very attractive*. Blocks were organised by trait and voice identity (i.e., self, friend), and presented in a randomised order. Each block consisted of 32 trials - 16 stimuli per sound type (12 words; 4 sentences). On average, the task lasted 20 min. The procedure was identical for both participant groups apart from the labelling of the identities (*You* and *Your Friend* for experimental participants, who heard recordings and clones of these familiar voice identities; *Voice A* and *Voice B* for control participants).

2.3.4. Task B – similarity to real voice

For this task, experimental participants rated the similarity of the

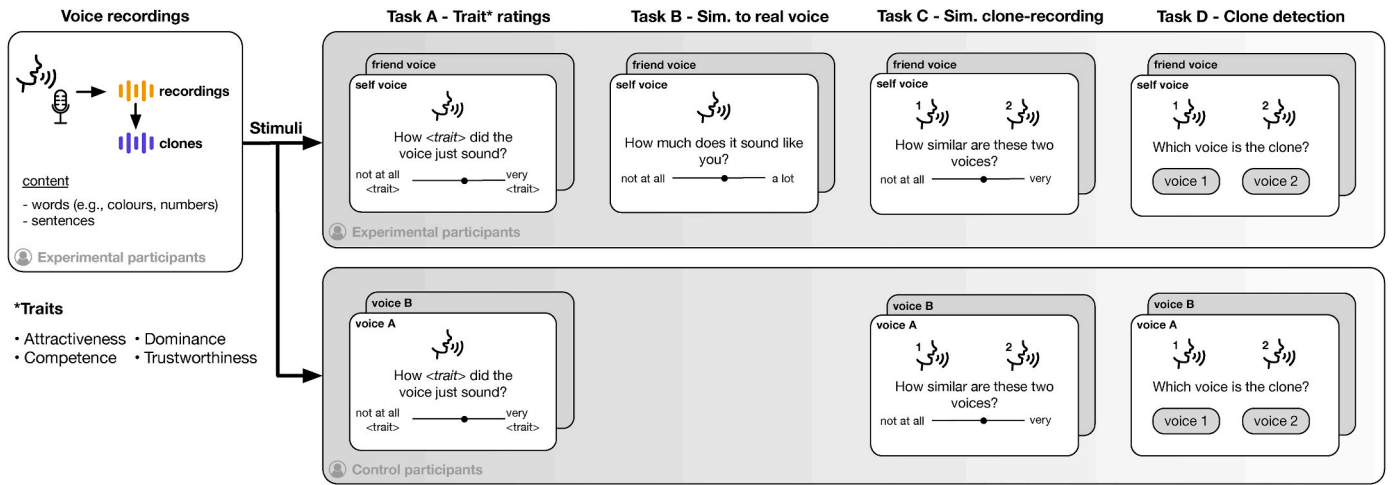


Fig. 1. Overview of the experimental design.

voice stimuli in relation to the real voices from which they were generated. In each trial, participants listened to a voice sample and responded to the question “How much does it sound like you/your friend?” using a slider ranging from 0 (not at all) to 100 (a lot). Blocks were organised by voice identity and presented in a randomised order. Each block consisted of 32 trials – 16 stimuli per sound type. On average, the task lasted 6 min. Control participants did not take part in this task as they were not familiar with the identities they heard in the study.

2.3.5. Task C – similarity clone-recording

In this task, all participants rated the similarity between voice clones and voice recordings from the same speaker. In each trial, participants listened to a voice clone and a voice recording from the same speaker and with the same word content, and responded to the question “How similar are these two voices?” using a slider ranging from 0 (not at all) to 100 (very). The order of the sounds (i.e., clone, recording) was randomised per trial. Blocks were organised by identity and presented in a randomised order. Each block consisted of 16 trials. On average the task lasted 3 min.

2.3.6. Task D – clone detection

Here, experimental participants were requested to identify the clone within pairs of recorded and cloned voice samples. In each trial, participants listened to a pair of voice samples, including one clone and one recording of the same word or sentence. They then responded to the question “Which one is the clone?” by clicking an onscreen button on the left (for the first sound) or right (for the second sound) of the screen. The order of the sounds (i.e., clone, recording) was randomised per trial. Blocks were organised by voice identity and block order was randomised. Each block consisted of 16 trials. On average, the task lasted 2 min. The voice identities were presented to experimental participants as *You* and *Your Friend*, and to control participants as *Voice A* and *Voice B*.

For each task, we presented vigilance tests at random locations within the sequences of trials. For these tests, participants had to select a number between 1 and 6 that was announced to them beforehand.

2.4. Data analysis

We analysed four types of dependent variable based on participants’ responses: trait ratings on four scales (task A), ratings of similarity to the real voice (Task B), ratings of similarity between clones and recordings (Task C), and sensitivity to clone detection (Task D).

For Tasks A, C and D, we assessed the impacts of self-relevance and familiarity on voice perception in two separate analyses. Specifically, the analysis of self-relevance (i.e., self vs. friend) only included data

from the experimental group (i.e., familiar listeners for whom the two specific identities were meaningful), while the analysis of familiarity included both experimental and control groups (i.e., familiar and unfamiliar). In the rest of the paper, we will refer to the experimental group as familiar listeners, and the control group as unfamiliar listeners.

Due to the acoustic, prosodic, and linguistic differences between words and sentences, we report separate analyses for these two types of materials.

2.4.1. Task A

The first model included trialwise ratings from familiar listeners only, with *Voice Identity* (self, friend) and *Sound Type* (recording, clone) as fixed factors and participant as a random intercept. The second model included ratings from familiar and unfamiliar listeners, with *Familiarity* and *Sound Type* as fixed factors and participant as a random intercept:

$$\text{Trait} \sim \text{Voice Identity} * \text{Sound type} + (1|\text{Participant})$$

$$\text{Trait} \sim \text{Familiarity} * \text{Sound type} + (1|\text{Participant})$$

The models were repeated for each trait scale: Attractiveness, Trustworthiness, Dominance, and Competence.

2.4.2. Task B

This analysis included trialwise ratings from familiar listeners only, with *Voice Identity* (self, friend) and *Sound Type* (recording, clone) as fixed factors and participant as a random intercept.

$$\text{Similarity}_{\text{real voice}} \sim \text{Voice Identity} * \text{Sound type} + (1|\text{Participant})$$

2.4.3. Tasks A and B

Combining data from Tasks A and B for familiar listeners only, we investigated the effects of *Voice Identity* (self, friend) and *Similarity* on trait evaluation of cloned voice samples.

$$\text{Trait} \sim \text{Voice Identity} * \text{Similarity}_{\text{real voice}} + (1|\text{Participant})$$

2.4.4. Task C

Similarly to Task B, the analysis of *Similarity* between voice clones and voice recordings was examined with two models. The first model included ratings from familiar listeners only, with *Voice Identity* as a fixed factor and participant as a random intercept. The second model included ratings from both familiar and unfamiliar listeners, with *Familiarity* as a fixed factor and participant as a random intercept.

$$\text{Similarity}_{\text{clone-recording}} \sim \text{Voice Identity} + (1|\text{Participant})$$

$$\text{Similarity}_{\text{clone-recording}} \sim \text{Familiarity} + (1|\text{Participant})$$

2.4.5. Task D

Sensitivity to clone detection (d') was first assessed by comparing the mean d' values from familiar listeners and unfamiliar listeners with 0 – a value indicating the inability to distinguish clones from recordings, or chance level – with one-sample t-tests. Then we investigated the impact of *Voice Identity* and *Familiarity* with two mixed models. The first model included ratings from familiar listeners only with *Voice Identity* as a fixed factor and participant as a random intercept. The second model included ratings from familiar and unfamiliar listeners, with *Familiarity* and *Sound Type* as fixed factors and participant as a random intercept.

$$d' \sim \text{Voice Identity} + (1|\text{Participant})$$

$$d' \sim \text{Familiarity} + (1|\text{Participant})$$

To calculate unbiased d' scores, we classified participants' correct identification of the clone when it was the first sound (out of two) as *Hit*, and their incorrect identification of the clone when it was the second sound as *False Alarm*. Then, we applied a log-linear correction to the score for cases involving 100 % *Hits* or 0 % *False alarms* (Stanislaw and Todorov, 1999).

For each model, we tested the significance of the interactions and the effects by performing likelihood ratio tests with the *afex* package in R (version 4.4.1). Depending on the significance of the factors and their interactions, we ran post hoc pairwise comparisons using *emmeans* with a Bonferroni correction.

3. Results

As a result of the vigilance tests, no participant was excluded for the analysis.

Due to the smaller number of samples in the sentence dataset and the similarity of trends between word and sentence results, we focus here on the responses to words. See the Supplementary Material for a detailed analysis of the sentence data.

First, we present the analyses of Task D (clone detection), Task B (similarity to the real voice ratings), and Task C (similarity clones vs. recordings) to illustrate listener sensitivities to cloning technology and its outcomes. Second, we present results related to trait perception from clones and recordings (Task A). Finally, we report results of the combined analysis of Task A and Task B data.

3.1. Sensitivity to voice cloning

Fig. 2(A) illustrates listener sensitivity to discriminate voice clones from voice recordings. One-sample t-tests against 0 revealed that for d' values from familiar listeners were significantly above chance level when listening to either their own voice or their friend's voice (self: $M = 2.13$, $CI = [1.84, 2.41]$, $t(46) = 15.10$, $p < .001$; friend: $M = 1.91$, $CI = [1.60, 2.22]$, $t(46) = 12.36$, $p < .001$), indicating successful discrimination. In contrast, unfamiliar listeners did not exhibit d' values significantly different from 0 ($M = -0.14$, $CI = [-0.50, 0.21]$, $t(46) = -0.81$, $p = .420$), suggesting an inability to reliably differentiate clones and recordings. Importantly, we note that individual performances covered the whole range of identification accuracy, from highly accurate identification to consistent misidentification of recordings as clones. Mixed model analysis revealed that there was no significant effect of *Voice Identity* ($\chi^2(1) = 1.13$, $p = .289$) in the responses provided by familiar listeners only. However, in an analysis of all listeners there was a main effect of *Familiarity* ($\chi^2(1) = 76.74$, $p < .001$) revealing that familiar listeners were significantly better than unfamiliar listeners at distinguishing recordings from clones.

Fig. 2(B) illustrates how familiar listeners rated voice clones and voice recordings for perceived similarity to the real voices of themselves and their friend. Although there was a significant interaction between *Voice Identity* and *Sound Type* ($\chi^2(1) = 4.91$, $p = .027$), post hoc comparisons indicated that perceived similarity was significantly lower for clones than recordings for both voice identities ($ps < .001$), while differences between the voice identities self and friend were non-significant for both recordings and clones.

Fig. 2(C) illustrates how listeners rated the similarity of pairs of voice clones and voices recordings. We found no significant main effect of *Voice Identity* ($\chi^2(1) = 0.49$, $p = .482$) in the responses provided by familiar listeners only. Additionally, there was no main effect of *Familiarity* ($\chi^2(1) = 1.11$, $p = .292$) in the responses provided by all listeners.

3.2. Social evaluations of recorded and cloned voices

Fig. 3 illustrates significant and non-significant interactive effects of the model fixed factors on mean ratings of the four traits – Attractiveness, Competence, Dominance, and Trustworthiness – for voice clones and voice recordings. Fig. 3(A) plots the interactive effects of *Voice Identity* and *Sound Type* on trait ratings provided by the experimental group (familiar listeners). We observed a significant two-way interaction between *Voice Identity* and *Sound Type* for Attractiveness but not for

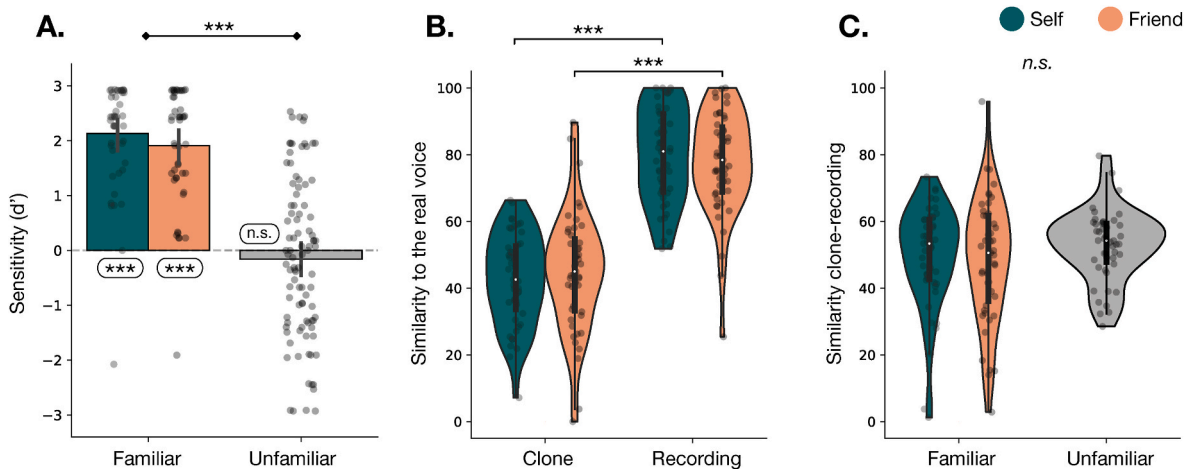


Fig. 2. A. Mean sensitivity (d') measures from the clone spotting task (Task D) as a function of self-relevance and *Familiarity*. Error bars indicate 95 % confidence intervals of the means. The top bar indicates the significance of the *Familiarity* effect. Circled information indicates the significance of the difference of the mean with 0. B. Similarity to the real voice as evaluated by experimental participants (i.e., familiar listeners) in Task B. C. Similarity between voice clones and voice recordings from the same speaker. *: $p < .05$ - **: $p < .01$ - ***: $p < .001$.

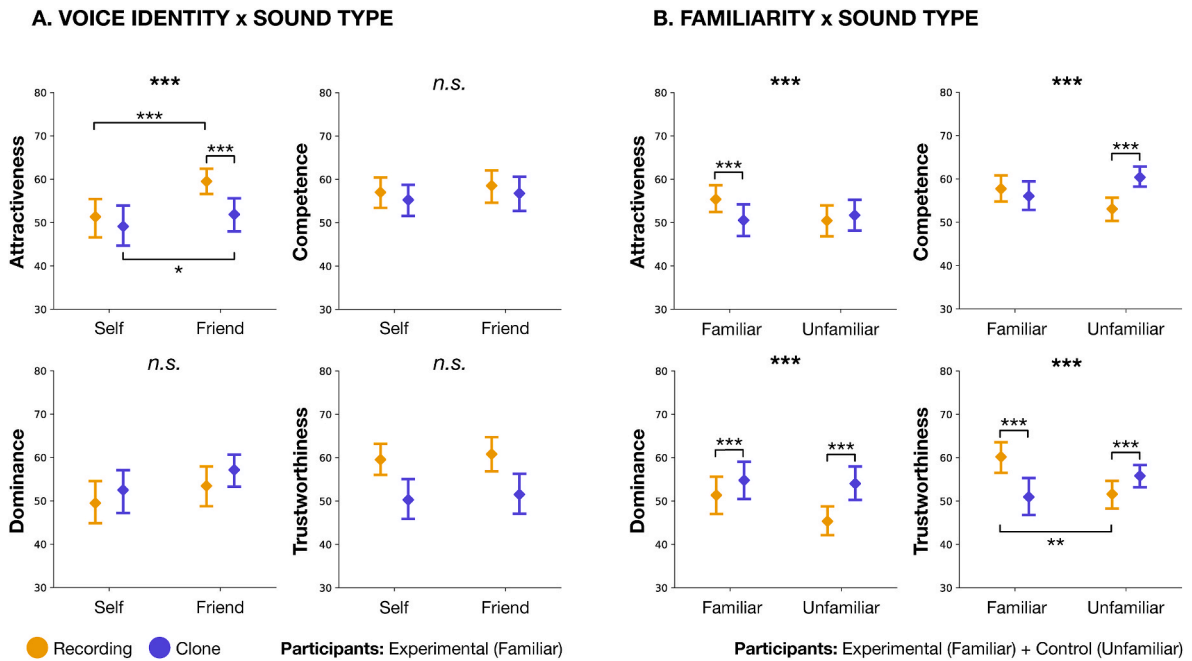


Fig. 3. Mean ratings for Attractiveness, Competence, Dominance, and Trustworthiness on voice samples of words (Task A). In all panels, the uppermost annotation indicates the significance of the two-way interaction; where applicable, other annotations indicate significant post hoc pairwise comparisons. Significant main effects are reported in the main text. **A.** Trait ratings from the Experimental group (Familiar listeners) as a function of *Sound type* (Recording, Clone) and *Voice identity* (Self, Friend). **B.** Trait ratings from the Experimental (Familiar listeners) and Control group (Unfamiliar listeners) as a function of *Sound type* (Recording, Clone) and *Voice familiarity* (Familiar, Unfamiliar). Error bars indicate 95 % confidence intervals of the means. *: $p < .05$ - **: $p < .01$ - ***: $p < .001$.

the other traits. There were significant main effects of *Voice Identity* and *Sound Type* for Competence (friend > self; recording > clone) and for Dominance (friend > self; clone > recording), and a significant main effect of *Sound Type* only for Trustworthiness (recording > clone). Fig. 3 (B) shows the significant interactive effects of *Familiarity* with *Sound Type* on all trait ratings. See Table 1 for the full list of significant main effects and interactions, along with post hoc pairwise comparisons for each trait and each model.

3.3. Social evaluations and similarity to the real voice

We observed non-significant interactions between *Voice Identity* and *Similarity_{real voice}* for all of the trait scales. There was a main effect of *Similarity_{real voice}* for Attractiveness ($\chi^2(1) = 39.8$, $p < .001$), Competence ($\chi^2(1) = 18.0$, $p < .001$), Dominance ($\chi^2(1) = 4.09$, $p = .043$), and Trustworthiness ($\chi^2(1) = 59.6$, $p < .001$). Fig. 4 shows the contribution of *Similarity_{real voice}* for each trait along with a summary of the corresponding statistics.

4. Discussion

4.1. Main findings

We present a detailed examination of the impacts of state-of-the-art voice cloning technology on listeners' perception and social evaluation of voice stimuli. While both familiar and unfamiliar listeners gave higher Dominance ratings to voice clones than voice recordings, the groups otherwise dissociated in terms of whether they rated clones or recordings more highly on Attractiveness, Competence, and Trustworthiness. Within this, familiar listeners additionally tended to afford higher ratings to the voice of their friend, compared to their own. These observations are contextualised within overall differences in listeners' sensitivity to cloning: familiar listeners were highly accurate at distinguishing clones from recordings, while unfamiliar listeners showed highly variable performance at the individual level. Furthermore, familiar listeners' social evaluations of cloned stimuli were dependent

on their perceived similarity to the "real" voice of the original speaker. Overall, our findings indicate that state-of-the-art voice synthesis can now create highly naturalistic and humanlike stimuli that, for unfamiliar listeners, can be both indistinguishable from human recordings yet simultaneously distinct in character from them, while having multifaceted impacts on familiar listeners with pre-existing representations of the original speakers.

4.2. The impact of familiarity and self-relevance on clone detection

Results from the clone detection task showcase the remarkable advancements in voice cloning technology. At the group level, our findings replicate the results of other recent studies reporting that, on average, AI-generated clones of unfamiliar voices are perceived to be equally human to recordings of the corresponding human speakers (Barrington & Farid, 2024; Lavan, Irvine, et al., 2024). We note that the ability of familiar listeners to readily and accurately distinguish voice clones from voice recordings may be, to some extent, task based: Given the nature of the experiment, they anticipated hearing one of two well-known voice identities, enabling them to compare each sound to their mental representation of the original speaker. Any deviation from that representation – such as differences in voice quality, speech rate, accent, and more – could have acted as an indicator of a voice clone. Indeed, we observed large effects of *Sound Type* (i.e., clone vs. recording) in familiar listeners' rating of the similarity between the heard sound and the "real" voice of the speaker, where the recordings were judged to be much more similar than the clone. We did not observe any systematic effects of self-relevance (self vs. friend) on clone detection or similarity judgements. This indicates that, at least in the context of evaluating voice identity, there is no apparent advantage or disadvantage for the self-voice despite it being experienced differently from during speech (i.e., by air-conduction rather than additionally via bone-conduction; see Orepic et al., 2023).

Table 1

– Mixed model analyses for trait evaluations on voice clones and voice recordings for (a) familiar listeners only (factors: *Voice Identity*, *Sound Type*) and (b) all listeners (factors: *Familiarity*, *Sound Type*).

(a)	Trait	Effect	Post hoc
Attractiveness	Interaction	$\chi^2(1) = 12.46, p < .001$	Friend-Record > Self-Record ($p < .001$) Friend-Clone > Self-Clone ($p = .023$) Friend-Record > Friend-Clone ($p < .001$) Friend > Self
Competence	Interaction	$\chi^2(1) = 0.00, p = .953$	
	Voice ID	$\chi^2(1) = 5.58, p = .018$	Record > Clone
	Sound Type	$\chi^2(1) = 5.43, p = .020$	
Dominance	Interaction	$\chi^2(1) = 0.01, p = .904$	Friend > Self
	Voice ID	$\chi^2(1) = 30.21, p < .001$	Clone > Record
	Sound Type	$\chi^2(1) = 19.63, p < .001$	
Trustworthiness	Interaction	$\chi^2(1) = 0.02, p = .878$	Record > Clone
	Voice ID	$\chi^2(1) = 3.06, p = .080$	
	Sound Type	$\chi^2(1) = 138.6, p < .001$	
(b)	Trait	Effect	Post hoc
Attractiveness	Interaction	$\chi^2(1) = 37.06, p < .001$	Familiar-Record > Familiar-Clone ($p < .001$)
Competence	Interaction	$\chi^2(1) = 70.97, p < .001$	Unfamiliar-Record > Unfamiliar-Clone ($p < .001$)
Dominance	Interaction	$\chi^2(1) = 24.97, p < .001$	Familiar-Clone > Familiar-Record ($p < .001$) Unfamiliar-Clone > Unfamiliar-Record ($p < .001$)
Trustworthiness	Interaction	$\chi^2(1) = 155.79, p < .001$	Familiar-Record > Familiar-Clone ($p < .001$) Unfamiliar-Clone > Unfamiliar-Record ($p < .001$) Familiar-Record > Unfamiliar-Record ($p < .001$)

The Post hoc column reports pairwise comparison t-tests in the case of a significant interaction or fixed effect. In cases with a significant interaction, main effects are not reported; non-significant post hoc tests are not reported. *Record* = Recording.

4.3. Familiarity shapes social perception of voice clones

Across ratings of four perceived social traits from all participants, we found repeated evidence of the effect of *Familiarity* in interaction with

Sound Type. The profiles varied across traits: for Attractiveness, familiar listeners rated recordings higher than clones while unfamiliar listeners' ratings showed no effect; for Competence, familiar listeners rated recordings and clones equivalently while the unfamiliar listeners rated clones more highly than recordings; for Dominance, both groups gave higher ratings to clones but the effect was more pronounced in unfamiliar listeners; for Trustworthiness, familiar listeners gave higher ratings for the recordings while the unfamiliar listeners rated the clones more highly. In the research literature on social evaluation of voices and faces (McAlee et al., 2014; Todorov et al., 2008), percepts of attractiveness and trustworthiness are associated with an underlying dimension that is associated with a concept of warmth (Cuddy et al., 2008; Lan et al., 2022). Dominance and competence are typically associated with a dimension orthogonal to warmth, often labelled after one of these traits (Anderson & Kilduff, 2009; Cuddy et al., 2008). Here, the overall profile of responses indicates that for familiar listeners, recorded voices are perceived as higher in warmth than clones, but not necessarily in Competence/Dominance. In contrast, for unfamiliar listeners, clones are clearly perceived as higher in Competence/Dominance, as well as higher in warmth.

For unfamiliar listeners, voice clones generated an equivalent sense of humanness to human recordings on average, but conveyed distinct social traits. Surprisingly, where previous studies have reported lower ratings of warmth-related traits such as Likeability and Trustworthiness for artificial voices relative to human recordings (Bruder et al., 2023, pp. 170–171; Cabral et al., 2017; Kühne et al., 2020), we observed enhanced ratings on the warmth dimension. This has implications for how listeners might evaluate cloned voices in everyday situations. While the literature has explored the anthropomorphic effects of adding voices to machines (Eysel et al., 2012; Festerling & Siraj, 2022), there is less focus on the significance of convincingly humanlike artificial voices that sound *more appealing* and *more able* than real humans (i.e., higher Trustworthiness, Competence, and Dominance). According to the stereotype content model (Cuddy et al., 2008; Fiske et al., 2007), a combination of high perceived warmth and competence in others provokes feelings of admiration, which may lead to behaviours including both active and passive facilitation of the other (i.e., “acting for” and “acting with”). In the context of our study, the generated voice clones might have been more effective in gaining human listener trust and fostering stronger acceptance of the voice's expertise or authority (Abercrombie, Cercas-Curry, Dinkar et al., 2023).

Familiar listeners evaluated voice clones as lower in warmth than human recordings, but provided only partial evidence for enhanced competence as clones were only rated higher on the Dominance scale. Here, the findings are influenced by the listeners' familiarity with the two voice identities they expected to hear, and their high sensitivity to imperfect replicas of these familiar voices. Familiar listeners were very successful in distinguishing clones from recordings, and thus their

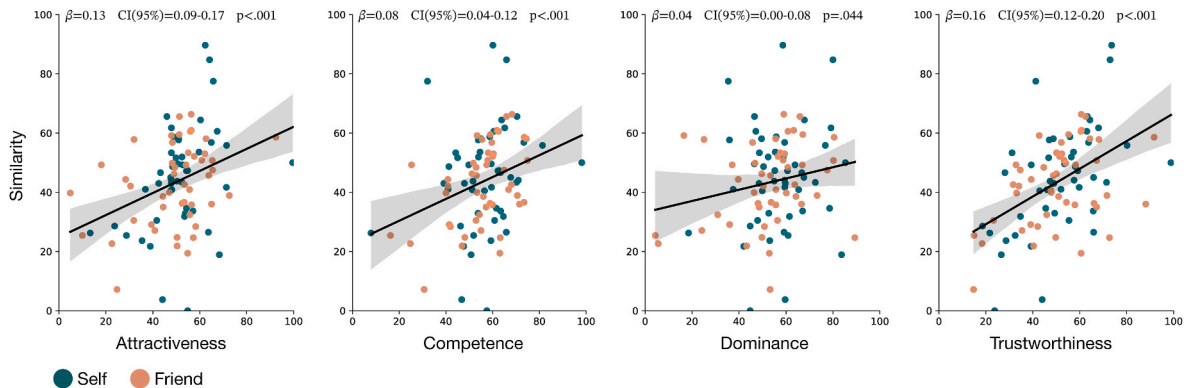


Fig. 4. Trait ratings as a function of Similarity to the real voice ratings provided by the Experimental group (Familiar listeners only). β : Estimate, $CI(95\%)$: 95 % confidence intervals.

overall lower ratings of clones on the warmth dimension may reflect their awareness that the voice clones are inauthentic replicas of themselves and their friend (see Pisanski & Reby, 2021 on detection of deception through voice modulation). While Dominance appears less affected by low authenticity (see Figs. 2 and 3), the ambiguous trend of Competence for clones compared to recordings may result from opposing influences: reduced authenticity weakening perceptions of the notionally positive attribute of Competence (Christoforakos et al., 2021), while having less of an effect on the arguably more neutrally-valenced quality of Dominance.

4.4. The interplay of self-relevance and familiarity

A closer assessment of familiar listeners' responses revealed unexpected effects of familiarity and self-relevance on the social evaluations of clones and recordings. First, despite clear interactive effects of *Familiarity* and *Sound Type* on social evaluations, post hoc tests revealed only one simple effect of *Familiarity*, specifically for ratings of Trustworthiness of recorded voices. While the perceptible inauthenticity of clones might have impacted their social evaluation, we expected a familiarity preference for recorded voices on one or more traits: for example, several authors have reported enhanced attractiveness ratings for recordings of personally familiar and self-relevant voices (self, friend, and self-similar identities) compared with unfamiliar voices (Hughes & Harrison, 2013; Peng et al., 2019, 2020). It may be that such effects are more detectable in within-subjects designs, where listeners rate both familiar and unfamiliar identities in the same tasks. Second, we found numerous effects of self-relevance (self vs. friend) on the familiar listeners' ratings of both recordings and clones, with listeners affording higher ratings of Attractiveness, Competence, and Dominance to their friend's cloned and recorded voice compared to their own voice. Our findings for Attractiveness contradict recent evidence for self-voice enhancement relative to a friend's voice (Peng et al., 2020). Instead, they align with claims that the experience of hearing a recording of the self-voice is relatively disliked due to its mismatch with the multimodal experience of the self-voice during speaking (Holzman et al., 1967; Lee et al., 2005; Naunheim et al., 2023). Here, neither the recordings nor the clones of the self-voice replicated the perceptual experience of self-speech, potentially making them less appealing than the friend's voice that is typically experienced as an audio signal. Finally, social desirability bias may also play a role here, where participants' natural tendency toward self-deception (i.e., viewing the self favourably) was over-ridden by impression management (i.e., not wanting to appear egotistical to the researchers; Graeff, 2005). Similarly, Hughes and Harrison (2013) have claimed that the self-enhancement effect for voices depends on listeners being unaware of the presence of their voice within a task. However, the strong predictive relationship between greater perceived similarity to the "real" voice and high ratings on the warmth dimension in our study indicates that, even with explicit awareness that recordings and clones originated from the self, listeners overall endorsed the notion that "what sounds like me sounds good".

4.5. Principal implications

What are the implications of our findings for potential use cases of a model producing convincing voice clones of a known person? First, we note that familiar listeners' mean ratings of clones for Attractiveness, Competence, and Trustworthiness were all in the upper half of the scale. This suggests that cloned voices were evaluated rather positively on average. Second, our findings on similarity and trait ratings suggest that listeners encountering a clone of someone they know will likely be very sensitive to its authenticity, and where authenticity is lacking it may negatively impact users' engagement with that voice. Third, we didn't see any evidence for differences in the detectability or perceived similarity of self-voice clones compared to clones of the friend's voice. This suggests that evaluations of voice similarity for cloning use cases can be

treated equivalently across different personally-relevant identities. However, there may be mixed outcomes when a cloned voice represents the self in spoken communication (e.g., with AAC devices), as Attractiveness ratings from familiar listeners suggest that friends and family may have more favourable responses to a cloned voice than the user themselves. With this in mind, it will be instructive to further investigate the trade-offs of self-likeness (how much the voice clone resembles the original voice) versus self-representativeness (i.e., how well the clone conveys the identity of the speaker) in voice cloning use cases, especially if uncanny effects emerge. For example, it may be preferable to use a voice identity conveying key aspects of the user's personal identity (such as accent, gender, and delivery style) without attempting to replicate their specific vocal identity (Sutton et al., 2019, pp. 1–14; Zhang et al., 2021).

4.6. Research limitations and future works

The inspection of the individual participant data from our clone detection task (see Fig. 2(A)) reveals substantial variability in performance: while some control participants were systematically selecting a human voice recording as the clone (suggesting a "hyperrealism" effect; Miller et al., 2023; Nightingale & Farid, 2022; Tucciarelli et al., 2022), other unfamiliar listeners performed with high accuracy (suggesting that there are perceptible cues that can be used to detect AI-generated audio). This variability, which was much more pronounced in the control group than in the experimental group, may stem from the demographic diversity of our control participants, particularly regarding age or gender. Some recent work has identified possible demographic factors that might affect observers' sensitivities to AI-generated stimuli, including age, experience with AI, and critical thinking (Herrmann, 2023; Hulzebosch et al., 2020; Pehlivanoglu et al., 2024). Future work with larger samples of human and cloned voice identities should examine the impacts of demographic factors on the perception of artificial voices in more detail.

In our experiments, we used a generative speech model that was not fine-tuned to the characteristics of the original speakers. Thus, it does not achieve the best performance for the voice cloning task in term of likeness to the real voice. A model better fitted to the cloning of one voice identity would produce clones with higher perceived similarity to the original voice, potentially neutralising or even reversing the negative evaluation of clones relative to recordings observed here. This idea is supported by the finding that greater perceived similarity to the real voice predicted higher ratings on all four trait scales (with a weaker predictive relationship for Dominance). However, it remains unclear whether models achieving near-perfect similarity would uncover an "uncanny valley", where social evaluations of these vocal identities would be penalised due to experiences of eeriness (Kühne et al., 2020; Romportl, 2014). Future work could explore the relationship between clone resemblance and social evaluation in a more controlled manner, for example by parametrically varying the perceived similarity of voice samples to the original speaker (e.g. using voice morphing). On the other hand, the speech synthesis process of the model we used may have enhanced social qualities in generated voice clones as an unintended by-product of models optimised for speech intelligibility and naturalness (i.e., evoking perceptual stereotypes or halo effects). Future work should investigate whether these findings generalise across state-of-the-art models, and how these social evaluations might impact listener decision-making and behaviours (e.g. financial investment; Torre et al., 2018, pp. 1–6).

When considering the overall advantage of familiar over unfamiliar listeners here, we have contrasted high personal familiarity of the self and a friend in the experimental group with a total lack of presumed familiarity with those identities in the control group. Previous research has demonstrated a perceptual advantage for both familiar voices and the self-voice over unfamiliar voices at an individual level, but in this study, we were unable to directly compare the evaluation of familiar or

self-relevant voices to unfamiliar voices within the same individuals. Therefore, future studies should investigate this relationship using within-subject designs, where listeners rate both familiar and unfamiliar identities within the same tasks. In the same vein, it has been shown that relatively short periods of lab training can lead to advantages in identity recognition (Kanber et al., 2022) and speech recognition (Har-shai Yahav et al., 2024; Holmes & Johnsrude, 2020). It would be of interest to explore whether lab training to recognise an identity from voice recordings improves discriminability of clones of the same identity and matches performance for familiar identities (see Domingo, Holmes, & Johnsrude, 2020; Holmes, To, & Johnsrude, 2021).

5. Conclusion

We have found that AI voice clones are already highly humanlike, with widespread impacts on the evaluation of their social traits. The current state-of-the-art for low-input, low-cost cloning used in the current study generates unfamiliar AI voices apt to increase the anthropomorphism of machines and machine-generated speech, with specific risks for increased human compliance and over-trusting. In contrast, while the clone of a familiar voice identity may fail to fool a listener in a phone scam, it may also fall short of being convincing in intentional use cases (e.g., assisted communication) if it lacks sufficient similarity to the original speaker's voice. Future research should first explore the perceptual and acoustic drivers influencing AI voice clone detection, second delineate the individual differences in listeners' skills and pre-conceptions toward artificial voices, and third characterise the relationships between the similarity of a voice clone to the original voice and its acceptance by listeners/users. For all of these, it will also be crucial to examine the effects in more ecologically valid contexts that take account of factors such as deep fake and misinformation base rates (i.e., for unfamiliar clone detection), as well as contextualised use cases with relevant stakeholders (e.g., synthetic voice selection and use by plwMND).

CRedit authorship contribution statement

Victor Rosi: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Emma Soopramanien:** Investigation. **Carolyn McGettigan:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Data availability

The data/code used for this article is available at <https://osf.io/prv6e/>.

Funding sources

This work was supported by a Leverhulme Research Leadership Award (RL-2016-013), a British Academy Mid-Career Fellowship (MCFSS23\230112), both awarded to C McGettigan, and a British Academy Postdoctoral Fellowship (PFSS24\240043) awarded to V Rosi.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.chbah.2025.100143>.

References

- Abercrombie, G., Curry, A. C., Dinkar, T., Rieser, V., & Talat, Z. (2023). Mirages. On anthropomorphism in dialogue systems. *The 2023 conference on empirical methods in natural language processing*. <https://openreview.net/forum?id=i65hZUPWuQ>.
- Anderson, C., & Kilduff, G. J. (2009). Why do dominant personalities attain influence in face-to-face groups? The competence-signaling effects of trait dominance. *Journal of Personality and Social Psychology*, 96(2), 491–503. <https://doi.org/10.1037/a0014201>
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1), 388–407. <https://doi.org/10.3758/s13428-019-01237-x>
- Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43(3), 761–770. <https://doi.org/10.3758/s13428-011-0075-y>
- Balas, B., & Pacella, J. (2017). Trustworthiness perception is disrupted in artificial faces. *Computers in Human Behavior*, 77, 240–248. <https://doi.org/10.1016/j.chb.2017.08.045>
- Barrington, S., & Farid, H. (2024). *People are poorly equipped to detect AI-powered voice clones* (No. arXiv:2410.03791). [arXiv. https://doi.org/10.48550/arXiv.2410.03791](https://doi.org/10.48550/arXiv.2410.03791)
- Baus, C., McAleer, P., Marcoux, K., Belin, P., & Costa, A. (2019). Forming social impressions from voices in native and foreign languages. *Scientific Reports*, 9(1), 414. <https://doi.org/10.1038/s41598-018-36518-6>
- Bruder, C., Breda, P., & Larrouy-Maestri, P. (2023). Attractiveness and social appeal of synthetic voices. *Proceedings of the 23rd conference of the European society for cognitive Psychology (ESCoP)*. <https://doi.org/10.17605/OSF.IO/9DYQE>
- Cabral, J. P., Cowan, B. R., Zibrek, K., & McDonnell, R. (2017). The influence of synthetic voice on the evaluation of a virtual character. *Interspeech*, 229–233. <https://doi.org/10.21437/Interspeech.2017-325>, 2017.
- Cercas Curry, A. C., Abercrombie, G., & Rieser, V. (2021). *ConvAbuse: Data, analysis, and Benchmarks for nuanced abuse Detection in conversational AI* (No. arXiv:2109.09483). [arXiv. https://doi.org/10.48550/arXiv.2109.09483](https://doi.org/10.48550/arXiv.2109.09483)
- Cercas Curry, A., & Rieser, V. (2018). #MeToo alexa: How conversational systems respond to sexual harassment. In M. Alfano, D. Hovy, M. Mitchell, & M. Strube (Eds.), *Proceedings of the second ACL workshop on ethics in natural language processing* (pp. 7–14). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0802>.
- Christoforakos, L., Gallucci, A., Surmava-Große, T., Ullrich, D., & Diefenbach, S. (2021). Can robots earn our trust the same way humans do? A systematic exploration of competence, warmth, and anthropomorphism as determinants of trust development in hri. *Frontiers in Robotics and AI*, 8, Article 640444. <https://doi.org/10.3389/frobt.2021.640444>
- Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. In *Advances in experimental social Psychology* (Vol. 40, pp. 61–149). Academic Press. [https://doi.org/10.1016/S0065-2601\(07\)00002-0](https://doi.org/10.1016/S0065-2601(07)00002-0)
- Dinkar, T., Clavel, C., & Vasilescu, I. (2023). *Fillers in spoken language understanding: Computational and psycholinguistic perspectives* (No. arXiv:2301.10761). [arXiv. https://doi.org/10.48550/arXiv.2301.10761](https://doi.org/10.48550/arXiv.2301.10761)
- Domingo, Y., Holmes, E., & Johnsrude, I. S. (2020). The benefit to speech intelligibility of hearing a familiar voice. *Journal of Experimental Psychology: Applied*, 26(2), 236–247. <https://doi.org/10.1037/xap0000247>
- Eysel, F., Kuchenbrandt, D., Bobinger, S., de Ruiter, L., & Hegel, F. (2012). 'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 125–126. <https://doi.org/10.1145/2157689.2157717>
- Festerling, J., & Siraj, I. (2022). Anthropomorphizing technology: A conceptual review of anthropomorphism research and how it relates to children's engagements with digital voice assistants. *Integrative Psychological and Behavioral Science*, 56(3), 709–738. <https://doi.org/10.1007/s12124-021-09668-y>
- Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2), 77–83. <https://doi.org/10.1016/j.tics.2006.11.005>
- Graeff, T. R. (2005). Response bias. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 411–418). Elsevier. <https://doi.org/10.1016/B0-12-369398-5/00037-2>
- Graux, J., Gomot, M., Roux, S., Bonnet-Brilhaut, F., & Bruneau, N. (2015). Is my voice just a familiar voice? An electrophysiological study. *Social Cognitive and Affective Neuroscience*, 10(1), 101–105. <https://doi.org/10.1093/scan/nsu031>
- Guldner, S., Lavan, N., Lally, C., Wittmann, L., Nees, F., Flor, H., et al. (2024). Human talkers change their voices to elicit specific trait percepts: Brief report. *Psychonomic Bulletin & Review*, 31, 209–222. <https://doi.org/10.3758/s13423-023-02333-y>, 1 vom: Feb.
- Har-shai Yahav, P., Sharaabi, A., & Zion Golumbic, E. (2024). The effect of voice familiarity on attention to speech in a cocktail party scenario. *Cerebral Cortex*, 34(1), Article bhad475. <https://doi.org/10.1093/cercor/bhad475>
- Herrmann, B. (2023). The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults. *International Journal of Speech Technology*, 26(2), 395–415. <https://doi.org/10.1007/s10772-023-10027-y>
- Holmes, E., & Johnsrude, I. S. (2020). Speech spoken by familiar people is more resistant to interference by linguistically similar speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(8), 1465. <https://doi.org/10.1037/xlm0000823>
- Holmes, E., To, G., & Johnsrude, I. S. (2021). How long does it take for a voice to become familiar? Speech intelligibility and voice recognition are differentially sensitive to

- voice training. *Psychological Science*, 32(6), 903–915. <https://doi.org/10.1177/0956797621991137>
- Holzman, P. S., Berger, A., & Rousey, C. (1967). Voice confrontation: A bilingual study. *Journal of Personality and Social Psychology*, 7(4, Pt.1), 423–428. <https://doi.org/10.1037/h0025233>
- Hughes, S. M., & Harrison, M. A. (2013). I like my voice better: Self-enhancement bias in perceptions of voice attractiveness. *Perception*, 42(9), 941–949. <https://doi.org/10.1068/p7526>
- Hulzebosch, N., Ibrahim, S., & Worring, M. (2020). Detecting CNN-generated facial Images in real-world scenarios (No. arXiv:2005.05632). *arXiv*. <https://doi.org/10.48550/arXiv.2005.05632>
- IEEE Recommended Practice for Speech Quality Measurements. (1969). IEEE transactions on audio and electroacoustics. *IEEE Transactions on Audio and Electroacoustics*, 17(3), 225–246. <https://doi.org/10.1109/TAU.1969.1162058>
- Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, 24(10), 1995–2004. <https://doi.org/10.1177/0956797613482467>
- Kanber, E., Lavan, N., & McGettigan, C. (2022). Highly accurate and robust identity perception from personally familiar voices. *Journal of Experimental Psychology: General*, 151, 897–911. <https://doi.org/10.1037/xge0001112>
- Kirk, N. W., & Cunningham, S. J. (2024). Listen to yourself! Prioritization of self-associated and own voice cues. *British Journal of Psychology*, (n/a), 1018. <https://doi.org/10.1111/bjop.12741>. n/a.
- Klofstad, C. A., Anderson, R. C., & Nowicki, S. (2015). Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. *PLoS One*, 10(8), Article e0133779. <https://doi.org/10.1371/journal.pone.0133779>
- Kühne, K., Fischer, M. H., & Zhou, Y. (2020). The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. Evidence from a subjective ratings study. *Frontiers in Neuroinformatics*, 14. <https://www.frontiersin.org/articles/10.3389/fninf.2020.593732>
- Lan, M., Peng, M., Zhao, X., Chen, H., Liu, Y., & Yang, J. (2022). Facial attractiveness is more associated with individual warmth than with competence: Behavioral and neural evidence. *Social Neuroscience*, 17(3), 225–235. <https://doi.org/10.1080/17470919.2022.2069152>
- Lavan, N. (2023a). How do we describe other people from voices and faces? *Cognition*, 230, Article 105253. <https://doi.org/10.1016/j.cognition.2022.105253>
- Lavan, N. (2023b). The time course of person perception from voices: A behavioral study. *Psychological Science*, 34(7), 771–783. <https://doi.org/10.1177/09567976231161565>
- Lavan, N. (2024). Studying person perception from voices: Creating common ground by looking beyond accuracy. *The Cognitive Psychology Bulletin*. <https://doi.org/10.53841/bpsocg.2024.1.9.40>
- 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. <https://doi.org/10.3758/s13423-018-1497-7>
- Lavan, N., Irvine, M., Rosi, V., & McGettigan, C. (2024). Voice deep fakes sound realistic but not (yet) hyperrealistic. *OSF*. <https://doi.org/10.31234/osf.io/jqg6e>
- Lavan, N., Rinke, P., & Scharinger, M. (2024). The time course of person perception from voices in the brain. *Proceedings of the National Academy of Sciences*, 121(26), Article e2318361121. <https://doi.org/10.1073/pnas.2318361121>
- Lee, M., Drinnan, M., & Carding, P. (2005). The reliability and validity of patient self-rating of their own voice quality. *Clinical Otolaryngology*, 30(4), 357–361. <https://doi.org/10.1111/j.1365-2273.2005.01022.x>
- Mahrholz, G., Belin, P., & McAleer, P. (2018). Judgements of a speaker's personality are correlated across differing content and stimulus type. *PLoS One*, 13(10), Article e0204991. <https://doi.org/10.1371/journal.pone.0204991>
- Maurer, D., & Landis, T. (2009). Role of bone conduction in the self-perception of speech. *Folia Phoniatrica et Logopaedica*, 42(5), 226–229. <https://doi.org/10.1159/000266070>
- McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. *PLoS One*, 9(3), Article e90779. <https://doi.org/10.1371/journal.pone.0090779>
- McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., et al. (2015). librosa: Audio and music signal analysis in python. *SciPy*, 18–24. <https://www.academia.edu/download/40296500/librosa.pdf>
- McGettigan, C. (2015). The social life of voices: Studying the neural bases for the expression and perception of the self and others during spoken communication. *Frontiers in Human Neuroscience*, 9. <https://doi.org/10.3389/fnhum.2015.00129>
- McGettigan, C., Bloch, S., Bowles, C., Dinkar, T., Lavan, N., Reus, J., et al. (2024). Voice cloning: Psychological and ethical implications of intentionally synthesising familiar voice identities. *OSF*. <https://doi.org/10.31234/osf.io/29jyq>. <https://osf.io/29jyq/download>
- Mileva, M., & Lavan, N. (2023). Trait impressions from voices are formed rapidly within 400 ms of exposure. *Journal of Experimental Psychology: General*, 152(6), 1539–1550. <https://doi.org/10.1037/xge0001325>
- Mileva, M., Tompkinson, J., Watt, D., & Burton, A. M. (2020). The role of face and voice cues in predicting the outcome of student representative elections. *Personality and Social Psychology Bulletin*, 46(4), 617–625. <https://doi.org/10.1177/0146167219867965>
- Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., & Dawel, A. (2023). AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychological Science*, 34(12), 1390–1403. <https://doi.org/10.1177/09567976231207095>
- Naunheim, M. R., Puka, E., & Huston, M. N. (2023). Do you like your voice? A population-based survey of voice satisfaction and voice enhancement. *The Laryngoscope*, 133(12), 3455–3461. <https://doi.org/10.1002/lary.30822>
- Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8), Article e2120481119. <https://doi.org/10.1073/pnas.2120481119>
- Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., & Sorokowska, A. (2017). Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic Bulletin & Review*, 24(3), 856–862. <https://doi.org/10.3758/s13423-016-1146-y>
- Orepic, P., Kannape, O. A., Faivre, N., & Blanke, O. (2023). Bone conduction facilitates self-other voice discrimination. *Royal Society Open Science*, 10(2), Article 221561. <https://doi.org/10.1098/rsos.221561>
- Payne, B., Addelee, A., Rieser, V., & McGettigan, C. (2024). Self-ownership, not self-production, modulates bias and agency over a synthesised voice. *Cognition*, 248, Article 105804. <https://doi.org/10.1016/j.cognition.2024.105804>
- Payne, B., Lavan, N., Knight, S., & McGettigan, C. (2021a). Perceptual prioritization of self-associated voices. *British Journal of Psychology*, 112(3), 585–610. <https://doi.org/10.1111/bjop.12479>
- Payne, B., Lavan, N., Knight, S., & McGettigan, C. (2021b). Perceptual prioritization of self-associated voices. *British Journal of Psychology*, 112(3), 585–610. <https://doi.org/10.1111/bjop.12479>
- Pehlivanoglu, D., Zhu, M., Zhen, J., Gagnon-Roberge, A., Kern, R., Woodard, D., et al. (2024). Is this real? Susceptibility to deepfakes in machines and humans. *OSF*. <https://doi.org/10.31219/osf.io/etxzw>
- Peng, Z., Hu, Z., Wang, X., & Liu, H. (2020). Mechanism underlying the self-enhancement effect of voice attractiveness evaluation: Self-positivity bias and familiarity effect. *Scandinavian Journal of Psychology*, 61(5), 690–697. <https://doi.org/10.1111/sjop.12643>
- Peng, Z., Wang, Y., Meng, L., Liu, H., & Hu, Z. (2019). One's own and similar voices are more attractive than other voices. *Australian Journal of Psychology*, 71(3), 212–222. <https://doi.org/10.1111/ajpy.12235>
- Pinheiro, A. P., Sarzedas, J., Roberto, M. S., & Kotz, S. A. (2023). Attention and emotion shape self-voice prioritization in speech processing. *Cortex*, 158, 83–95. <https://doi.org/10.1016/j.cortex.2022.10.006>
- Romportl, J. (2014). Speech synthesis and uncanny valley. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, speech and dialogue* (pp. 595–602). Springer International Publishing. https://doi.org/10.1007/978-3-319-10816-2_72
- Rosi, V., Payne, B., & McGettigan, C. (2024). Effects of self-similarity and self-generation on the perceptual prioritisation of voices. *OSF*. <https://doi.org/10.31234/osf.io/aq4tn>
- Scott, S., & McGettigan, C. (2016). The voice: From identity to interactions. In D. Matsumoto, H. C. Hwang, & M. G. Frank (Eds.), *APA handbook of nonverbal communication* (pp. 289–305). American Psychological Association. <https://doi.org/10.1037/14669-011>
- Shiramizu, V. K. M., Lee, A. J., Altenburg, D., Feinberg, D. R., & Jones, B. C. (2022). The role of valence, dominance, and pitch in perceptions of artificial intelligence (AI) conversational agents' voices. *Scientific Reports*, 12(1), Article 22479. <https://doi.org/10.1038/s41598-022-27124-8>
- Sutton, S. J., Foulkes, P., Kirk, D., & Lawson, S. (2019). Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. *Proceedings of the 2019 CHI conference on human factors in computing systems*. <https://doi.org/10.1145/3290605.3300833>
- Sweet, H. (1890). *A primer of spoken English*. Clarendon Press.
- Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior*, 33(3), 210–216. <https://doi.org/10.1016/j.evolhumbehav.2011.09.004>
- Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, 12(12), 455–460. <https://doi.org/10.1016/j.tics.2008.10.001>
- Torre, I., Goslin, J., White, L., & Zanotto, D. (2018). Trust in artificial voices: A 'congruency effect' of first impressions and behavioural experience. *Proceedings of the technology. Mind, and Society*. <https://doi.org/10.1145/3183654.3183691>
- Tucciarelli, R., Vehar, N., Chandaria, S., & Tsakiris, M. (2022). On the realism of people who do not exist: The social processing of artificial faces. *iScience*, 25(12), Article 105441. <https://doi.org/10.1016/j.isci.2022.105441>
- Wilson, S., & Moore, R. K. (2017). Robot, alien and cartoon voices: Implications for speech-enabled systems. *Proceedings of the 1st international workshop on vocal interactivity in-and-between humans, animals and robots*.
- Zhang, L., Jiang, L., Washington, N., Liu, A. A., Shao, J., Fournay, A., et al. (2021). Social media through voice: Synthesized voice qualities and self-presentation. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), 161:1–161:21. <https://doi.org/10.1145/3449235>