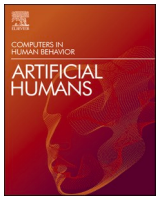
Computers in Human Behavior: Artificial Humans 6 (2025) 100211

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



Attractive synthetic voices

Camila Bruder [a,*], Pamela Breda [b], Pauline Larrouy-Maestri [a]

[a] Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
[b] University of Applied Arts, Vienna, Austria

ARTICLE INFO

Keywords:
TTS
Text-to-speech
Attractiveness
Emotional prosody
Artificial intelligence
AI voices

ABSTRACT

With recent advances in Artificial Intelligence (AI), synthetic voices have become increasingly prevalent in our everyday soundscape. This study examined listeners' perception of human and neural Text-To-Speech (TTS) voices. In an online experiment, 75 participants listened to different versions of a short utterance spoken by eight different voices (half human, half TTS), each presented in four expressed emotions (neutral, happy, sad, angry). For each stimulus, participants rated voice attractiveness and willingness to interact, and selected the perceived emotion from a forced-choice list. In a second part, participants were asked to classify each voice as human or AI-generated. Results revealed that participants were often "fooled" by the TTS voices, misidentifying them as human. Voice ratings were influenced by the perceived emotion regardless of the voice type, with happy-sounding voices rated more positively than those perceived as sad or angry. However, TTS voices were rated as less attractive and socially appealing overall, though with large individual differences. These findings indicate that TTS voices are approaching human ones in how they are perceived by listeners, highlighting progress in their naturalness.

1. Introduction

Voices are an intrinsic part of human communication. Beyond conveying semantic meaning (based on "what" is said), speech conveys information through the voice, based on "how" something is said. Voices can cue a speaker's body size, health and age (Pisanski et al., 2014, 2016), convey emotional states (Banse & Scherer, 1996; Larrouy--Maestri et al., 2024; Scherer, 2021; van Rijn & Larrouy-Maestri, 2023), intent (Hellbernd & Sammler, 2016), and personality traits (e.g., Goupil et al., 2021; McAleer et al., 2014; Scherer, 1978). A speaker's voice thus plays a key, multifaceted role in social interaction and interpersonal dynamics.

Until the early 21st century, most people interacted exclusively with human voices and were not used to listening to synthetically-generated ones. However, recent technological advances, particularly through the use of deep neural networks and generative models, have led to remarkable improvements in the quality, accessibility, and pervasiveness of synthetic voices. Currently, deep neural networks trained with large amounts of data can "learn" the combination of expressive elements in the human voice, recognizing patterns in the data, and use that information to synthesise new, naturalistic-sounding voices (Mu et al., 2021). Neural Text-to-Speech (TTS) voices (i.e., generated with deep neural networks) can be easily generated based on a text prompt and are accessible to a broad range of users through cost-effective, user-friendly online platforms. These platforms typically offer numerous options of voices in different languages and/or accents, in some cases with multiple profiles for different use cases (e.g., narration, conversational) and emotional profiles (e.g., happy, terrified - see Murf AI, 2025). These voices are used in entertainment channels through voiceovers accompanying videos, podcasts, and audiobooks, as well as in education, customer services, automated systems and voice assistants (e.g., Cortana, Siri, Google Assistant). In 2022, over 80 % of people in the UK and US reported being familiar with voice technology to some extent, compared to 62 % in Germany (Vixen Labs, 2022). In the US alone, the number of voice assistant users reached 150 million in 2024, and continues to grow, with users aged 25–49 as the most frequent users (eMarketer, 2024). Synthetic voices thus increasingly permeate shared spaces and shift our soundscape in unprecedented ways.

* Corresponding author.
E-mail addresses: cajmila@gmail.com (C. Bruder), plm@ae.mpg.de (P. Larrouy-Maestri).

https://doi.org/10.1016/j.chbah.2025.100211
Received 26 July 2025; Received in revised form 25 September 2025; Accepted 26 September 2025
Available online 3 October 2025

Considering the increasing pervasiveness of synthetic voices, situations where people are not sure if they are listening to a human- or a synthetically-generated voice (or interacting with a human versus an AI agent) become increasingly common. It is thus of utmost importance to explore how such voices are perceived in the first place, in order to establish, based on empirical evidence, whether further research should focus on the potential consequences of the widespread presence of synthetic speech.

Studies indicate that earlier TTS voices (i.e., resulting from concatenative and/or formant-based synthesis approaches) were consistently evaluated less favorably than human voices (Abdulrahman & Richards, 2022; Mullennix et al., 2003; Stern et al., 1999) and were perceived as more cognitively demanding to process (Luce, 1982; Roring et al., 2007). While the likeability of synthetic voices has been increasing along with their perceived human-likeness (Baird et al., 2018; Kühne et al., 2020; Zhu et al., 2022), studies suggest that neural TTS voices are still perceived less favorably than human ones (Baird et al., 2018; Gessinger et al., 2022, 2023; Herrmann, 2023; Kühne et al., 2020; Le Maguer et al., 2024). For instance, Gessinger et al. showed that Amazon's Alexa TTS voice in US-English as well an in German (in its standard, neutral profile) was rated as less human-like, less natural, less comforting, and less warm than a control human voice, by both German and US participants (Gessinger et al., 2022, 2023).

On the other hand, voiceover providers claim that current neural TTS voices are "realistic", "human-like", "natural-sounding", or "barely distinguishable from an authentic human voice" (e.g., ElevenLabs, 2024; Murf AI, 2025; Speechify, n.d.) – though such claims are not supported by openly available empirical data. Developers often report subjective Mean Opinion Scores (MOS), an average of listener evaluations, and/or Comparative MOS (CMOS), which entail pairwise comparisons between recordings. For instance, Tan et al. (2024) reported that their NaturalSpeech system has already achieved human-like quality, with no statistically significant difference between MOS and CMOS of their system and those of natural speech (i.e., human recordings). While these are impressive achievements, it is important to note a major limitation of MOS and CMOS as measures of preference: they are based on arithmetic means, which assume homogeneity among subjects (Streijl et al., 2016; Xu et al., 2011), even though no measure of inter-rater reliability is usually reported, thus ignoring individual differences across listeners. This is a significant, though largely overlooked, caveat, given that recent studies indicate a substantial influence of individual differences (i.e., preferences that are not shared across participants) in the perception of social traits such as vocal attractiveness (Lavan & Sutherland, 2024) and in the overall aesthetic enjoyment ("liking") of speaking and singing voices (Bruder et al., 2024). Another important issue is the lack of adherence to standards in speech system evaluation. Although the International Telecommunication Union (ITU, 1994, 1996, 2016) provides clear guidelines, including recommended test types and detailed procedures and reporting practices, recent evidence shows that these standards are often overlooked, with insufficient reporting of instructions given to participants (e.g., whether the focus was on quality or naturalness), the scale resolution (e.g., half- or full-point increments), or listener characteristics (Kirkland et al., 2023). Moreover, with the impressive recent advances in the quality of neural TTS, studies have stressed the need for new evaluation methods capable of capturing the increasingly subtle and localized differences between synthesized and natural speech (Le Maguer et al., 2024; Perrotin et al., 2025; Wagner et al., 2019).

In addition to research on synthetic speech evaluation, a substantial body of work has examined how humans perceive voices. This literature provides extensive insights into the psychoacoustic, biological, and social bases of voice perception in general, and of voice attractiveness in particular. Evolutionary accounts of voice attractiveness mostly frame it within a sexual selection framework (e.g., "the desirability of vocalizations to potential mates" – Hill & Puts, 2016), linking it to acoustic cues such as a voice's fundamental frequency (Collins, 2000; Collins & Missing, 2003; Feinberg et al., 2005; see Pisanski & Feinberg, 2018 for a review) or body traits (Hughes et al., 2004). However, there is not one clear and consistent definition across studies. The concept of voice attractiveness is rather context-dependent and seems to encompass multiple types of attraction (e.g., political charisma, business leadership, nonsexual attraction, romantic desirability – Rosenberg & Hirschberg, 2021; Trouvain et al., 2021). Importantly, more nuanced accounts of voice attractiveness have been proposed, highlighting influences such as the degree to which a voice conforms to community speech norms (Babel et al., 2014), or the valence of the spoken text (Jones et al., 2008). Presumably, there is some overlap between the concepts of voice attractiveness studied in psychology and the notion of voice pleasantness used in speech evaluation research – in the sense that an attractive voice is likely also pleasant. In the ITU-T P.85 recommendation, voice pleasantness is defined to participants as "your attitude to the voice and whether or not you found it pleasant to listen to" (International Telecommunication Union, 1994). However, research suggests that the two concepts, though related, are not identical. For instance, Petty et al. (2022) reported that gay men rated female voices as pleasant but not attractive (while rating male voices as both pleasant and attractive).

Individual differences play a central role in shaping how voices are evaluated. Age, in particular, has been identified as an important source of variability. For instance, Herrmann (2023) found that older participants were less likely to recognize synthetic voices (specifically, Google's Wavenet TTS) and rated them as sounding more natural and human-like compared to younger participants. Similarly, Müller et al. (2022) reported higher accuracy of recognition of synthetic voices (from an audio deepfake dataset) among younger participants and native speakers, while finding no significant effect of self-reported IT experience. Cohn et al. (2020) explored the role of individuals' cognitive processing style in the perception of Amazons' Alexa TTS voice. They used the Autism Quotient (AQ) survey (Baron-Cohen et al., 2001), which was designed to quantify the extent of autistic-like traits in adults of normal intelligence in a non-clinical setting, but also captures variation within neurotypical populations (though interpretation of this scale in non-clinical settings is controversial, see Taylor et al., 2020). With neurotypical university students as participants, they found that subjects with higher AQ scores (indicating stronger autistic-like traits) were more likely to provide distinct ratings of likeability and human-likeness for the Alexa and human voices (lower for Alexa than human).

A closely related question concerns how subjects respond to increasingly expressive TTS voices. Gessinger et al. (2022, 2023) manipulated the expressiveness of voices in the direction of "happiness", making clear to participants if they were listening to a human or a TTS voice (specifically, Amazon's Alexa voice in English and German). They reported that both German and US participants rated stimuli as having higher arousal with increasing "happiness" manipulation, both for human and TTS voices – while the pattern for valence ratings was less straightforward. These broadly comparable trends in arousal ratings suggest that participants may have responded to the expressive cues in a similar manner, regardless of the voice's origin. Further insight comes from studies of user interactions with voice chatbots. Zhu et al. (2022) manipulated the expressiveness of a bot's responses (with Alexa's US-English voice), and reported that listeners were sensitive to the addition of "expressiveness" to the bot's voice, but this did not affect listeners' ratings of the bot's human-likeness, likability, or of how

engaging participants thought the conversation was. On the other hand, the perception of these interactions by independent participants was indeed modulated by expressivity (i.e., increase in likeability ratings with increasing perceived human-likeness, especially in the condition featuring expressive prosody). These findings support the need for further research on the specific role of expressivity in user experience and behavior (Aylett et al., 2021).

In this study, rather than asking participants to judge naturalness or quality, we focused on two dimensions of voice perception that are likely related but distinct: attractiveness and social appeal. We aimed to test: a) whether subjects can (still) discriminate neural TTS voices from human ones; and b) whether neural TTS voices already sound as attractive and socially appealing as human ones. To achieve greater variability in the stimulus set and promote engagement with the voices, we included neural TTS (and human) voices with different emotional profiles – though it should be noted that the accuracy of emotional profile classification, while informative, was not our main objective. In line with previous reports of individual differences in attitudes toward artificial agents (e.g., Cohn et al., 2020; Herrmann, 2023; Müller et al., 2022; Zellou et al., 2021), we also examined the role of participants' age and cognitive processing style in their perception of TTS voices. Note that, throughout this paper, we use the terms "synthetic" (in the sense of not natural or not produced by a human vocal apparatus), "AI" (to refer to voices generated using artificial intelligence), and "TTS voices" interchangeably despite the technical distinctions that could be made between them.

2. Method

2.1. Participants

Seventy-five participants (39 self-reported as male, 34 self-reported as female, one non-binary, one undisclosed), with an average age of 39.2 years old ($SD = 12.1$, range: 19–76) were recruited using Prolific (www.prolific.com) and paid an hourly rate of 9£. Inclusion criteria were to be located in the USA, to have English as first language, and to have an approval rate of 95 % or higher in previous studies in the Prolific platform. Informed consent was obtained from all participants prior to data collection.

2.2. Material

For the human voices, we selected a subset of audio recordings from the RAVDESS dataset (Livingstone & Russo, 2018), which consists of recording of actors speaking and singing the same sentences with different emotional profiles, in two different emotional intensities (i.e., normal and strong). We used the speech condition from actors number 01, 02, 11 and 12 (two males, two females), and the sentence "Kids are talking by the door," in the "normal" intensity, with four emotional profiles: calm, happy, sad, and angry. For the TTS voices, we generated (in December 2022) the same sentence (i.e., "Kids are talking by the door") with two different TTS providers. From the platform murf (https://murf.ai), we used the voices "Nate" and "Ava" (male and female, respectively), each with four emotional profiles: friendly, cheerful, sad, and angry. From the platform lovo's Voice Lab (beta version; https://voicelab.lovo.ai), we used the voices "Rick" and "Kim" (male and female, respectively), with four emotional profiles: narrative, admiration, tired and furious. A total of 32 stimuli was thus included, comprising eight different voices (four human- and four computer-generated; half male, half female), each with four different emotional profiles. Note that the labels of emotional profiles offered by the TTS providers did not match the ones from RAVDESS perfectly, but we matched the valence and arousal levels as much as possible while selecting stimuli. RAVDNESS audio files were slightly edited in Audacity to remove breathing sounds at beginning of sentences, to better match them to the TTS voices. We used the software To Audio Converter (version 1.0.16–1059) to normalize all stimuli to the same loudness level of −14 Loudness Units relative to Full Scale (LUFS). We followed the EBU R128 standard (European Broadcasting Union, 2023), which uses an integrated loudness measure that aligns better with human perception of loudness than RMS or peak normalization. All stimuli are openly available at https://osf.io/4qfaz/.

2.3. Procedure

The experimental procedure was ethically approved by the Ethics Council of the Max Planck Society (No. 2017_12). The online experiment was implemented in Labvanced (Finger et al., 2017). The task, illustrated in Fig. 1, was completed on average in 15.6 min ($SD = 6.3$).

The experiment began with general instructions informing participants that, in a voice perception experiment, they would listen to short recordings and be asked to rate them in different scales. Participants were asked to use headphones and adjust the volume to a comfortable level. In Part 1 (Fig. 1, left), the instructions read: *'Click on the "Play" button to listen to the voices as many times as you would like, and rate the voices on the presented scales'*. Participants then answered three questions on each trial: *'How attractive is this voice?'* and *'How much would you like to interact with this person?'* (both on 7-point scales ranging from 1 = not at all to 7 = a lot); and *'How does this voice sound?'* with response options *happy, sad, angry,* or *none of the above*. The visual arrangement of the rating questions was counterbalanced across participants. After answering all questions, participants proceeded to the next trial. All 32 stimuli were presented in randomized order within a single block. After this block, a random subset of eight stimuli was presented again, to be used exclusively for intra-rater agreement analysis. Part 2 of the experiment (Fig. 1, right) began with the question: *'Did you suspect some of these voices were generated with a computer?'* (response options: *yes* or *no*). This was followed by the instructions: *"In the next page, please sort the voices based on your impression of if they were human or generated with a computer (using artificial intelligence - AI). To do that, click on the play button to listen to the voice as many times as you'd like and then choose if it is 'human' or 'AI'. If you are not sure, take a guess."* On the following screen, all 32 stimuli were displayed simultaneously in three columns (with column order counterbalanced across participants). As before, participants could listen to each stimulus as many times as they liked before choosing between human and AI. After completion of Part 2 (Fig. 1, right), participants reported information about age, gender, and sexual identity/orientation and mother tongue, indicated whether they performed the task conscientiously (open-ended response), and had the opportunity to leave final comments. Finally, participants completed the short version of the self-report Autism-Spectrum Quotient (AQ10) (Allison et al., 2012). While originally developed as a diagnostic screening tool for autism, the questionnaire has also been employed to assess individual differences in cognitive processing styles (e.g., Cohn et al., 2020). The AQ-10 consists of ten questions, with two items for each of five domains: Attention to Detail, Attention Switching, Communication, Imagination, and Social Skills. Responses are provided on a 4-point Likert scale: Definitely Agree, Slightly Agree, Slightly Disagree, and Definitely Disagree. However, it is important to note that the AQ10 has been criticized for limited psychometric robustness in non-clinical populations (Taylor et al., 2020); thus, its use in the present study is exploratory and any findings should be interpreted with caution.

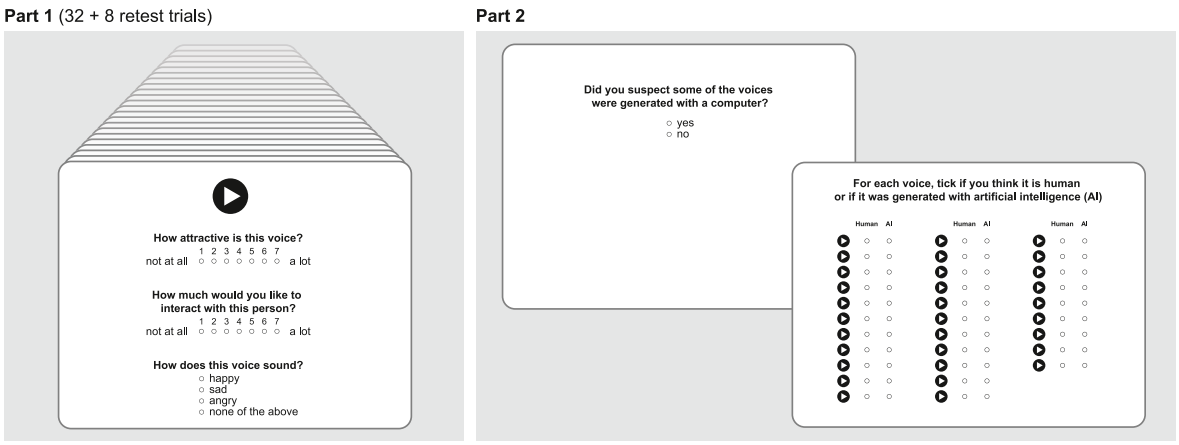


Fig. 1. Illustration of the experimental procedure. In Part 1, participants provided ratings of attractiveness and willingness to interact on 7-point scales ranging from 1 (not at all) to 7 (a lot), and indicated the perceived emotional profile (forced-choice response among "happy," "sad," "angry," or "none of the above"). In Part 2, participants first indicated if they suspected some of the voices were generated with a computer, to then indicate if they thought each voice was human or generated with a computer (using artificial intelligence - AI).

2.4. Statistical analyses

All analyses were performed using R Statistical Software (version 4.1.2; R Core Team, 2021) and R Studio (version 2022.7.1.554; RStudio Team, 2022). The analyses reported here were preregistered (https://osf.io/g4sq2).

2.4.1. Attractiveness and willingness to interact

To test whether voice attractiveness and willingness to interact ratings differed between human and TTS voices, we analyzed trial-level rating data using generalized linear mixed-effects models implemented with the lmer function from the lme4 package in R (Bates et al., 2015), which is well-suited to handle the nested and repeated measures nature of our data. These models included fixed effects for voice type (human vs. TTS), the perceived emotional profile[1] (neutral, happy, sad, and angry), and their interaction, as well as participant-level covariates such as Autism Quotient (AQ) scores and age, including an interaction term between age and voice type.[2] To account for variability among individuals and stimuli, we included random intercepts for participants and stimulus items, with stimuli nested within voice, and allowed participants to vary in their responses to voice type and perceived emotion through random slopes. We applied contrast coding, setting 'human' and 'neutral' as reference levels for voice type and perceived emotion, respectively, to facilitate interpretation of the model estimates. Model predictions and associated confidence intervals (CI) were extracted using the ggeffects package (Lüdecke, 2018), enabling us to visualize the estimated effects while accounting for the complexities of the mixed-effects models. The syntax of both models was: lmer(rating ~ voice type*perceived emotion + AQ + age + age:voice type + (1 + perceived emotion + voice type | participant) + (1 | voice/stimulus)).

2.4.2. Emotional profile classification

To compare the accuracy of emotional profile classification between human and synthetic voices, we compared the proportion of correct responses for the two categories in two ways: 1) with all emotions pooled together, using a one sample Z-test of proportion (two tailed); and 2) separately for each emotion, with four planned pairwise comparisons. We used the Holm method (Holm, 1979) to control for these multiple comparisons, and report adjusted p-values, while keeping the significance threshold at the conventional alpha level of 0.05. Note that these results are reported for thoroughness and transparency, with the limitation that the emotional profile labels from different voice-over providers did not perfectly match the human ones, thus not allowing for a strictly fair comparison.

2.4.3. Intra- and inter-rater reliability

To assess inter- and intra-rater reliability, we used Krippendorff's alpha, which is a generalization of several well-known reliability coefficients, widely applicable across different measurement scales (Krippendorff, 2004, 2011). This measure is robust to missing data and supports various data types. In this study, we applied the interval metric for quantitative measures such as attractiveness and willingness to interact, and the nominal metric for categorical data such as classification of emotional profile and voice type. Krippendorff's alpha values range from 0 (no agreement beyond chance) to 1 (perfect agreement). We interpreted these values according to the guidelines of Landis and Koch (1977), where values below 0.20 indicate slight agreement; 0.21 to 0.40 fair agreement; 0.41 to 0.60 moderate agreement; 0.61 to 0.80 substantial agreement; and values above 0.81 almost perfect agreement. The kripp.alpha function from the irr package was used for the calculations. We used Krippendorff's alpha to estimate the extent of agreement among participants' ratings (inter-rater reliability), and to assess self-agreement over repeated trials (intra-rater or test-retest reliability). For the latter, we focused on the eight repeated trials presented at the end of the first experimental block. Additionally, we report the

[1] Though the contrasted emotional profiles aimed at adding variability in the sample, we also examined their potential role in the attractiveness/willingness to interact with the voices.

[2] This model diverged from the tentative one proposed in the preregistration (rating ~ voice type * Emotion + AQ + (1 + voice type | Subject) + (1 | Stimulus)) in the structure of the random effects and the inclusion of participants' age as a predictor, as well as in the fact that we ultimately concluded that it makes more sense to predict (perceived) attractiveness and willingness to interact ratings from the perceived emotional profile than from the "ground truth" emotional profile. This choice was both conceptually motivated (as attractiveness and social appeal are subjective impressions) and supported by model comparisons: models equivalent to the ones reported here using the ground-truth emotional profile instead of participants' answers of perceived emotional profile had convergence issues and showed substantially poorer fit, as evaluated through model comparisons based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). In these model comparisons, lower values indicate better fit to the data while accounting for model complexity. For attractiveness ratings, AIC was 7913.4 in the preregistered model (vs. 7689.3 in the reported model), and BIC was 8080.7 (vs. 7856.6). For willingness to interact ratings, AIC was 7784.6 in the preregistered model (vs. 7302.2 in the reported model), and BIC was 7951.9 (vs 7469.5). The preregistered models are fully reported in the online supplementary materials at https://osf.io/4qfaz/.

Pearson correlation of test-retest responses for each participant. Note that no repeated trials were included in the voice classification task during the second part of the experiment, preventing estimation of intra-rater reliability for this particular task.

2.4.4. Human vs AI classification

According to the preregistration, to test if participants could distinguish TTS from human voices, we compared the proportion of correct classification for human and TTS voices using a one sample z-test of proportions. We also computed unbiased hit rates (separately for human and AI voices) according to Wagner (1993). These are defined as "the joint probability that a stimulus category is correctly identified given that it is presented at all and that a response is correctly used given that it is used at all" (Wagner, 1993, p. 3). We also followed Wagner (1993) in computing an unbiased chance-level performance estimate as the joint probability of the co-occurrence by chance of a stimulus and response of a corresponding category. Additionally, to explore the influence of the perceived emotion[3] on the accuracy of classification, we fit a mixed effects logistic regression predicting the accuracy of participants responses (at the trial level, coded as 0 = incorrect response, 1 = correct response) from the voice type (Human or AI) and its interaction with the perceived emotion, as well as participants' AQ score and age, including the interaction between age and voice type. We included random intercepts for subjects and stimuli items, with voices nested in stimuli items; and by-participant random slopes for the effect of voice type (convergence issues did not allow including random slopes for the effect of perceived emotion as well – the model syntax was: accuracy ~ voice type *perceived emotion + AQ + age + age:voice type + (1 + voice type | participant) + (1 | voice/stimulus))). We used contrast coding in the same way mentioned in the lmer models mentioned above. Model predictions were exported using the ggeffects function from the ggpredict R package (Lüdecke, 2018), with the argument bias_correction = TRUE specified, following the package developers' recommendation to adjust for the bias inherent in back-transforming from the logit to the probability scale.

3. Results

3.1. Limited inter-rater agreement despite substantial test-retest agreement

Analysis of repeated trials at the end of Part 1 showed substantial test-retest agreement within participants, according to Landis and Koch's (1977) interpretation guidelines. Krippendorff's α[4] was .71 for attractiveness ratings, .77 for willingness-to-interact ratings, and .67 for emotional profile classification. The overall test-retest Pearson correlation, computed individually for each participant, had an average value of .56 (SD = . 34) for attractiveness ratings and of .62 (SD = . 34) for willingness to interact ratings. These results indicate that most participants were self-consistent when evaluating the same stimulus a second time. Conversely, inter-rater agreement was low for both attractiveness and willingness to interact ratings, with Krippendorff's α values of .18 and .21, respectively. Inter-rater agreement was somewhat higher for emotional profile classification (α = .40). In Part 2, the inter-rater reliability for classifying voices as human or AI-generated was fair, with Krippendorff's α of .38.

The limited inter-rater agreement, despite substantial intra-rater consistency, suggests that participants differed systematically in how they evaluated the voices, but did not respond randomly – in other words, it points to large individual differences in how participants evaluated the voices.

3.2. Higher recognition of human voices

At the end of the first part of the experiment, participants were asked whether they suspected that some of the voices they had heard were computer-generated. Seventy-six percent of participants responded "yes" (and 24 % responded "no"). Fig. 2A presents the accuracy of classification for each of the presented voices. Classification accuracy was higher for human voices (85.6 %) than synthetic ones (55.2 %; χ^2 = 264.8, $p < .001$), and both were higher than chance: χ^2 = 606.3 for human voices and χ^2 = 12.6 for AI voices (both adj. $ps < .001$).[5] Individual accuracy ranged from .38 to 1 (M = .86, SD = .16) for human voices and from .19 to 1 (M = .55, SD = .16) for AI voices, indicating substantial variability in classification ability across participants (Supplementary Fig. S1).

As illustrated in Fig. 1C, the mixed-effects logistic model predicting accuracy of responses from the voice type, the perceived emotion, participants' AQ scores, and participants' age revealed a significant main effect of voice type, with lower accuracy of classification for synthetic voices (model estimates: Human: .85, 95 % confidence interval (CI) = [.56, .85]; AI: .54, 95 % CI = [.46, .58]; odds ratio = −.04, i.e., the odds of correctly identifying a TTS voice as AI-generated are about 96 % lower compared to identifying a human voice as human). There was no main effect of perceived emotion, but for voices perceived as angry, there was a significant interaction with the voice type. Specifically, the accuracy for AI voices was even lower when the voice was perceived as angry, indicating that angry AI voices were the most challenging for participants to classify correctly (Fig. 2D). There was no significant effect of cognitive style (as measured by the AQ score) on classification accuracy, but a small main effect of age emerged, with older participants showing lower accuracy (odds ratio = 0.95, i.e., each additional year of age was associated with a 5 % decrease in the odds of responding correctly). However, this effect did not differ between human and AI voices, as indicated by the non-significant interaction between age and voice type. For both types of voice, the average accuracy of classification per participant was negatively correlated with participants' age ($r(73)$ = −.43, $p < .001$ – see Fig. 2B for scatterplots of this relationship separately for AI and human voices). Note that this model achieved low prediction based on its fixed effects (marginal R^2 = .22 conditional R^2 = .56), and inspection of intraclass correlations shows that more variance was captured by participants' random intercepts (ICC = 0.30) than by stimulus items (nested within voices, ICC = 0.19) or by voices themselves (ICC = 0.06). We also exploratory included participants' response about suspecting that some of the voices were computer-generated as a fixed effect in the model reported above (coded as a binary variable: 0 = no, 1 = yes). This variable was not a significant predictor, and its inclusion did not significantly improve model fit, as indicated by a likelihood ratio test, $\chi^2(1) = 0.75$, $p = .39$. In other words, suspecting that some of the voices presented in Part 1 were AI-generated did not predict accuracy in the voice classification task proposed in Part 2 of the experiment.

[3] An alternative model using the *ground-truth* emotional profile instead of participants' *perceived* emotional profile showed a slightly better fit. The alternative model led to AIC = 2050.9 (vs. 2057.1 in the reported model) and BIC = 2143.2 (vs. 2149.4), and yielded a somewhat different pattern of results: voice type was only marginally significant ($p = .062$), and there were significant interactions between voice type and emotional profile for both happy and angry voices. However, given our conceptual focus on participants' perceptions and for consistency across analyses we report the perceived-emotional profile model in the main text. The alternative model is reported in our online supplementary materials at https://osf.io/4qfaz/.

[4] Note that for these data, Krippendorff's alpha values are similar to Cohen's Kappa.

[5] Unbiased hit rates (Wagner, 1993) yielded similar patterns: 56 % for human and 44 % for synthetic voices, above their respective unbiased chance levels of 33 % and 17 %.

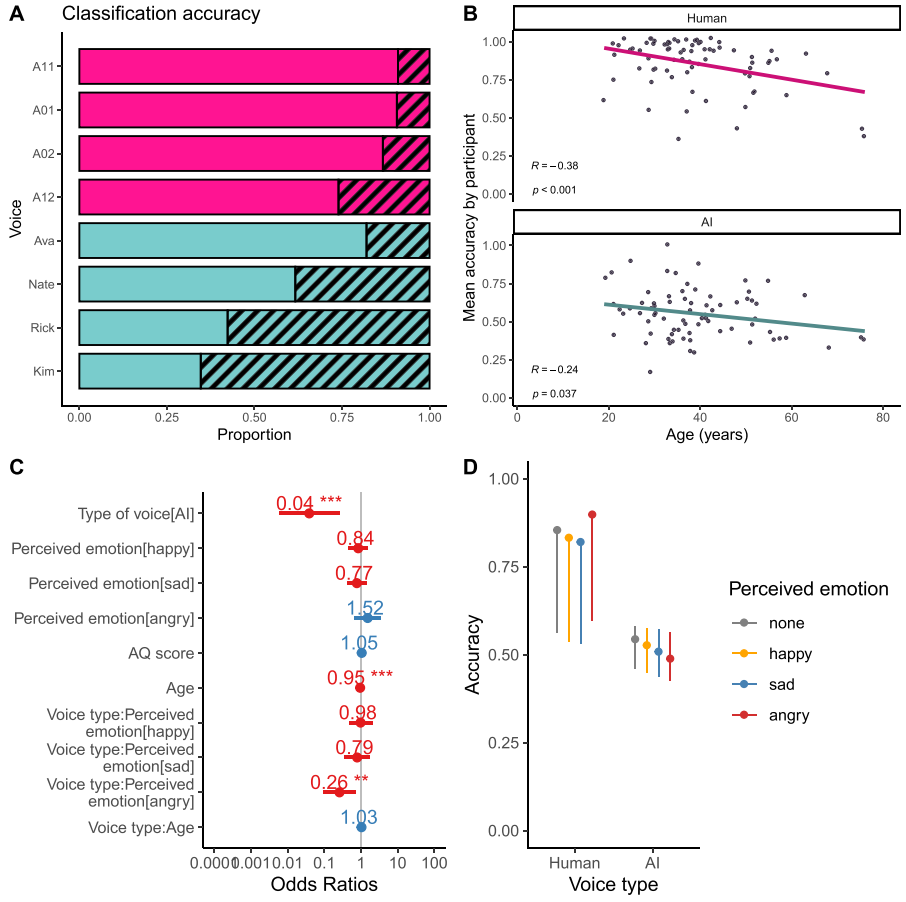


Fig. 2. Listeners' perception of human and TTS voices. A) Classification accuracy for human and AI-generated voices (proportion of correct responses). Human voices (A11, A01, A02, and A12) are shown in pink, and AI voices (Ava, Nate, Rick, and Kim) in blue. The proportion of correct responses is shown in full color, and the proportion of incorrect responses is shown with a stripped pattern. **B)** Relationship between participants' age and their average accuracy of classification, separately for human (top) and AI-generated (bottom) voices. Each dot represents one participant ($N = 75$). **C)** Estimates from the mixed-effects logistic regression model predicting accuracy of classification (Human vs. AI) reported as odds ratios. The reference group was "Human" for voice type and "none of the above" for perceived emotion. For example, an odds ratio of −0.04 for TTS voice implies that the odds of correctly identifying a voice as AI-generated are 96 % lower compared to identifying a human voice correctly. Additionally, the odds ratio of 0.95 for age indicates a 5 % decrease in the odds of responding correctly for every additional year of age. **D)** Model-based estimates for the interaction between voice type and perceived emotion. Error bars represent 95 % confidence intervals.

3.3. Similar patterns for attractiveness and willingness to interact ratings

Attractiveness and willingness to interact ratings led to overall similar patterns of results (Fig. 3, top).[6] In both cases, the distribution of ratings was wide, with ratings covering the whole response scale, indicating considerable variability in participants' responses. Please see Supplementary Fig. S2 for a visualization of attractiveness and willingness to interact ratings for each voice.

The proposed linear mixed models predicting attractiveness or willingness to interact ratings from the voice type, the perceived emotion, participants' AQ scores, and participants' age are represented in Fig. 3C and D (also see Supplementary Table S2). These models revealed a significant main effect of voice type, with participants rating human voices as more attractive and more socially appealing than synthetic voices (model estimates for attractiveness: Human: 4.28, 95 % confidence interval (CI) = [3.81–4.74], AI: 3.45, 95 % CI = [2.97–3.92]; for willingness to interact: Human: 4.10, 95 % CI = [3.69–4.52], AI: 3.49, 95 % CI = [3.06–3.91]). For both attractiveness and willingness to interact, there was a significant main effect of perceived emotion, with stimuli perceived as happy receiving higher attractiveness/willingness to interact ratings than the reference group of "none of the above"; and stimuli perceived as sad and angry receiving lower ratings. In other words, stimuli with negative valence were "penalized" and received lower ratings. The penalizing effect for angry voices was more pronounced in human than AI voices, as shown by the interaction between the voice type and the perceived emotion for stimuli perceived as angry. Note that, to check whether the position of the stimuli within the experimental block impacted ratings, we included the trial number as a covariate in our mixed-effects models. For attractiveness ratings, there was a small but significant effect (estimate = 0.008, SE = 0.003, $t = 3.23$, $p = .001$), suggesting a modest upward drift in ratings across trials. Crucially, this trend does not alter the main findings, with nearly identical fixed-effect estimates across the two specifications, and unchanged significance patterns. For willingness-to-interact ratings, the trial number was non-significant (estimate = −0.001, SE = 0.002, $t = -0.40$, $p = .69$). Please see Supplementary Fig. S3 for the confusion matrices displaying the accuracy of emotional profile classification, which was overall higher for human than AI voices (proportion correct: human: 67 %, AI: 51.8 %, $p < .001$).

Although participants handled synthetic and human voices similarly,

[6] Ratings of attractiveness and willingness to interact were highly correlated with each other — based on average ratings per stimulus across participants: $r(30) = .96$. A linear mixed model predicting willingness to interact ratings from attractiveness ratings explained 71 % of the variance in the dependent variable (syntax: lmer(willingness_interact ~ attractiveness + (1|participant) + (1|voice/stimulus)); marginal $R^2 = .66$, conditional $R^2 = .71$).

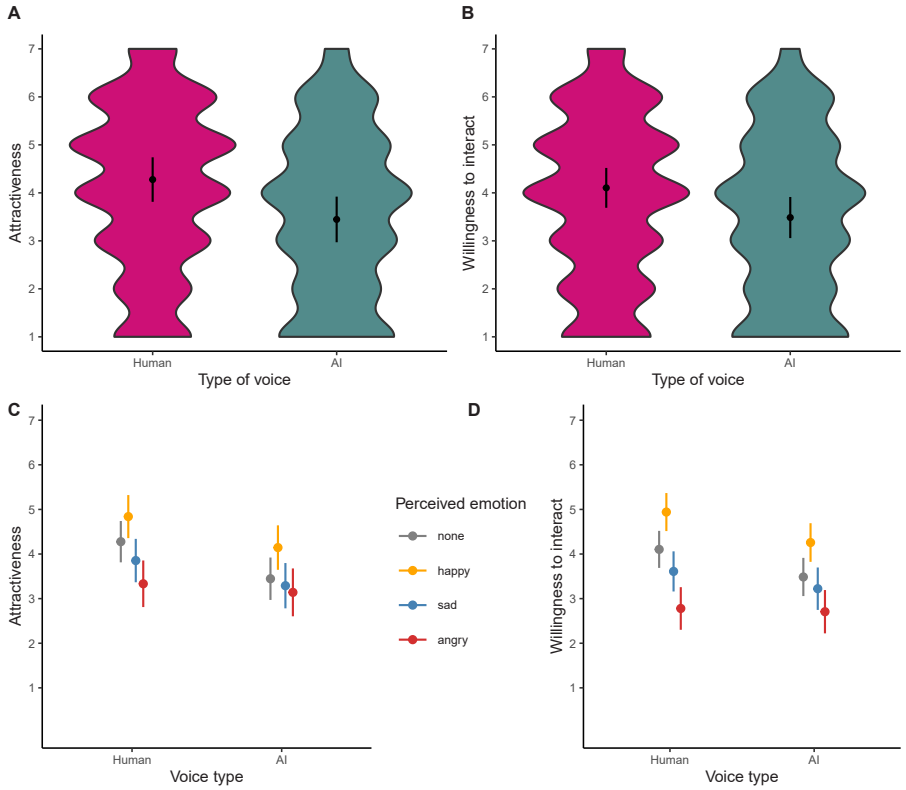


Fig. 3. Attractiveness and willingness to interact ratings. Upper panels: The violin plots depict the distribution of raw ratings of attractiveness (A) and willingness to interact (B) with the proposed voices. Dots and error bars represent model-based estimates and 95 % confidence intervals for the main effect of voice type. **Bottom panels:** Model-based estimates for the interaction between voice type and perceived emotion in the models predicting attractiveness (C) and willingness to interact ratings (D). Please see Supplementary Table S2 for the complete model output.

with higher ratings for positively valenced voices, they still rated the human voices as more attractive and socially appealing than the synthetic ones. Notably, the models explained only a modest proportion of the variance in ratings, as shown in Supplementary Table S2 (marginal $R^2 = 0.11$ and conditional $R^2 = 0.60$ for the attractiveness model; marginal $R^2 = 0.18$ and conditional $R^2 = 0.66$ for the willingness to interact model). Considerable individual differences were evident, as indicated by the low inter-rater agreement and the substantial variance captured by random intercepts for participants (intraclass correlation coefficients of 0.34 for attractiveness and 0.37 for willingness to interact ratings; Supplementary Table S2). This suggests that additional factors influence the evaluation of voice attractiveness and willingness to interact. Whereas cognitive style, as indirectly measured by the Autism Spectrum Quotient (AQ), did not significantly affect ratings, we observed a small but significant effect of age on willingness to interact ratings. Specifically, willingness to interact ratings increased by approximately 0.02 points for each additional year of age, regardless of voice type (i.e., with no interaction with the voice type).

4. Discussion

In this study, we examined participants' perception of voices without mentioning that some of them were generated with a computer. The results support that human voices were perceived as more attractive and more socially appealing than synthetic ones, albeit with large individual differences in participants' rating behavior. Despite this difference, participants seemed to handle human and synthetic voices in a similar way, giving higher ratings to voices they perceived as happy and lower ratings to voices they perceived as sad or angry. Considering that most participants indicated suspecting some of the voices were generated with a computer, one could interpret these findings as agreeing with the Computers Are Social Actors framework (CASA; Nass et al., 1994; Nass & Lee, 2000), according to which subjects apply social behaviors from human-human interaction when interacting with a computer, provided the system has human-like attributes (e.g., speech). It is interesting to note, however, that suspecting some of the voices were AI-generated did not aid participants in distinguishing between human and AI voices. While the proportion of correct classification was 86 % for human voices, it was only 55 % for AI voices, indicating that participants struggled to distinguish synthetic from human voices, with synthetic voices often "fooling" participants. Crucially, this effect was stronger for synthetic voices perceived as angry. We speculate that this relates to the novelty of emotional TTS voices to participants, who may still expect synthetic voices to sound more "robotic" (or, at least, "non-emotional"). Importantly, the similarity in the emotional confusion matrices (see Supplementary Fig. S3) suggests that emotional TTS voices are approaching human ones also in terms of the paralinguistic information they convey. This points to real progress toward greater naturalness of TTS voices (with naturalness defined as resemblance to a real human voice – Nussbaum et al., 2025).

Interestingly, we found a similar pattern of results for voice attractiveness and willingness to interact ratings. The high correlation between these two types of ratings is in line with previous literature on (general) voice perception. Listeners can quickly form impressions of person characteristics based on voices. According to Lavan et al. (2024), impressions of attractiveness, trustworthiness, educatedness and professionalism are highly correlated and seem to emerge at a later point in the neural processing chronometry than the perception of physical characteristics such as age and gender.

There were large individual differences in participants' overall perception of human and AI-generated voices. Among factors driving these differences, we observed a role of age, with older participants

showing reduced accuracy in distinguishing between human and AI-generated voices. This is in line with Herrmann (2023) and Müller et al. (2022). However, this lower sensitivity among older participants affected human and synthetic voices equally, which contrasts with previous research suggesting lower sensitivity of older participants to synthetic voices, potentially linked to their limited experience and exposure to synthetic voices and/or reduction in auditory sensitivity, especially in higher frequencies (Herrmann, 2023; Müller et al., 2022). Additional data on participants' exposure to or attitudes toward technologies such as voice assistants (e.g., Siri, Alexa) would help clarify to what extent the lower accuracy of synthetic voice recognition in our study reflects a lack of familiarity with synthetic voices or may be attributable to typical age-related auditory decline. We also observed a small but significant effect of age in willingness to interact ratings, with older participants more willing to interact with the voices, regardless of voice type. Research suggests that the psychological trait of agreeableness increases significantly from age 31 to 60 for both men and women (Srivastava et al., 2003), and points to age-related differences in social goals (Carstensen et al., 1999). We speculate that the higher willingness to interact ratings given by older participants could reflect a combination of socioemotional motivation for connection and personality shifts toward greater agreeableness in later life. However, without measuring attitudes toward technology or social openness explicitly, this remains a hypothesis worth investigating in future research. In our search for factors underlying individual differences in perception of synthetic voices, we also tested participants' Autism Quotient scores, as an indication of different cognitive styles. However, this was not a significant predictor of attractiveness or willingness to interact with the voices, nor did it influence participants' classification accuracy. Further research is needed to identify the factors driving individual variations in responses to synthetic voices.

More generally, the observation of large individual differences, despite self-consistency when evaluating the voices, is in line with the idiosyncrasies reported in the evaluation of other aspects of the voice (Bruder et al., 2024; Lavan & Sutherland, 2024). To acknowledge and address such individual differences in voice perception (including TTS voice perception), it would be valuable to systematically report measures of inter-rater reliability. In addition, analyses should go beyond average preference scores like MOS, which flatten variability and assume voice perception is unidimensional and shared. This idea agrees with recent reports calling for more nuanced methods for speech synthesis evaluation, especially considering advances in the quality of neural TTS (Le Maguer et al., 2024; Perrotin et al., 2025; Wagner et al., 2019). In the context of speech synthesis evaluation, it has been argued that many of the questions posed to participants are inherently multidimensional and could benefit from decomposition for greater clarity. For example, Hinterleitner et al. (2013) recommended that evaluation designs explicitly consider multiple perceptual dimensions, such as naturalness of voice, prosodic quality, fluency and intelligibility, disturbances, and calmness. In line with this reasoning, we focused on two specific dimensions of voice perception that are likely related but not interchangeable: attractiveness and social appeal of TTS voices. More broadly, we believe that the issues highlighted by Nussbaum et al. (2025) regarding naturalness – namely, conceptual underspecification, heterogeneous operationalization, limited exchange between research domains, and insufficient grounding in voice perception theory – also apply to these other constructs, indicating the need for more precise definitions and cross-disciplinary dialogue in TTS evaluation.

Our findings also invite reflection on the potential risk of over-standardization – or a lack of consideration for variability – in the production of TTS voices. Generative models primarily learn and replicate average or stereotypical patterns from training data, which may lead to a growing sameness and stereotypically in synthetic voices. How this homogenization will shape our future soundscape remains unclear. Given that human perception and cognition are shaped by our interactions with and exposure to the environment (e.g., statistical learning in language acquisition and pitch sequence learning – Larrouy-Maestri et al., 2013; Romberg & Saffran, 2010), increased exposure to voices lacking variability could have social and cognitive consequences that would need to be specifically investigated.

Some limitations of our study should be acknowledged. First, our stimulus set included only one sentence and eight voices, each presented in four emotional profiles. While this allowed for a well-controlled stimulus set in terms of semantic and lexical content, using a more varied stimulus set (e.g., with additional voices from different providers, a wider range of utterances, varied semantic content, and different accents or communicative intentions) would improve the generalizability of our findings. Further, extensive work has emphasized the importance of considering experimental context and intended use cases when evaluating and comparing speech systems (e.g., Clark et al., 2019; Edlund et al., 2024; Wagner et al., 2019). Because our stimuli were presented 'out of context', it remains an open question how well the present findings would generalize to specific situations. Moreover, using more recent synthetic voices (our stimuli were generated in December 2022) could better reflect the current state of rapidly evolving generative AI technologies. Further insight could also be gained by using voice cloning or conversion systems to generate synthetic voices that match original human voices, allowing for more direct comparisons. Future research may also explore how listeners' perception of synthetic voices shifts over time, particularly in terms of perceived human-likeness/naturalness, attractiveness, and social appeal, as voice synthesis tools continue to improve.

5. Conclusion

This study demonstrates that, while human voices are still perceived as more attractive and socially appealing, modern emotional TTS voices are approaching human ones in how listeners respond to them. The fact that TTS voices "fooled" participants points to substantial progress in the expressive quality and naturalness of these systems. The marked individual differences in participants' responses highlight the need for more nuanced evaluation methods that go beyond average preferences, as well as further exploration of listener diversity, to create more inclusive and context-sensitive technologies that reflect the complexity of human voice perception.

CRediT authorship contribution statement

Camila Bruder: Conceptualization, Data curation, Methodology, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing, Project administration. **Pamela Breda:** Conceptualization, Writing – review & editing. **Pauline Larrouy-Maestri:** Conceptualization, Methodology, Supervision, Writing – review & editing.

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.

Acknowledgements

We thank Klaus Frieler for invaluable discussions on data analysis, and Janniek Wester and Talip Ata Aydin for their insightful feedback and stimulating discussions. We also thank David Poeppel, Melanie Wald-Fuhrmann, and the INHABIT Artist-in-Residence program of the Max Planck Institute for Empirical Aesthetics for their support in developing this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbah.2025.100211.

Data availability

All data, materials, and analysis code are openly available in the project folder on the Open Science Framework (OSF) at https://osf.io/4qfaz.

References

Abdulrahman, A., & Richards, D. (2022). Is natural necessary? Human voice versus synthetic voice for intelligent virtual agents. *Multimodal Technologies and Interaction, 6*(7), 51. https://doi.org/10.3390/mti6070051

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist in 1,000 Cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(2), 202–212.e7. https://doi.org/10.1016/j.jaac.2011.11.003

Aylett, M. P., Clark, L., Cowan, B. R., & Torre, I. (2021). Building and designing expressive speech synthesis. In B. Lugrin, C. Pelachaud, & D. Traum (Eds.), *The handbook on socially interactive agents* (1st ed., pp. 173–212). ACM. https://doi.org/10.1145/3477322.3477329.

Babel, M., McGuire, G., & King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PLoS One, 9*(2), Article e88616. https://doi.org/10.1371/journal.pone.0088616

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 2018* (pp. 2863–2867). https://doi.org/10.21437/Interspeech.2018-1093

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 5–17. https://doi.org/10.1023/A:1005653411471

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Bruder, C., Frieler, K., & Larrouy-Maestri, P. (2024). Appreciation of singing and speaking voices is highly idiosyncratic. *Royal Society Open Science, 11*(11), Article 241623. https://doi.org/10.1098/rsos.241623

Carstensen, L. L., Isaacowitz, D. M., & Charles, S. T. (1999). A theory of Socioemotional Selectivity. *American Psychologist, 54*(3), 165–181.

Clark, R., Silen, H., Kenter, T., & Leith, R. (2019). Evaluating long-form Text-to-Speech: Comparing the ratings of sentences and paragraphs (No. arXiv:1909.03965). *arXiv*. https://doi.org/10.48550/arXiv.1909.03965

Cohn, M., Sarian, M., Predeck, K., & Zellou, G. (2020). Individual variation in language attitudes toward voice-AI: The role of listeners' autistic-like traits. In *Interspeech 2020* (pp. 1813–1817). https://doi.org/10.21437/Interspeech.2020-1339

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour, 60*(6), 773–780. https://doi.org/10.1006/anbe.2000.1523

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour, 65*(5), 997–1004. https://doi.org/10.1006/anbe.2003.2123

Edlund, J., Tånnander, C., Le Maguer, S., & Wagner, P. (2024). Assessing the impact of contextual framing on subjective TTS quality. In *Interspeech 2024* (pp. 1205–1209). https://doi.org/10.21437/Interspeech.2024-781

ElevenLabs. (2024). How to make text-to-speech sound less robotic. ElevenLabs Blog. eMarketer. (2024). Forecast 2024 https://www.emarketer.com.

European Broadcasting Union. (2023). R 128 – Loudness normalisation and permitted maximum level of audio signals. https://tech.ebu.ch/docs/r/r128.pdf.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour, 69*(3), 561–568. https://doi.org/10.1016/j.anbehav.2004.06.012

Finger, H., Goeke, C., Standvoß, K., & König, P. (2017). LabVanced: A unified JavaScript framework for online studies. In *International conference on computational social science. International conference on computational social science, cologne*. https://www.labvanced.com/static/2017_IC2S2_LabVanced.pdf.

Gessinger, I., Cohn, M., Cowan, B. R., Zellou, G., & Möbius, B. (2023). Cross-linguistic emotion perception in human and TTS voices. In *Interspeech 2023* (pp. 5222–5226). https://doi.org/10.21437/Interspeech.2023-711

Gessinger, I., Cohn, M., Zellou, G., & Möbius, B. (2022). Cross-cultural comparison of gradient emotion perception: Human vs. Alexa TTS voices. In *Interspeech 2022* (pp. 4970–4974). https://doi.org/10.21437/Interspeech.2022-146

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., & Aucouturier, J.-J. (2021). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature Communications, 12*(1), 861. https://doi.org/10.1038/s41467-020-20649-4

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language, 88*, 70–86. https://doi.org/10.1016/j.jml.2016.01.001

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

Hill, A. K., & Puts, D. A. (2016). Vocal attractiveness. In V. Weekes-Shackelford, T. K. Shackelford, & V. A. Weekes-Shackelford (Eds.), *Encyclopedia of evolutionary psychological science* (pp. 1–5). Springer International Publishing. https://doi.org/10.1007/978-3-319-16999-6_1880-1.

Hinterleitner, F., Norrenbrock, C. R., & Möller, S. (2013). Is intelligibility still the main problem? A review of perceptual quality dimensions of synthetic speech. In *Th ISCA speech synthesis workshop*.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://www.jstor.org/stable/4615733.

Hughes, S. M., Dispenza, F., & Gallup, G. G. (2004). Ratings of voice attractiveness predict sexual behavior and body configuration. *Evolution and Human Behavior, 25*(5), 295–304. https://doi.org/10.1016/j.evolhumbehav.2004.06.001

International Telecommunication Union. (1994). *ITU-T recommendation P.85, "A method for subjective performance assessment of the quality of speech output devices"*.

International Telecommunication Union. (1996). *ITU-T recommendation P.800, "Methods for subjective determination of transmission quality."*.

International Telecommunication Union. (2016). *ITU-T recommendation P.800.2, "Mean opinion score interpretation and reporting."*.

Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., & Vukovic, J. (2008). Integrating cues of social interest and voice pitch in men's preferences for women's voices. *Biology Letters, 4*(2), 192–194. https://doi.org/10.1098/rsbl.2007.0626

Kirkland, A., Mehta, S., Lameris, H., Henter, G. E., Szekely, E., & Gustafson, J. (2023). Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. In *12th ISCA Speech Synthesis Workshop (SSW2023)* (pp. 41–47). https://doi.org/10.21437/SSW.2023-7

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (Second). Sage.

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability. https://repository.upenn.edu/asc_papers/43.

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 Neurorobotics, 14*, Article 593732. https://doi.org/10.3389/fnbot.2020.593732

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Larrouy-Maestri, P., Leybaert, J., & Kolinsky, R. (2013). The benefit of musical and linguistic expertise on language acquisition in sung material. *Musicae Scientiae, 17*(2), 217–228. https://doi.org/10.1177/1029864912473470

Larrouy-Maestri, P., Poeppel, D., & Pell, M. D. (2024). The sound of emotional prosody: Nearly 3 decades of research and future directions. *Perspectives on Psychological Science*. , Article 17456916231217722. https://doi.org/10.1177/17456916231217722

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

Lavan, N., & Sutherland, C. A. M. (2024). Idiosyncratic and shared contributions shape impressions from voices and faces. *Cognition, 251*, Article 105881. https://doi.org/10.1016/j.cognition.2024.105881

Le Maguer, S., King, S., & Harte, N. (2024). The limits of the Mean Opinion Score for speech synthesis evaluation. *Computer Speech & Language, 84*, Article 101577. https://doi.org/10.1016/j.csl.2023.101577

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS One, 13*(5), Article e0196391. https://doi.org/10.1371/journal.pone.0196391

Luce, P. A. (1982). Comprehension of fluent synthetic speech produced by rule. *Journal of the Acoustical Society of America, 71*(S1). https://doi.org/10.1121/1.2019658. S96–S96.

Lüdecke, D. (2018). Ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software, 3*(26), 772. https://doi.org/10.21105/joss.00772

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

Mu, Z., Yang, X., & Dong, Y. (2021). Review of end-to-end speech synthesis technology based on deep learning (No. arXiv:2104.09995). *arXiv*. http://arxiv.org/abs/2104.09995.

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

Müller, N. M., Pizzi, K., & Williams, J. (2022). Human perception of audio deepfakes. In *Proceedings of the 1st international workshop on deepfake detection for audio multimedia* (pp. 85–91). https://doi.org/10.1145/3552466.3556531

Murf, A. I. (2025). Murf: AI voice generator. https://murf.ai.

Nass, C., & Lee, K. M. (2000). Does computer-generated speech manifest personality? An experimental test of similarity-attraction. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 329–336). https://doi.org/10.1145/332040.332452

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Conference companion on human factors in computing systems - CHI '94* (p. 204). https://doi.org/10.1145/259963.260288

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

Perrotin, O., Stephenson, B., Gerber, S., Bailly, G., & King, S. (2025). Refining the evaluation of speech synthesis: A summary of the Blizzard Challenge 2023. *Computer Speech & Language, 90*, Article 101747. https://doi.org/10.1016/j.csl.2024.101747

Pisanski, K., & Feinberg, D. R. (2018). *The interdisciplinary nature of voice attractiveness research*.

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., … Feinberg, D. R. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour, 95*, 89–99. https://doi.org/10.1016/j.anbehav.2014.06.011.

Pisanski, K., Jones, B. C., Fink, B., O'Connor, J. J. M., DeBruine, L. M., Röder, S., & Feinberg, D. R. (2016). Voice parameters predict sex-specific body morphology in men and women. *Animal Behaviour, 112*, 13–22. https://doi.org/10.1016/j.anbehav.2015.11.008.

R Core Team. (2021). *R: A language and environment for statistical computing (Vienna, Austria)*. https://www.R-project.org/.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *WIREs Cognitive Science, 1*(6), 906–914. https://doi.org/10.1002/wcs.78

Roring, R. W., Hines, F. G., & Charness, N. (2007). Age differences in identifying words in synthetic speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 49*(1), 25–31. https://doi.org/10.1518/001872007779598055

Rosenberg, A., & Hirschberg, J. (2021). Prosodic aspects of the attractive voice. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice attractiveness* (pp. 17–40). Singapore: Springer. https://doi.org/10.1007/978-981-15-6627-1_2.

RStudio Team. (2022). RStudio: Integrated development environment for R. Boston, MA http://www.rstudio.com.

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology, 8*(4), 467–487. https://doi.org/10.1002/ejsp.2420080405

Scherer, K. R. (2021). Comment: Advances in studying the vocal expression of emotion: Current contributions and further options. *Emotion Review, 13*(1), 57–59. https://doi.org/10.1177/1754073920949671

Speechify. (n.d.). Speechify: Text to speech with real human like voices. https://speechify.com/blog/text-to-speech-with-real-human-like-voices/.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84*(5), 1041–1053. https://doi.org/10.1037/0022-3514.84.5.1041

Stern, S. E., Mullennix, J. W., Dyson, C., & Wilson, S. J. (1999). The persuasiveness of synthetic speech versus human speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 41*(4), 588–595. https://doi.org/10.1518/001872099779656680

Streijl, R. C., Winkler, S., & Hands, D. S. (2016). Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives. *Multimedia Systems, 22*(2), 213–227. https://doi.org/10.1007/s00530-014-0446-1

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Zhao, S., Qin, T., Soong, F., & Liu, T.-Y. (2024). *NaturalSpeech*: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(6), 4234–4245. https://doi.org/10.1109/TPAMI.2024.3356232

Taylor, E. C., Livingston, L. A., Clutterbuck, R. A., & Shah, P. (2020). Psychometric concerns with the 10-item Autism-Spectrum Quotient (AQ10) as a measure of trait autism in the general population. *Experimental Results, 1*, e3. https://doi.org/10.1017/exp.2019.3

Trouvain, J., Weiss, B., & Barkat-Defradas, M. (2021). Voice attractiveness: Concepts, methods, and data. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice attractiveness* (pp. 3–16). Singapore: Springer. https://doi.org/10.1007/978-981-15-6627-1_1.

van Rijn, P., & Larrouy-Maestri, P. (2023). Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nature Human Behaviour, 7*(3), 386–396. https://doi.org/10.1038/s41562-022-01505-5

Vixen Labs. (2022). Voice Consumer Index 2022. https://vixenlabs.co/wp-content/uploads/2022/06/VixenLabs_VoiceConsumerIndex2022.pdf.

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior, 17*(1), 3–28. https://doi.org/10.1007/BF00987006

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Henter, G. E., Maguer, S. L., Malisz, Z., Székely, É., & Tånnander, C. (2019). Speech synthesis Evaluation—State-of-the-art assessment and suggestion for a novel research program. In *10th ISCA workshop on speech synthesis (SSW 10)* (pp. 105–110). https://doi.org/10.21437/SSW.2019–19

Xu, J., Xing, L., Perkis, A., & Jiang, Y. (2011). On the properties of mean opinion scores for quality of experience management. In *2011 IEEE international symposium on multimedia* (pp. 500–505). https://doi.org/10.1109/ISM.2011.88

Zellou, G., Cohn, M., & Ferenc Segedin, B. (2021). Age- and gender-related differences in speech alignment toward humans and voice-AI. *Frontiers in Communication, 5*, Article 600361. https://doi.org/10.3389/fcomm.2020.600361

Zhu, Q., Chau, A., Cohn, M., Liang, K.-H., Wang, H.-C., Zellou, G., & Yu, Z. (2022). Effects of emotional expressiveness on voice chatbot interactions. In *Proceedings of the 4th conference on conversational user interfaces* (pp. 1–11). https://doi.org/10.1145/3543829.3543840