The vocal uncanny valley: Deviation from typical organic voices best explains uncanniness.

Alexander Diel (diela@cardiff.ac.uk)
Cardiff University
Michael Lewis
Cardiff University

Article

Keywords: uncanny valley, voice processing, pathological voice, voice distortion, text-to-speech, deviation from familiarity

Posted Date: April 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2784067/v1

Abstract

The *uncanny valley* describes the negative evaluation of near humanlike artificial entities. Previous research with synthetic and real voices failed to find an uncanny valley of voices. This may have been due to the selection of stimuli. In Experiment 1 ($n = 50$), synthetic, normal, and deviating voices (distorted and pathological) were rated on uncanniness and human likeness and categorized as human or non-human. Results showed a non-monotonic function when the uncanniness was plotted against human likeness indicative of an uncanny valley. However, the shape could be divided into two monotonic functions based on voice type (synthetic vs deviating). Categorization ambiguity could not predict voice uncanniness but moderated the effect of realism on uncanniness. Experiment 2 ($n = 35$) found that perceived organicness of voices significantly moderated the effect of realism on uncanniness, while attribution of mind or animacy did not. Results indicate a vocal uncanny valley re-imagined as monotonic functions of two types of deviations from typical human voices. While voices can fall into an uncanny valley, synthetic voices successfully escape it. Finally, the results support the account that uncanniness is caused by deviations from familiar categories, rather than categorical ambiguity or the misattribution of mind or animacy.

1. Introduction

Artificial humanlike entities with imperfect human appearance are evaluated negatively, a phenomenon called *uncanny valley* [1, 2, 3, 4]. The relationship between human likeness and likability or uncanniness, typically measured by rating scales, is defined as a polynomial function consisting of a gradual increase of likability with increasing human likeness and a drop into the negative at near human likeness [1, 4]. The uncanny valley remains a pressing issue in human-machine interaction, yet the underlying cognitive mechanisms remain unclear.

1.1 The vocal uncanny valley

The uncanny valley has been observed in the context of android appearance and behaviour and their mismatch with voices [5, 6]. However, previous research has consistently failed to find a 'vocal uncanny valley' when isolated voice stimuli were used: likability increased with a voice's human likeness [7, 8, 9, 10, 11]). However, except for one study [9], all researchers investigating a vocal uncanny valley have used exclusively synthetic voices and/or fully human voice stimuli. There are four explanations on why an uncanny valley of voices may not have been found: 1) a vocal uncanny valley does not exist; 2) stimulus selection has sufficient range but lacks stimuli that fall into the valley; 3) stimulus selection does not extent into the valley and stops before the drop [12]; 4) stimulus selection begins at the valley and ends at full human likeness. These explanations urge different implications for the design of artificial voices: If an uncanny valley of voices has not yet been reached, technological development may yet lead to its emergence. If, on the other hand, today's synthetic voices already overcome an uncanny valley or if a vocal uncanny valley does not exist, then this particular issue can be disregarded for the design of artificial voices.

1.2 Uncanniness and deviation from typical variation

Multiple theories on the uncanny valley predict the effect to be transferred to voice stimuli. Atypical stimuli may be uncanny because of their strong deviation from typical appearances [13, 14]. Sensitivity to deviations is increased for more familiar stimulus categories, as is the case on human-specific stimuli like faces [15, 16, 17, 18]. Deviating stimuli may recruit additional processing need, decreasing aesthetic evaluation processing disfluency [19]. Alternatively, stimuli belonging to more familiar categories may have more solidified or strict predictive patterns, increasing the likelihood of prediction errors [20, 21]. In any case, humanlike yet deviating voices would suffer from aesthetic devaluation.

Certain disorders related to the vocal tract, like vocal fold paresis, Reinke's Edema, or muscle tension dysphonia, can lead to changes in the voice. Pathological voices are more likely to be categorized as atypical [22, 23, 24] and are evaluated more negatively across various social dimensions compared to healthy voices [25, 26, 27, 28]. In analogy, previous research has suggested that dysmorphic, diseased, or very unattractive faces are perceived as uncanny or creepy [17, 29]. Thus, pathological voices, similarly to disfigured faces, may fall into an uncanny valley as highly realistic yet deviating stimuli.

1.3 Uncanniness and categorization difficulty

Stimuli difficult to categorize may fall into an uncanny valley [3, 13, 30, 31]. Categorization difficulty may decrease likability due to processing disfluency [32, 33] or cognitive conflict [34]. As categorization theories do not depend on stimulus domain, categorical ambiguity should thus also predict the uncanniness of voices.

2. Experiment 1

2.1 Research question and hypotheses

The aim of the experiment is to investigate the existence of a vocal uncanny valley using (manipulated) natural voices, synthetic voices, and pathological voices. In addition, it is investigated whether the uncanniness of voices can be explained by deviation from familiar categories or categorical ambiguity

First, the role of familiarity is investigated by comparing the effects of distortion on uncanniness for very familiar (human) voices and less familiar (cat) voices. Hypothesis 1 is thus:
1. Distortion Of Human Voices Increases Uncanniness More Than Distortion Of Cat Voices.

Furthermore, a vocal uncanny valley is replicated, including artificially distorted and naturally pathological voice stimuli. It is tested whether a vocal uncanny valley exists in principle but is successfully avoided by contemporary synthetic voices. Hypotheses 2 and 3 are thus:

2. A monotonic function of human likeness can best explain the uncanniness of synthetic and natural voices.

3. A non-monotonic function akin to an uncanny valley can best explain the uncanniness of synthetic, natural, distorted, and pathological voices.

Finally, it is investigated whether ambiguity in categorizing a voice as either human or non-human can best explain the uncanniness ratings. Categorization ambiguity is operationalized as 1. Categorization reaction time, and 2. Categorization uncertainty, i.e., the inconsistency of categorizations across participants. Hypothesis 4 is thus:

4. Categorization Reaction Time And Categorization Uncertainty Predict Uncanniness Ratings Of Voices.

2.2 Methods

2.2.1 Participants

Power analysis revealed that $n = 50$ participants are sufficient to exceed a power of $1\text{-}\beta = 0.8$ with a six-voice-conditions within-subject design and a standard effect size of $d = 0.5$ [56]. Participants were Psychology students at the Cardiff University School of Psychology, recruited via the Experimental Management System (EMS). Participants were on average 19 years old ($SD_{age} = 1.05$), 37 identified as female, 11 as male, one as other, and one preferred not to say. Participants were compensated with 4 credits equivalent to the advertised compensation of a 60 minutes online study.

2.2.2 Stimuli

Ten typical and 15 pathological voices were taken from the Perceptual Voice Qualities Database (PVQD) [57]. Specifically, the 15 pathological voices with the highest subjective severity ratings were selected as stimuli. Specific pathologies included Reinke's Edema (x3), lesions (x3), vocal fold paralysis (x3), muscle tension dysphonia (x2), ulcerative laryngitis, adductor spasmodic dysphoria, and one unrecorded pathology. Ten distorted voice variants were created by using the STRAIGHT software, specifically by multiplying the normal voices' fundamental frequencies by 1000 [58]. In addition, 10 normal cat meowing sounds were selected from www.freesound.org, and 10 distorted variants were created with STRAIGHT by multiplying the fundamental frequency by 1000. Finally, 15 synthetic voices were selected from various sources: Four mechanical sounds were taken from www.freesound.org, five voices from IBM Watson, three voices by Azure Microsoft TTS, Microsoft Sam, one voice created by a Stephen Hawking Voice Generator, and one generic Google TTS voice.

Fifteen pathological and synthetic voices were selected instead of 10 (as in the other conditions) because both conditions were expected to be more heterogenous and thus would need a higher stimulus number to be adequately statistically represented.

Because the fifteen pathological voices consisted of 9 female and 6 male voices, the same ratio was selected for typical voice counterparts. As distorted voices were created by manipulating the typical voices, the same ratio was present for those. For synthetic voices, six were artificial female voices, five were artificial male voices, and four were mechanical sounds. Voice accent was not controlled.

All stimuli were shortened to be around 5 seconds in length. For standardization, all typical, pathological, distorted, and synthetic voices (except for the mechanical sounds) were expressing the same sentences. The spoken sentences, “The blue spot is on the key again. How hard did he hit him?”, were used as basic sentences in the PVQD database and recreated for synthetic voices. More details on the voice stimuli are shown in *Table A1*.

2.2.3 Rating Task

For the rating task, participants had to rate each sound based on three items: eerie/uncanny, strange/weird, and humanlike/realistic. Items ranged from the extremes of 0 to 100 and participants could choose to place the slider on any point of the item. Voices were presented in a random order for each participant and were replayed for each item. Participants had an unlimited amount of time responding to the items. Because uncanniness and human likeness are here understood as subjective experiences and assessments, the terms were presented with minimal information to the participants to gauge their own interpretations. An example screenshot of the rating task is presented in Figure A1.

The eerie/uncanny and strange/weird items were combined into an uncanniness index by calculating the means of the items for each stimulus. Analogous item combinations have been used in previous research with reliable consistency [1, 59, 60].

2.2.4 Categorization Task

For the categorization task, all sounds except the normal and distorted cat meow voices were used. For each presented sound, participants had to do a two-alternative forced choice task on whether the voice was humanlike or not. Participants first heard 2 seconds of the sound before the choice text appeared, at which point participants had the ability to decide by pressing either the left or right key on their keyboard. Participants were instructed to be as accurate and fast as possible.

2.2.5 Procedure

The whole procedure was conducted online. After giving informed consent and filling out a demographic questionnaire asking for participants’ gender and age, participants were redirected to the experiment. They first went through the rating task followed by the categorization task.

The human likeness ratings were used to operationalize the x-axis of the uncanny valley function. Meanwhile, human categorization responses (both reaction times and response inconsistencies) were used as indicators of categorical ambiguity.

2.2.6 Analysis, ethics statement, and data availability

Analysis was conducted via R. Linear mixed models were used to control for participants, as well as analyses of variance (ANOVAs) and linear regressions. Data cleaning was conducted by removing all outlier (1.5*IQR) uncanniness, human likeness, and categorization reaction time ratings for each stimulus. A total of 17 values were removed. The experiment was approved by the Cardiff University School of Psychology Ethics Committee in October 2021 (reference number: EC.21.09.14.6411G). All

methods were performed in accordance with the Declaration of Helsinki and informed consent was collected from all participants. at the datasets generated and analysed during the current studies and the analysis scriptsare available on OSF: https://osf.io/7xs6j. Original versions of the voice samples can be downloaded from the PVQD website.

2.3 Results

The *eerie/uncanny* and *strange/weird* items were combined into an *uncanniness* index with a Cronbach's alpha of $\alpha = .79$, indicating acceptable, almost good construct validity.

2.3.1 Voice distortion: human vs cat

A within-subject 2x2 ANOVA was conducted with distortion (normal vs distorted) and species (cat vs human) as factors of uncanniness. The analysis showed main effects of both distortion ($F(1,48) = 567.02$, $p < .001$, $d = .77$) and species ($F(1,48) = 51.84$, $p < .001$, $d = .20$), as well as an interaction between these two ($F(1,48) = 47.35$, $p < .001$, $d = .15$). The interaction is visualized in Fig. 1.

Follow-up *p*-adjusted post-hoc Tukey tests showed that distortion increased the uncanniness of both cat ($t(1825) = 33$, $p_{adj} < .001$, $d = 2.16$, CI[0.8, 3.52]) and human voices ($t(1825) = 52.48$, $p_{adj} < .001$, $d = 3.43$, CI[1.27, 5.59]). Furthermore, normal human voices were significant less uncanny than cat voices ($t(1825) = 21.328$, $p_{adj} < .001$, $d = 1.39$, CI[0.51, 2.27]), but not in the distortion conditions ($t(1825) = 1.82$, $p_{adj} = .19$, $d = 0.12$, CI[-0.03, 0.27]). Thus, the same distortion procedure increased the uncanniness of human voices more than the uncanniness of cat voices. Hypothesis 1 is thus supported.

2.3.2 An uncanny valley of voices

An uncanny valley of voice stimuli was investigated using a linear mixed model with human likeness ratings as fixed effects and participants and stimuli as random effects on uncanniness. Cat sounds were excluded from the analysis to focus on humanlike and mechanical voices. Results show that a cubic term ($t(1637) = -5.51$, $p < .001$, $R^2_{adj} = .67$) could explain the variance better than a linear term ($\chi^2 = 57.57$, $p < .001$) or a quadratic term ($\chi^2 = 30.27$, $p < .001$). The model is plotted in Fig. 2. As can be seen in the plot, confidence intervals in the curves' "valleys" do not overlap with the confidence intervals of the curves' maxima. Taken together with the significant cubic term, a non-monotonic relationship explains the relationship between uncanniness and human likeness across voice categories.

In a second step, distorted and pathological voices were removed and the analysis was redone. The results show that again, a cubic term ($t(26000) = -2.86$, $p = .004$, $R^2_{adj} = .56$) could better explain the variance than a linear ($\chi^2 = 47.62$, $p < .001$) or quadratic term ($\chi^2 = 8.16$, $p = .004$). The function, depicted in Fig. 3, however does not reflect an uncanny valley plot. To complement the function, a second plot with distorted, normal, and pathological voices was plotted as well. Given that at no point in the functions in Fig. 3, the confidence intervals seem to significantly decrease, but only increase with increasing realism, both functions indicate monotonic relationships between uncanniness and human likeness when the

data depicted in Fig. 2 is divided based on different voice categories. Thus, a non-monotonic relation between uncanniness and human likeness seems to result from a combination of multiple monotonic functions. Thus, hypotheses 2 and 3 are supported.

2.3.3 Categorization difficulty as a predictor of voice uncanniness

A linear mixed model with reaction time as a fixed effect and stimuli and participants as random effects on uncanniness showed that reaction time could not predict voice uncanniness ratings ($t(2207) = 1.29$, $p = .197$). The data is plotted in Fig. 4.

Voice categorization data was transformed into a *voice certainty* variable by coding participants' *non-human* categorizations as 0 and *human* categorizations as 1, then using the absolute values after subtracting the averaged categorizations for each stimulus by 0.5. *Voice certainty* thus reflects a variable ranging from 0 (50:50 categorization as human and non-human across participants) to 0.5 (consistent categorization as either human or non-human across participants) to be used as an operationalization of consistent categorization.

Because the transformed data was already aggregated across participants for each stimulus, a linear regression model was used to investigate the effect of categorization certainty on uncanniness. The results show that categorization certainty could not predict voice uncanniness ratings ($t(50) = 0.15$, $p = .88$), and the data is visualized in Fig. 5.

The figures indicate that while distorted voices were both ambiguous and uncanny (compared to synthetic and normal voices which were neither ambiguous nor uncanny), pathological voices seemed to be uncanny yet consistently categorized as human (Fig. 6). This has been mostly confirmed by post-hoc tests: While distorted voices were not more ambiguous and uncanny than synthetic (ambiguous: $t(46) = -4.553$, $p < .001$; uncanny: $t(46) = 9.192$, $p < .001$) or human voices (ambiguous: $t(46) = -3.197$, $p = .008$; uncanny: $t(46) = 11.59$, $p < .001$), pathological voices were more uncanny than synthetic ($t(46) = 8.03$, $p < .001$) and human voices ($t(46) = 10.69$, $p < .001$), while not being more ambiguous (synthetic: $t(46) = 1.29$, $p = .475$; human: $t(46) = -1.05$, $p = .621$). Thus, hypothesis 4 was not supported.

2.3.4 Human categorization as a moderator of human-deviation on uncanniness

The model plotted in Fig. 2 indicates a *W*-shaped relationship with "two valleys". Such a relationship may be a consequence of choosing different categories of voices which interact differently with human likeness to affect uncanniness. The effect of voice type could be moderated by a variable influencing the perception of a decrease in realism (or closeness to the human norm) on uncanniness. Hence, a third variable may underlie the observed data by moderating a linear relationship between human likeness and uncanniness. As the uncanny valley has been linked to perceptions of markers of death and disease

avoidance [2, 4], the effect may be linked to the perception of organic appearance. ,. Thus, a perceived high “organicness” of a voice may increase the sensitivity of uncanniness towards deviations from human likeness, potentially due to evolutionary disease avoidance mechanism. Although “organicness” has not been measured in the experiment, the categorization of a voice as human may indicate how organic it was perceived to be, as both distorted and synthetic voices were categorized as non-human, while pathological and normal voices were categorized as human. Categorization of a stimulus as human may increase the effect of deviation on uncanniness: Hence, the slope from fully synthetic to human voices would be less steep than for (partially artificial) distorted to human voices, which would be again less steep than the slope for (fully organic) pathological to human voices (see black lines in Fig. 1). A post-hoc linear regression analysis has thus been conducted for the interaction between categorization response (human vs non-human) and human likeness on uncanniness. The results show main effects of response ($t(46) = 10.011$, $p < .001$), human likeness (($t(46) = -8.922$, $p < .001$), and an interaction between these two (($t(46) = -6.163$, $p < .001$; $R^2_{adj} = .80$). Thus, a moderated linear relationship between human likeness, uncanniness, and categorization as “human” is indicated.

2.4 Discussion

2.4.1 Voice distortion and familiarity

Voice distortion created by multiplying the fundamental frequency by 1000 increased the uncanniness of both human cat voices. The increase was stronger for human compared to cat voices. A higher degree of familiarity to a voice category may sensitize uncanniness caused by deviation.

Differences in fine details between human voices carry vital information about spoken messages and characteristics and states of the speaker [23, 24]. The recognition of analogous information is less important for the perception of cat voices. Thus, the degree of familiarity (and change sensitivity) in humans is lower for cat compared to human sounds. Higher uncanniness sensitivity for human compared to cat voices can thus be explained by higher familiarity to typical voice patterns and sensitivity to deviations from these patterns.

2.4.2 An uncanny valley of voices

A function with only synthetic and normal human voices showed that a linear relationship between human likeness and uncanniness akin to previous research [7, 8, 9, 10, 11]. However, adding voices that are either deliberately distorted or naturally deviating produces a non-monotonic function of uncanniness and human likeness. Especially when excluding distorted voices, the curve would be akin to an *N*-shaped uncanny valley plot, and the pathological voices would lie within an uncanny valley akin to the prediction of dead bodies falling into an uncanny valley [4]. Such an interpretation would favour explanations of the uncanny valley related to mortality salience or disease avoidance [2].

Previous researchers have noted that an uncanny valley could occur at any point at a graph, allowing multiple valley-shaped functions, potentially due to a multicausal emergence of the effect [17, 37, 38, 39].

An uncanny valley may not necessarily occur on just one area on the human likeness axis, and polynomial functions more complex than an N-shaped curve may occur depending on the stimuli selected, as in this study.

2.4.3 Categorization ambiguity does not predict uncanniness

Categorization ambiguity has been proposed to underlie the uncanny valley effect [30, 31, 34]. This study failed to find evidence for the categorization ambiguity hypothesis: Neither categorization reaction time nor categorization response consistency could predict uncanniness ratings. While distorted voices were both uncanny and difficult to categorize, pathological voices were not. Categorization ambiguity may correlate with stimulus deviations when stimuli are incremental morphs between two easily categorizable stimuli [31] and thus may be uncanny due to their deviation. However, certain stimuli can be uncanny despite being easy to categorize [3, 17]. Thus, uncanniness cannot be explained solely by categorization ambiguity.

2.4.4 A mediator of uncanniness?

A significant interaction between human categorization and human likeness was found that could explain uncanniness better than a polymonial model of human likeness. Categorization as human sensitized the effect of deviation on uncanniness. As, dehumanization can decrease the uncanniness of androids [40], categorization as human may activate a stricter evaluation of stimuli based on their proximity to the human norm. As a humanization manipulation can affect the specialized processing of faces [41, 42], an increase of humanization (and human categorization) may also further sensitize the detection of configural deviations and thus uncanniness. Similarly, as mind perception increases configural processing [43], it may also increase the sensitivity to deviations and thus uncanniness when a stimulus is perceived as both having a mind and deviates from the norm of appearance.

However, human likeness and categorization choice was highly correlated in this study, decoupling human likeness (or deviation) from human categorization (or humanization) would be required, which however should be difficult given the conceptual similarity of these concepts.

3. Experiment 2

3.1 Research Question and hypotheses

The aim of Experiment 2 is to investigate a potential third variable that may moderate a monotonic effect of human likeness on uncanniness. Several candidates for this third variable were explored.

3.1.1 Pathogen avoidance: Perception of organic voice

Uncanniness may be a response to the detection of indicators of contagious disease [2, 44]. Disease indicators may appear as physical anomalies or deviations co-occurring with pathology or physical

disabilities [45, 46]. As disease threat is only relevant for organic material, the perception of an entity being organic (vs synthetic) should then increase negative response towards norm deviation in a stimulus. Meanwhile, a voice recognized as inorganic should pose no disease-related threat even despite deviating from the norm.

1. Perception Of Organicness Moderates The Relation Between Human Likeness And Uncanniness Across Voice Categories

3.1.2 Mind attribution and animacy

Uncanniness may be elicited when human qualities like mind or animacy are attributed to non-human entities [47, 48]. Thus, less humanlike voices not perceived as having a mind or being animate should not elicit uncanniness, while deviating voices which appear to have a mind or to be animate should be uncanny.

2. Attribution Of Mind Moderates The Relation Between Human Likeness And Uncanniness Across Voice Categories

3. Perception Of Animacy Moderates The Relation Between Human Likeness And Uncanniness Across Voice Categories

3.2 Method

3.2.1 Participants

According to a power analysis, $n = 35$ participants are sufficient to exceed a power of $1-\beta = 0.8$ for a within-subject design with a standard effect size of $d = 0.5$ [56]. Participants were Psychology students at the Cardiff University School of Psychology, recruited via the Experimental Management System (EMS). Participants were on average 19.26 years old ($SD_{age} = 1.29$), 34 identified as female and one as male.

3.2.2 Stimuli

Per category (distorted, normal, pathological, synthetic), five stimuli were selected from Experiment 1. In addition, variation of distortion degree was created for distorted and pathological voices: For distorted voices, fundamental frequencies of normal (base) voices were increased by 250, 500 and 750, in addition to the present distorted voices with an increase by the value of 1000. These distortion levels were created to simulate an incremental increase of distortions starting with the normal counterparts. As the goal of the experiment is to investigate a moderated linear function of uncanniness, an incremental increase of distortion may reflect a linear function for one value of the moderator variable. For pathological voices, additional sets of five voices were selected based on the level of perceived severity ratings as reported in

the PVQD [57]. The five most severe pathological voices were selected for the severity rating limits of 100, 75, 50, and 25. Spoken sentences were the same as in Experiment 1. The stimuli are summarized in *Table A2*.

3.2.3 Procedure: Rating Task

The experiment consisted only of a rating task conducted online. The rating task was identical to the one in Experiment 1, except participants rated each voice based on the items eerie, strange, and humanlike only, in addition to its perceived animacy, mind attribution, and organicness. The additional items were presented the same way as the previous ones described in Experiment 1.

3.2.4 Analysis, ethics statement, and data availability

Analysis was conducted via R. Linear mixed models were used to control for participants, as well as analyses of variance (ANOVAs) and linear regressions. Data cleaning was conducted by removing all outlier (1.5*IQR) uncanniness, human likeness, and categorization reaction time ratings for each stimulus. A total of 13 values were removed. All methods were performed in accordance with the Declaration of Helsinki and informed consent was collected from all participants. at the datasets generated and analysed during the current studies and the analysis scriptsare available on OSF: https://osf.io/7xs6j.

3.3 Results

Eerie and *strange* items were combined into an *uncanniness* index with a Cronbach's alpha of $\alpha = .8$, indicating good consistency.

3.3.1 Moderating effects

Linear mixed models with human likeness and either animacy, mind attribution, or organicness as fixed effects stimuli and participants as random effects showed that the interaction between human likeness and animacy ($t(1762) = -3.568$, $p < .001$, $R^2_{adj} = .58$), mind attribution ($t(1856) = 2.824$, $p = -005$, $R^2_{adj} = .57$), or organicness ($t(1690) = -2.539$, $p = .011$, $R^2_{adj} = .58$) each significantly predicted uncanniness.

To test whether a moderated function can explain uncanniness better than a quadratic function of human likeness, the linear moderator models were tested against a quadratic human likeness function. A quadratic human likeness model was able to predict uncanniness ($t(6366) = -4.065$, $p < .001$, $R^2_{adj} = .56$). Model comparisons showed that only the model with organicness fitted the data significantly better than the quadratic human likeness model ($\chi^2 = 20.184$, $p < .001$). Replacing a quadratic human likeness term with either animacy or mind perception did not change model fit. Thus, a moderated linear function of organicness and human likeness could explain the results better than a quadratic function of human likeness.

Relations between uncanniness and the other variables are depicted in *Figures A2* to *A5*.

3.3.2 Differences between voice types

P-adjusted Tukey tests on differences between voice categories showed that distorted voices were more uncanny than normal ($t(56) = 6.789$, $p_{adj} < .001$) and synthetic voices ($t(56) = 7.097$, $p_{adj} < .001$). However, while distorted voices were perceived as less animate ($t(55) = -9.825$, $p_{adj} < .001$) and as having less mind ($t(55) = -9.725$, $p_{adj} < .001$) compared to normal voices, they did not differ from synthetic voices.

3.4 Discussion

3.4.1 “Uncanny valley” as a moderated linear function

A third variable of organicness moderates a linear relationship between human likeness and uncanniness. A moderating function may appear as an increase of the slope with increasing organicness: While distinctively artificial voices can deviate from the human norm without suffering from uncanniness, deviations in organic-sounding voices may quickly become unnerving, for example due to the threat of contamination from infected organic entities (MacDorman & Ishiguro, 2006).

However, all tested predictors were highly intercorrelated, and correlated highly with human likeness. Thus, it is not clear whether organicness itself is the third variable, or whether the third variable can be better described by a different construct.

3.4.2 Animacy and mind perception

Previous research aimed to explain the uncanny valley phenomenon through the attribution of humanlike characteristics like animacy or mind onto visibly artificial or inanimate stimuli (e.g., Stein & Ohler, 2017). However, the present results suggest that voice uncanniness also occurs for deviating voices clearly perceived as animate or having a mind (i.e., pathological voices). Meanwhile, artificially distorted voices perceived as inanimate or lacking mind were still uncanny. These results cannot be explained by misattribution of human qualities onto artificial entities.

4. General Discussion

4.1 Uncanny valley of voices

In two experiments, non-monotonic relationships between uncanniness and human likeness for voices were observed, although the function differs from a typical uncanny valley function [4]. The cognitive processing underlying the uncanny valley effect may be analogous across visual and auditory domains. Distinct face and voice variants elicit stronger activity in neural substrates specific to these categories [49, 50, 51], which may indicate increased processing need. Increased processing need may in turn decrease the aesthetic appeal of a stimulus [33]. Alternatively, a higher familiarity with a face or voice category may sensitize to errors or deviations, leading to prediction error signals [20, 21].

4.2 Synthetic voices and the uncanny valley

Synthetic voices did not fall into the valley of the function and instead were allocated around it Hence, modern TTS synthetisation can successfully replicate human voices. In fact, participants consistently rated one of the Watson voices to be about as humanlike as typical human voices (however, the same voice was ambiguously categorized with a 53% human categorization rate). Thus, synthetic voices manage to overcome the uncanny valley while visual synthetic replications of humans (i.e., androids) often do not.

It may be easier to replicate a synthetic voice than a synthetic face without errors: Synthetic voice replication can rely on recorded natural voices while synthetic faces must be artificially reconstructed. Alternatively, as human identity discrimination ability is more sensitive to faces than to voices [52], visual human processing may also be more sensitive to deviations compared to auditory human processing, making errors in design more apparent and appalling.

In general, the results affirm current technology of artificial voice: While a vocal uncanny valley exist, today's artificial voices manage to overcome it.

4.3 Theories on the uncanny valley

The present results conflict with two existing theories on the uncanny valley: That uncanniness is caused by either 1) categorical ambiguity or categorization difficulty, or 2) by misattribution of human qualities onto nonhuman entities. While distorted voices in Experiment 1were both uncanny and categorically ambiguous, pathological voices were uncanny despite being clearly categorized as human. In Experiment 2, distorted voices were uncanny despite having less mind or animacy attributed to them than normal voices, and with no differences compared to synthetic voices. Furthermore, pathological voices were uncanny in both experiments, contrasting the misattribution theory's prediction that uncanniness is caused by non-human entities.

The present data can be better explained by a deviation-from-familiarity account [15, 16]: both distorted and pathological voices are uncanny because they deviate from the pattern of human voice that has been experienced throughout life. Categorical ambiguity can correlate with stimulus uncanniness as categorically ambiguous stimuli [31] also deviate from typical appearance. Similarly, mind attribution can enhance configural processing of faces [43], which in turn may sensitize the negative evaluation of deviations [15]. Thus, mind attribution may increase uncanniness by sensitizing to deviations [40, 54, 55]. The interaction between attribution of human qualities, degree of configural processing, and uncanniness sensitivity can be explored in future research.

4.4 A moderated monotonic function of uncanniness

Rather than being a non-monotonic, valley-shaped function, the uncanny valley may consist of two or more monotonic functions with different slopes (e.g., one for an increase of likability from synthetic to full human variants, and one for a decrease of uncanniness from deviating or abnormal to typical humanlike variants). To test this, both experiments have investigated a moderated linear function of uncanniness.

Experiment 1 found that a moderated linear function could predict uncanniness, and Experiment 2 found that it could explain uncanniness better than a non-linear function of human likeness. Although the specific moderating variables differed between experiments, both “human” categorization and perceived organicness increased the effect of deviation on uncanniness. However, both variables also highly correlated with human likeness.

The investigated moderator variables are evolutionarily sensible: Disease avoidance may underlie the uncanny valley effect [44], and markers of infectious disease are expressed as changes from typical (human) appearance or behaviour [45]. Given that the threat of infection is present only in organic entities, avoidance of deviating organic or human entities should be effective for minimizing risk of infection. Meanwhile deviating yet clearly inorganic entities pose no threat of infection.

Alternatively, the increased uncanniness for less humanlike stimuli in organic entities or those categorized as hu man may be due to a higher level of perceptual experience with naturally humanlike stimuli: Perceptual expertise with a stimulus category increases the uncanniness of deviating exemplars [16].

4.5 Limitations and future directions

Interpretations of test results on a moderated linear function of the uncanny valley are limited due to the intercorrelation between the predictors. As multicollinearity cannot be excluded, the exact relationship between the predictor variables and uncanniness remains unclear. Future research may aim to tackle this problem using decorrelated predictors.

The use of linguistic content in the stimuli adds additional dimensions which could have influenced the results. For the difference between distortion effects on human and cat voices, a reduced intelligibility of the human voices but not cat voices due to distortion may have been a reason for the increased uncanniness for distorted human voices. Similarly, as distorted and pathological voices could be less intelligible, the additional processing need for these voices could have been a cause of uncanniness.

5. Conclusion

Contrary to previous research, this work reaffirms the existence of a vocal uncanny valley when summing over a range of disparate stimuli types. Modern synthetic voices successfully escape a vocal uncanny valley. Multiple theories on the uncanny valley have been tested, favouring deviation-based and disease avoidance accounts over categorical ambiguity or the perception of animacy. Furthermore, the results indicate that uncanniness of voices is best explained by a moderator of human categorization or perception of organicness on the effect of human likeness on uncanniness. This leads to multiple monotonic changes that can be observed as a valley across a single range.

References

1. Diel, A., Weigelt, S. & MacDorman, K. F. A meta-analysis of the Uncanny Valley's independent and dependent variables. ACM Transactions on Human-Robot Interaction 11, 1–33 (2021).
2. MacDorman, K. F. & Ishiguro, H. The uncanny advantage of using androids in cognitive and Social Science Research. Interaction Studies 7, 297–337 (2006).
3. Mathur, M. B. *et al.* Uncanny but not confusing: Multisite study of perceptual category confusion in the Uncanny Valley. Computers in Human Behavior 103, 21–30 (2020).
4. Mori, M., MacDorman, K. & Kageki, N. The uncanny valley [from the field]. IEEE Robotics & Automation Magazine 19, 98–100 (2012).
5. Meah, L. F. & Moore, R. K. The uncanny valley: A focus on misaligned cues. Social Robotics 256–265 (2014). doi:10.1007/978-3-319-11973-1_26
6. Mitchell, W. J. *et al.* A mismatch in the human realism of face and voice produces an Uncanny Valley. *i-Perception* 2, 10–12 (2011).
7. Baird, A. *et al.* The perception and analysis of the likeability and human likeness of synthesized speech. *Interspeech* 2018 (2018). doi:10.21437/interspeech.2018-1093
8. Baird, A. *et al.* The perception of vocal traits in synthesized voices: Age, gender, and human likeness. Journal of the Audio Engineering Society 66, 277–285 (2018).
9. Kimura, M. & Yotsumoto, Y. Auditory traits of "own voice". PLOS ONE 13, (2018).
10. Kühne, K., Fischer, M. H. & Zhou, Y. 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, (2020).
11. Romportl, J. Speech synthesis and Uncanny Valley. Text, Speech and Dialogue 595–602 (2014). doi:10.1007/978-3-319-10816-2_72
12. Mara, M., Appel, M. & Gnambs, T. Human-like robots and the Uncanny Valley: A meta-analysis of user responses based on the Godspeed Scales. (2021). doi:10.31234/osf.io/d4qc3
13. Chattopadhyay, D. & MacDorman, K. F. Familiar faces rendered strange: Why inconsistent realism drives characters into the Uncanny Valley. Journal of Vision 16, 7 (2016).
14. Käsyri, J., Förger, K., Mäkäräinen, M. & Takala, T. A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. Frontiers in Psychology 6, (2015).
15. Diel, A. & Lewis, M. Familiarity, orientation, and realism increase face uncanniness by sensitizing to facial distortions. Journal of Vision 22, 14 (2022).
16. Diel, A. & Lewis, M. The deviation-from-familiarity effect: Expertise increases uncanniness of deviating exemplars. PLOS ONE 17, (2022).
17. Diel, A. & MacDorman, K. F. Creepy cats and strange high houses: Support for configural processing in testing predictions of nine Uncanny Valley theories. Journal of Vision 21, 1 (2021).
18. Jung, N.-ri, Lee, M.-ji & Choi, H. The uncanny valley effect for celebrity faces and celebrity-based avatars. Korean Society for Emotion and Sensibility 25, 91–102 (2022).

19. Reber, R., Schwarz, N. & Winkielman, P. Processing fluency and aesthetic pleasure: Is Beauty in the perceiver's processing experience? Personality and Social Psychology Review 8, 364–382 (2004).
20. Friston, K. J. & Kiebel, S. Predictive coding: A free-energy formulation. Predictions in the Brain 231–246 (2011). doi:10.1093/acprof:oso/9780195395518.003.0076
21. Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J. & Frith, C. The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. Social Cognitive and Affective Neuroscience 7, 413–422 (2011).
22. Kreiman, J. E., Auszmann, A. & Gerratt, B. R. What does it mean for a voice to be “normal?” The Journal of the Acoustical Society of America 143, 1820–1820 (2018).
23. Kreiman, J. & Gerratt, B. R. Difference Limens for vocal aperiodicities. The Journal of the Acoustical Society of America 113, 2328–2328 (2003).
24. Kreiman, J., Gerratt, B. R., Precoda, K. & Berke, G. S. Individual differences in voice quality perception. Journal of Speech, Language, and Hearing Research 35, 512–520 (1992).
25. Altenberg, E. P. & Ferrand, C. T. Fundamental frequency in monolingual English, bilingual English/russian, and bilingual English/Cantonese young adult women. Journal of Voice 20, 89–96 (2006).
26. Amir, O. & Levine-Yundof, R. Listeners' attitude toward people with dysphonia. Journal of Voice 27, (2013).
27. Eadie, T. L., Rajabzadeh, R., Isetti, D. D., Nevdahl, M. T. & Baylor, C. R. The effect of information and severity on perception of speakers with adductor spasmodic dysphonia. American Journal of Speech-Language Pathology 26, 327–341 (2017).
28. Schroeder, S. R., Rembrandt, H. N., May, S. & Freeman, M. R. Does having a voice disorder hurt credibility? Journal of Communication Disorders 87, 106035 (2020).
29. Corradi, G., la Rosa, A. O., Ingram, G. & Villacampa, J. The creepy, the bad and the ugly: Exploring perceptions of moral character and social desirability in Uncanny Faces. (2021). doi:10.31234/osf.io/3rvgb
30. Cheetham, M., Pavlovic, I., Jordan, N., Suter, P. & Jancke, L. Category processing and the human likeness dimension of the Uncanny Valley Hypothesis: Eye-tracking data. Frontiers in Psychology 4, (2013).
31. Yamada, Y., Kawabe, T., & Ihaya, K. Categorization difficulty is associated with negative evaluation in the “Uncanny Valley” phenomenon. Japanese Psychological Research 55, 20–32 (2012).
32. Carr, E. W., Hofree, G., Sheldon, K., Saygin, A. P. & Winkielman, P. Is that a human? categorization (dis)fluency drives evaluations of agents ambiguous on human-likeness. Journal of Experimental Psychology: Human Perception and Performance 43, 651–666 (2017).
33. Winkielman, P., Schwarz, N. & Nowak, A. 5. affect and processing dynamics. Emotional Cognition 111–135 (2002). doi:10.1075/aicr.44.05win

34. Weis, P. P. & Wiese, E. Cognitive conflict as possible origin of the Uncanny Valley. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, 1599–1603 (2017).
35. Kreiman, J. & Sidtis, D. Foundations of Voice Studies. (2011). doi:10.1002/9781444395068
36. Sidtis, D. & Kreiman, J. In the beginning was the familiar voice: Personally familiar voices in the evolutionary and contemporary biology of Communication. Integrative Psychological and Behavioral Science 46, 146–159 (2011).
37. Bartneck, C., Kanda, T., Ishiguro, H. & Hagita, N. My robotic doppelgänger - a critical look at the Uncanny Valley. *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication* (2009). doi:10.1109/roman.2009.5326351
38. Hanson, D. Exploring the Aesthetic Range for Humanoid Robots. Proceedings of the ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science, (2006).
39. Kim, B., de Visser, E. & Phillips, E. Two uncanny valleys: Re-evaluating the uncanny valley across the full spectrum of real-world human-like robots. Computers in Human Behavior 135, 107340 (2022).
40. Yam, K. C., Bigman, Y. & Gray, K. Reducing the uncanny valley by dehumanizing humanoid robots. Computers in Human Behavior 125, 106945 (2021).
41. Fincher, K. M. & Tetlock, P. E. Perceptual dehumanization of faces is activated by Norm Violations and facilitates norm enforcement. Journal of Experimental Psychology: General 145, 131–146 (2016).
42. Fincher, K. M., Tetlock, P. E. & Morris, M. W. Interfacing with faces: Perceptual Humanization and dehumanization. Current Directions in Psychological Science 26, 288–293 (2017).
43. Deska, J. C., Lloyd, E. P. & Hugenberg, K. Facing humanness: Facial width-to-height ratio predicts ascriptions of humanity. (2017). doi:10.31234/osf.io/5ak6b
44. MacDorman, K. F. & Entezari, S. O. Individual differences predict sensitivity to the uncanny valley. Interaction Studies 16, 141–172 (2015).
45. Schaller, M., Park, J. & Faulkner, J. Prehistoric dangers and contemporary prejudices. European Review of Social Psychology 14, 105–137 (2003).
46. Workman, C. I. *et al.* Morality is in the eye of the beholder: The neurocognitive basis of the “anomalous-IS-bad” stereotype. (2020). doi:10.31234/osf.io/mz75u
47. Gray, K. & Wegner, D. M. Feeling Robots and human zombies: Mind perception and the Uncanny Valley. Cognition 125, 125–130 (2012).
48. Stein, J.-P. & Ohler, P. Venturing into the Uncanny Valley of mind—the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. Cognition 160, 43–50 (2017).
49. Andics, A. *et al.* Neural mechanisms for voice recognition. NeuroImage 52, 1528–1540 (2010).
50. Latinus, M., McAleer, P., Bestelmeyer, P. E. G. & Belin, P. Norm-based coding of voice identity in human auditory cortex. Current Biology 23, 1075–1080 (2013).
51. Loffler, G., Yourganov, G., Wilkinson, F. & Wilson, H. R. FMRI evidence for the neural representation of faces. Nature Neuroscience 8, 1386–1391 (2005).

52. Barsics, C. G. Person recognition is easier from faces than from voices. Psychologica Belgica 54, 244–254 (2014).
53. Deska, J. C., Lloyd, E. P. & Hugenberg, K. Facing humanness: Facial width-to-height ratio predicts ascriptions of humanity. (2017). doi:10.31234/osf.io/5ak6b
54. Müller, B. C., Gao, X., Nijssen, S. R. & Damen, T. G. I, robot: How human appearance and mind attribution relate to the perceived danger of robots. International Journal of Social Robotics 13, 691–701 (2020).
55. Yin, J., Wang, S., Guo, W. & Shao, M. More than appearance: The uncanny valley effect changes with a robot's mental capacity. Current Psychology (2021). doi:10.1007/s12144-021-02298-y
56. Cohen, G. *et al.* Statistics problems and solutions. The Statistician 37, 347 (1988).
57. Walden, P. R. Perceptual Voice Qualities Database (PVQD): Database characteristics. Journal of Voice 36, (2022).
58. Kawahara, H. *et al.* Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. 2008 *IEEE International Conference on Acoustics, Speech and Signal Processing* (2008). doi:10.1109/icassp.2008.4518514
59. Ho, C.-C. & MacDorman, K. F. Measuring the uncanny valley effect. International Journal of Social Robotics 9, 129–139 (2016).
60. Kätsyri, J., Mäkäräinen, M. & Takala, T. Testing the 'uncanny valley' hypothesis in semirealistic computer-animated film characters: An empirical evaluation of natural film stimuli. International Journal of Human-Computer Studies 97, 149–161 (2017).

Figures

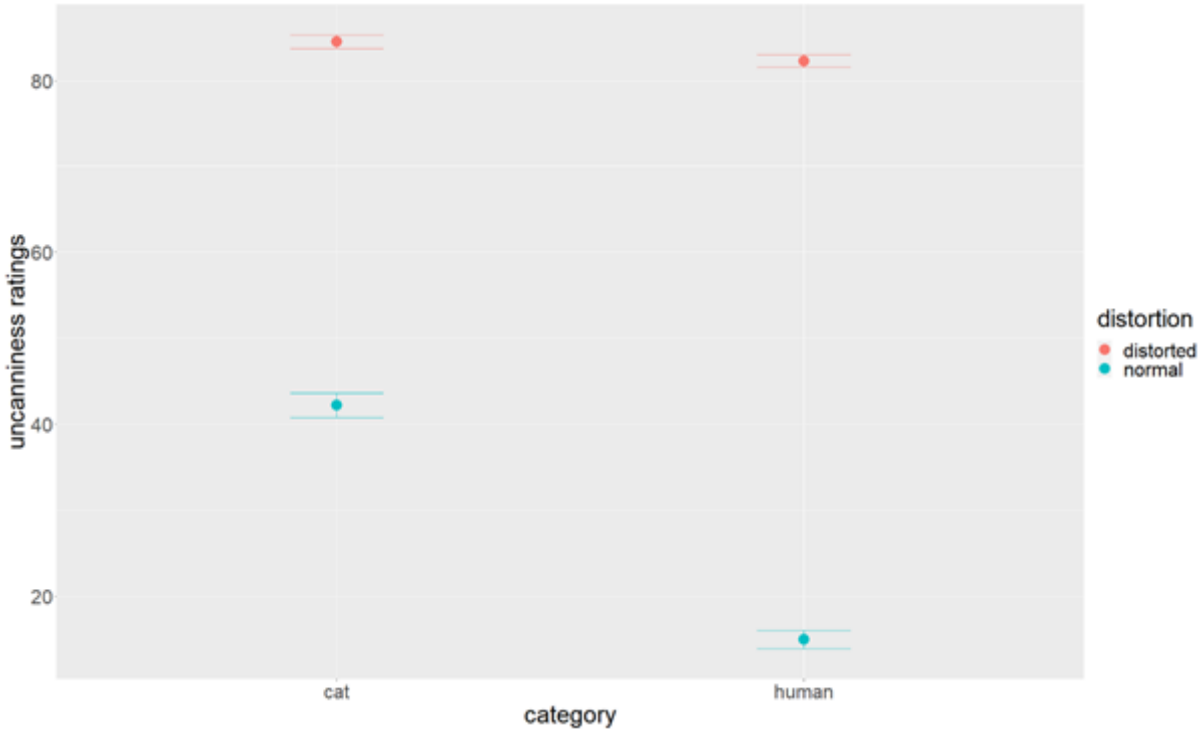

Figure 1

Uncanniness ratings across species (category) and distortion condition. Error bars represent standard errors.

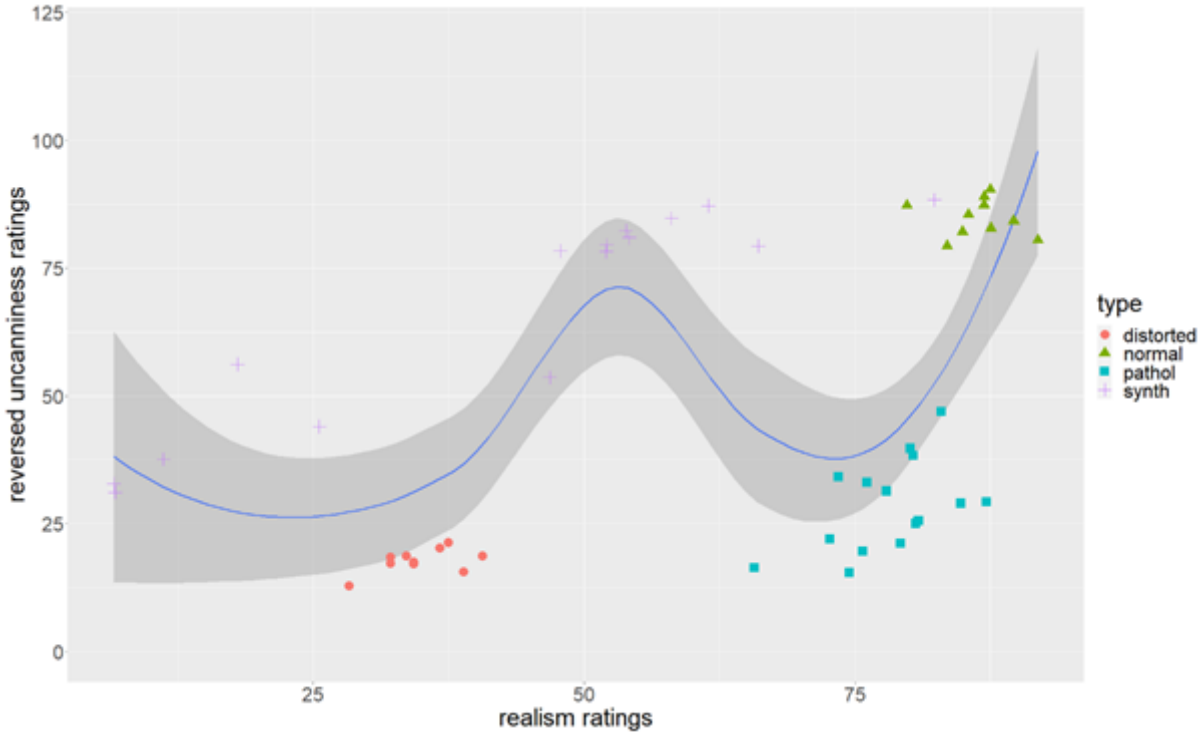

Figure 2

Reversed uncanniness ratings plotted as a function of human likeness ratings across the four voice conditions. Each point represents a stimulus. The grey area represents the 95% confidence interval.

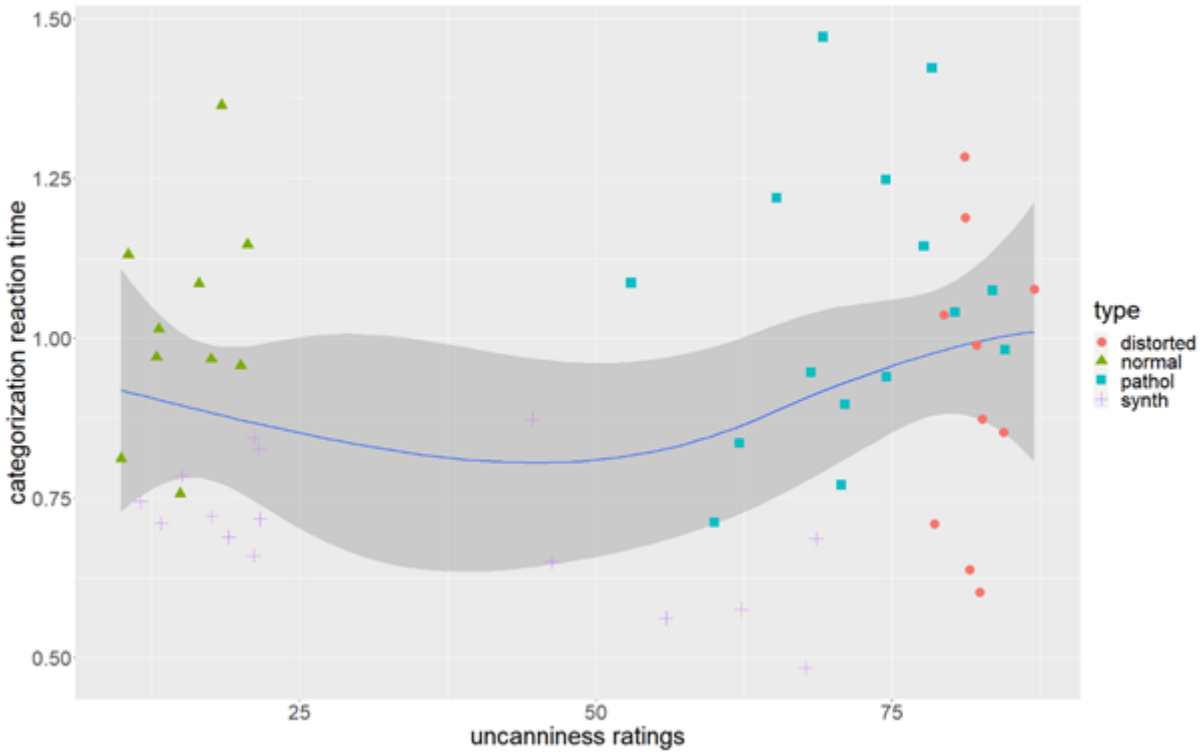

Figure 3

Reversed uncanniness ratings when only normal human and synthetic voices are used (left), or when only normal, distorted, and pathological voices are used (right). Grey areas represent 95% confidence intervals.

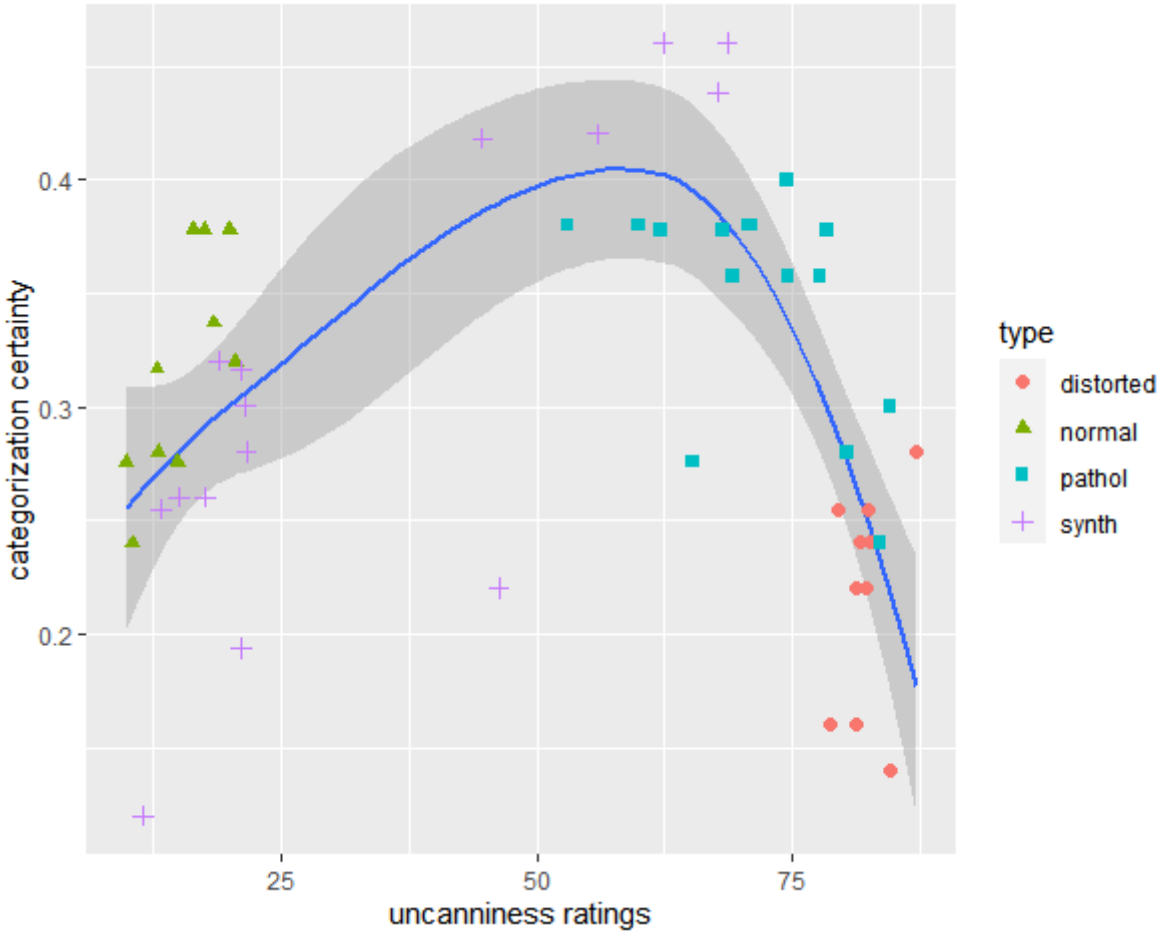

Figure 4

Categorization reaction time plotted against uncanniness ratings. No relation between the variables was found. Each point represents a stimulus.

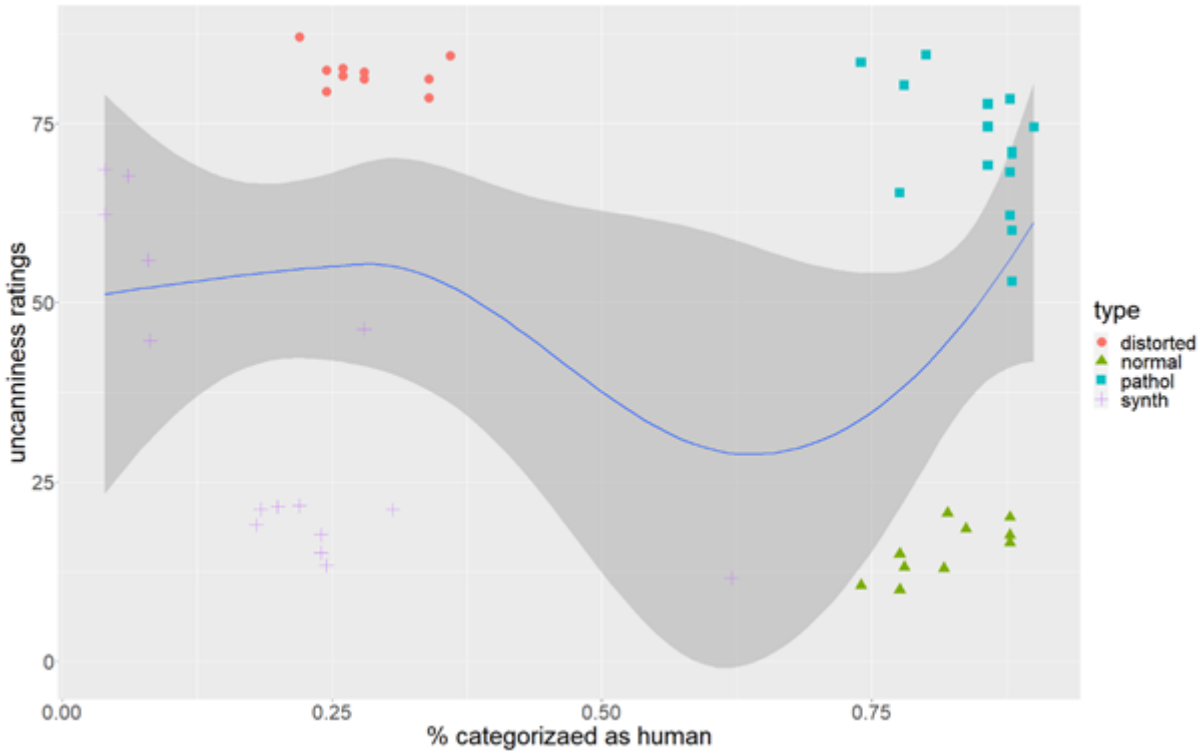

Figure 5

Voice certainty (0 = inconsistent categorization, 0.5 = consistent categorization across participants) plotted against uncanniness ratings. Each point represents a stimulus.

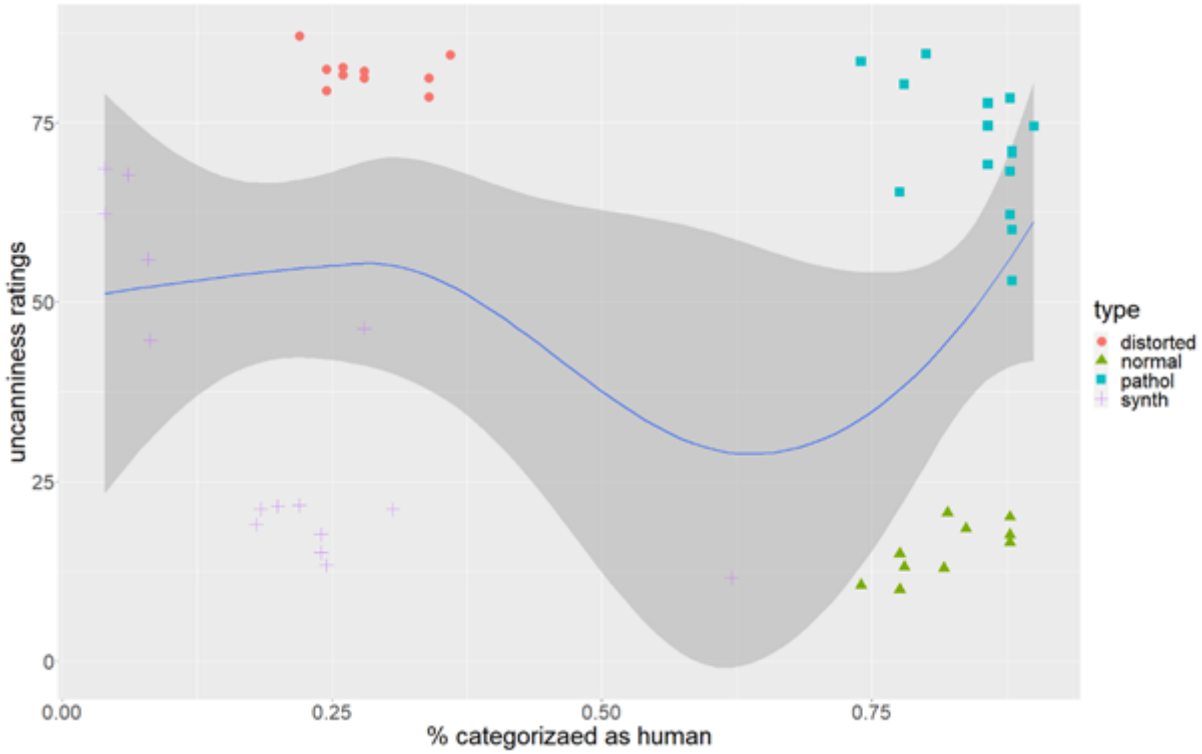

Figure 6

Uncanniness ratings plotted against the percentage of “human” categorization for each stimulus. The plot indicates differences between voice conditions in both variables.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx