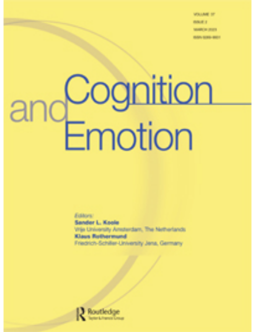

Routledge
Taylor & Francis Group

Cognition and Emotion

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pcem20

Perceived naturalness of emotional voice morphs

Christine Nussbaum, Manuel Pöhlmann, Helene Kreysa & Stefan R. Schweinberger

To cite this article: Christine Nussbaum, Manuel Pöhlmann, Helene Kreysa & Stefan R. Schweinberger (2023): Perceived naturalness of emotional voice morphs, Cognition and Emotion, DOI: 10.1080/02699931.2023.2200920

To link to this article: https://doi.org/10.1080/02699931.2023.2200920

Published online: 27 Apr 2023.

COGNITION AND EMOTION
https://doi.org/10.1080/02699931.2023.2200920

RESEARCH ARTICLE

Perceived naturalness of emotional voice morphs

Christine Nussbaum [a,b], Manuel Pöhlmann [a], Helene Kreysa [a,b] and Stefan R. Schweinberger [a,b,c]

[a]Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Germany; [b]Voice Research Unit, Friedrich Schiller University, Jena, Germany; [c]Swiss Center for Affective Sciences, University of Geneva, Switzerland

ABSTRACT

Research into voice perception benefits from manipulation software to gain experimental control over acoustic expression of social signals such as vocal emotions. Today, parameter-specific voice morphing allows a precise control of the emotional quality expressed by single vocal parameters, such as fundamental frequency (F0) and timbre. However, potential side effects, in particular reduced naturalness, could limit ecological validity of speech stimuli. To address this for the domain of emotion perception, we collected ratings of perceived naturalness and emotionality on voice morphs expressing different emotions either through F0 or Timbre only. In two experiments, we compared two different morphing approaches, using either neutral voices or emotional averages as emotionally non-informative reference stimuli. As expected, parameter-specific voice morphing reduced perceived naturalness. However, perceived naturalness of F0 and Timbre morphs were comparable with averaged emotions as reference, potentially making this approach more suitable for future research. Crucially, there was no relationship between ratings of emotionality and naturalness, suggesting that the perception of emotion was not substantially affected by a reduction of voice naturalness. We hold that while these findings advocate parameter-specific voice morphing as a suitable tool for research on vocal emotion perception, great care should be taken in producing ecologically valid stimuli.

ARTICLE HISTORY
Received 22 June 2022
Revised 3 April 2023
Accepted 5 April 2023

KEYWORDS
Naturalness; parameter-specific voice morphing; vocal emotions; F0; Timbre

1. Introduction

The human voice is a powerful transmitter of emotions, which are expressed through its acoustic properties (Scherer, 1986). The functional role of vocal parameters such as fundamental frequency contour (F0), timbre, amplitude/intensity and temporal features in the expression and perception of different emotions has been extensively studied (Banse & Scherer, 1996; Juslin & Laukka, 2003), but findings are mostly based on correlational data and do not allow causal inferences (Arias et al., 2021). Recently, however, technical and computational progress has led to the development of voice manipulation tools allowing experimental control over the acoustic properties of voices (Kawahara & Skuk, 2019). In *parameter-specific voice morphing*, a parameter of voice A is combined with another parameter of voice B. For example, one can resynthesise a voice with a happy F0 contour together with the timbre information of a non-emotional voice, resulting in a voice which expresses happiness only via F0.

While this technology offers exciting prospects in determining the acoustic correlates of socio-emotional signals in voices, it comes with a central caveat: these manipulations may lead to profound acoustic distortion, making them sound unnatural and less human-like. By naturalness, we understand the voice stimulus to be perceived as a plausible outcome of the human speech production system. To date, it is unclear how an impression of naturalness in voices is formed, how it may be affected by voice manipulations, and how it

CONTACT Christine Nussbaum christine.nussbaum@uni-jena.de Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Am Steiger 3/Haus 1, 07743 Jena, Germany; Stefan R. Schweinberger stefan.schweinberger@uni-jena.de Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Am Steiger 3/Haus 1, 07743 Jena, Germany

interacts with the perception of vocal emotions. In two experiments exploring the perception of naturalness in parameter-specific voice morphs, we investigated these open questions. In what follows, we will first discuss the potentials and caveats of parameter-specific voice morphing. Then, we outline insights into voice naturalness across different research domains, motivating the design of our experiments.

1.1. The potentials and limits of parameter-specific voice morphing in vocal emotional research

Voice morphing can be used to study relevant questions in the field of vocal emotion perception, including perceptual adaptation (Bestelmeyer et al., 2010), categorical vs. dimensional representations of emotions (Giordano et al., 2021) and the perceptual consequences of emotional caricaturing (Whiting et al., 2020). A special form of voice morphing constitutes the *parameter-specific manipulation* of different acoustic cues, which has been used to study how this facilitates the perception of vocal age, gender, and identity (Kawahara & Skuk, 2019; Skuk et al., 2020; Skuk & Schweinberger, 2014), as well as – recently – vocal emotion (Nussbaum, Schirmer, & Schweinberger, 2022; Nussbaum, von Eiff, et al., 2022). In most cases, the main focus has been on the functional role of F0 (perceived as voice pitch) and timbre (perceived as voice quality, and formally defined as "the difference between two voices of identical F0, intensity and temporal structure", ANSI (1973)). For emotional stimuli, the relative importance of F0 and timbre differs as a function of emotion category, but overall F0 seems to be more important for the perception of emotional quality, at least in the normal-hearing population (Nussbaum, Schirmer, & Schweinberger, 2022). In individuals using cochlear implants, by contrast, von Eiff et al. (2022) observed a greater reliance on timbre cues. Further, timbre seems to play a predominant role in emotional adaptation (Nussbaum, von Eiff, et al., 2022), similar to findings on voice gender adaptation (Skuk et al., 2015). Interestingly, these findings are in contrast to Hubbard and Assmann (2013), who found F0 to be more important than timbre in gender and emotion adaptation, based on the absence of effects in a F0-removed condition. Both Skuk et al. (2015) and Nussbaum, von Eiff, et al. (2022) argued that this discrepancy could be explained by a lack of naturalness in Hubbard and Assmanns (2013) F0-removed condition, which might have eliminated the adaptation effects. It therefore seems essential that parameter-specific voice morphing results in natural sounding stimuli, by which we understand them to constitute a *plausible outcome of the human speech production system*. In fact, many studies using voice manipulation explicitly comment on the naturalness of their stimulus material (Grichkovtsova et al., 2012; Nussbaum, von Eiff, et al., 2022; Skuk et al., 2015), though this is usually based on subjective listening impression only. Vocal emotions, however, are often characterised by acoustic extremes, and this could result in reduced naturalness to a degree that compromises stimulus validity, thus calling for an objective validation.

1.2. Perspectives on voice naturalness across different research domains

Due to different perspectives and motivations, *voice naturalness* is not uniformly defined across research contexts. First, in *speech-language pathology*, naturalness forms an important rehabilitation outcome in conditions such as stuttering, dysarthria, Parkinson's disease, developmental communication disorders, and speech prostheses (Anand & Stepp, 2015; Coughlin-Woods et al., 2005; Eadie & Doyle, 2002; Klopfenstein et al., 2020; Mackey et al., 1997; Martin et al., 1984; Meltzner & Hillman, 2005; Péron et al., 2015; Yorkston et al., 1999, 1990). In these rehabilitative contexts, it is usually defined as a quality of voice that allows individuals to express their wants and needs efficiently, appropriately, and socially adequately (Klopfenstein et al., 2020). Note that this conceptualisation has a subjective component with a strong dependency on the vocal expectations of listeners (Klopfenstein et al., 2020). Second, in *human-robot-interaction*, research is driven by the observation that robots and computers can be perceived as social actors (Nass et al., 1994), whose likability and human-likeness are important factors for user satisfaction and acceptance (Gong, 2008; McGinn & Torre, 2019; Mitchell et al., 2011; Schweinberger et al., 2020). To the extent that perceptions of naturalness correspond to those of human-likeness, one of the major challenges in the auditory domain is the creation of synthesised speech that sounds natural (Baird, Jørgensen, et al., 2018a; Baird, Parada-Cabaleiro, et al., 2018b; Mayo et al., 2011; Nusbaum et al., 1997; Yamasaki et al., 2017). Finally, a similar conceptualisation can be found in studies addressing the *methodology of voice research,* which rely on the ecological validity of the stimulus materials used to study human voice perception (Alku et al., 1999; Burton &

Blumstein, 1995; Kawahara & Skuk, 2019). The empirical operationalisation of voice naturalness varies across studies. Most often, naturalness was quantified based on listener ratings on Likert scales (cf. Klopfenstein et al., 2020). Alternatively, listeners chose the more natural option out of a stimulus pair or classified voices as human vs. robotic (e.g. Mayo et al., 2011; Nusbaum et al., 1997). In several studies on human-robot-interaction, naturalness was even experimentally manipulated (e.g. Gong, 2008; McGinn & Torre, 2019; Mitchell et al., 2011).

Despite the conceptual and empirical heterogeneity of these different perspectives, a surprisingly consistent picture emerges concerning the acoustic features that are associated with perceived naturalness in voices. There is ample evidence that *fundamental frequency variation* is linked to perceived naturalness (Anand & Stepp, 2015; Baird, Parada-Cabaleiro, et al., 2018b; Ilves & Surakka, 2013; Vojtech et al., 2019). For example, when comparing different speech synthesis methods, Baird, Parada-Cabaleiro, et al. (2018b) found a relationship between perceived human-likeness (i.e. naturalness) and F0 variation showing that voices with higher F0 variation are generally rated as more natural than those with lesser variation. Likewise, Anand and Stepp (2015) found F0 variation and naturalness to be highly correlated in patients with Parkinson's disease. Another important determinant of voice naturalness seems to be the covariation of F0 and formant frequencies, which was observed in recorded human speech (Assmann & Katz, 2000). In a subsequent experiment, frequency-shifted speech samples were rated as more natural when they followed this relationship, while utterances were judged to be less natural the more they deviated from it (Assmann et al., 2006). Further, synthetic voices which contain microvariations such as jitter and shimmer are perceived as more natural than those without (Yamasaki et al., 2017). Finally, several studies reported that a low speech rate was associated with a decline in perceived naturalness (Klopfenstein et al., 2020; Mackey et al., 1997; Vojtech et al., 2019; Yorkston et al., 1990).

Despite these initial insights into the acoustic determinants of voice naturalness, little is known about their interplay with vocal emotion perception. The few studies that have investigated the effects of voice naturalness on the processing of lexical emotional content (Ilves et al., 2011; Ilves & Surakka, 2013) have emphasised the importance of vocal naturalness for conveying emotional messages. However, to the best of our knowledge, the interaction of voice naturalness with emotional prosody has never been explicitly assessed; a gap we aim to fill with the following two experiments using emotional pseudowords.

1.3. Aims of the present studies

In the present studies, we investigated the perceived naturalness and emotionality of parameter-specific voice morphs containing emotional information in either F0 or timbre only, while the other parameter is held at an emotionally non-informative level. In the F0-condition, the emotional F0-contour is combined with the timbre of the non-emotional reference stimulus, and vice versa in the Timbre-condition. This procedure inevitably results in a mismatch between fundamental frequency and formant frequencies, which has been reported to be an important acoustic feature related to voice naturalness (Assmann et al., 2006). Accordingly, we predicted that the F0 and the Timbre conditions would be perceived as less natural compared to a Full condition comprising both parameters. We further considered F0 variance as an important factor of perceived naturalness, based on the literature discussed above. When creating the parameter-specific voice morphs using neutral voices as non-emotional reference, we noted that a neutral voice quality in these recordings was expressed by a monotonous voice with limited F0 variance. This could have a particularly detrimental effect on the Timbre-morphs, where the emotional timbre is combined with the monotonous F0 of the neutral voices. We therefore employed a second morphing approach, using an emotional average as reference, exhibiting more F0 variance. Both neutral voices and averaged emotions have in common that they are assumed to be non-informative with respect to the given emotional quality. We predicted that the naturalness of the Timbre condition could be improved by using the emotional average as reference. In Experiment 1, all stimuli were rated regarding perceived naturalness and emotionality. In Experiment 2, participants chose the more natural-sounding option of two voices that differed only with regard to the morphing reference, to allow a direct comparison of these approaches.

2. Experiment 1

2.1. Method

2.1.1 Stimuli

Original Audio Recordings. The original audio recordings from a database of vocal actor portrayals were provided by Sascha Frühholz, similar to the ones used in Frühholz et al. (2015). For voice morphing, we used three pseudowords (/belam/, /molen/, /loman/) expressing happiness, pleasure, fear, sadness, and produced in an emotionally neutral voice by 4 male and 4 female speakers. Stimuli were validated with a rating study to ensure that they conveyed the intended emotions sufficiently (for more details, see https://osf.io/sybrd/).

Voice Morphing. Using the TANDEM-STRAIGHT software (Kawahara et al., 2013, 2008), we created morphing trajectories between each emotion and a reference stimulus of the same speaker and pseudoword, generating resynthesised vocal samples on these trajectories via weighted interpolation of the originals. Of importance, TANDEM-STRAIGHT allows independent interpolation of five different parameters: (1) F0-contour, (2) timing, (3) spectrum-level, (4) aperiodicity, and (5) spectral frequency; the latter three parameters constitute *timbre* (for a more detailed description see Kawahara and Skuk (2019)).

For the purposes of this study, three different morph types (Morph Types) were created (see Figure 1):

Full-Morphs were stimuli with all TANDEM-STRAIGHT parameters taken from the emotional stimulus (corresponding to 100% from the emotion and 0% from reference), except for the timing parameter, which was always taken from the reference (corresponding to 0% emotion and 100% reference). **F0-Morphs** were stimuli with the F0-contour taken from the emotional version, but timbre and timing taken from the reference. **Timbre-Morphs** were stimuli with all timbre parameters taken from the emotional version, but F0 and timing from the reference. Note that the timing was kept constant across all conditions to allow a pure comparison of F0 vs. timbre.

As reference stimuli, we used two different options (reference types): we either used the neutral expression or an emotional average of the four emotions. Accordingly, we assumed that both reference types would be uninformative with respect to the expression of one of the four emotions, even if they did not necessarily sound fully non-emotional.

In addition to the Morph Types, the reference stimuli were included in the rating study. In total, this resulted in 8 (speakers) × 3 (pseudowords) × 4 (emotions) × 3 (morphing conditions) × 2 (reference types) + 48 reference (8 speakers × 3 pseudowords × 2 reference types) = 624 stimuli.

Using PRAAT (Boersma, 2018) (2018), we normalised all stimuli to a root-mean-square of 70 dB SPL (duration M = 751 ms, Min = 411 ms, Max = 968 ms, SD = 138 ms). A summary of the acoustic properties of the resulting stimuli can be found in Table 1. Stimulus examples can be found on https://osf.io/jzn63/.

2.1.2 Data collection and participants

Data were collected online via PsyToolkit (Stoet, 2010, 2017). Participants were required to use a computer with a physical keyboard and headphones. As browser, we recommended Google Chrome, and excluded Safari for technical reasons. Participants had to be between 18 and 40 years old, speak German as their native language, have normal hearing abilities and ensure a quiet environment for the duration of the study. Data collection took place from February to April 2021. All participants had to provide informed consent before completing the experiment, and data were collected completely anonymised. To avoid fatigue, each participant rated only stimuli of one of the pseudowords. Average duration of the experiment was about 30 min. Participants who completed the experiment were compensated with course credit. The experiment was in line with the ethical guidelines of the German Society of Psychology (DGPs). All methodological aspects of this research were covered by an approval from the local ethics committee of the Friedrich Schiller University Jena (Reg.-Rr. FSV 19/045).

Prior to data collection, we conducted a power-analysis using the R-package 'Superpower' (Lakens & Caldwell, 2019) with a medium effect size f = .23, an alpha level of .05 and a power of .80 for the interaction of Morph Type and Reference Type on the naturalness ratings, resulting in a required sample size of 16. Since participants rated only stimuli from one pseudoword each, we decided to collect 16 participants per pseudoword, resulting in a total required sample size of 48. This would allow detection of a small effect if data could be collapsed across pseudowords (f = .13). The online experiment was accessed by approximately 100 participants, of whom 59 contributed complete data. Of these, eight datasets (13.6%) had to be removed (four participants reported that the sounds were not played properly, three admitted in the post-experimental questionnaire that they had responded randomly, one had a native language other than German). Thus, the final sample consisted of 51 participants (40 females, 11 males, aged 19–31 years [M = 21.49; Mdn = 21; SD = 2.65], with 16/18/17 per pseudoword). Three participants reported a minor hearing problem such as occasional tinnitus,

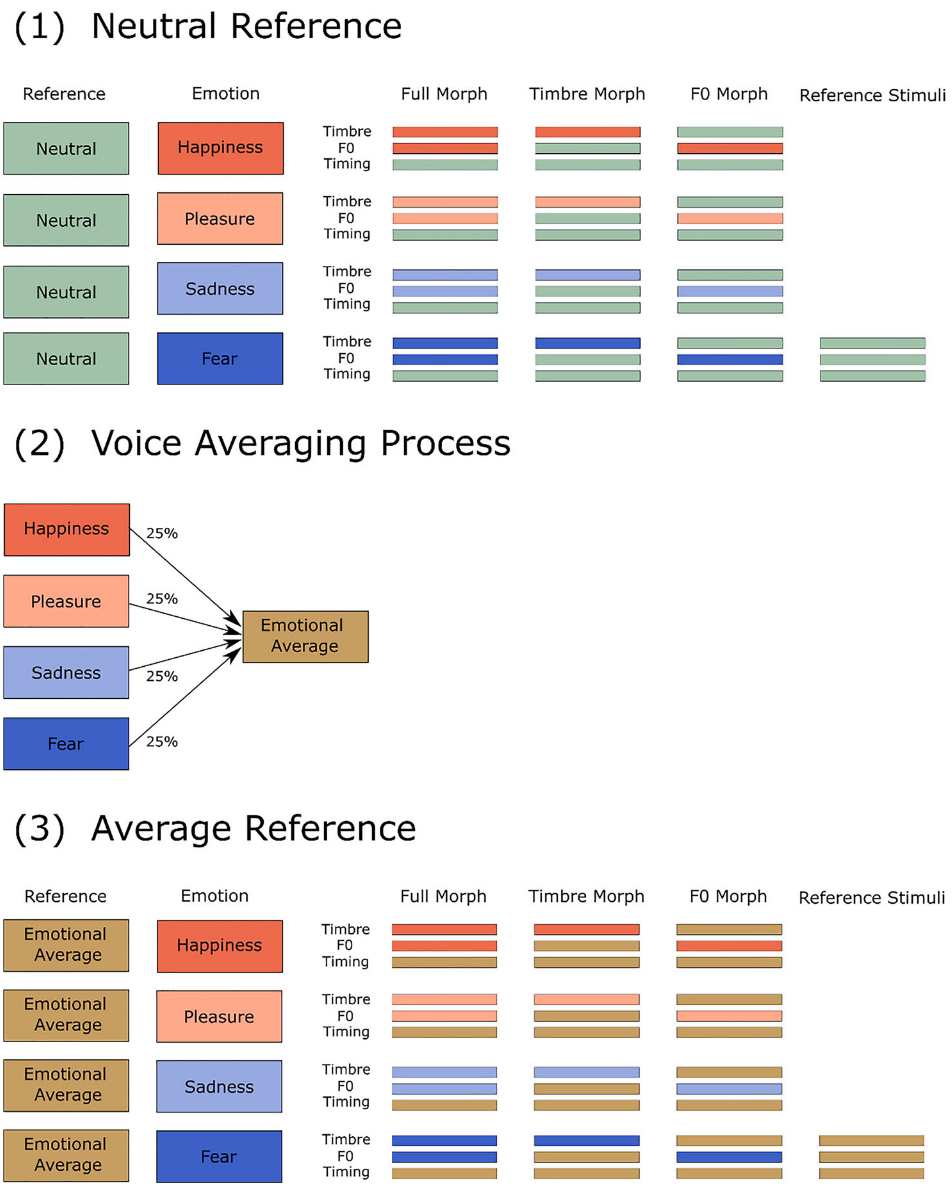

Figure 1. Illustration of parameter-specific voice morphs based on two different references. Note. (1) Morphing matrix for stimuli with an actor portrayal of "neutral" as reference. (2) Schematic depiction of the voice averaging process. (3) Morphing matrix for stimuli with averaged voices as reference.

but since they reported that they were not limited in their hearing, they were included in the analysis.

2.1.3 Design

Prior to the two rating tasks, participants entered demographic information such as age and gender.

Ratings of perceived Naturalness. Participants were assigned to one of the three pseudoword-conditions randomly and instructed to rate the naturalness of each voice they heard. They were informed that "natural" in the context of this study meant that "the voices sound human/natural and do not sound distorted or robotic in any way" [German original: "dass sich die Stimmen tatsächlich menschlich/natürlich anhören und nicht auf irgendeine Art verzerrt oder robotisch klingen"]. The participants entered their ratings via keyboard on a 6-point Likert scale with the endpoints 1 = very inauthentic/robotic and

Table 1. Acoustic properties of the stimulus material used in this study.

MType	Ref	F0 Mean	F0 SD	F0 Glide	FormDisp	HNR
Female						
Full	AVG	260	42	−39	1082	20
Full	NEU	258	41	−34	1083	20
F0	AVG	260	42	−39	1095	21
F0	NEU	258	41	−34	1054	19
Tbr	AVG	247	25	−37	1077	20
Tbr	NEU	197	11	1	1075	20
Male						
Full	AVG	173	36	−43	1045	16
Full	NEU	173	36	−47	1041	16
F0	AVG	173	36	−43	1045	16
F0	NEU	173	36	−47	972	15
Tbr	AVG	158	21	−43	1037	16
Tbr	NEU	110	4	0	1041	15

Notes: All acoustical parameters were adapted from (McAleer et al., 2014) and extracted using Praat software (Boersma, 2018) and the F0 contour information from the TANDEM-STRAIGHT object in Matlab (MATLAB, 2020). *F0 Glide* = F0End − F0Start; *Formant Dispersion (FormDisp)* = ratio between consecutive formant means (from F1 to F4, maximum formant frequency set to 5.5 kHz, window length 0.025 s); *HNR* (harmonics-to-noise ratio) was extracted with the cross-correlation method (mean value; time step = 0.01 s; min pitch = 75 Hz; silence threshold = 0.1, periods per window = 1.0). Full = full morphs, F0 = F0 morphs, Tbr = Timbre morphs, AVG = average reference, NEU = neutral reference.

6 = very human. Note that we specifically opted for a Likert-Scale without a midpoint, as midpoints have been shown to be interpreted inconsistently by raters (Nadler et al., 2015). After 8 practice trials with different stimuli, all 208 voice stimuli were presented in randomised order in two blocks of 104 trials each, and participants could take a short break in between. Each trial started with a green fixation cross and after 300 ms the rating stimulus was played. Then, a screen with the 6-point scale was presented and participants had to enter a response within 5000 ms after stimulus onset. If no answer was given in that time, participants were prompted to respond faster by a slide with red lettering for 500 ms. Otherwise, only a black screen was shown, before the next trial started.

Ratings of perceived Emotionality. After completion of the naturalness ratings, the same stimuli were rated for emotionality on a rating scale from 1 = very negative to 6 = very positive. Note that on this rating scale, 1–3 corresponds to negative and 4–6 to positive valence with different intensity levels. The procedure was identical to the naturalness ratings, except that the voice was played 500 ms (instead of 300 ms) after presentation of the green fixation cross, due to a programming error. Stimuli were presented in a different randomised order.

Post-experimental questions. After the experiment, participants were asked whether all sounds were played and whether they had understood the instructions. Furthermore, they could comment on the task and indicate whether they had developed a certain strategy.

2.1.4 Data processing and analysis

Trials of omission (< 0.01%) were removed. Data were analysed using R Version 4.1.0 (R Core Team, 2021). Analyses of Variance (ANOVAs) and correlational analyses were performed on averaged rating data, whereas cumulative link mixed models (calculated with the “ordinal” Package in R, Christensen, 2015) were used to model ratings of single trials. Please note that we interpret our findings based on effect sizes rather than significance values only, in line with recent recommendations (Cumming, 2014; Fritz et al., 2012). Due to the novelty of the design, our approximate power analysis resulted in a somewhat overpowered design, making even small effects ($d < 0.4$) appear significant, which we nevertheless treated as negligible. Preprocessed data, analysis scripts, stimulus examples and supplemental materials can be found in the associated OSF repository (https://osf.io/jzn63/).

2.2. Results

2.2.1 Perceived naturalness

Naturalness ratings were averaged across speakers and the reference stimuli (average and neutral) were excluded for the first analysis. Mean ratings were analysed with a mixed-effects 3 × 4 × 3 × 2 ANOVA with the between-subject factor pseudoword (/belam/, /molen/, /loman/) and the within-subject factors Emotion (happiness, pleasure, fear, and sadness), Morph Type (Full, F0, and Timbre) and Reference Type (NEU and AVG). A summary of all significant main effects and interactions is displayed in Table 2.[1]

Post-hoc tests on the main effect of **Pseudoword** revealed that the pseudoword /molen/ was perceived as less natural than the other two ($|ts(32.69)| \geq 2.46$, $ps \leq .019$), which did not differ ($t(30) = 0.05$, $p = .962$; Ms = 3.27 ± 0.08, Ms = 3.28 ± 0.10, Ms = 2.90 ± 0.11, for /belam/, /loman/, and /molen/, respectively). There was a prominent main effect of **Morph Type**, but crucially, this was qualified by an interaction of **Morph Type x Reference Type** (Figure 2, A). A comparison of the different Morph Types separately for the two Reference Types revealed the following pattern: With the neutral reference, Timbre-Morphs were rated as substantially less natural than the other two ($|ts(50)| \geq 15.23$, $ps \leq .001$, $ds \geq 2.15$ [1.65 2.65]), whereas Full and F0 differed only marginally ($t(50) = 1.95$, $p = .057$, $d = 0.28$ [−0.01 0.56]). Note that in the Timbre-Morphs, the F0 information is contributed by the reference stimuli, in this case neutral. With the average refence, both Timbre- and F0-Morphs were rated as more unnatural than the Full-Morphs ($|ts(50)| \geq 12.85$, ps $\leq .001$, $ds \geq 1.82$ [1.36 2.26]). However, the difference between them was very small ($t(50) = 2.43$, $p = .019$, $d = 0.34$ [0.06 0.63]). For the same interaction, a comparison of the different Reference Types within each Morph Type revealed that in F0-morphs, stimuli with neutral as reference were perceived as more natural ($|t(50)| = 9.23$, $p \leq .001$, $d \geq 1.31$ [0.93 1.69]), whereas in Timbre- and Full-morphs, stimuli with averaged emotions as reference were perceived as more natural ($ts(50)| \geq 4.32$, $p \leq .001$, $d \geq 0.61$ [0.31 0.91]).

The prominent interaction of **Morph Type × Emotion** suggested that while in all Emotions Timbre was rated as more unnatural than the other two, this effect was most pronounced for happiness. For the detailed statistical report including the other main effects and interactions, please refer to [https://osf.io/jzn63/]. Finally, a planned comparison between the two reference conditions revealed that the neutral one was rated as substantially more natural than the averaged emotions ($t(50) = 7.53$, $p < .001$, $d = 1.06$ [0.71 1.41], refer to Figure 2, B).

2.2.2 Perceived emotionality

Similar to the naturalness ratings, mean ratings of emotionality were analysed with a mixed-effects 3 × 4 × 3 × 2 ANOVA with the between-subject factor pseudoword (/belam/, /molen/, /loman/) and the within-subject factors Emotion (happiness, pleasure, fear, and sadness), Morph Type (Full, F0, and Timbre) and Reference Type (NEU and AVG), refer to Table 2. The main effects of **Morph Type** and **Emotion** were qualified by a prominent interaction (Figure 3, A). While in Full-Morphs, emotionality ratings were widespread, this range was reduced in F0-Morphs and almost absent in the timbre condition. Thus, emotions could be much better discriminated in the F0 compared to the Timbre condition, suggestion that F0 is more effective in signalling emotional quality, although not as informative as Full Morphs. Further, F0 seems important for the differentiation in the valence dimension, since on average, emotionality ratings in the timbre condition were all slightly negative (i.e. below 3.5), even for the positive emotions happiness and pleasure. Crucially, this pattern was not further qualified by the Reference Type, suggesting it does not depend on the morphing reference used. For the detailed statistical report including the other main effects and interactions, see [https://osf.io/jzn63/]. A planned comparison between the two reference conditions revealed that the averaged emotions were rated a bit more negative than the neutral ones, but this effect was small ($(|t(50) = 2.72$, $p = .009$, $d = 0.38$ [0.10 0.67], refer to Figure 3, B).

2.2.3 Relationship between perceived naturalness and emotionality

To assess whether the perception of naturalness and emotionality was linked, we averaged both ratings across participants to correlate mean ratings of each stimulus. We found no relationship, r (624) = -.043, p = .279, suggesting that we observe both natural and unnatural stimuli across all levels of emotional quality (Figure 4, A). However, since emotional valence and intensity are combined in our emotionality rating, we ran a second analysis in which we were interested in the link between naturalness and emotional intensity. To this end, we recoded our rating scores such that responses of 1 or 6

Table 2. Results of the 3 × 4 × 3 × 2 mixed-effects ANOVAs on mean ratings of Naturalness and Emotionality.

			Naturalness			Emotionality		
	df1	df2	F	p	η_p^2 [95%-CI]	F	p	η_p^2 [95%-CI]
Pseudoword (Pw)	2	48	4.63	.014	.16 [.01 .34]		-	
Morph Type	2	96	257.78	<.001	.84 [.79 .88]	79.22	<.001	.62 [.51 .71]
Emotion	3	144	54.41	<.001	.53 [.43 .62]	174.69	<.001	.78 [.73 .83]
Pw × Emotion	6	144	3.36	0.007	.12 [.02 .20]		-	
Morph Type × Reference Type	2	96	104.05	<.001	.68 [.58 .75]	9.50	<.001	.17 [.05 0.29]
Morph Type × Emotion	6	288	120.01	<.001	.71 [.66 .75]	121.54	<.001	.72 [.67 .76]
Reference Type × Emotion	3	144	10.42	<.001	.18 [.07 .28]	19.24	<.001	.29 [.17 .40]
Pw x Morph Type × Emotion	12	288	2.92	.001	.11 [.02 .15]		-	
Pw x Morph Type × Reference Type	4	96		–		3.11	.024	.12 [.00 .22]

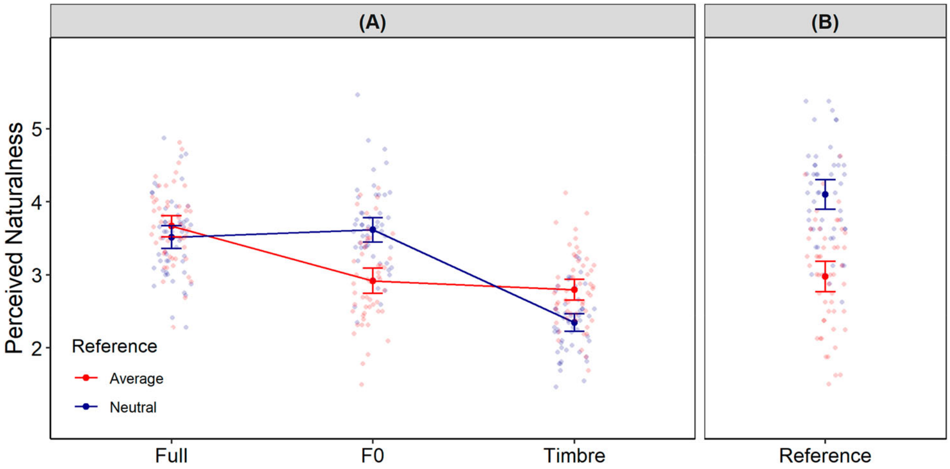

Figure 2. Interaction of Morph Type and Reference Type on perceived Naturalness. Note. Whiskers represent 95%-confidence intervals. Dots represent individual participants' data.

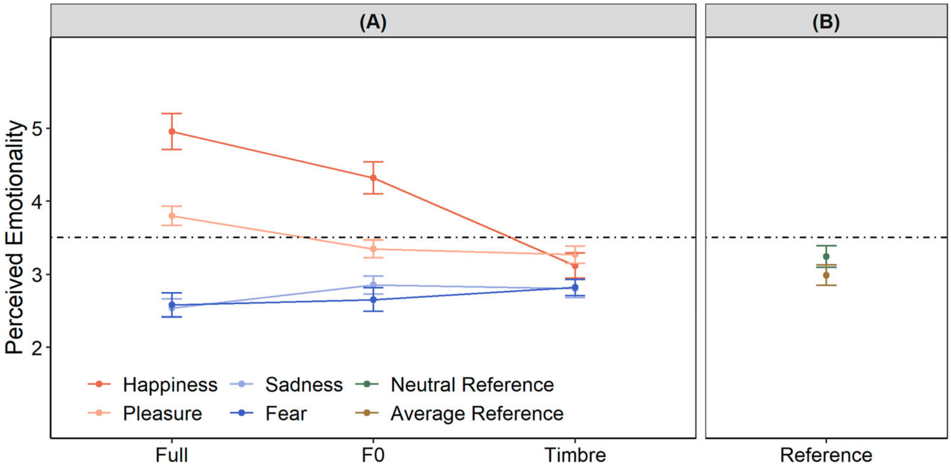

Figure 3. Interaction of Morph Type and Emotion on perceived Emotionality. Note. The dashed line marks the midpoint of the rating scale at 3.5. Whiskers represent 95%-confidence intervals.

corresponded to high intensity (= 3), 2 or 5 to medium intensity (= 2), and 3 or 4 to low intensity (= 1). However, there was no correlational relationship either, r (624) < .001, p = .988 (Figure 4, B). Thus, we concluded that perceived naturalness and perceived emotional quality/intensity were not related in our stimuli.

2.2.4 Link between ratings and acoustic properties of the stimuli

For modelling, the influence of acoustic properties on the perception of naturalness and emotionality in the stimuli (including neutral and average reference stimuli), the standardised predictor variables $F0_{Mean}$, $F0_{SD}$, $F0_{Glide}$, Formant Dispersion (FormDisp) and Harmonic to Noise Ratio (HNR) were chosen to calculate two cumulative link mixed models with the syntax.

$$\text{Rating} \sim F0_{Mean} + F0_{SD} + F0_{Glide} + \text{FormDisp} + \text{HNR} + (1|\text{Participant}) + (1|\text{SpID})$$

on ratings of naturalness and emotionality separately. We included random intercepts for both participant and speaker as the model fit was better compared to models with one of these random intercepts only (cf. OSF). The results are summarised in Table 3. In short, all included parameters seemed to play a

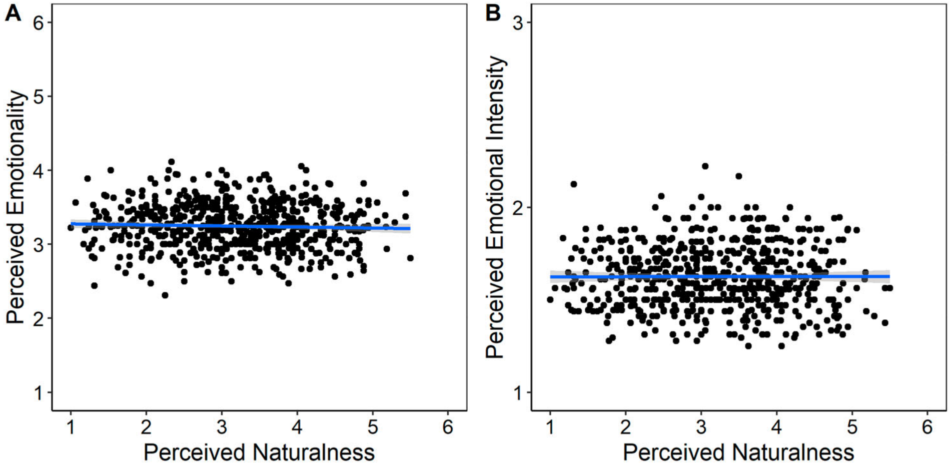

Figure 4. Mean ratings of perceived naturalness, emotionality (A), and emotional intensity (B). Note. Data points represent mean ratings of individual stimuli averaged across participants. The blue line illustrates the linear regression, the shaded grey area around it the standard error.

Table 3. Results of the regression analyses using cumulative link mixed models.

	Naturalness				Emotionality			
	β	SE	z	*p*	β	SE	z	*p*
F0 Mean	**−0.551**	0.044	−12.59	< .001	**−0.608**	0.047	−12.88	< .001
F0 SD	**0.596**	0.035	17.21	< .001	**1.182**	0.040	29.77	< .001
F0 Glide	**−0.228**	0.021	−10.85	< .001	**−0.275**	0.022	−12.38	< .001
FormDisp	**−0.070**	0.027	−2.64	.008	0.012	0.027	0.44	0.663
HNR	**0.387**	0.031	12.68	< .001	**−0.232**	0.031	−7.44	< .001

significant role for both ratings, except for Formant Dispersion in the context of emotionality ratings. For both ratings, the biggest effect was observed for F0 variability.

2.3. Short summary

In Experiment 1, we showed that perceived naturalness was affected by the choice of the morphing reference, whereas perception of emotionality was not. In fact, we did not find evidence for a relationship between perception of naturalness and emotionality. In a regression analysis, most of the vocal parameters we took into consideration predicted ratings of both naturalness and emotionality, with the biggest effect observed for F0 variability.

3. Experiment 2

In a second Experiment, we aimed to replicate and expand the findings of Experiment 1 with a different paradigm. In a two-alternative forced-choice task, participants listened to pairs of corresponding stimuli (same speaker, emotion, pseudoword and morph-type) which only differed in the morphing reference. Their task was to decide which sample sounded more natural. This provided a direct comparison of morphing approaches.

3.1. Method

3.1.1 Stimuli

Stimuli were identical to Experiment 1.

3.1.2 Data collection and participants

Data were collected online via PsyToolkit (Stoet, 2010, 2017) from May to July 2021, with the same general conditions and inclusion criteria as in Experiment 1. Average duration of the experiment was about 35 min. The online experiment was accessed by approximately 65 participants of whom 34 contributed complete data. Of these, six datasets (17.4%) had to be removed (two participants reported that the sounds were not played properly, three exceeded the age range of 18-40, one had >5% trials of omission). Thus, the final sample consisted of 28 participants (14 females, 14 males, aged 18–30 years [M = 22.39; Mdn = 22; SD = 2.75]).

3.1.3 Design

Each trial started with a fixation cross for 500 ms. Afterwards, a black screen with two loudspeaker symbols labelled "1" and "2" appeared. Then the first sound was played, visually highlighted with the first sound symbol turning green. After an inter-stimulus interval of 750 ms, the second sound was played, with the other sound symbol turning green. The participants decided via keypress (f = 1, j = 2) which voice sounded more natural, in a time window of 5000 ms after the second stimulus offset. Within trials, the two stimuli were of the same speaker, emotion, pseudoword and morph type, and differed only in reference category (AVG/NEU). Trials with the neutral and average reference stimuli were included as well. Whether AVG or NEU was presented first was randomised. After 6 practice trials with different stimuli, all 312 voice pairs were presented in randomised order in four blocks of 52 trials each, and participants could take short breaks between blocks.

3.1.4 Data processing and analysis

Trials of omission (< 0.01%) were removed. To keep the analysis parallel to Experiment 1, we transformed data to display the response tendency as the proportion of "average sounds more natural"-responses, and then analysed them with an ANOVA and subsequent t-tests. Note that we additionally pursued an alternative analysis approach by running a logistic regression on individual trial data. This analysis resulted in a virtually identical pattern of effects and is reported on https://osf.io/jzn63/.

3.2. Results

Responses were averaged across speakers and pseudoword and trials with reference stimuli (average/neutral) were excluded for the first analysis. A 3 × 4 repeated-measures ANOVA on the response tendency revealed main effects of both factors Morph Type, $F(2, 54) = 45.34$, $p < .001$, $\eta^2 = .63$, 95%-CI [0.46 0.73]; and Emotion $F(3, 81) = 11.39$, $p < .001$, $\eta^2 = .30$, 95%-CI [0.13 0.43]. Post-hoc analyses revealed that in Full and Timbre morphs, average-referenced stimuli were perceived as more natural, whereas in F0 morphs the neutral option was chosen more often ($|ts(27)| \geq 2.51$, $ps \leq .019$, $|ds| \geq 0.48$ [0.08 0.88], see Figure 5, A). This pattern was further supported by a planned comparison against 0.5 as the point without a response tendency ($|ts(27)| \geq 3.09$, $ps \leq .004$, $|ds| \geq 0.60$ [0.18 1.00]). In the trials comparing average and neutral stimuli directly, neutral stimuli were chosen more often (t-test against 0.5: $t(27) = -2.28$, $p = .031$, $|d| = 0.44$ [0.04 0.83], see Figure 5, B). This represents a full replication of the pattern found in Experiment 1 (refer to Figure 2). The main effect of Emotion was mainly driven by happiness, which was perceived as more natural with average reference, and sadness, which was perceived more natural with neutral reference ($M_{Happiness} = .55 \pm 0.02$; $M_{Pleasure} = .51 \pm 0.02$; $M_{Fear} = .50 \pm 0.02$; $M_{Sadness} = .47 \pm 0.02$; detailed analysis on https://osf.io/jzn63/).

3.3. Short summary

Experiment 2 employed a two alternative-forced choice task to provide a conceptual replication of Experiment 1 regarding the perception of naturalness as a function of morphing reference. For Full and Timbre morphs, average-referenced emotional voices were perceived as more natural than neutral-referenced emotional voices, whereas the opposite was found for F0 morphs.

4. Discussion

The present experiment explored a number of important determinants of the perception of naturalness and emotionality in voices. Specifically, we investigated how the impression of naturalness in voices is formed, how it can be affected by different voice manipulations (especially those related to parameter-specific voice morphing), and how it can interact with the perception of vocal emotions. In line with our hypotheses, we observed that voice manipulation affected perceived naturalness, presumably due to an inherent mismatch between fundamental frequency contour and timbre features. Perceived naturalness was also strongly affected by fundamental frequency variation: On the one hand, naturalness could be tremendously improved in the Timbre condition by using the average emotion as reference, which expressed much more F0 variation than the neutral one. On the other hand, a regression analysis revealed F0 variation to be an important predictor of both naturalness and emotionality ratings. Most importantly, we found no evidence that emotionality ratings were affected by a lack of stimulus naturalness, suggesting that stimuli like the ones used here are

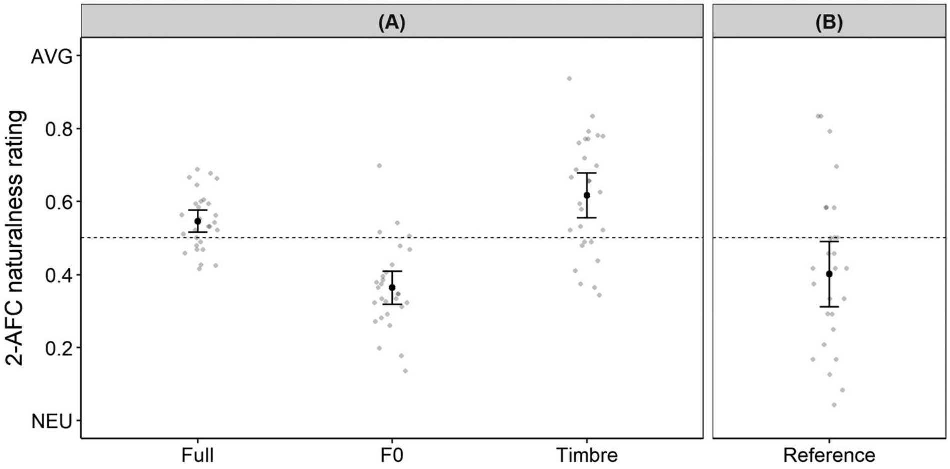

Figure 5. Response tendency towards the more natural reference category as a function of Morph Type Note. The dashed lined represents the 0.5 point with no response tendency. Whiskers represent 95%-confidence intervals. Grey dots represent individual participants' data. 2-AFC = two alternative-forced choice task.

valid for vocal emotional research. In what follows, we discuss how these findings relate to (a) the role of naturalness in emotion perception, (b) the possible existence of an uncanny valley for voices, and (c) the potentials and limits of emotional voice morphing.

4.1. The role of naturalness in the perception of emotion and other social signals

Although the communication of emotion is limited to humans and living creatures, emotional processing per se is not. In 1944, Heider and Simmel (1944) presented a short film with geometrical figures moving on the screen and asked participants to describe it. Intriguingly, most of the participants provided a description of animated beings with personalities, backstories, and emotions. One figure was consistently perceived as aggressive and angry, whereas another was perceived as frightened. This shows that humans attribute human traits and emotions to non-living objects. In fact, our brain displays a strong tendency to pick up and process emotions, even in highly artificial settings (Hortensius et al., 2018; Spatola & Wudarczyk, 2021). This property is deliberately employed for improving communication with non-human actors, such as robots (Crumpton & Bethel, 2016). Thus, emotional processing may not depend on naturalness or human-likeness. Our data fit into this line of argumentation, by showing that the processing of emotionality was remarkably unrelated to the perceived naturalness of voices. It is noteworthy that in the facial domain, Calder et al. (2000) observed a comparable pattern using emotional caricatures: With increasing caricaturing level, faces were rated as more emotionally intense, despite being perceived as less natural. Based on these findings, one could assume that emotional processing can suppress any disruptive effects of unnaturalness or artificial circumstances.

However, both theoretical considerations and conflicting empirical evidence suggest that this might not be entirely true: Models of both face and voice perception suggest that voices and faces are "special" to the brain, in the sense that they recruit neural resources which are not recruited by other types of stimuli (Belin et al., 2011; Young & Bruce, 2011). With stimulus material deviating profoundly in human-likeness, recruitment of these networks might be disrupted. Evidence from the domain of face recognition that computer-generated faces do not fully tap face expertise (Crookes et al., 2015) could indicate that the same might hold for emotional processing. Indeed, studies using electroencephalography (EEG) suggest that naturalness and emotionality interact at the neural level for both voices and faces. For example, Schirmer and Gunter (2017) manipulated "voiceness" by using spectrally rotated versions of non-verbal exclamations and observed modulations of several ERP components (N100, P200, LPP), which partly interacted with emotional processes. Similarly, manipulation of "face-realism" affected the N170 components and the LPP, and also interacted with the processing of facial expression (Schindler et al., 2017). Further, the human voice is perceived as more expressive and likeable than an expressive synthetic one (Cabral et al., 2017; Ilves & Surakka, 2013). Finally, adaptation paradigms using sine tones or F0-removed stimuli as adaptors fail to elicit reliable adaptation aftereffects in voices (Hubbard & Assmann, 2013; Schweinberger et al., 2008), presumably due to their unnatural/non-human quality.

Taken together, these findings imply that naturalness of the stimulus materials does play a role in emotional processing. Yet, the circumstances under which somewhat unnatural stimuli still allow a direct generalisation to perception of real human voices remain unclear. As it stands, emotional processing can to some degree disregard unnatural features but is likely not completely detached from them. Thus, it remains the responsibility of researchers to give this matter explicit consideration for specific voice stimulus sets. For the voice stimuli used in the present study, naturalness does not seem to play a crucial role for emotional processing.

4.2. Is there an uncanny valley for voices?

A question related to the interplay of naturalness and emotion is the existence of an uncanny valley for voices. The uncanny valley, originally proposed by Mori in 1970 (Mori et al., 2012) has been described as a sudden drop in the likeability of humanoid robots that almost approach, but do not entirely reach a human-like appearance. This almost human-like quality is assumed to evoke a sudden feeling of eeriness, which would most certainly affect the processing of emotionality in voices. For the present investigation, this is especially relevant, since emotional voice morphs are resynthesised from human voices, and thus could fall into this “almost human-like” gap which is assumed to evoke the uncanny valley phenomenon. However, empirical evidence for the uncanny valley effect across modalities is scarce and inconsistent (Kätsyri et al., 2015). So far, it has been observed for static and dynamic visual depictions of robots, as well as for a mismatch of human-likeness between the auditory and the visual channel (Mitchell et al., 2011; Schweinberger et al., 2020). For voices, there is no evidence for an uncanny valley so far, and previous studies only found a linear relationship between human-likeness and likeability (Baird, Parada-Cabaleiro, et al., 2018b). Although our data cannot speak to this directly, the absence of a relationship between naturalness and emotionality ratings may indirectly support the notion that a strong effect of naturalness on listeners’ own feelings towards the presented voices is rather unlikely.

4.3. Emotional voice morphing – a tool of unlimited possibilities?

In the past, research linking voice acoustics to socio-emotional signals was predominantly based on correlational inference. This has improved through the development of voice manipulation tools, such as voice morphing (Kawahara & Skuk, 2019). While offering exciting research prospects, the degrees of freedom allowed by this method are both tempting and intimidating, especially when morphing vocal emotional utterances: It is possible to morph between two emotions of choice (Nussbaum, von Eiff, et al., 2022), or to morph one emotion with respect to a reference, which in turn can be non-emotional (i.e. neutral) or emotionally ambiguous (i.e. average). If an emotional average is used, consideration should be given to the emotions that enter into this average: An average comprised of the six basic emotions (Ekman, 1992) would sound different from the one used in the present experiments, composed of two negative and two positive emotions. Further, voice averaging itself constitutes a special form of voice morphing which is still in its infancy and technically very challenging (Kawahara & Skuk, 2019). The more voices enter an average, the more prone it is to stimulus artefacts such as reduced aperiodicities and higher harmonics-to-noise ratios (Bruckert et al., 2010). This could make the average sound less natural than original human recordings, a pattern we observed in both Experiments, when our averages were compared to neutral voices (cf. Figures 2 and 5, B). Further, one can not only interpolate between voices, but also extrapolate and thus create emotional caricatures (Whiting et al., 2020). Finally, morphing allows a parameter-specific manipulation of the voice, as for F0 and timbre in the present study. While undeniably powerful, all these options carry a potential to affect empirical findings to a substantial degree, making them hard to compare across studies – an important caveat when designing and interpreting voice morphing studies.

For faces, Calder et al. (2000) demonstrated that perception of emotional caricatures was comparable when they were created with respect to a neutral, an averaged or a different emotional face. For voices, the present data also confirm that the perception of emotion was not substantially affected by the choice of morphing reference. Still, many of our methodological choices are likely to have impacted on our results: First, our design was limited to four emotions balanced in valence and we included only these to create the emotional average. Second, we specifically focused on the contrast of F0 and timbre as vocal parameters. We found that F0 played a larger role than timbre in emotion discrimination, in line with previous research in the normal-hearing population (Nussbaum, Schirmer, & Schweinberger, 2022). However, we would not claim that this would necessarily generalise to a different set of emotions. Third, we showed that even though different voice morphing approaches did not affect emotional ratings, they affected perceived naturalness. In both experiments, F0 morphs were perceived as more natural with neutral reference, but Timbre and Full morphs were perceived as more natural with average reference. The effect of the average reference on the Timbre morphs was predicted, because of its increased F0 variation (Baird, Parada-Cabaleiro, et al., 2018b; Vojtech et al., 2019). More importantly however, perceived naturalness between F0 and Timbre was comparable in the average-referenced condition only, thus excluding differences in naturalness as a potential confound when comparing the two. This clearly advocates the average-referenced approach as more suitable for research contrasting these two parameters, since naturalness and emotional processing may interact at the neural level, as discussed above. Altogether, the present investigation demonstrates both the potentials and the pitfalls of emotional voice morphing and encourages an explicit consideration of its methodological subtleties.

4.4. Directions for future research

The present investigation only provides a starting point in understanding the role of naturalness in the context of voice perception and emotional voice processing. For example, without further investigation, we can only speculate how stimulus naturalness might affect emotion perception of emotions other than the ones that were studied in our experiments, or actual classification performance instead of emotionality ratings. Further, in the present study, we focused on emotional prosody in pseudowords. However, emotions are frequently expressed in the context of longer utterances of coherent speech, or non-verbally through vocal expressions such as cries, laughter, moans or other short exclamations (Pell et al., 2015). As these non-verbal vocalisation are not speech-embedded, their acoustic features can be very distinct, and their processing differs from emotional prosody with regard to the time-course and underlying neural structure (Paulmann & Kotz, 2018; Pell et al., 2015). Therefore, the present findings may not generalise across all types of vocal emotional expression, leaving room for further exploration.

While several studies comment on the acoustic quality of their stimulus material (Grichkovtsova et al., 2012; Nussbaum, von Eiff, et al., 2022; Skuk et al., 2015), objective research efforts to validate stimulus material with respect to such aspects remain sparse. In this context, it is important to note that perceived naturalness may not be a function of physical stimulus properties alone but can also be affected by perceptual exposure and adaptation. For instance, it is well-known that a sufficient degree of adaptation to highly unnatural (e.g. spatially expanded or compressed) faces can make subsequent faces of the same distortion appear far more natural (Kloth et al., 2017; Webster & Maclin, 1999). Future research will have to elucidate the psychological and neuronal mechanisms by which perceptual experience with morphed stimuli (by experimental participants, but potentially also by researchers who are in daily contact with such stimuli) may affect perceptions of naturalness. More generally, with the present experiments, we hope to inspire more research on naturalness and its impact on the processing of different social signals in the vocal and facial domain. This could offer insight into the processing of both human and non-human signals, making valuable contributions to psychological models of person perception as well as human-robot interaction, and, for example, inform speech synthesis technologies for patients with voice disabilities or laryngectomy (e.g. Yamagishi et al., 2012).

5. Summary and conclusion

In two experiments, we explored the impact of parameter-specific voice morphing on the perception of naturalness and emotionality. We compared Full, F0 and Timbre morphs of emotions based on two different morphing references, neutral and average. In line with our hypotheses, we found that parameter-specific voice morphing affected perceived naturalness. In F0 morphs, stimuli with neutral as reference were perceived as more natural, while Timbre and Full morph stimuli were perceived as more natural with averaged emotions as reference. Crucially, naturalness of F0 and Timbre morphs was comparable only in the average-reference condition, making this form of reference more suitable for future research. Finally, we found no relationship between ratings of emotionality and naturalness. This suggests that perceived emotionality was not extensively affected by a lack of stimulus naturalness and that parameter-specific voice morphing is thus a suitable tool for vocal emotional research.

Note

1. Note that the number of participants is slightly unequal for each pseudoword (16/18/17). Therefore, we ran a second analyses where we randomly excluded three participants to have equal group size, resulting in an identical pattern of effects.

Acknowledgements

The original voice recordings that served as a basis for creating our stimulus material were provided by Sascha Frühholz. We are grateful to all participants of the study. We thank Laura Luther, Sven Kachel, and Annett Schirmer for helpful contributions to this manuscript, and Andrea E. Kowallik for contributing to the validation of voice stimuli used in this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

CN has been supported by the German National Academic Foundation ("Studienstiftung des Deutschen Volkes"). Experiment 1 was conducted by MP in partial fulfilment of a Bachelor degree. Experiment 2 was conducted by Laura Luther in partial fulfilment of a Bachelor degree.

Data availability

Supplemental figures and tables, analysis scripts, and preprocessed data can be found on the associated OSF repository (https://osf.io/jzn63/).

Credit author statement

Christine Nussbaum – Conceptualisation, Methodology, Software, Visualisation, Formal analysis, Writing – Original Draft, Supervision.

Manuel Pöhlmann – Conceptualisation, Methodology, Formal analysis, Visualisation, Writing – Review & Editing.

Helene Kreysa – Methodology, Supervision, Writing – Review & Editing, Supervision.

Stefan R. Schweinberger – Conceptualisation, Writing – Review & Editing, Supervision.

ORCID

Christine Nussbaum http://orcid.org/0000-0003-2718-2898
Manuel Pöhlmann http://orcid.org/0000-0002-1062-1201
Helene Kreysa http://orcid.org/0000-0001-7163-7023
Stefan R. Schweinberger http://orcid.org/0000-0001-5762-0188

References

Alku, P., Tiitinen, H., & Näätänen, R. (1999). A method for generating natural-sounding speech stimuli for cognitive brain research. *Clinical Neurophysiology*, *110*(8), 1329–1333. https://doi.org/10.1016/S1388-2457(99)00088-7

Anand, S., & Stepp, C. E. (2015). Listener perception of monopitch, naturalness, and intelligibility for speakers with Parkinson's disease. *Journal of Speech, Language, and Hearing Research*, *58*(4), 1134–1144. https://doi.org/10.1044/2015_JSLHR-S-14-0243

ANSI. (1973). Terminology. In *Psychoacoustical. S3. 20* (pp. 61–67). American National Standards Institute, Psychoacoustical.

Arias, P., Rachman, L., Liuni, M., & Aucouturier, J. J. (2021). Beyond correlation: Acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review*, *13*(1), 12–24. https://doi.org/10.1177/1754073920934544

Assmann, P. F., Dembling, S., & Nearey, T. M. (2006). Effects of frequency shifts on perceived naturalness and gender information in speech. In *INTERSPEECH*. Symposium conducted at the meeting of Citeseer.

Assmann, P. F., & Katz, W. F. (2000). Time-varying spectral change in the vowels of children and adults. *The Journal of the Acoustical Society of America*, *108*(4), 1856–1866. https://doi.org/10.1121/1.1289363

Baird, A., Jørgensen, S. H., Parada-Cabaleiro, E., Cummings, N., Hantke, S., & Schüller, B. (2018a). The perception of vocal traits in synthesized voices: Age, gender, and human likeness. *Journal of the Audio Engineering Society*, *66*(4), 277–285. https://doi.org/10.17743/jaes.2018.0023

Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummings, N., & Schüller, B. (2018b, September 2). The perception and analysis of the likeability and human likeness of synthesized speech. In *Interspeech* 2018 (pp. 2863–2867). ISCA. 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

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 711–725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

Bestelmeyer, P. E. G., Rouger, J., DeBruine, L. M., & Belin, P. (2010). Auditory adaptation in vocal affect perception. *Cognition*, *117*(2), 217–223. https://doi.org/10.1016/j.cognition.2010.08.008

Boersma, P. (2018). *Praat: Doing phonetics by computer* [Computer program]: Version 6.0.46, retrieved January 2020 from http://www.praat.org/.

Bruckert, L., Bestelmeyer, P. E. G., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., Kawahara, H., & Belin, P. (2010). Vocal attractiveness increases by averaging. *Current Biology*, *20*(2), 116–120. https://doi.org/10.1016/j.cub.2009.11.034

Burton, M. W., & Blumstein, S. E. (1995). Lexical effects on phonetic categorization: The role of stimulus naturalness and stimulus quality. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(5), 1230–1235. https://doi.org/10.1037/0096-1523.21.5.1230

Cabral, J. P., Cowan, B. R., Zibrek, K., & McDonnell, R. (2017). The influence of synthetic voice on the evaluation of a virtual character. In *Interspeech* 2017 (pp. 229–233). ISCA. https://doi.org/10.21437/Interspeech.2017-325

Calder, A. J., Rowland, D., Young, A. W., Nimmo-Smith, I., Keane, J., & Perrett, D. I. (2000). Caricaturing facial expressions. *Cognition*, *76*(2), 105–146. https://doi.org/10.1016/S0010-0277(00)00074-3

Christensen, R. H. B. (2015). Package 'ordinal'. *Stand*, *19*, 2016.

Coughlin-Woods, S., Lehman, M. E., & Cooke, P. A. (2005). Ratings of speech naturalness of children ages 8-16 years. *Perceptual and Motor Skills*, *100*(2), 295–304. https://doi.org/10.2466/pms.100.2.295-304

Crookes, K., Ewing, L., Gildenhuys, J. D., Kloth, N., Hayward, W. G., Oxner, M., Pond, S., & Rhodes, G. (2015). How well do computer-generated faces tap face expertise? *PLoS One*, *10*(11), e0141353. https://doi.org/10.1371/journal.pone.0141353

Crumpton, J., & Bethel, C. L. (2016). A survey of using vocal prosody to convey emotion in robot speech. *International Journal of Social Robotics*, *8*(2), 271–285. https://doi.org/10.1007/s12369-015-0329-4

Cumming, G. (2014). The New statistics. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *Journal of Speech, Language, and Hearing Research*, *45*(6), 1088–1096. https://doi.org/10.1044/1092-4388(2002/087)

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, *99*(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18. https://doi.org/10.1037/a0024338

Frühholz, S., Klaas, H. S., Patel, S., & Grandjean, D. (2015). Talking in fury: The cortico-subcortical network underlying angry vocalizations. *Cerebral Cortex*, *25*(9), 2752–2762. https://doi.org/10.1093/cercor/bhu074

Giordano, B. L., Whiting, C., Kriegeskorte, N., Kotz, S. A., Gross, J., & Belin, P. (2021). The representational dynamics of perceived voice emotions evolve from categories to dimensions. *Nature Human Behaviour*, *5*(9), 1203–1213. https://doi.org/10.1038/s41562-021-01073-0

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*(4), 1494–1509. https://doi.org/10.1016/j.chb.2007.05.007

Grichkovtsova, I., Morel, M., & Lacheret, A. (2012). The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, *54*(3), 414–429. https://doi.org/10.1016/j.specom.2011.10.005

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243. https://doi.org/10.2307/1416950

Hortensius, R., Hekele, F., & Cross, E. S. (2018). The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(4), 852–864. https://doi.org/10.1109/TCDS.2018.2826921

Hubbard, D. J., & Assmann, P. F. (2013). Perceptual adaptation to gender and expressive properties in speech: The role of fundamental frequency. *The Journal of the Acoustical Society of America*, *133*(4), 2367–2376. https://doi.org/10.1121/1.4792145

Ilves, M., & Surakka, V. (2013). Subjective responses to synthesised speech with lexical emotional content: The effect of the naturalness of the synthetic voice. *Behaviour & Information Technology*, *32*(2), 117–131. https://doi.org/10.1080/0144929X.2012.702285

Ilves, M., Surakka, V., & Vanhala, T. (2011). The effects of emotionally worded synthesized speech on the ratings of emotions and voice quality. In (pp. 588–598). Springer. https://doi.org/10.1007/978-3-642-24600-5_62

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814. https://doi.org/10.1037/0033-2909.129.5.770

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). 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*, 390. https://doi.org/10.3389/fpsyg.2015.00390

Kawahara, H., Morise, M., & Skuk, V. G. (2013). Temporally variable multi-aspect N-way morphing based on interference-free speech representations. *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1–10). https://doi.org/10.1109/APSIPA.2013.6694355

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3933–3936). https://doi.org/10.1109/ICASSP.2008.4518514

Kawahara, H., & Skuk, V. G. (2019). Voice morphing. In S. Frühholz & P. Belin (Eds.), *The Oxford handbook of voice perception* (pp. 685–706). Oxford University Press.

Klopfenstein, M., Bernard, K., & Heyman, C. (2020). The study of speech naturalness in communication disorders: A systematic review of the literature. *Clinical Linguistics & Phonetics*, *34*(4), 327–338. https://doi.org/10.1080/02699206.2019.1652692

Kloth, N., Rhodes, G., & Schweinberger, S. R. (2017). Watching the brain recalibrate: Neural correlates of renormalization during face adaptation. *Neuroimage*, *155*, 1–9. https://doi.org/10.1016/j.neuroimage.2017.04.049

Lakens, D., & Caldwell, A. R. (2019). *Simulation-based power-analysis for factorial ANOVA designs*. https://doi.org/10.31234/osf.io/baxsf

Mackey, L. S., Finn, P., & Ingham, R. J. (1997). Effect of speech dialect on speech naturalness ratings: A systematic replication of Martin, Haroldson, and Triden (1984). *Journal of Speech, Language, and Hearing Research*, *40*(2), 349–360. https://doi.org/10.1044/jslhr.4002.349

Martin, R. R., Haroldson, S. K., & Triden, K. A. (1984). Stuttering and speech naturalness. *Journal of Speech and Hearing Disorders*, *49*(1), 53–58. https://doi.org/10.1044/jshd.4901.53

MATLAB. (2020). *version 9.8.0 (R2020a)*. The MathWorks Inc.

Mayo, C., Clark, R. A. J., & King, S. (2011). Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis. *Speech Communication*, *53*(3), 311–326. https://doi.org/10.1016/j.specom.2010.10.003

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PLoS One*, *9*(3), e90779. https://doi.org/10.1371/journal.pone.0090779

McGinn, C., & Torre, I. (2019, March 11–14). Can you tell the Robot by the Voice? An exploratory study on the role of voice in the perception of robots. In 2019 *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 211–221). IEEE. https://doi.org/10.1109/HRI.2019.8673305

Meltzner, G. S., & Hillman, R. E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language, and Hearing Research*, *48*(4), 766–779. https://doi.org/10.1044/1092-4388(2005/053)

Mitchell, W. J., Szerszen, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & Macdorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *I-Perception*, *2*(1), 10–12. https://doi.org/10.1068/i0415

Mori, M., Macdorman, K. F., & Kageki, N. (2012). The Uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology*, *142*(2), 71–89. https://doi.org/10.1080/00221309.2014.994590

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence - CHI '94*. ACM Press.

Nusbaum, H. C., Francis, A. L., & Henly, A. S. (1997). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology*, *2*(1), 7–19. https://doi.org/10.1007/BF02215800

Nussbaum, C., Schirmer, A., & Schweinberger, S. R. (2022). *Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates*. In press.

Nussbaum, C., Schirmer, A., & Schweinberger, S. R. (2022). Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates. *Social Cognitive and Affective Neuroscience*, *17*(12), 1145–1154. https://doi.org/10.1093/scan/nsac033

Nussbaum, C., von Eiff, C. I., Skuk, V. G., & Schweinberger, S. R. (2022). Vocal emotion adaptation aftereffects within and across speaker genders: Roles of timbre and fundamental frequency. *Cognition*, *219*, 104967. https://doi.org/10.1016/j.cognition.2021.104967

Paulmann, S., & Kotz, S. A. (2018). The electrophysiology and time course of processing vocal emotion expressions. In S. Frühholz, P. Belin, S. Frühholz, P. Belin, & K. R. Scherer (Eds.), *The Oxford handbook of voice perception* (pp. 458–472). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198743187.013.20

Pell, M. D., Rothermich, K., Liu, P., Paulmann, S., Sethi, S., & Rigoulot, S. (2015). Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. *Biological Psychology*, *111*, 14–25. https://doi.org/10.1016/j.biopsycho.2015.08.008

Péron, J., Cekic, S., Haegelen, C., Sauleau, P., Patel, S., Drapier, D., Vérin, M., & Grandjean, D. (2015). Sensory contribution to vocal emotion deficit in Parkinson's disease after subthalamic stimulation. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior*, *63*, 172–183. https://doi.org/10.1016/j.cortex.2014.08.023

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*(2), 143–165. https://doi.org/10.1037/0033-2909.99.2.143

Schindler, S., Zell, E., Botsch, M., & Kissler, J. (2017). Differential effects of face-realism and emotion on event-related brain potentials and their implications for the uncanny valley theory. *Scientific Reports*, *7*(1), 45003. https://doi.org/10.1038/srep45003

Schirmer, A., & Gunter, T. C. (2017). Temporal signatures of processing voiceness and emotion in sound. *Social Cognitive and Affective Neuroscience*, *12*(6), 902–909. https://doi.org/10.1093/scan/nsx020

Schweinberger, S. R., Casper, C., Hauthal, N., Kaufmann, J. M., Kawahara, H., Kloth, N., Robertson, D. M., Simpson, A. P., & Zäske, R. (2008). Auditory adaptation in voice perception. *Current Biology*, *18*(9), 684–688. https://doi.org/10.1016/j.cub.2008.04.015

Schweinberger, S. R., Pohl, M., & Winkler, P. (2020). Autistic traits, personality, and evaluations of humanoid robots by young and older adults. *Computers in Human Behavior*, *106*, 106256. https://doi.org/10.1016/j.chb.2020.106256

Skuk, V. G., Dammann, L. M., & Schweinberger, S. R. (2015). Role of timbre and fundamental frequency in voice gender adaptation. *The Journal of the Acoustical Society of America*, *138*(2), 1180–1193. https://doi.org/10.1121/1.4927696

Skuk, V. G., Kirchen, L., Oberhoffner, T., Guntinas-Lichius, O., Dobel, C., & Schweinberger, S. R. (2020). Parameter-Specific morphing reveals contributions of timbre and fundamental frequency cues to the perception of voice gender and Age in cochlear implant users. *Journal of Speech, Language, and Hearing Research*, *63*(9), 3155–3175. https://doi.org/10.1044/2020_JSLHR-20-00026

Skuk, V. G., & Schweinberger, S. R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of Speech, Language, and Hearing Research*, *57*(1), 285–296. https://doi.org/10.1044/1092-4388(2013/12-0314)

Spatola, N., & Wudarczyk, O. A. (2021). Ascribing emotions to robots: Explicit and implicit attribution of emotions and perceived robot anthropomorphism. *Computers in Human Behavior*, *124*, 106934. https://doi.org/10.1016/j.chb.2021.106934

Stoet, G. (2010). Psytoolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Vojtech, J. M., Noordzij, J. P., Cler, G. J., & Stepp, C. E. (2019). The effects of modulating fundamental frequency and speech rate on the intelligibility, communication efficiency, and perceived naturalness of synthetic speech. *American Journal of Speech-Language Pathology*, *28*(2S), 875–886. https://doi.org/10.1044/2019_AJSLP-MSC18-18-0052

von Eiff, C. I., Skuk, V. G., Zäske, R., Nussbaum, C., Frühholz, S., Feuer, U., Guntinas-Lichius, O., & Schweinberger, S. R. (2022). Parameter-Specific morphing reveals contributions of timbre to the perception of vocal emotions in cochlear implant users. *Ear & Hearing*, *43*(4), 1178–1188. https://doi.org/10.1097/AUD.0000000000001181

Webster, M. A., & Maclin, O. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review*, *6*(4), 647–653. https://doi.org/10.3758/BF03212974

Whiting, C. M., Kotz, S. A., Gross, J., Giordano, B. L., & Belin, P. (2020). The perception of caricatured emotion in voice. *Cognition*, *200*, 104249. https://doi.org/10.1016/j.cognition.2020.104249

Yamagishi, J., Veaux, C., King, S., & Renals, S. (2012). Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, *33*(1), 1–5. https://doi.org/10.1250/ast.33.1

Yamasaki, R., Montagnoli, A., Murano, E. Z., Gebrim, E., Hachiya, A., Lopes da Silva, J. V., Behlau, M., & Tsuji, D. (2017). Perturbation measurements on the degree of naturalness of synthesized vowels. *Journal of Voice*, *31*(3), 389.e1–389.e8. https://doi.org/10.1016/j.jvoice.2016.09.020

Yorkston, K. M., Beukelman, D. R., Strand, E. A., & Hakel, M. (1999). *Management of motor speech disorders in children and adults*. Austin, TX: Pro-ed.

Yorkston, K. M., Hammen, V. L., Beukelman, D. R., & Traynor, C. D. (1990). The effect of rate control on the intelligibility and naturalness of dysarthric speech. *Journal of Speech and Hearing Disorders*, *55*(3), 550–560. https://doi.org/10.1044/jshd.5503.550

Young, A. W., & Bruce, V. (2011). Understanding person perception. *British Journal of Psychology*, *102*(4), 959–974. https://doi.org/10.1111/j.2044-8295.2011.02045.x