Cognition 219 (2022) 104967

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Vocal emotion adaptation aftereffects within and across speaker genders: Roles of timbre and fundamental frequency

Christine Nussbaum *, Celina I. von Eiff, Verena G. Skuk, Stefan R. Schweinberger **

Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Germany

ARTICLE INFO

Keywords:
Vocal emotion adaptation
Timbre
Fundamental frequency (F0)
Parameter-specific voice morphing
Gender-correspondence

ABSTRACT

While the human perceptual system constantly adapts to the environment, some of the underlying mechanisms are still poorly understood. For instance, although previous research demonstrated perceptual aftereffects in emotional voice adaptation, the contribution of different vocal cues to these effects is unclear. In two experiments, we used parameter-specific morphing of adaptor voices to investigate the relative roles of fundamental frequency (F0) and timbre in vocal emotion adaptation, using angry and fearful utterances. Participants adapted to voices containing emotion-specific information in either F0 or timbre, with all other parameters kept constant at an intermediate 50% morph level. Full emotional voices and ambiguous voices were used as reference conditions. All adaptor stimuli were either of the same (Experiment 1) or opposite speaker gender (Experiment 2) of subsequently presented target voices. In Experiment 1, we found consistent aftereffects in all adaptation conditions. Crucially, aftereffects following timbre adaptation were much larger than following F0 adaptation and were only marginally smaller than those following full adaptation. In Experiment 2, adaptation aftereffects appeared massively and proportionally reduced, with differences between morph types being no longer significant. These results suggest that timbre plays a larger role than F0 in vocal emotion adaptation, and that vocal emotion adaptation is compromised by eliminating gender-correspondence between adaptor and target stimuli. Our findings also add to mounting evidence suggesting a major role of timbre in auditory adaptation.

1. Introduction

Emotional signals in human speech form an important part of our daily life and are fundamental for successful vocal communication. Therefore, it is not surprising that humans are strikingly good at recognizing emotions in vocal signals and seem to have somewhat culture-independent representations of vocal patterns signaling discrete emotions (Juslin & Laukka, 2003; Laukka & Elfenbein, 2021). At the same time, the system is highly flexible and constantly adapts to the perceptual configuration of the environment (Pérez-González & Malmierca, 2014; Stilp, 2020; Webster et al., 2005). A hallmark of this flexibility is the phenomenon of *adaptation*: prolonged exposure to a stimulus feature leads to a decreased response of dedicated neuron populations, which subsequently fire more vigorously when the stimulus feature changes (Grill-Spector et al., 2006). Along the auditory processing pathway, this form of stimulus-specific adaptation can be found as early as in the inferior colliculus, but has been most widely reported in the auditory cortex of both humans and animals (for a review, see Pérez-González & Malmierca, 2014). Note that we differentiate adaptation from habituation, which is commonly understood as a simple form of learning. Behaviorally, perceptual adaptation manifests in profound *contrastive aftereffects*, which take the form of a bias towards perceiving stimulus features opposite of the adapted stimulus quality. Initially reported for basic stimulus qualities such as color or motion (Mather et al., 1998), such contrastive aftereffects were later demonstrated for a variety of complex social stimuli as well: In vision, adaptation was found for distorted faces (Webster & Maclin, 1999), facial gender, ethnicity (Webster et al., 2004), identity, (Leopold et al., 2001), age (Schweinberger et al., 2010), emotion (Fox & Barton, 2007; Webster et al., 2004), or eye gaze (Jenkins et al., 2006). In the auditory domain, adaptation was found for voice identity (Zäske et al., 2010), gender (Hubbard & Assmann, 2013; Schweinberger et al., 2008; Skuk et al., 2015; Zäske

* Correspondence to: C. Nussbaum, Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Leutragraben 1, 07743 Jena, Germany.

** Correspondence to: S.R. Schweinberger, Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Am Steiger 3/Haus 1, 07743 Jena, Germany.

E-mail addresses: christine.nussbaum@uni-jena.de (C. Nussbaum), stefan.schweinberger@uni-jena.de (S.R. Schweinberger).

https://doi.org/10.1016/j.cognition.2021.104967
Received 18 February 2021; Received in revised form 22 October 2021; Accepted 23 November 2021
Available online 4 December 2021

et al., 2009), age (Zäske et al., 2013; Zäske & Schweinberger, 2011) and emotion (Bestelmeyer, Rouger, et al., 2010; Skuk & Schweinberger, 2013). In the domain of identity perception, a recent paper indicates that individual differences in voice adaptation are linked to voice perception skills (Bestelmeyer & Mühl, 2021). Overall, the above findings suggest that adaptation forms a relevant and general mechanism of the human perceptual system.

1.1. Adaptation to vocal emotion

Many recent studies on vocal emotion adaptation use voice morphing technology (Kawahara et al., 2008; Kawahara & Skuk, 2019), which is an efficient tool to control for emotional information in the voice (e.g. Whiting et al., 2020). Bestelmeyer, Rouger, et al. (2010) were the first who showed consistent aftereffects for vocal emotions. After prolonged exposure to angry /a/−vowels, participants classified stimuli from a morphed angry-fearful continuum as more fearful, and vice versa after exposure to fearful sounds. According to a framework by Schirmer and Kotz (2006), vocal emotion processing follows multiple steps, with an initial analysis of acoustic cues, followed by allocation of emotional significance and higher-order cognitive processes such as evaluative judgements. Adaptation to vocal emotion occurs on all these stages. Using functional magnetic resonance imaging, Bestelmeyer et al. (2014) showed that low-level acoustic analyses and abstract emotional representations could be dissociated by neuronal adaptation in distinct brain areas. Further evidence for adaptation on later integrative stages has been provided by cross-modal (face-to-voice) and cross-domain (voice-to-music) aftereffects (Bowman & Yamauchi, 2017; Pye & Bestelmeyer, 2015; Skuk & Schweinberger, 2013). In event-related potential (ERP) studies on vocal gender and identity (Schweinberger et al., 2011; Zäske et al., 2009), adaptation induced attenuations of both early frontocentral N1 and P2 components and a later parietal P3 component. For vocal emotion processing, the N1 has been linked to low-level acoustic cue analysis in auditory areas, whereas the P2 presumably reflects higher level integrative analysis, including emotional saliency (Paulmann & Kotz, 2018). Although no study to date directly tested how these ERP components are affected by adaptation in vocal emotions, it seems plausible to assume that adaptation would occur at both early acoustic-bound and later integrative stages.

1.2. Role of different acoustic parameters

Different emotional states are associated with distinct physiological changes in the whole vocal production system (including muscular tension, vocal fold vibration, vocal tract alteration, blood flow or heart rate), resulting in different patterns of acoustic cues (Banse & Scherer, 1996; Juslin & Laukka, 2003; Scherer, 1986). For example, fearful voices may sound high-pitched and trembling, angry ones may sound load and harsh (Brück et al., 2011), whereas smiling may result in a bright voice quality due to increased formant frequencies (Tartter, 1980). In fact, all vocal parameters related to fundamental frequency (F0, perceived as pitch), amplitude (perceived as loudness), timing (e.g. speech rate) or timbre (voice quality) seem to be important for vocal emotion recognition (Juslin & Laukka, 2003). For voice adaptation, however, the role of different acoustic cues is less well understood. Whereas a seminal study by Schweinberger et al. (2008) suggested that voice adaptation depends on high-level information but not on F0/pitch, Hubbard and Assmann (2013) came to the opposite conclusion that F0 is a crucial parameter for voice adaptation. As a limitation, neither of these studies used adaptor stimuli that sounded like human voices, to test the contribution of F0. Skuk et al. (2015) therefore readdressed this question for vocal gender with a different approach, by selectively testing the relative importance of F0 and timbre while keeping the respective other cues constant at a gender-uninformative level. Intriguingly, while both parameters were equally important for voice classification, timbre played a much larger role for vocal gender adaptation than F0. Recently, Piazza et al. (2018) suggested that timbre plays a crucial role in adaptation to a variety of natural sounds, including human voices, environmental sounds, and musical instruments. By definition, timbre reflects a combination of several parameters (e.g. formant frequencies or spectral energy distribution), as it is "the difference between two voices of identical F0, intensity and temporal structure" (ANSI, 1973). Based on an absence of aftereffects following adaptation to a single aspect of timbre, Piazza et al. (2018) argued that timbre adaptation depends on integrating all timbral features into a holistic percept of an auditory object. Thus, while the role of acoustic cues for vocal emotion adaptation remains unclear, evidence from other acoustic domains suggest a predominant role of timbre.

1.3. The present study

In the present study, we aimed at comparing the role of timbre and F0 for vocal emotion adaptation, using an approach very similar to Skuk et al. (2015). Using the voice-morphing software TANDEM-STRAIGHT (Kawahara et al., 2008; Kawahara et al., 2013), we generated stimuli with controlled acoustics, with some adaptor stimuli conveying specific emotional information either in F0 or timbre parameters only. Importantly, we kept the other parameters from the voice at a non-informative intermediate morph-level, preserving the "human-likeness" of the stimuli. Based on the studies from other domains reviewed above, we predicted that timbre would play a larger role than F0 for vocal emotion adaptation.

From an ecological perspective, we also considered that (unlike age or gender) emotions are dynamic social signals requiring constant monitoring within communication. This is because emotion can change within an ongoing interaction, and if this is the case, it may signal that something needs immediate attention (Young et al., 2020). To investigate the interactive processing of speaker gender and emotion, we assessed parameter-specific adaptation effects both in a same-gender (Experiment 1) and a cross-gender (Experiment 2) design. To date, only one study tested cross-categorical aftereffects of adaptation to speaker age and gender (Zäske et al., 2013), but did not control for F0 or other acoustic features separately. While adaptation to gender was unaffected by a change in age between adaptor and target voice, adaptation effects to vocal age were reduced by a change of vocal gender. Whether vocal emotion adaptation is affected by gender-correspondence remains unclear but is of particular interest here because a change of speaker gender between adaptor and target entails substantial acoustic changes that require recalibration (Gelfer & Mikos, 2005). Female and male voices differ both in terms of F0 and timbre. Specifically, female voices are higher in pitch, have higher formant frequencies and a breathier quality (Gelfer & Mikos, 2005; Klatt et al., 1990), and exhibit enhanced F1-F2 formant vowel spaces compared to men (e.g. Eichhorn et al., 2018). Although gender-related acoustic changes are irrelevant to the task, the degree to which emotion adaptation will be preserved in the cross-gender design should inform us about the functional locus of adaptation aftereffects. To the extent that these effects operate at the level of acoustic cues of F0 or timbre, which both differ substantially between male and female voices, we considered that they should be compromised or abolished in the cross-gender design. Conversely, to the extent that these effects operate at the level of more abstract emotion categories, they should be preserved in this situation.

2. Experiment 1

2.1. Method

2.1.1. Listeners

Due to the novelty of the present design and the resulting lack of information on effect sizes from previous research, we conducted an a-priori Power analysis for repeated measures ANOVAs (number of measurements = 3 adaptor morph types) with a medium effect size

(f = 0.25), an α = .05, and a desired power of .80, using G-Power 3 (Faul et al., 2007), which yielded a required minimum sample size of 28. We collected data from 36 participants, all students of the Friedrich Schiller University of Jena, native German speakers, without neurological, psychiatric, or hearing impairments. All were compensated with course credit. Data from six participants had to be removed ($n = 1$ being a non-native German speaker, $n = 2$ exceeding the <2.5%-criterion for missing trials in the adaptation task, $n = 3$ due to a technical maladjustment during data collection). The final sample used for data analysis consisted of 30 participants (15 females, 15 males, aged 18 to 26 years [M = 21.57; Mdn = 21.50; SD = 2.30], 5 left-handed).

2.1.2. Stimuli

2.1.2.1. Original audio recordings. We selected original audio recordings from a database of vocal actor portrayals provided by Sascha Frühholz from the Department of Cognitive and Affective Neuroscience of the University of Zurich, that were similar to the ones used in Frühholz et al. (2015). To create the stimuli for the present study, we used four pseudowords (/molen/, /namil/, /loman/, /belam/), spoken by 4 speakers (2 male, 2 female) in a fearful and angry emotion.

2.1.2.2. Voice morphing. We created pairwise morphs between fearful and angry expressions of the same speaker and pseudoword using TANDEM-STRAIGHT (Kawahara et al., 2008; Kawahara et al., 2013). This allowed resynthesizing of voices on the fear/anger-continuum, with independent control of different speech parameters. For a more detailed description and interpretation of the individual TANDEM-STRAIGHT parameters, please refer to the overview by Kawahara and Skuk (2019).

Target stimuli. Morphed stimuli were created in 7 target morph levels (tML), encompassing equidistant 10% steps from 20/80 (anger/fear, in %) to 80/20. In total, all combinations of 4 (speakers) x 4 (pseudowords) x 7 (tML) resulted in 112 target stimuli.

Adaptor stimuli. Four types of morphed stimuli were created as adaptors: (1) **Full adaptors**, with all TANDEM-STRAIGHT parameters taken from fearful or angry utterances, with only time kept at the intermediate (50/50) level. (2) **F0 adaptors**, with the F0-contour of either fear or anger, but with timbre (encompassing the TANDEM-STRAIGHT parameters *formant frequencies, spectral level information*, and *aperiodicity* in conjunction) and time (resulting from an interpolation of time anchor positions) kept at the intermediate morph level. (3) **Timbre adaptors**, comprising timbre information of anger or fear, but with F0 and time kept at the intermediate morph level. (4) **Ambiguous adaptors**, with all parameters kept at the intermediate morph level. Please refer to Fig. 1 for a visualization of spectral properties and fundamental frequency contour of representative stimulus examples.

This resulted in 7 adaptation conditions: adaptor morph type (aMType: Full, Tbr, F0) x adaptor emotion (aEmo: fea, ang) plus the ambiguous condition (amb). In summary, all combinations of 4 (speakers) x 4 (pseudowords) x 7 (adaptation conditions) resulted in 112 adaptor stimuli. Note that ambiguous adaptors were used to create a baseline condition in which no systematic aftereffects were expected. Time was kept at the intermediate morph level in all adaptation conditions, to avoid any influence of this parameter on the adaptation aftereffect.

Using Praat software (Boersma, 2018), all morphed stimuli were root-mean-square normalized to 70 dB SPL. Please refer to Table 1 for stimulus characteristics of targets and adaptors.

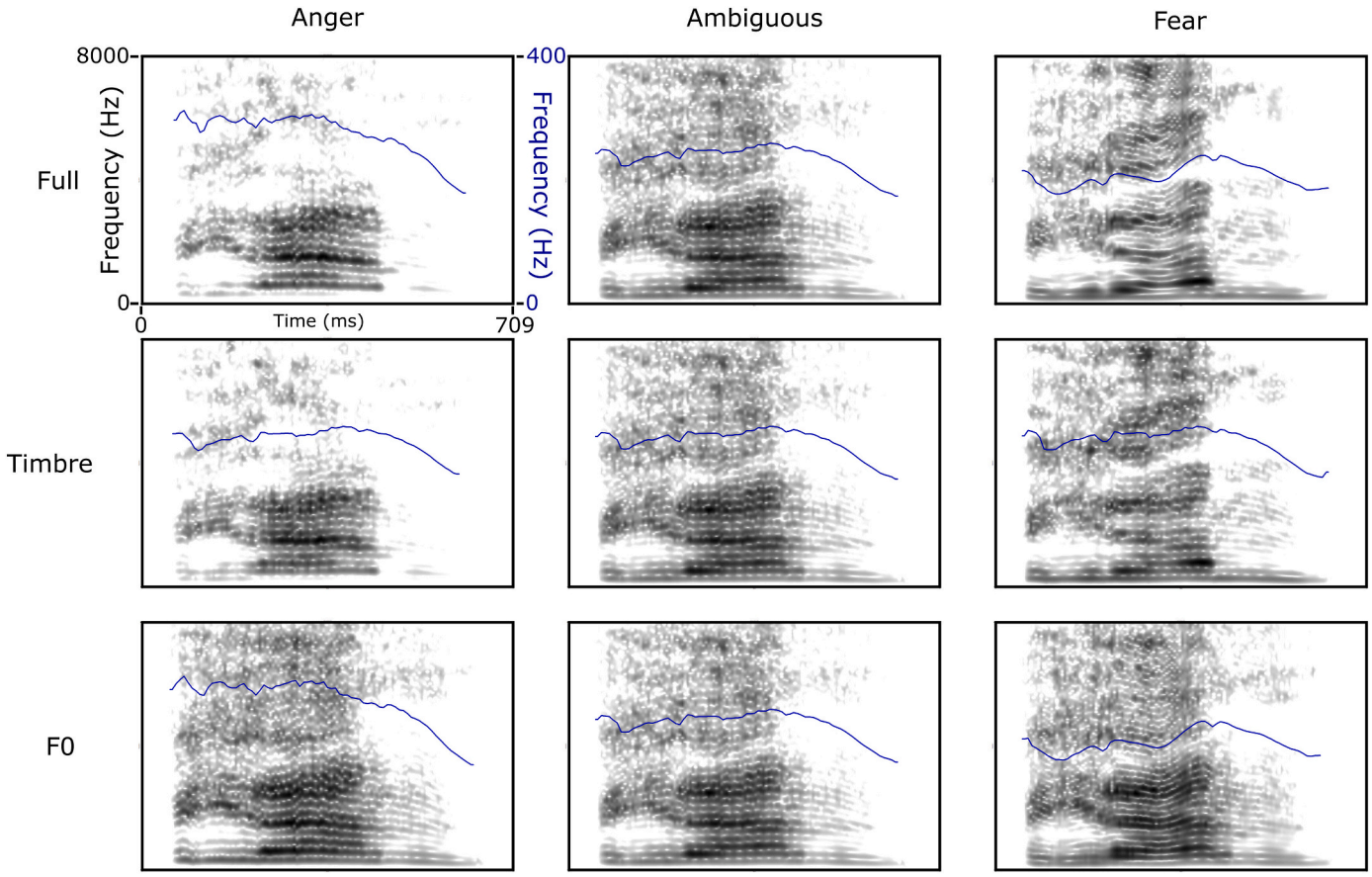

Fig. 1. Spectrogram and fundamental frequency contour of representative adaptor stimulus examples.
Note. Adaptor stimulus examples of a female speaker uttering the pseudoword /belam/. Note different frequency scales for the spectrogram (in greyscale and black, left in top left panel) and the fundamental frequency contour line (in blue, right in top left panel). The ambiguous stimulus is identical in all three rows.

Table 1
Stimulus characteristics of stimuli used as continuum endpoints.

	Female speakers		Male speakers		Paired t.test for ang vs. fea [a]	
	ang	fea	ang	fea	*t(15)*	*p*
F0 Mean	366	288	264	205	3.69	0.002 **
F0 SD	67.1	19.0	45.6	26.8	6.50	0.001 ***
F0 Intonation	259	79	171	95	6.99	0.001 ***
F0 Glide	−60	3	−41	12	−3.08	0.007 **
Formant Dispersion	910	1043	819	1074	−5.61	0.001 ***
Alpha Ratio	1.0	2.3	1.1	2.28	−11.16	0.001 ***
HNR	12.8	21.3	7.3	19	−11.44	0.001 ***
Jitter	0.004	0.004	0.011	0.004	−3.20	0.006 **
Shimmer	0.036	0.013	0.063	0.013	5.72	0.001 ***
Duration	885	760	731	854	0.02	0.9883

Note. All acoustical parameters were adapted from McAleer et al. (2014) and extracted using Praat software (Boersma, 2018). For F0 extraction, pitch ranges were set to 170–600 Hz for female and 100–370 Hz for male stimuli. *F0 Intonation* = $F0_{max} - F0_{min}$; *F0 Glide* = $F0_{End} - F0_{Start}$; *Formant Dispersion* = ratio between consecutive formant means (from F1 to F4, maximum formant frequency set to 5 kHz, window length 0.025 s); *Alpha ratio* (a measure of the spectral slope) = ratio of mean energy within low (0–1 kHz) and high frequencies (1–5 kHz), computed from the long-term average spectrum; *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), *Jitter* as a measure of local F0 variation (shortest period = 0.0001 s; longest period = 0.02 s; max. period factor = 1.3); *Shimmer* as a measure of local amplitude variation (shortest period = 0.0001 s; longest period = 0.02 s; max. period factor = 1.3; max. Amplitude factor = 1.6). Fea = fear, ang = anger.

[a] Including both male and female speakers.

2.1.3. Design and procedure

After a short audio-test on frequencies between 0.5 and 8 kHZ (Cotral Audiotest; www.cotral.de), the listening-experiment, programmed using *E*-Prime 2.0 (Psychology Software Tools, Inc, 2012), started. Instructions were presented on a computer screen to minimize interference from the experimenter's voice, and auditory stimuli were presented with Sennheiser™ HD212Pro circum-aural headphones (www.sennheiser.com). Up to three participants were tested simultaneously in a larger testing room with identical computers separated by folding screens. The experiment consisted of two parts: an adaptor-classification task and an adaptation task.

2.1.3.1. Adaptor classification task. The purpose of the adaptor classification task was to obtain information about the explicit perception of the voice stimuli that were later used as adaptors in the adaptation task. Participants classified voice stimuli as either angry or fearful by pressing the corresponding "D" and "L" keys. Key assignment was counterbalanced across participants. Each trial started with a green fixation cross, which was presented for 500 ms and then replaced by a green question mark, presented simultaneously with voice stimulus onset. Responses were recorded in a time window starting with voice onset and ending 2000 ms following voice offset. If no response had been entered in the time window (error of omission), the last trial slide (500 ms) comprised a feedback screen prompting for faster response; otherwise, a black screen was shown instead. All 112 stimuli were presented once in randomized order. To acquaint listeners with the experimental procedure, we used 32 practice trials before the experimental trials, with stimuli that were not shown thereafter.

2.1.3.2. Adaptation task. Directly after completion of the adaptor classification task, the adaptation task followed. Participants completed 14 adaptation blocks (7 adaptation conditions, blocked for speaker sex). These blocks were presented in pseudorandomized order with a few constraints: the adaptation blocks of the same parameter appeared after another. Order of emotion adaptation was counterbalanced across participants: Half of the participants always completed the fearful blocks for a given parameter before the angry blocks of the same parameter, with the reversed pattern for the other half, respectively. Within the emotion conditions, however, the order of presentation of speaker sex blocks was randomized.

The trial procedure was the following: each block consisted of an adaptation and a response phase (Fig. 2). During the adaptation phase, 16 adaptor stimuli (2 speakers × 4 pseudowords × 2 presentations) were presented in random order with a red fixation cross on the computer screen and participants were instructed to listen attentively. After the adaptation phase, the response phase started, as signaled to the participants by a message ("Now it starts") in red font for 3000 ms. Here, trial procedure and response collection were as in the adaptor classification task described above, except that every fourth trial, a red fixation cross appeared between trials and another top-up adaptor was presented. This was done to periodically refresh the adaptation level during the response phase, while keeping experimental duration within practical limits. Note that previous research demonstrated the efficiency of single top-up stimuli (Jenkins et al., 2006; Schweinberger et al., 2007). Crucially, both adaptors from the preceding adaptation phase and top-up adaptors were always of the *same gender* as the target voices in a given response phase. However, note that for any given target voice, the preceding top-up adaptors were always from a different speaker and different pseudoword. Participants completed a short practice phase with 16 adaptor and 32 target trials prior to the main adaptation block, using stimuli that were not used thereafter. Individual self-paced breaks were allowed between blocks of 56 trials. A total number of 784 (56 trials × 14 blocks) adaptation trials were presented in the adaptation task. Total duration of the experiment was between 60 and 70 min.

Post-experimental questionnaire. After the adaptation experiment, participants completed a computerized version of the Autism Quotient Questionnaire (AQ, Baron-Cohen et al., 2001; Freitag et al., 2007). Analysis of the AQ was fully explorative and can be found on the associated OSF Repository (https://osf.io/qzj6d/).

The experiment was in line with the ethical guidelines of the German Society of Psychology (DGPs) and all participants gave informed consent prior to participation.

2.1.4. Data collection and analysis

Both errors of omission and reaction times (RTs) < 200 ms were excluded from the data. Data was analyzed using R Version 4.0. (R Core Team, 2020). Analysis of variances (ANOVA) were performed with the R-package "ez" (Lawrence & Lawrence, 2016), using epsilon corrections for heterogeneity of covariances throughout (Huynh & Feldt, 1976). Where appropriate, two-sided paired sample *t*-tests were computed to follow up on significant interactions, and Bonferroni corrections to adjust α-levels were applied where necessary (Abdi, 2007).

To analyze adaptation aftereffects, we fitted Cumulative Gaussian functions to responses along target continua. Cumulative Gaussians are defined by two parameters: the mean, or the point of subjective equality (PSE), marks the point on the x-axis at which the function crosses 0.5 on the y-axis. In the present context, this indicates, for each condition, the morph level at which participants were equally likely to give an 'angry'- or a 'fearful'-response. The second parameter, the standard deviation (SD), directly reflects the slope, with smaller SDs corresponding to steeper slopes. As we expected adaptation to cause a shift in the fitted cumulative functions along the x-axis, we considered the PSE as the primary dependent variable.

2.2. Results

Here, we only report results that were of primary interest for the purpose of this study. Further documents, including supplemental figures and tables, analysis scripts (including response times), and raw data can however be found on the associated OSF Repository (https://osf.io/qzj6d/).

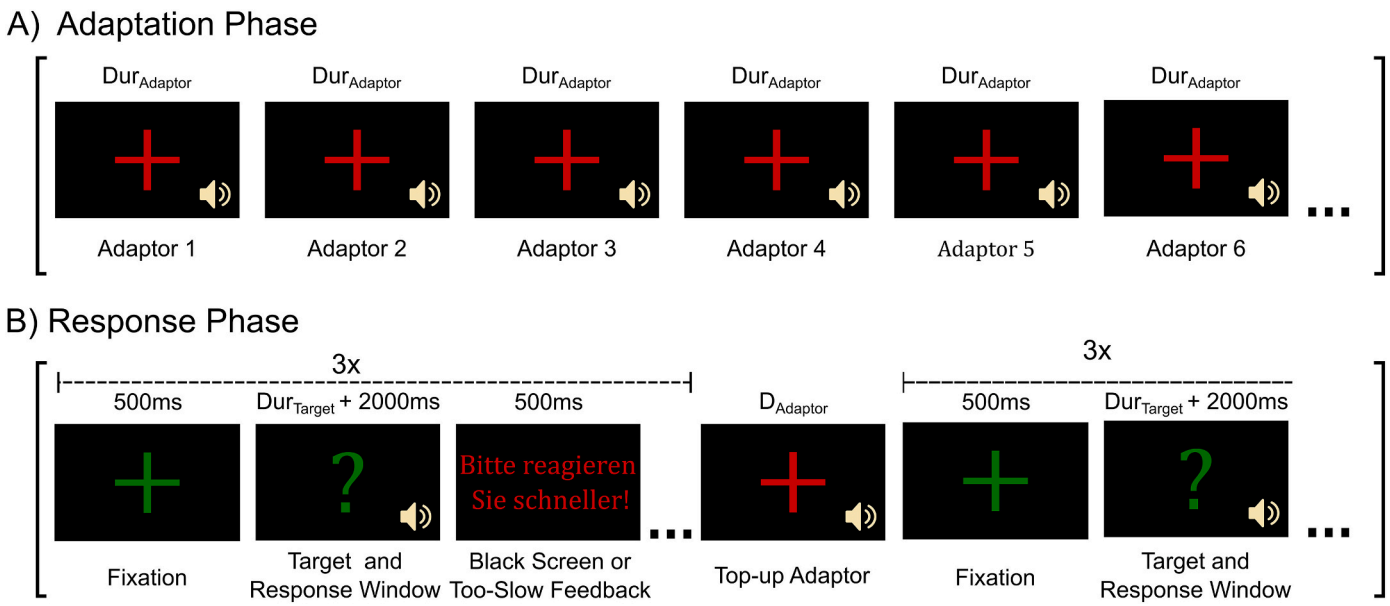

Fig. 2. Experimental trial design of the Adaptation Task, A) adaptation phase and B) response phase.
Note. Dur = duration.

2.2.1. Adaptor classification

The adaptor classification task provided some information about participants' explicit perception of the stimuli that were later used as adaptors in the adaptation experiment. Note that while the ambiguous adaptation condition provides an important reference, it cannot be included in an orthogonal 2 × 3 design with factors Emotion (ang, fea) and Morph Type (Full, Tbr, F0), since morph type is meaningless when all parameters are set to 50/50. Accordingly, we analyzed data in a two-step way: First, we performed an ANOVA, excluding the ambiguous adaptation condition. Second, we included all adaptor conditions, including the ambiguous ones, averaged across the factor Morph Type.

We analyzed the **proportion of 'angry'-responses** to the stimuli in an initial 3 × 2 × 2 × 2 mixed-effects ANOVA with the within-subject factors Morph Type (MType: Full, F0, Tbr), Emotion (Emo: ang, fea), and Speaker Sex (SpSex: m, f), and the between-subject factor Listener Sex (LSex: male, female). Please refer to Table 2 for a summary of significant main effects and interactions. For descriptive data, please refer to Fig. 3.

As expected, a prominent main effect of **Emo** confirmed that angry stimuli were perceived as more angry than fearful ones (Ms ± SEMs = 0.78 ± 0.02 and 0.07 ± 0.01, respectively). The main effect of **SpSex** indicated that male speakers were perceived as angry more often than female speakers (0.46 ± 0.01 and 0.39 ± 0.02, respectively). Most importantly, there was a significant interaction of **Emo x MType**. The effect of MType was a significant for both fearful and angry stimuli when tested separately; $F(2, 58) = 31.14$, $p < .001$, $\eta p^2 = .518$, $\varepsilon_{HF} = .827$ and $F(2, 58) = 44.03$, $p < .001$, $\eta p^2 = .603$, respectively, reflecting differences in how consistently voices with different morph types could be classified as fearful and angry. For fearful stimuli, post-hoc comparisons revealed significant differences between all three morph types, $|ts(29)| \geq 2.36$, $ps \leq .025$. Full morphs were classified being closest to zero, corresponding to highly consistent fearful classification, followed by Tbr and finally by F0. Similarly, the analysis for angry stimuli revealed significant differences between all three morph types, $|ts(29)| \geq 4.98$, $ps < .001$. Again, Full morphs were perceived as most angry, followed by Tbr, and F0. For descriptive data and all further post-hoc tests, please refer to the materials on https://osf.io/qzj6d/.

In order to compare the ambiguous condition with the other conditions, we computed an ANOVA with factors Emo (fea, amb, ang), SpSex (m, f), and LSex (male, female). As expected, we found the main effect of Emo; $F(2, 56) = 474.34$, $p < .001$, $\eta p^2 = .944$. Both fearful and angry adaptors differed significantly from ambiguous adaptors, $|ts(29)| \geq 10.55$, $p < .001$, indicating that these were indeed perceived as ambiguous. At the same time, a t-test against guessing rate (.5), $t(29) = -6.21$, $p < .001$, showed that the perception of ambiguous stimuli displayed a bias towards fearfulness, in line with the finding that angry classification rates for angry Full Morphs also remained below 100% (cf. Fig. 3).

In summary, all fearful and angry adaptors were classified correctly well above chance, as expected, but classification performance was also affected by morph type, such that full morphs were classified best. By comparison, classification performance was reduced for timbre, and was

Table 2
Adaptor classification task: results of the 3 × 2 × 2 x 2 mixed-effects ANOVA of Experiments 1 and 2.

	Experiment 1					Experiment 2				
Effect	F	df1	df2	p	ηp^2	F	df1	df2	p	ηp^2
MType	6.78	2	56	.002	.195	11.14	2	54	< .001	.292
Emo	925.70	1	28	< .001	.971	710.50	1	27	< .001	.963
SpSex	24.69	1	28	< .001	.469	34.20	1	27	< .001	.559
LSex x SpSex	4.33	1	28	.047	.134	3.21	1	27	.084	.106
MType x Emo	76.33	2	56	< .001	.732	59.27	2	54	< .001	.687
MType x SpSex	3.41	2	56	.040	.108	7.78	2	54	.001	.224
Emo x SpSex	22.13	1	28	< .001	.441	23.14	1	27	< .001	.462

Note. Mixed-effects ANOVA with the within-subject factors Morph Type (MType: Full, F0, Tbr), Emotion (Emo: ang, fea), and Speaker Sex (SpSex: m, f), and the between-subject factor Listener Sex (LSex: male, female) for Experiment 1 and 2.

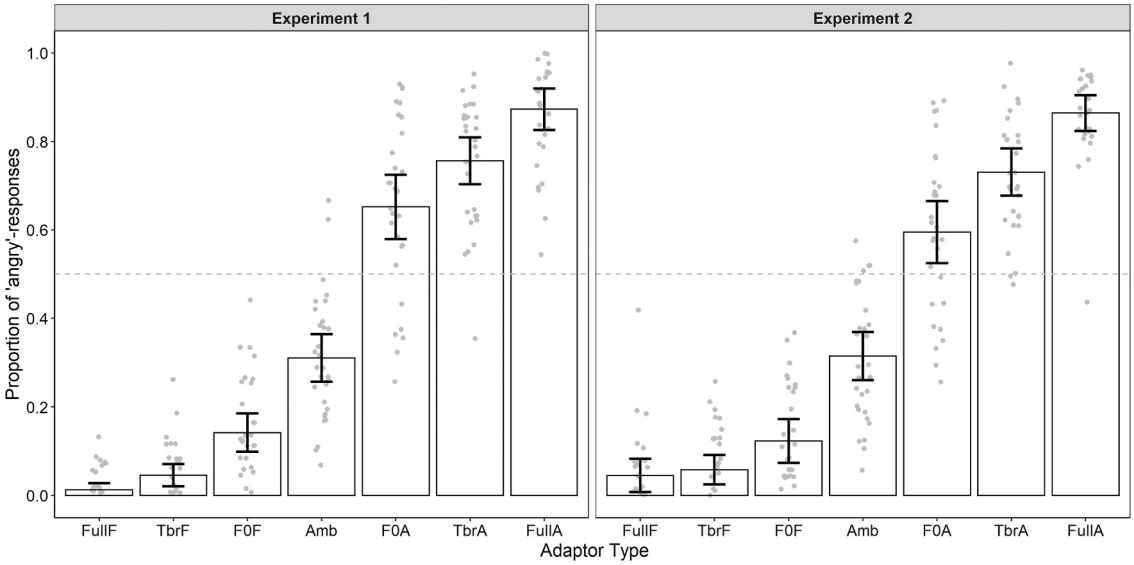

Fig. 3. Adaptor Classification Task: Descriptive data (Ms and CIs) of the seven different adaptor stimulus types for Experiment 1 and 2.
Note. The dotted line represents the guessing rate (proportion of 'angry responses = 0.5). The grey dots represent individual participants' data. Full = full morphs, Tbr = timbre morphs, F0 = F0 morphs, F = fear, A = anger, Amb = ambiguous.

even reduced to a greater extent for F0 morphs.

2.2.2. Adaptation experiment

2.2.2.1. Cumulative Gaussian fits. Cumulative Gaussian functions were fitted for each participant and adaptation condition (2 aEmo × 3 MType + amb), resulting in a total number of 210 fits. On average, fits were excellent in terms of R^2 (M = 0.94, SD = 0.04, range: 0.66–0.99, $N = 210$). Note that only 3 out of 210 fits had an $R^2 < 70$.

Again, the analysis followed a two-step rationale: First, we analyzed data without the ambiguous adaptation condition, permitting a factorial design of the ANOVA. Second, we tested all aMType x aEmo combination against the ambiguous condition in a planned comparison.

PSEs were analyzed in a 2 × 3 ANOVA with factors aEmo (fea, ang) and aMType (Full, Tbr, F0), see Fig. 4. The analysis revealed the expected main effect of **aEmo** ($F(1, 29) = 115.18$, $p < .001$, $\eta p^2 = .799$), with greater PSEs after angry compared to fearful adaptation (Ms = 57.44 ± 1.33 and 50.61 ± 1.15, respectively). Accordingly, to classify a voice as angry, participants needed more anger in a target voice after angry compared to fearful adaptation reflecting the expected bias towards fear. Crucially, we found an interaction of **aEmo x aMType**, $F(2, 58) = 15.33$, $p < .001$, $\eta p^2 = .346$, and there was no main effect of **aMType**, $F(2, 58) = 0.05$, $p = .952$, $\eta p^2 = .012$.

2.2.2.2. Magnitude of adaptation aftereffects. To investigate the interaction of aEmo x aMType, the PSE shifts (adaptation aftereffects) for each of the three adaptor morph types were computed by subtracting the PSEs of the fearful adaptation condition from the PSEs of the angry adaptation condition (details in Table 3). All these effects differed significantly from zero ($ts(29) \geq 2.81$, $ps \leq .009$).

Importantly, F0 adaptation was significantly smaller when compared to both Full and Tbr adaptation, $t(29) = -5.72$, $p < .001$, $d = 1.06$ and $t(29) = -3.37$, $p = .002$, $d = 0.56$, respectively. Conversely, Tbr adaptation was only marginally smaller than Full adaptation, $t(29) = -1.97$, $p = .058$, $d = 0.35$, indicating that Tbr was almost as effective as the Full condition in eliciting aftereffects. Fig. 4 and Table 3 also suggests

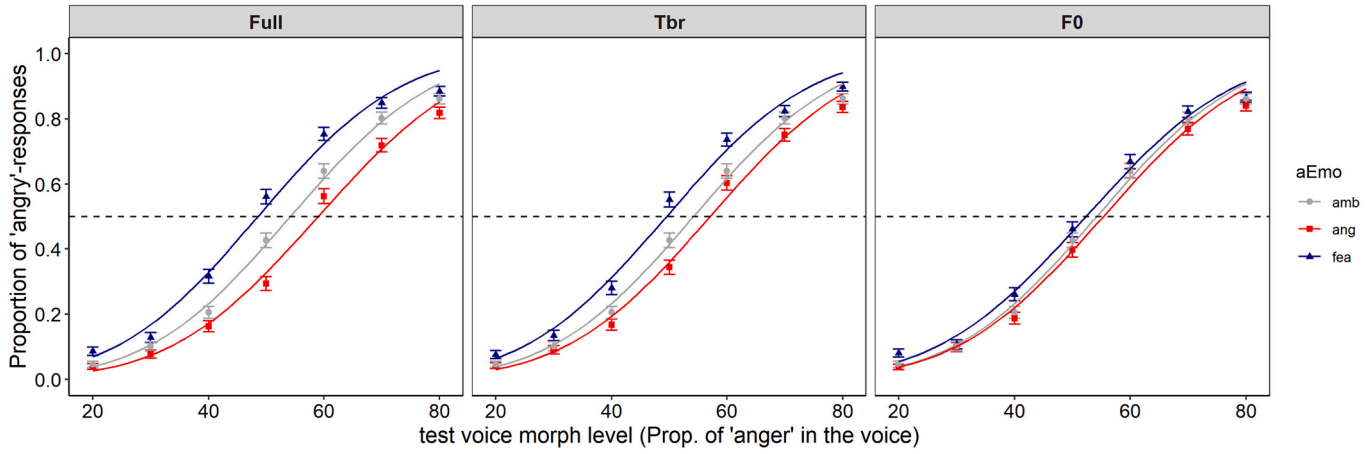

Fig. 4. Experiment 1: Cumulative Gaussians on the proportion of 'angry'-responses.
Note. Comparing the reference (amb) with effects of the six adaptation conditions separately for each adaptor morph type: A) Full, B) Tbr, and C) F0. The reference curve is the same in all three sub-plots and always fell between the curves of the angry and fearful adaptation condition for each of the adaptor morph types. aEmo = adaptor Emotion, ang = anger, fea = fea, amb = ambiguous.

Table 3
Descriptive statistics of the adaptation aftereffect size.

	Experiment 1				Experiment 2			
Aftereffect	M	SD	SE	Cohens d [95%-CI]	M	SD	SE	Cohens d [95%-CI]
Full	10.2	5.61	1.02	1.82 [1.23–2.41]	3.96	7.27	1.35	0.55 [0.16–0.93]
Tbr	7.49	5.78	1.05	1.30 [0.80–1.78]	2.73	5.86	1.09	0.47 [0.09–0.84]
F0	2.72	5.29	0.97	0.51 [0.13–0.89]	1.38	6.20	1.15	0.22 [−0.14–0.58]

Note. Adaptation aftereffect size computed by subtracting the PSEs of fearful adaptation condition from the PSEs of the angry adaptation condition, for Experiment 1 and 2.

efficient combination of timbre and F0 information for Full adaptors. Overall, the largest adaptation aftereffect was elicited by the Full condition, followed closely by timbre, while F0 elicited substantially smaller (but still significant) aftereffects.

2.2.2.3. Comparison with ambiguous adaptors. In order to compare the different adaptation conditions with a reference, planned comparisons were carried out between each adaptation condition (aMType x aEmo) and the ambiguous condition. Whereas all Full and Tbr conditions differed significantly from the ambiguous condition, F0 did not, although displaying the same numerical pattern ($|ts(29)| \geq 2.92$, $p \leq .007$ for Full and Tbr conditions and $|ts(29)| \leq 1.58$, $p \geq .125$ for F0 conditions, all against Amb respectively).

2.2.2.4. Analysis of slope (SD) data. In analogy to the PSEs, we performed a 2×3 ANOVA with the within-subject factors aEmo and aMType on the slope (SD) data as well, revealing no significant main effects and interactions. This suggests that the slope of the function, and thus the shape of the curve displaying the transition from voices perceived as fearful to angry was unaffected by the experimental conditions.

2.3. Short summary

In Experiment 1, we found consistent aftereffects for all three adaptor morph types, but with profound differences in effect size: the biggest effect was found for the full adaptation condition, followed by timbre and F0. Crucially, average effects in the timbre condition were more than twice in magnitude compared to the F0 condition and were only marginally smaller than full adaptation effects. Moreover, the sum of timbre and F0 effects approximately matched the size of the full effect, suggesting an additive nature of timbre and F0 for vocal adaptation. Compared to the ambiguous condition, full and timbre adaptors elicited significant PSE shifts, whereas F0 adaptors did not.

3. Experiment 2

In Experiment 2, we used a cross-gender design to investigate interactive processing of speaker gender and emotion in vocal emotion adaptation. Our aim was not only to estimate the degree to which adaptation would be preserved or compromised in full emotion adaption, but also to reveal whether the pattern of parameter-specific effects observed in Experiment 1 would potentially transfer to cross-gender adaptation. To the degree to which vocal emotion adaptation operates at the level of more abstract emotion categories rather than their expression in acoustic cues that differ tremendously between male and female voices such as mean F0 and formant frequencies, we considered that aftereffects should be preserved in a cross-gender design.

3.1. Method

3.1.1. Listeners

We collected data from 33 new participants, with the same inclusion criteria as in Experiment 1. Data from four participants had to be removed, because they exceeded the criterion (<2.5%) for missing trials in the adaptation blocks. Thus, the final sample used for data analysis consisted of 29 participants (15 females, 14 males, aged 18 to 28 years [M = 21.51; Mdn = 21; SD = 2.78], 2 left-handed).

3.1.2. Stimuli, procedure, and analysis

Stimuli, procedure, and analysis were identical to Experiment 1, with the exception that in the adaptation task, the adaptors presented in the adaptation phase as well the top-up adaptors were always of the *opposite gender* as the target stimuli in the response phase. Note that the adaptor classification task was identical in Experiment 1 and 2.

3.2. Results

3.2.1. Adaptor classification experiment

Since the adaptor classification task in Experiment 2 was identical to Experiment 1, the analysis was carried out analogously. As can be seen in Table 2, the results of the mixed-effects ANOVA were highly analogous to results of Experiment 1. In short, Fig. 3 illustrates that the adaptor classification data from Experiment 2 fully replicated the pattern seen in Experiment 1. For the full statistical analysis, please refer to https://osf.io/qzj6d/.

3.2.2. Adaptation experiment

3.2.2.1. Cumulative Gaussian fits. As in Experiment 1, Cumulative Gaussian functions were fitted for each participant and adaptation condition (2 Emo × 3 MType + amb), resulting in a total number of 203 fits. On average, fits were excellent in terms of R^2 (M = 0.92, SD = 0.09, range: 0.32–0.99, $N = 203$). Note that only 6 out of 203 fits had an $R^2 < 70$.

PSEs were analyzed in a 2×3 ANOVA with factors aEmo (fea, ang) and aMType (Full, Tbr, F0). The analysis revealed the expected main effect of **aEmo**, $F(1, 28) = 17.219$, $p < .001$, $\eta p^2 = .381$, with greater PSEs after angry than fearful adaptation ($Ms = 57.30 \pm 1.71$ and 54.6 ± 1.60, respectively). Unlike in Experiment 1 however, the interaction of **aEmo x aMType**, did not reach significance, $F(2, 56) = 1.09$, $p = .345$, $\eta p^2 = .037$. Overall, the adaptation aftereffect was massively reduced in Experiment 2 compared to Experiment 1 (please refer to Fig. 5).

3.2.2.2. Magnitude of the adaptation aftereffect. Although the interaction of aEmo x aMType did not reach significance, we calculated adaptation aftereffects for the three morph types, to compare the results with those of Experiment 1 (Table 3). The magnitude of adaptation aftereffects was greatly reduced in Experiment 2 compared to Experiment 1, but the numerical pattern appeared to be similar; with Full eliciting

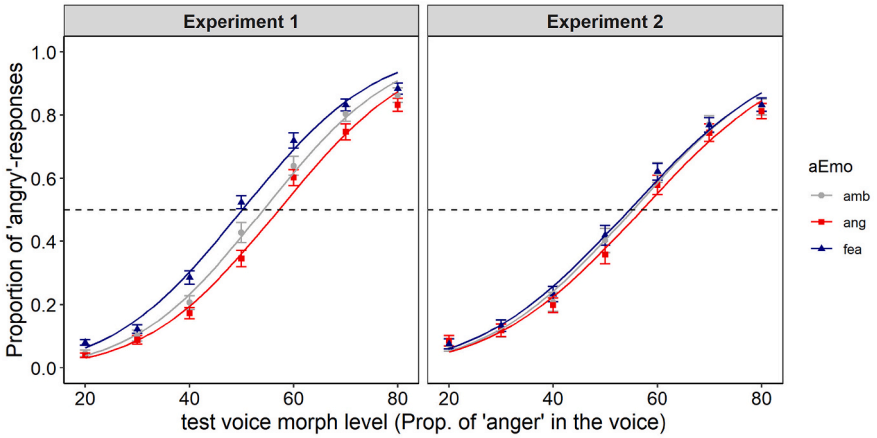

Fig. 5. Adaptation task: comparison of the adaptation aftereffect (averaged across adaptor morph type) of Experiments 1 and 2.
Note. aEmo = adaptor Emotion, ang = anger, fea = fear, amb = ambiguous.

the largest adaptation effect, followed by Tbr and F0. Also, aftereffects in the Full and Tbr conditions differed significantly from zero, *ts* $(28) \geq 2.51$, *p*-value $\leq .018$, whereas F0 did not, $t(28) = 1.20$, *p*-value $= .240$. As a caveat, in the absence of main effects or interactions involving aMType, it is difficult to exclude the possibility that the results reflect random variation, although it may be more likely that they reflect low statistical power in the context of a small adaptation effect size in Experiment 2.

3.2.2.3. Comparison with ambiguous adaptors. As in Experiment 1, we tested all aEmo x aMType combinations against the ambiguous condition. Significant differences were found for two combinations only: for angry full and angry timbre adaptation, $|ts(28)| \geq 2.06, p \leq .048$). For all the others, no difference was found, $|ts(28)| \leq 0.98, p \geq .33$.

3.2.2.4. Analysis of slope (SD) data. As in Experiment 1, a 2×3 ANOVA with the within-subject factors aEmo and aMType on the slope (SD) data revealed no significant main effects and interactions.

3.3. Short summary

Overall, Experiment 2 revealed significant vocal emotion aftereffects even in a cross-gender design. However, effect sizes appeared to be substantially reduced relative to the within-gender design in Experiment 1. Adaptation aftereffects did not differ significantly as a function of morph type, although displaying a similar numerical pattern.

4. General discussion

The present experiments on vocal emotion adaptation considered that an effective perceptual system needs to detect subtle emotional changes on a very rapid time scale within a communication partner, but potentially also across speakers (Young et al., 2020). Imagine for example a sports event with a very aroused and joyful atmosphere, where a single horrified or angry shout by one individual saliently pops out of the crowded auditory soundscape, possibly due to potential injury or misconduct of an athlete. Adaptive coding is an effective means to enhance the system's sensitivity to such changes in the acoustic environment via flexible recalibration. Perceptual adaptation has been framed as a mechanism for "fitting the mind to the world" (Clifford & Rhodes, 2005) and is an important prerequisite for adequate attention allocation and response behavior in social encounters. The present study demonstrates that timbre can play a larger role than F0 in vocal emotion adaptation. Within speaker gender, timbre adaptation elicited contrastive aftereffects of almost three times the size compared to F0 adaptation. At the same time, emotion adaptation effects were substantially reduced when speaker gender changed between adaptors and targets. Overall, these findings indicate interactive processing of emotion and speaker gender in vocal emotion adaptation and suggests that adaptation-induced recalibration is disrupted when substantial changes in acoustic characteristics are present.

4.1. The role of timbre and F0 in vocal emotion adaptation

In Experiment 1, we found timbre to play a larger role than F0 in vocal emotion adaptation. These findings are in line with Skuk et al. (2015), who found very similar results for adaptation to vocal gender and with Piazza et al. (2018), who suggested timbre as the critical parameter for auditory adaptation by showing rapid and robust aftereffects for a variety of natural sounds. However, the present results conflict with Hubbard and Assmann (2013), who argued that F0 is crucial for vocal emotion adaptation, after they failed to observe aftereffects in a F0-removed condition. This discrepancy can be resolved by considering differences in the stimulus material: instead of removing a vocal parameter, we kept it at a task-uninformative morph-level. This way it was ensured that the vocal stimuli encompassed all necessary features of a human voice and would be perceived as a possible outcome of the human vocal production system. Potentially, this lack of "human-likeness" could be responsible for the absence of adaptation effects in the F0-removed condition in Hubbard and Assmann's (2013) study, not the missing F0 per se. Skuk et al. (2015) made a very similar point for vocal gender adaptation. The finding that timbre plays an important role for vocal emotion adaptation seems plausible, considering the demands on our perceptual system for everyday vocal emotion perception: During a social encounter, expressed emotion can change on a moment-to-moment-manner, within and across speakers (Young et al., 2020). Hence, the auditory stream needs to be monitored in a way that allows rapid detection of emotion relevant changes; within a variety of voice modification related to other factors such as speech. Since pitch changes constantly as a function of suprasegmental speech prosody (e.g. to distinguish a question from a statement), timbre may be the more diagnostic cue for emotional changes, making timbre adaptation a very efficient means to accomplish this task.

4.2. The role of gender correspondence for vocal emotion adaptation

Compared to Experiment 1, the dramatically reduced cross-gender aftereffects in Experiment 2 suggest that vocal gender interacts with emotion processing. Zäske et al. (2013) reported a similar pattern for speaker age and speaker gender: vocal age aftereffects were reduced but still present in a cross-gender condition. Further, our results are reminiscent of dependencies in the processing of facial expression and facial

identity (Campbell & Burke, 2009; Ellamil et al., 2008; Fox & Barton, 2007; Schweinberger & Soukup, 1998; Vida & Mondloch, 2009) and, importantly, of facial emotional expression and gender (Bestelmeyer, Jones, et al., 2010). Of interest, one study showed that it was possible to simultaneously induce opposite aftereffects for male and female faces that varied on an anger-fear continuum (Bestelmeyer, Jones, et al., 2010). Such findings have been interpreted to demonstrate an interdependent processing of facial emotion and gender. It was even suggested that simultaneous opposite aftereffects for two categories of faces supports the existence of independent representational spaces for these categories (Jaquet, Rhodes, & Hayward, 2007). As a qualification, the magnitude of simultaneous opposite aftereffects for male and female stimuli is typically smaller compared to simultaneous concordant adaptation, suggesting both gender-specific and gender-independent components to contribute to the overall aftereffect (Schweinberger et al., 2010).

As an important distinction, vocal gender is an extralinguistic cue, which signals relatively stable characteristics of the speaker, whereas vocal emotion is a paralinguistic cue, which signals situation-specific characteristics and varies enormously between utterances, as well as within and across speakers (Schweinberger et al., 2014). Theoretical frameworks of voice processing suggested that extra- and paralinguistic cues are processed in distinct but interacting neural networks (Belin et al., 2011). Here, we offer two potential explanations for the moderation effect of vocal gender on the emotion aftereffect: (1) effects of acoustical proximity or (2) differences in emotion processing for male and female voices (beyond acoustic differences).

4.2.1.1. Acoustical proximity. Male and female voices differ on a variety of acoustic features including both F0 and timbre, due to morphological differences in the vocal production system and presumably as a result of social learning (Schweinberger et al., 2014; Skuk & Schweinberger, 2014). Thus, in our Experiment 2, a change of speaker gender between adaptors and targets therefore resulted in a dramatic change of acoustic cues that were irrelevant for the task. Thus, vocal emotion adaptation may be affected by the auditory similarity between adaptors and targets, irrespective of speaker gender. For timbre-only adaptation, Piazza et al. (2018) found aftereffects to be robust to moderate pitch changes, but to decrease with more substantial pitch differences, until they were only marginally measurable at 9 semitones difference. In our sample, F0 of male and female speakers differed by about 90 Hz, which corresponds to an average change of 5–6 semitones. Further, they tended to differ in jitter and the harmonics-to-noise ratio. These acoustic dissimilarities may have driven the aftereffect reduction observed in the cross-gender compared to the gender-corresponding experiment, suggesting that vocal emotion adaptation depends to a substantial degree on the acoustic features of the sounds.

4.2.1.2. Differences in emotional processing of male and female voices. As an alternative explanation, emotions expressed by male and female voices may be processed in a qualitatively different manner, based on different internal representations of e.g. “male expression of anger” and “female expression of anger”. In the classification task, male voices were consistently more often classified as angry and females more often as fearful, but it remains unclear whether this indeed represents differential internal representations, or whether the speakers were just more efficient in expressing the respective emotion. According to a developmental model proposed by Brody (2000), gender differences in emotional expression occur as a result of biologically based predispositions and culturally mediated socializations. As a result, women may be generally more emotional expressive, and especially so in positive and internalizing negative emotions (e.g. fear), whereas men show less emotions overall, but display rather externalizing emotions (e.g. anger); a pattern indeed observed in a meta-analysis on children (Chaplin & Aldao, 2013). Therefore, it seems plausible to assume gender-specific representation of emotional displays to account for these differences (Bestelmeyer, Jones, et al., 2010), very similar to the idea of gender-specific “voice spaces” for the representation of speaker identity (Latinus et al., 2013).

Because acoustical proximity and gender-correspondence are confounded, the present research only allows speculation about the mechanisms underlying the gender effect observed for vocal emotions. To resolve this, one would need an experimental design where acoustically identical adaptors would be reliably perceived as female or male in different experimental conditions. While this is certainly a challenge, one could try a paradigm with double adaptation, in which a gender adaptation paradigm is used to induce the reliable impression of either a male or female voice in the very same androgynous emotional stimuli, which are in turn used for a subsequent emotion adaptation experiment. For a manipulation following this logic, refer to Rhodes et al. (2010). In the future, the possibilities of parameter-specific voice morphing could be used to create such stimuli with controlled acoustics and then compare adaptation conditions within and across perceived speaker gender.

4.3. The role of context in vocal emotion perception and beyond

The present findings add to a substantial body of evidence illustrating that vocal emotion perception is embedded within a given context and thus relative rather than absolute in nature, an idea that is also incorporated in current models on vocal emotion perception (Frühholz & Schweinberger, 2021; Grandjean, 2020). Contextual influence can be observed on different time scales. In adaptation, emotional perception is influenced by recent and preceding events. Other forms of contextual influence operate simultaneously: Vocal emotion perception is influenced by the musical or non-musical background (Liuni et al., 2020), semantic speech content (Bliss-Moreau et al., 2010), or input from other modalities (Baart & Vroomen, 2018; de Gelder & Vroomen, 2000). While these are examples that operate on relatively short times scales, contextual influences on emotional processing can also be observed over longer time scales. The own-culture advantage in vocal emotion recognition indicates that expression and perception of emotions are, to some extent at least, shaped by learning within the context of a given culture (Barrett, 2017; Gendron et al., 2014; Laukka & Elfenbein, 2021).

Contextual dependence is not specific to vocal emotions, but may form a general perceptual mechanism in the auditory domain and beyond (Pérez-González & Malmierca, 2014; Webster & MacLeod, 2011). For example, auditory context effects in speech and speech prosody in particular are well documented (Cole, 2015; King & Walker, 2020; Stilp, 2020). A key feature of these context effects seems to be contrast enhancement (Stilp, 2020), allowing the auditory system to magnify perceptual differences in order to increase its’ sensitivity towards meaningful changes in the auditory stream. Some forms of contrast enhancement interact with long-term learning processes, such as in the formation of language-specific phonemic categories through perceptual narrowing (Vihman, 2017).

As it stands, audition involves change detection, which is partly achieved through contrast enhancement. Against this background, the present study contributes to closing a gap, by showing that recalibration of emotion perception through adaptation operates both on the acoustic and a more abstract emotional level.

4.4. Limitations and further research

The present study has a few limitations that may serve as excellent starting points for further research. Obviously, our findings are limited to angry and fearful emotions only and one might argue, that they are both characterized by similar F0 contours (Brück et al., 2011; Juslin &

Laukka, 2003; Scherer, 1986). Consequently, it may be assumed that this limited variance in F0 between the two emotions would lead to a decrease in the aftereffect size in the F0 condition – a result that could be entirely different with other emotions. This is valid criticism and further research should address this point. Nonetheless, in the present set of stimuli substantial and systematic differences in F0 contours between angry and fearful utterances were found in the Mean, SD, Intonation and Glide (cf. Table 1).

It may be noted that the predominant role of timbre compared to F0 was already visible in the classification data of the adaptor stimuli. Classification was more disrupted in the F0 condition, where timbre was uninformative. One could therefore argue that the decrease in aftereffect size in the timbre and even more in the F0 condition compared to the Full condition was not due to the manipulation of different parameters per se, but rather to the ambiguity introduced by this manipulation. Any stimulus manipulation increasing emotional ambiguity could therefore affect the adaptation aftereffect. However, this seems unlikely, because adaptor perception and subsequent adaptation effects have diverged in previous studies: Bestelmeyer, Rouger, et al. (2010) used emotional caricatures, which were perceived as more emotionally expressive but did not result in greater adaptation effects. Likewise, Skuk et al. (2015) observed F0 and timbre to be equally important for gender classification but obtained a predominant role of timbre in the subsequent adaptation experiment.

In the future, valuable insight into the differential effects of timbre and F0 information could be provided by ERP effects of emotion adaptation. Previous electrophysiological studies reported an adaptation-induced attenuation of the N1 and P2 components (Schweinberger et al., 2011; Zäske et al., 2009). Other researchers suggested that the N1 is more sensitive to F0 contour whereas the P2 likely reflects processing of spectral information (Chartrand et al., 2008; Schröger, 2007). As a perspective, it might therefore be possible to dissociate the role of different vocal parameters on emotion adaptation at the brain level.

4.5. Summary and conclusion

Here, we showed for the first time that adaptation to vocal emotion is predominantly driven by timbre compared to F0 information when adaptation occurs within speaker gender. Across speaker gender, aftereffects are still visible but dramatically reduced, which hampers differentiation between parameter-specific effects. These results provide evidence that adaptation to vocal emotions interacts with the processing of vocal gender. Explanatory factors could be either the acoustic dissimilarity between speaker of opposite genders or different vocal processing mechanisms for emotions expressed by males and females. This study establishes the important role timbre in vocal emotion adaptation and adds to a growing body of evidence for the importance of timbre for auditory adaptation in other domains.

Declaration of Competing Interest

The authors declare no conflicts of interests.

Acknowledgements

Experiment 1 was conducted by C.N. in partial fulfilment of the requirements for a Master's thesis. The data for Experiment 2 were collected by Sarah Eckert in partial fulfilment of a Bachelor degree. We are grateful to Sascha Frühholz for providing us with original voice recordings that served as a basis for creating the stimulus material for these experiments. We thank Helene Kreysa for helpful suggestions on the manuscript. We thank all participants of the study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104967. Supplemental figures and tables, analysis scripts, and raw data can be found on the associated OSF Repository (https://osf.io/qzj6d/).

References

Abdi, H. (2007). Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics, 3*, 103–107.

ANSI. (1973). *Terminology, psychoacoustical. S3. 20.* Terminology, New York: American National Standards Institute, Psychoacoustical.

Baart, M., & Vroomen, J. (2018). Recalibration of vocal affect by a dynamic face. *Experimental Brain Research, 236*(7), 1911–1918. https://doi.org/10.1007/s00221-018-5270-y.

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

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5–17.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

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., Jones, B. C., DeBruine, L. M., Little, A. C., Welling, L. L. M., Bestelmeyer, P. E. G., Jones, B. C., DeBruine, L. M., & Welling, L. L. M. (2010). Face aftereffects suggest interdependent processing of expression and sex and of expression and race. *Visual Cognition, 18*(2), 255–274. https://doi.org/10.1080/13506280802708024.

Bestelmeyer, P. E. G., Maurage, P., Rouger, J., Latinus, M., & Belin, P. (2014). Adaptation to vocal expressions reveals multistep perception of auditory emotion. *The Journal of Neuroscience, 34*(24), 8098–8105. https://doi.org/10.1523/JNEUROSCI.4820-13.2014.

Bestelmeyer, P. E. G., & Mühl, C. (2021). Individual differences in voice adaptability are specifically linked to voice perception skill. *Cognition, 210*, Article 104582. https://doi.org/10.1016/j.cognition.2021.104582.

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.

Bliss-Moreau, E., Barrett, L. F., & Owren, M. J. (2010). I like the sound of your voice: Affective learning about vocal signals. *Journal of Experimental Social Psychology, 46*(3), 557–563. https://doi.org/10.1016/j.jesp.2009.12.017.

Boersma, P. (2018). Praat: Doing phonetics by computer [computer program]. version 6.0.46, retrieved January 2020 from http://www.Praat.Org/.

Bowman, C., & Yamauchi, T. (2017). Processing emotions in sounds: Cross-domain aftereffects of vocal utterances and musical sounds. *Cognition & Emotion, 31*(8), 1610–1626. https://doi.org/10.1080/02699931.2016.1255588.

Brody, L. R. (2000). The socialization of gender differences in emotional expression: Display rules, infant temperament, and differentiation. *Gender and Emotion: Social Psychological Perspectives, 2*(11), 122–137.

Brück, C., Kreifelts, B., & Wildgruber, D. (2011). Emotional voices in context: A neurobiological model of multimodal affective information processing. *Physics of Life Reviews, 8*(4), 383–403. https://doi.org/10.1016/j.plrev.2011.10.002.

Campbell, J., & Burke, D. (2009). Evidence that identity-dependent and identity-independent neural populations are recruited in the perception of five basic emotional facial expressions. *Vision Research, 49*(12), 1532–1540. https://doi.org/10.1016/j.visres.2009.03.009.

Chaplin, T. M., & Aldao, A. (2013). Gender differences in emotion expression in children: A meta-analytic review. *Psychological Bulletin, 139*(4), 735. https://doi.org/10.1037/a0030737.

Chartrand, J.-P., Peretz, I., & Belin, P. (2008). Auditory recognition expertise and domain specificity. *Brain Research, 1220*, 191–198. https://doi.org/10.1016/j.brainres.2008.01.014.

Clifford, C. W. G., & Rhodes, G. (2005). *Fitting the mind to the world: Adaptation and aftereffects in high-level vision*. Oxford University Press.

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience, 30*(1–2), 1–31. https://doi.org/10.1080/23273798.2014.963130.

Eichhorn, J. T., Kent, R. D., Austin, D., & Vorperian, H. K. (2018). Effects of aging on vocal fundamental frequency and vowel formants in men and women. *Journal of Voice, 32*(5), 644. e1-644. e9 https://doi.org/10.1016/j.jvoice.2017.08.003.

Ellamil, M., Susskind, J. M., & Anderson, A. K. (2008). Examinations of identity invariance in facial expression adaptation. *Cognitive, Affective, & Behavioral Neuroscience, 8*(3), 273–281.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Fox, C. J., & Barton, J. J. S. (2007). What is adapted in face adaptation? The neural representations of expression in the human visual system. *Brain Research, 1127*(1), 80–89. https://doi.org/10.1016/j.brainres.2006.09.104.

Freitag, C. M., Retz-Junginger, P., Retz, W., Seitz, C., Palmason, H., Meyer, J., Rösler, M., & von Gontard, A. (2007). Evaluation der deutschen version des autismus-spektrum-quotienten (aq) - die kurzversion aq-k. *Zeitschrift für Klinische Psychologie und Psychotherapie, 36*(4), 280–289. https://doi.org/10.1026/1616-3443.36.4.280.

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.

Frühholz, S., & Schweinberger, S. R. (2021). Nonverbal auditory communication - evidence for integrated neural systems for voice signal production and perception. *Progress in Neurobiology, 199,* Article 101948. https://doi.org/10.1016/j.pneurobio.2020.101948.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion, 14*(3), 289–311. https://doi.org/10.1080/026999300378824.

Gelfer, M. P., & Mikos, V. A. (2005). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of Voice, 19*(4), 544–554. https://doi.org/10.1016/j.jvoice.2004.10.006.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Cultural relativity in perceiving emotion from vocalizations. *Psychological Science, 25*(4), 911–920. https://doi.org/10.1177/0956797613517239.

Grandjean, D. (2020). Brain networks of emotional prosody processing. *Emotion Review, 13*(1), 34–43. https://doi.org/10.1177/1754073919898522

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences, 10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006.

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.

Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1*(1), 69–82. https://doi.org/10.3102/10769986001001069.

Jaquet, Emma, Rhodes, Gillian, & Hayward, William G. (2007). Opposite aftereffects for chinese and caucasian faces are selective for social category information and not just physical face differences. *The Quarterly Journal Of Experimental Psychology, 60*(11), 1457–1467. https://doi.org/10.1080/17470210701467870

Jenkins, R., Beaver, J. D., & Calder, A. J. (2006). I thought you were looking at me: Direction-specific aftereffects in gaze perception. *Psychological Science, 17*(6), 506–513. https://doi.org/10.1111/j.1467-9280.2006.01736.x.

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.

Kawahara, H., Morise, M., & Skuk, V. G. (2013). Temporally variable multi-aspect n-way morphing based on interference-free speech representations. In *IEEE international conference on acoustics, speech and signal processing*.

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. In *IEEE international conference on acoustics, speech and signal processing*.

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

King, A. J., & Walker, K. M. (2020). Listening in complex acoustic scenes. *Current Opinion in Physiology, 18*, 63–72. https://doi.org/10.1016/j.cophys.2020.09.001.

Klatt, D. H., Klatt, L. C., Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87*(2), 820–857. https://doi.org/10.1121/1.398894.

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology, 23*(12), 1075–1080. https://doi.org/10.1016/j.cub.2013.04.055.

Laukka, P., & Elfenbein, H. A. (2021). Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis. *Emotion Review, 13*(1), 3–11. https://doi.org/10.1177/1754073919897295.

Lawrence, M. A., & Lawrence, M. M. A. (2016). Package 'ez'. *R Package Version, 4*(0).

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience, 4*(1), 89–94. https://doi.org/10.1038/82947.

Liuni, M., Ponsot, E., Bryant, G. A., & Aucouturier, J. J. (2020). Sound context modulates perceived vocal emotion. *Behavioural Processes, 172*, Article 104042. https://doi.org/10.1016/j.beproc.2020.104042.

Mather, G. E., Verstraten, F. E., & Anstis, S. E. (1998). *The motion aftereffect: A modern perspective*. The MIT Press.

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

Paulmann, S., & Kotz, S. A. (2018). The electrophysiology and time course of processing vocal emotion expressions. *The Oxford Handbook of Voice Perception*, 459–472.

Pérez-González, D., & Malmierca, M. S. (2014). Adaptation in the auditory system: An overview. *Frontiers in Integrative Neuroscience, 8*, 19. https://doi.org/10.3389/fnint.2014.00019.

Piazza, E. A., Theunissen, F. E., Wessel, D., & Whitney, D. (2018). Rapid adaptation to the timbre of natural sounds. *Scientific Reports, 8*(1), 13826. https://doi.org/10.1038/s41598-018-32018-9w.

Psychology Software Tools, Inc. (2012). E-Prime 2.0. Retrieved from https://support.pstnet.com/.

Pye, A., & Bestelmeyer, P. E. G. (2015). Evidence for a supra-modal representation of emotion from cross-modal adaptation. *Cognition, 134*, 245–251. https://doi.org/10.1016/j.cognition.2014.11.001.

R Core Team. (2020). R: A language and environment for statistical computing. https://www.R-project.org/.

Rhodes, G., Lie, H. C., Ewing, L., Evangelista, E., & Tanaka, J. W. (2010). Does perceived race affect discrimination and recognition of ambiguous-race faces? A test of the sociocognitive hypothesis. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 36*(1), 217–223. https://doi.org/10.1037/a0017680.

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.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences, 10*(1), 24–30. https://doi.org/10.1016/j.tics.2005.11.009.

Schröger, E. (2007). Mismatch negativity: A microphone into auditory memory. *Journal of Psychophysiology, 21*(3–4), 138. https://doi.org/10.1027/0269-8803.21.34.138.

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., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zäske, R. (2014). Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(1), 15–25. https://doi.org/10.1002/wcs.1261.

Schweinberger, S. R., Kloth, N., & Jenkins, R. (2007). Are you looking at me? Neural correlates of gaze adaptation. *Neuroreport, 18*(7), 693–696. https://doi.org/10.1097/WNR.0b013e3280c1e2d2.

Schweinberger, S. R., & Soukup, G. R. (1998). Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. *Journal of Experimental Psychology: Human Perception and Performance, 24*(6), 1748. https://doi.org/10.1037/0096-1523.24.6.1748.

Schweinberger, S. R., Walther, C., Zäske, R., & Kovács, G. (2011). Neural correlates of adaptation to voice identity. *British Journal of Psychology, 102*(4), 748–764. https://doi.org/10.1111/j.2044-8295.2011.02048.x.

Schweinberger, S. R., Zäske, R., Walther, C., Golle, J., Kovács, G., & Wiese, H. (2010). Young without plastic surgery: Perceptual adaptation to the age of female and male faces // young without plastic surgery: Perceptual adaptation to the age of female and male faces. *Vision Research, 50*(23), 2570–2576. https://doi.org/10.1016/j.visres.2010.08.017.

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., & Schweinberger, S. R. (2013). Adaptation aftereffects in vocal emotion perception elicited by expressive faces and voices. *PLoS One, 8*(11), Article e81691. https://doi.org/10.1371/journal.pone.0081691.

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).

Stilp, C. (2020). Acoustic context effects in speech perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 11*(1), Article e1517. https://doi.org/10.1002/wcs.1517.

Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics, 27*(1), 24–27. https://doi.org/10.3758/bf03199901.

Vida, M. D., & Mondloch, C. J. (2009). Children's representations of facial expression and identity: Identity-contingent expression aftereffects // children's representations of facial expression and identity: Identity-contingent expression aftereffects. *Journal of Experimental Child Psychology, 104*(3), 326–345. https://doi.org/10.1016/j.jecp.2009.06.003.

Vihman, M. M. (2017). Learning words and learning sounds: Advances in language development. *British Journal of Psychology, 108*(1), 1–27. https://doi.org/10.1111/bjop.12207.

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature, 428*(6982), 557.

Webster, M. A., & MacLeod, D. I. A. (2011). Visual adaptation and face perception. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 366*(1571), 1702–1725. https://doi.org/10.1098/rstb.2010.0360.

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

Webster, M. A., Werner, J. S., & Field, D. J. (2005). Adaptation and the phenomenology of perception. In *Fitting the mind to the world: Adaptation and aftereffects in high level vision* (pp. 241–277).

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

Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences // face and voice perception: Understanding commonalities and differences. *Trends in Cognitive Sciences, 24*(5), 398–410. https://doi.org/10.1016/j.tics.2020.02.001.

Zäske, R., & Schweinberger, S. R. (2011). You are only as old as you sound: Auditory aftereffects in vocal age perception. *Hearing Research, 282*(1–2), 283–288. https://doi.org/10.1016/j.heares.2011.06.008.

Zäske, R., Schweinberger, S. R., Kaufmann, J. M., & Kawahara, H. (2009). In the ear of the beholder: Neural correlates of adaptation to voice gender. *The European Journal of Neuroscience, 30*(3), 527–534. https://doi.org/10.1111/j.1460-9568.2009.06839.x.

Zäske, R., Schweinberger, S. R., & Kawahara, H. (2010). Voice aftereffects of adaptation to speaker identity. *Hearing Research, 268*(1–2), 38–45. https://doi.org/10.1016/j.heares.2010.04.011.

Zäske, R., Skuk, V. G., Kaufmann, J. M., & Schweinberger, S. R. (2013). Perceiving vocal age and gender: An adaptation approach. *Acta Psychologica, 144*(3), 583–593. https://doi.org/10.1016/j.actpsy.2013.09.009.