AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2025, Vol. 154, No. 1, 279–293
https://doi.org/10.1037/xge0001672

Contrastive Adaptation Effects Along a Voice–Nonvoice Continuum

Zi Gao and Andrew J. Oxenham
Department of Psychology, University of Minnesota, Twin Cities

Adaptation to the environment is a universal property of perception across all sensory modalities. It can enhance the salience of new events in an ongoing background and helps maintain perceptual constancy in the face of variable sensory input. Several contrastive adaptation effects have been identified using sounds within the categories of human voice and musical instruments. The present study investigated whether such contrast effects can occur between voice and nonvoice stimulus categories. A 10-step continuum between "voice" (/a/, /o/, or /u/ vowels) and "instrument" (bassoon, horn, or viola) sounds was generated for each of the nine possible pairs. In each trial, an adaptor, either a voice or instrument, was played four times and was followed by a target from along the appropriate continuum. When trials with voice and instrumental adaptors were grouped into separate blocks, strong contrastive adaptation effects were observed, with the target more likely to be identified as a voice following instrumental adaptors and vice versa (Experiment 1). The effects were not observed for visual image adaptors (Experiment 2). The effects were somewhat larger when the adaptors and the target were presented to the same than to different ears, but significant adaptation was observed in both conditions, suggesting contributions of central mechanisms, following binaural integration (Experiment 3). The effect accumulated when the same type of adaptor was presented consecutively and persisted following the end of the adaptors (Experiment 4). The discovery of voice-nonvoice contrastive pairs opens the possibility of studying perceptual or neuronal voice selectivity while keeping acoustic features constant.

Public Significance Statement
Our study suggests that the perceptual boundary between voice and nonvoice is dependent on the preceding auditory context in a contrastive way. These context effects can build up when similar contexts are presented consecutively and can persist across time even when the context is no longer present. The results highlight the possibility of manipulating the perceptual identity of a sound as voice or nonvoice while keeping the acoustic signal the same, therefore addressing a common challenge in previous voice perception research.

Keywords: auditory adaptation, voice, morphing, context effects

Neural adaptation refers to a decrease in response after prolonged or repetitive stimulation (e.g., Zäske et al., 2010). Perceptually, it can manifest itself by way of aftereffects, where exposure to adaptor stimuli can bias perception toward opposite features of the adaptors. For example, after looking at a left-tilted line for an extended period of time, observers tend to perceive an upright line as right-tilted and vice versa (Gibson & Radner, 1937). Adaptation is crucial for acquiring information from varied sensory signals because it can facilitate the detection of changes or novelty in sensory signals and improve the efficiency of coding incoming information (for a review, see Pérez-González & Malmierca, 2014). It also allows for recalibrations in the mapping between sensory signals and perception based on the preceding context, an important step in performing challenging tasks like understanding accented speech (Clarke & Garrett, 2004).

Adaptation effects have been observed at multiple processing levels of different perceptual modalities. Early studies revealed contrastive aftereffects in the perception of many fundamental

This article was published Online First October 14, 2024.
Joseph Toscano served as action editor.
Zi Gao https://orcid.org/0009-0000-2508-4448
Andrew J. Oxenham https://orcid.org/0000-0002-9365-1157
Zi Gao is now at Ohio State Eye and Ear Institute, Columbus, Ohio, United States.
Materials used in this study and data are available for download at https://osf.io/zjq4s/. Experiments 1–3 and preliminary results of Experiment 4 were presented at the 182nd Meeting of the Acoustical Society of America in Denver, Colorado.

The study was supported by the National Institutes of Health (Grant R01 DC012262) awarded to Andrew J. Oxenham and the University of Minnesota Doctoral Dissertation Fellowship awarded to Zi Gao. The authors thank Penelope Corbett for assisting in data collection.

Zi Gao played a lead role in data curation, formal analysis, investigation, software, validation, visualization, and writing–original draft and an equal role in conceptualization, methodology, and writing–review and editing. Andrew J. Oxenham played a lead role in funding acquisition, project administration, resources, and supervision, a supporting role in formal analysis, and an equal role in conceptualization, methodology, and writing–review and editing.

Correspondence concerning this article should be addressed to Zi Gao, Ohio State Eye and Ear Institute, 915 Olentangy River Road, Columbus, OH 43212, United States. Email: gao00196@umn.edu

attributes of vision and audition, including sensitivity to light (Armington & Biersdorf, 1958), motion (Pantle, 1974), and sound intensity (Canévet et al., 1985). More recently, adaptation has been used as a tool to investigate face perception, a highly relevant and important social signal. Following the theoretical framework of the "face space" (Valentine, 1991), some of the dimensions along which facial representations differ from each other in the face space have been identified through adaptation studies. These dimensions include both global properties, like gender, and specific facial parameters, like eye–mouth distance (for a review, see Mueller et al., 2020).

In the auditory modality, adaptation effects have been observed for a variety of properties of the human voice. Early studies investigated whether and how neighboring sounds affect phoneme perception in speech. For example, following a sentence where the first formant was lowered in frequency, listeners tended to perceive an ambiguous test word as "bit" rather than "bet," implying an increase in the perceived first formant frequency (Broadbent & Ladefoged, 1960). Many of these effects in voice can be explained in terms of relatively low-level contrasts in spectral content and/or spectral motion (Wang & Oxenham, 2014). Holt et al. (2000) found that sine-wave tones, either with a steady frequency or modified to mimic the trajectory of a consonant formant, were sufficient to induce contrastive context effects in vowel perception, suggesting that language content is not necessary to elicit the effects. Contrastive context effects have also been found in Japanese quails (Lotto et al., 1997), suggesting that the effects do not require prior language knowledge. In addition, similar adaptation and contrast effects have been observed with nonvoice harmonic complex tones in the form of musical instrument sounds for both the adaptor and the target (Frazier et al., 2019; Lanning & Stilp, 2020), as well as synthetic inharmonic tone complexes (Byrne et al., 2013; Feng & Oxenham, 2015), confirming that adaptation is a general auditory process rather than a speech-specific mechanism.

Voice adaptation effects have also been observed in the perception of high-level paralinguistic information, such as gender (Schweinberger et al., 2008), speaker identity (Zäske et al., 2010), and emotion (Bestelmeyer et al., 2010). In these studies, adaptation to naturalistic stimuli from one of two categories (e.g., male and female) resulted in a contrastive shift in the perception of ambiguous sounds along a continuum between the two categories. It remains possible that these effects are caused at least in part by contrasts in low-level acoustic characteristics, as observed in earlier studies on phoneme perception (e.g., Broadbent & Ladefoged, 1960). However, Zäske et al. (2013) found that spatial attention is required for voice gender adaptation, whereas adaptation to phonetic features can persist without attention (Feng & Oxenham, 2018; Sussman, 1993), suggesting that the mechanisms of adaptation to high-level sound features, such as gender and speaker identity, are at least partially different from those governing adaptation to low-level features.

In addition to having been used to demonstrate a range of adaptation effects, the human voice has been hypothesized to hold a perceptually and neurally privileged position, with several studies reporting voice-sensitive brain regions (e.g., Agus et al., 2017; Belin et al., 2000) and others reporting perceptual voice advantage (Agus et al., 2012; Isnard et al., 2019; Weiss et al., 2015) and disadvantage (Gao & Oxenham, 2022; Hutchins et al., 2012) effects. It remains unclear, however, whether the perceptual boundary between voice and nonvoice categories is fixed or flexible and whether the neurally and perceptually relevant distinctions between voice and nonvoice sounds reflect low-level acoustic features or higher level perceptual categories. Adaptation effects between voice and nonvoice stimuli, if observed, could provide one way to address these questions, as adaptation offers a potential method to manipulate perceived categories without changing low-level acoustic features. Although adaptation effects have been studied using both voice and nonvoice sounds in the past, to our knowledge it is not known whether adaptation effects occur between these categories.

The present study was designed to test for the existence of perceptual adaptation effects between voice and nonvoice stimuli and to determine their time course. Tones played on musical instruments were selected as the nonvoice stimuli due to their spectral richness and harmonic structure (similar to voices) and because they have been used in previous studies of voice-selective processing (e.g., Agus et al., 2017; Weiss et al., 2015). Experiment 1 tested for the existence of adaptation between voice and nonvoice categories by having participants listen to either voice or musical instrument adaptors prior to the presentation of an ambiguous target, selected from a voice–instrument continuum, and then categorize the target as either a voice or instrument. To test for potential cross-modal effects, Experiment 2 replaced the voice and instrument adaptors with images of vowels and instruments. To help distinguish between peripheral and central auditory mechanisms underlying the adaptation effects, Experiments 3 and 4 tested the transfer of adaptation effects between the ears and measured the accumulation and persistence of adaptation by presenting trials in blocks of different lengths, based on adaptor type, or presenting them in random order. The results provide the first demonstration of robust contrastive adaptation effects between voice and nonvoice stimuli, which build up over time and seem to persist for up to a minute after the offset of the last adaptor.

Experiment 1

Experiment 1 served to determine whether contrastive adaptation effects between voice and nonvoice sounds exist, thereby extending previous findings on high-level auditory adaptation effects to a novel and important dimension. In this experiment, we measured such adaptation effects in terms of the shift in the point of subjective equality (PSE) between voice and instrumental sounds, following repeated exposure to voice or musical instrument adaptors, as compared with silence.

Method

Transparency and Openness

This study was not preregistered. Sound stimuli, experiment programs, de-identified data, and codes for data tidying and statistical analysis are available for download at https://osf.io/zjq4s/ (Gao & Oxenham, 2023). Data exclusions (if any) are reported in the Participants section of each experiment. Data were tidied and analyzed using MATLAB R2018b (MathWorks Inc., Natick, Massachusetts) and R v4.2.2 (R Core Team, 2022).

Because no previous study tested adaptation effects between voice and nonvoice, the required sample size was difficult to estimate. The current sample sizes were selected to emulate previous

studies on auditory adaptation effects with a similar paradigm, such as Schweinberger et al. (2008) and Zäske et al. (2010).

Participants

Thirty-five participants took part in this online experiment. Their ages ranged from 18 to 27 years ($M = 21.4$, $SD = 2.4$). Four identified as men and 31 as women. They identified their race/ethnicity as follows: 12 Asian, three Black, two multiracial, 18 White, five Hispanic or Latinx, and 30 non-Hispanic or Latinx. The participants had an average experience of 4.1 years (range = 0–18) playing musical instruments and/or receiving musical training. Two participants were excluded because their psychometric functions did not cross the 50% point, making their PSE unmeasurable and suggesting that they were unable to reliably categorize unambiguous voices and instrumental tones, leaving a total of 33 participants from whom data were analyzed. All participants reported having normal hearing. Participants were recruited through introductory psychology courses at the University of Minnesota. The study was approved by the University of Minnesota's institutional review board. All participants completed an online consent form prior to the study, and demographic information was collected following the consent form. Age was collected through a mandatory free-response box; gender, race, and ethnicity were collected using optional multiple-choice questions. All participants were awarded a digital gift card or extra course credit upon completion.

Stimuli

Three types of auditory stimuli were used in this experiment: voices, instrumental sounds, and intermediate sounds created by morphing voice and instrumental sounds. The voices were utterances of the vowels /a/, /o/, and /u/, pronounced by three female native speakers of American English. All three vowels were recorded by all three speakers in a sound attenuated booth at a 44.1-kHz sampling rate with a 16-bit resolution, using a PMD670 solid state recorder (Marantz, Carlsbad, California) and an ME64 stationary microphone (Sennheiser, Wedemark, Germany), resulting in a total of nine stimuli. The three instrumental stimuli were downloaded from the Philharmonia Orchestra online sound library (https://philharmonia.co.uk/resources/sound-samples/) and consisted of single notes played on bassoon, French horn, and viola by professional musicians. These three musical instruments were selected because they are each from different instrument families (woodwind, brass, and string), and their comfortable fundamental frequency (F0) ranges largely overlap with that of the human voice. All 12 stimuli (nine voices and three instrumental sounds) were trimmed to 686 ms, adjusted to the same pitch of G3 (196 Hz), scaled to the same root-mean-squared (rms) level, and gated off with a 30-ms raised-cosine ramp. The pitch G3 was selected because it was the closest to the mean F0 of the original voice stimuli prior to adjustments ($M = 200.8$ Hz, range = 192.4–207.2 Hz). The onsets of the stimuli were not altered to preserve the natural onset characteristics of voice and instrumental tones. Adjustments to the F0 were performed using Praat (Boersma & Weenink, 2021), and all other manipulations were performed in MATLAB R2018b (MathWorks Inc., Natick, Massachusetts).

Intermediate sounds were created by morphing voices with instrumental tones using Tandem-STRAIGHT (Kawahara & Matsui, 2003; Kawahara et al., 2008), a signal processing tool that can perform low-dimensional piecewise bilinear time-frequency mapping between two endpoint stimuli, creating a smooth continuum of ambiguous intermediate stimuli. For each possible pair of voice and instrumental tones (9 × 3 or 27 in total), a continuum was created, ranging in voice/instrument morphing proportions from 100%/0% to 0%/100% by steps of 10%. The spectrograms of example stimuli prior to morphing are shown in Figure 1, along with one example continuum ("/u/-horn").

Procedure

The experiment was conducted online using PsychoPy v2021.2.3 (Peirce et al., 2019) and could only be completed on computers, not tablets. Participants were instructed to complete the experiment in a quiet environment. Headphones were neither required nor screened for. A description of the experimental task was displayed at the beginning of the experiment and at the beginning of each block.

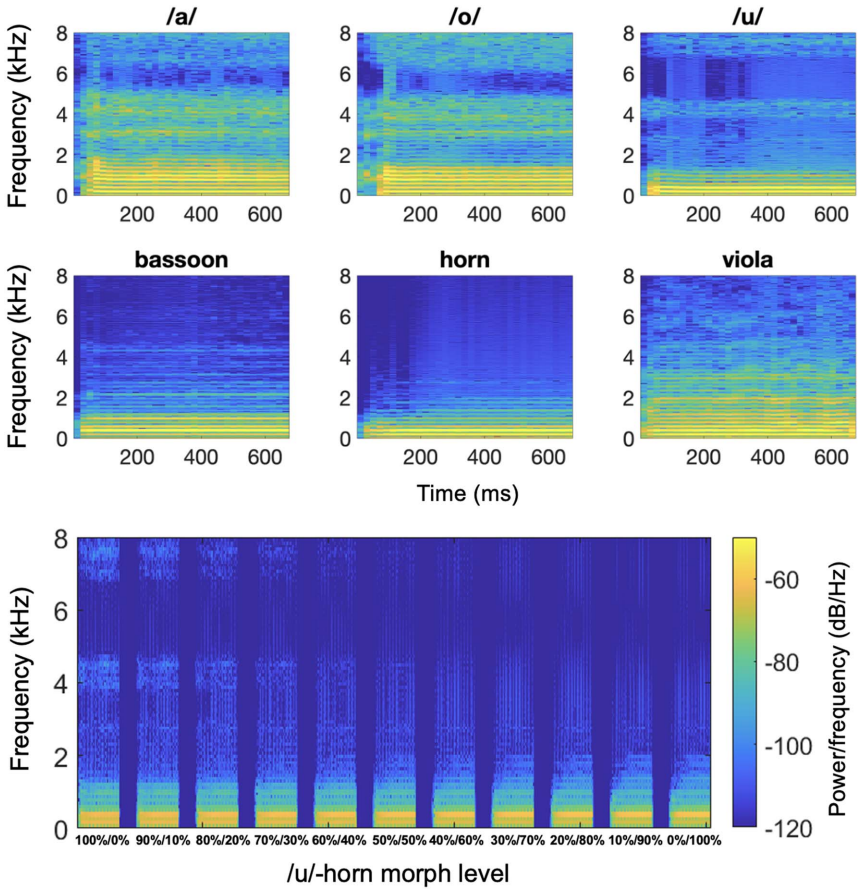
In each trial of the experiment (as shown in Figure 2), following a 600-ms fixation point displayed at the center of the screen, an adaptor was played four times with a 200-ms silent gap between each presentation. The last presentation of the adaptor was followed by a 700-ms silent gap, with the last 600 ms of the silence accompanied by a prompt screen, reminding the participant to categorize the target as either a voice or instrument (by pressing "V" or "I" on the keyboard, respectively). The target was then played, and the participants had up to 2.8 s to respond. The gaps between consecutive adaptors and the target, as well as the response time, were taken from the voice gender adaptation study of Schweinberger et al. (2008). The next trial started immediately after the response had been recorded. If the participant failed to respond within the designated time, a prompt of "please respond sooner" was displayed for 500 ms before the next trial commenced. These missed trials (2.2% of all trials) were removed prior to data analysis. Reaction times were collected to assist in removing the missed trials but were not further analyzed. Intermediate stimuli corresponding to 80%/20%, 70%/30%, 60%/40%, 50%/50%, 40%/60%, 30%/70%, and 20%/80% voice/instrument proportions were used as targets. The original voices (100%/0%) and instrumental tones (0%/100%) were used as adaptors.

The experiment consisted of three conditions: voice adaptor, instrumental adaptor, and silence, each having two blocks, for a total of six blocks. In each block, each speaker was randomly and uniquely assigned to one of the three vowels. Three vowels (each uttered by a different speaker) and three instruments resulted in nine vowel-instrument target pairs, each of which was presented at seven morphing levels, resulting in a total of 63 targets. With each target presented once per block, there were 63 trials per block. The adaptors were randomly selected with the constraint that the adaptor and the target within a given trial differed in instrument, vowel, and speaker to minimize any dependence of the results on specific spectral pattern similarities between adaptor and target and to help ensure generalizability across different voices and instruments. In the silent condition, the adaptors were simply replaced with silence of the same duration. The order of the trials and the blocks was randomized separately for each participant. Participants were permitted to take a break between blocks. The experiment took about 35 min to complete.

Prior to the experiment, participants completed a short practice block consisting of six trials. The practice trials served to familiarize

Figure 1

Spectrograms of Vowels, Instrumental Tones, and the "/u/-Horn" Continuum



Note. The vowels and morphed sounds shown are from one of the three speakers (Speaker 1). See the online article for the color version of this figure.

the participants with the task; they differed from the actual experiment in two ways: (a) The targets were either 100% voice or 100% instrument rather than the more ambiguous intermediate sounds and (b) feedback of "correct" or "wrong" was provided after participants responded to each test stimulus. The responses in the practice block were neither analyzed nor used as a screening criterion for the experiment. Participants were allowed to adjust the volume of the stimuli to a comfortable level in the practice block but were instructed not to adjust the volume further during the actual experiment.

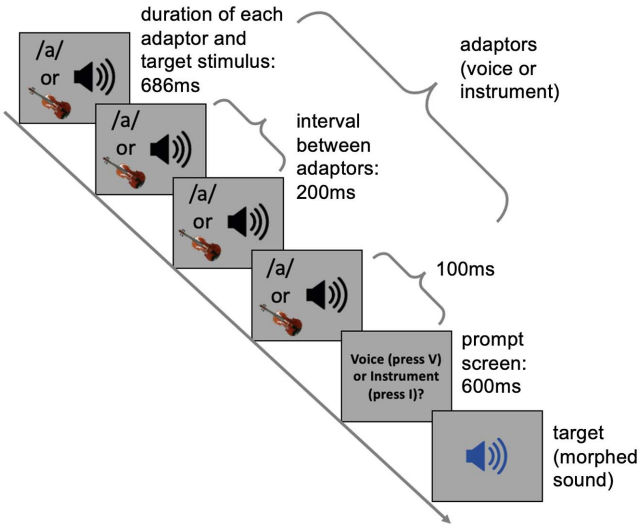
Results

The proportion of "voice" responses across different morph levels and adaptor conditions is shown in Figure 3. The mean data show clear adaptation effects, with instrument adaptors producing more "voice" responses and voice adaptors producing more "instrument" responses than in the silent control condition. To quantify the adaptation effect, the PSE was estimated for each participant, defined as the morphing level at which "voice" and "instrument" responses were equally likely. The PSE was estimated by fitting the proportions of each participant's "voice" responses across different morphing levels of the target following the same type of adaptor (i.e., voice, instrument, or silence) to a logit function in MATLAB 2018b and finding the midpoint (50%) of the function. A higher PSE implies that a larger proportion of voice in the morphed stimuli is required to induce a "voice" response, suggesting that the participants are less likely to perceive the morphed stimuli as a voice and more likely to perceive it as an instrumental sound. The PSEs for the three conditions are shown in Figure 4.

A repeated-measures analysis of variance (ANOVA) was performed in R v4.2.2 (R Core Team, 2022) with the PSE as the dependent variable and a within-subjects factor of adaptor type (voice, instrument, silence). A significant effect of adaptor type was observed, $F(2, 64) = 88.41$, $p < .0001$, partial eta-squared $\eta_p^2 = 0.73$, two-sided 90% confidence interval (CI) [0.64, 0.80]. Post hoc paired comparisons with Bonferroni correction (p values adjusted for three comparisons) showed significant differences between each of the three adaptor conditions ($p < .0001$ in all cases): The PSE was higher for voice adaptors than for silence and instrumental adaptors and was higher for silence than for instrumental adaptors. The pattern of results remained the same when participants' musical experience was added as a covariate in the analysis, suggesting no strong effect of musical training on the results.

Figure 2

Schematic Diagram of One Trial in Experiment 1

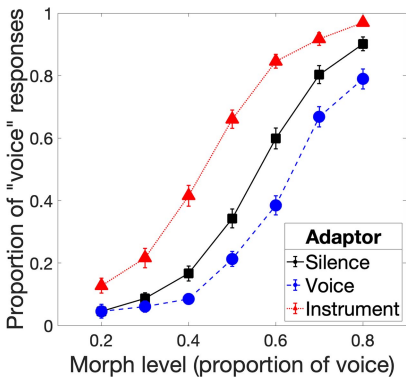


Note. The photos of the musical instruments are in the public domain and were acquired from Wikimedia: https://commons.wikimedia.org/wiki/File:Bratsche.jpg. See the online article for the color version of this figure.

Additional analyses were performed to explore the potential contribution of the identities of the adaptors. To test for the potential differences between different instrument adaptors, the PSE following each type of instrument adaptor (bassoon, horn, and viola) was calculated for each participant. A repeated-measures ANOVA comparing PSEs across the three adaptor instruments revealed no significant difference $F(2, 64) = 1$, $p = .374$, $\eta_p^2 = 0.03$, 90% CI [0.00, 0.11]. Similar repeated-measures ANOVAs revealed a significant effect on the PSE of the adaptor vowel identity, $F(2, 64) = 4.49$, $p = .015$, $\eta_p^2 = 0.12$, 90% CI [0.01, 0.24], driven by a greater tendency to respond "voice" following /o/ than /u/ ($p = .011$), and a marginally significant effect of adaptor speaker $F(2, 64) = 2.93$, $p = .061$, $\eta_p^2 = 0.08$, 90% CI [0.00, 0.19], driven by a greater tendency to respond "voice" following Speaker 3 than Speaker 2 ($p = .032$). However, because the target speaker, vowel, and instrument are always different from that of the adaptor in the same trial, our design does not allow us to clearly separate the effects of adaptors from targets.

Figure 3

Mean Proportion of "Voice" Responses Across Different Morph Levels and Sound Adaptor Types



Note. Error bars represent ±1 standard error of the mean. See the online article for the color version of this figure.

Discussion

Our hypothesis of contrastive adaptation effects between voice and instrumental sounds was supported by the observed shifts in the PSE following voice and instrumental adaptors: After listening to a voice adaptor, the participants were more likely to perceive the ambiguous target as an instrumental sound than when the adaptor was absent and vice versa. This result demonstrates that contrastive aftereffects can occur between categories (voice/instrument) in a similar manner as has previously been demonstrated for within-category contrasts using just voice (Schweinberger et al., 2008; Zäske et al., 2010) or instrument sounds (Lanning & Stilp, 2020). Because our adaptor stimuli were never the same speaker, instrument, or vowel as the target, it seems unlikely that the results are fully attributable to low-level adaptation due to repeated exposure to specific acoustic (e.g., spectral) characteristics. The results instead suggest that adaptation may be a domain-general mechanism in the categorization process of sound perception that can occur at different levels of auditory processing.

Experiment 2

Although adaptation effects between voice and instrumental sounds were observed in Experiment 1, it is unclear whether the presence of the auditory stimuli is a prerequisite of the effect or whether the abstract concepts of voice and instrument alone, conveyed via a different sensory modality, would be enough to elicit the effect. Experiment 2 investigated the potential role of cross-modal effects by replacing the sound adaptors of Experiment 1 with images of written vowels and photographed musical instruments. Visual adaptors like images and muted videos have been used in previous auditory adaptation studies to assess the contribution of abstract concepts to high-level auditory adaptation effects and have yielded mixed results: Neither images nor muted videos induced voice gender adaptation effects (Schweinberger et al., 2008), but weak effects were observed when muted videos were used as adaptors in speaker identity adaptation experiments (Zäske et al., 2010).

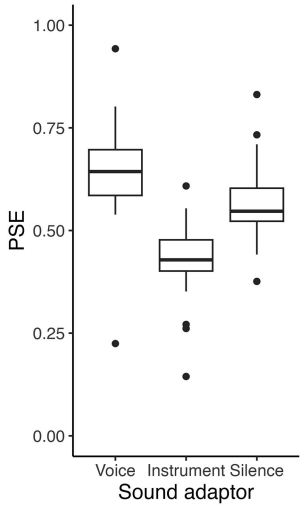
Method

Participants

Forty-seven participants took part in this online experiment. Their ages ranged from 18 to 71 years ($M = 21.4$, $SD = 7.7$). Ten identified as men, 36 as women, and one chose not to disclose gender. They identified their race/ethnicity as follows: 16 Asian, four Black, two multiracial, 25 White, three Hispanic or Latinx, and 44 non-Hispanic or Latinx. The participants had an average experience of 4.3 years (range = 0–21) playing musical instruments and/or receiving musical training. All participants reported having normal hearing. Participants were recruited through introductory psychology courses at the University of Minnesota. This study was approved by the University of Minnesota's institutional review

Figure 4

The Effects of Sound Adaptor Type on Point of Subjective Equality



Note. Larger PSEs correspond to a greater tendency to perceive ambiguous sounds as musical instruments. Black horizontal lines presented in bold denote median values, boxed areas are the interquartile ranges (IQR), and vertical lines denote the range of the values that are no further than 1.5IQR away from the boxed area. Individual outliers (defined as values more than 1.5IQR away from the boxed areas) are shown as dots. PSE = point of subjective equality.

board. All participants completed an online consent form prior to the study and were awarded a digital gift card or extra course credit upon completion.

Stimuli

In addition to the targets described in the Stimuli section of Experiment 1, this experiment involved six image adaptors: photographs of a bassoon, French horn, and viola and images of the spelled-out phonemes "/ah/," "/au/," and "/oo/," all cropped to a square of the same size (Figure 5). All images had a white background and were displayed over a gray (RGB = [128, 128, 128]) screen in the experiment. The height and width of the image were set to half of the screen height, so the actual display size of the image was dependent on the size of the participants' computer screen.

Procedure

The procedure was mostly the same as in Experiment 1, with two differences: (a) The voice and instrumental sound adaptors were replaced with the corresponding image adaptors, each of which flashed four times with the same onset and offset times as the sound adaptors in Experiment 1, and (b) the silent adaptor condition was removed to reduce the duration of the experiment to about 25 min.

Results

All missed trials (1.8% of all trials) were removed prior to data analysis. The mean proportion of "voice" responses and PSEs across conditions are shown in Figures 6 and 7, respectively. As the independent variable had two levels (rather than the three of Experiment 1), a paired-samples t test was performed in R v4.2.2. This test revealed no significant difference in PSEs between the vowel image adaptor and instrument image adaptor conditions, $t(46) = 1.01$, $p = .316$, Cohen's $d = 0.15$, 95% CI [−0.15, 0.44]. The pattern of results remained the same when participants' musical experience was added as a covariate in the analysis. Therefore, visual image adaptors failed to induce the kind of significant adaptation effects that we had observed in Experiment 1.

Discussion

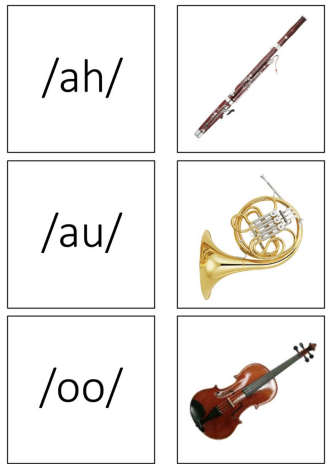
Combining the results from Experiments 1 and 2, adaptation effects between voice and instrumental sounds were observed when vowel utterances and instrumental tones were used as adaptors, but not images of vowel phonemes and instruments. This pattern of results suggests that the abstract concepts of voice and instrument, without the presence of auditory stimuli, are not sufficient to elicit adaptation effects between voice and instrument in the auditory domain. Therefore, it is unlikely that the effects observed in Experiment 1 reflect a modality-general mechanism for identifying human-sourced sensory stimuli; instead, these high-level auditory adaptation effects seem to rely on modality-specific mechanisms.

Experiment 3

The main aim of Experiment 3 was to investigate the loci of the auditory adaptation effects between voice and instrumental sounds observed in Experiment 1. Although we attempted to minimize the dependence on low-level acoustic features using a variety of vowels and instrumental tones in Experiment 1, voice and instrumental sounds still have some systematic differences in terms of spectral characteristics (e.g., voices usually have multiple spectral peaks,

Figure 5

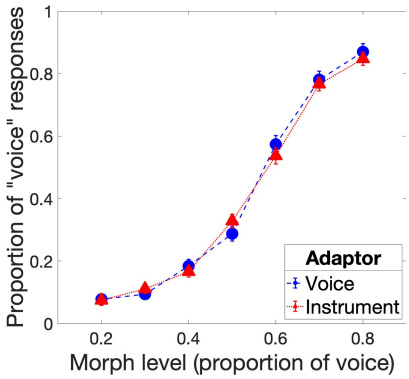
Image Adaptors: Vowel Syllables and Musical Instruments



Note. The photos of the musical instruments were acquired from Wikimedia under free licenses (bassoon and horn: CC BY-SA 4.0; viola: In the public domain). Bassoon: https://commons.wikimedia.org/w/index.php?curid=56208590; horn: https://commons.wikimedia.org/wiki/File:Yamaha_Horn_YHR-314II.tif; viola: https://commons.wikimedia.org/wiki/File:Bratsche.jpg. See the online article for the color version of this figure.

Figure 6

Mean Proportion of "Voice" Responses Across Different Morph Levels and Image Adaptor Types



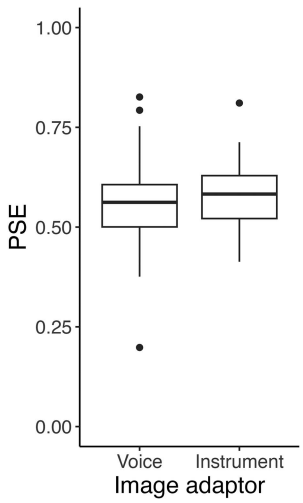
Note. Error bars represent ±1 standard error of the mean. See the online article for the color version of this figure.

corresponding to formants, whereas instrumental sounds do not). Therefore, it is still possible that the observed high-level adaptation effects were a result of auditory habituation following prolonged neural activation of the similar spectral patterns.

To pursue this possibility, Experiment 3 examined the loci of the adaptation effects using a dichotic listening paradigm, where the adaptors and the targets were presented to either the same or different ears. The rationale is that auditory inputs from different ears are initially processed quite independently in the two cochleae until convergence in the auditory brainstem and midbrain. Therefore, if the effects are still present when the stimuli are presented dichotically, then the effect does not depend fully on peripheral auditory mechanisms. If, on the other hand, the effect is reduced by dichotic presentation, that would indicate peripheral contributions to the effect and/or effects that depend on the

Figure 7

The Effects of Image Adaptor Type on Point of Subjective Equality



Note. Larger PSEs correspond to a greater tendency to perceive ambiguous sounds as musical instruments. PSE = point of subjective equality.

perceived spatial location of the stimuli (Feng & Oxenham, 2018; Holt & Lotto, 2002; Watkins, 1991). This paradigm has been widely used in previous studies of auditory and speech context effects based on spectral contrasts but has seldom been used in high-level adaptation effects like gender and speaker identity (Zäske et al., 2013).

Experiment 3 also served as a preliminary investigation into some of the timing aspects of the adaptation effects, specifically, whether the effects could accumulate over time. In Experiment 1, all trials in the same block always shared the same type of adaptor (i.e., either all voices or all instrumental sounds). It is possible that the adaptation effects accumulated across trials, resulting in a large effect at the block level but potentially much smaller effects after any individual trial. Experiment 3 examined this possibility by presenting the trials either in a blocked order (i.e., trials in the same block always share the same type of adaptors) or in a completely randomized order.

Method

Participants

Sixty-three participants took part in this in-person experiment. Their ages ranged from 18 to 28 years (M = 20.0, SD = 1.8). Twenty-three identified as men, 39 as women, and one chose not to disclose gender. They identified their race/ethnicity as follows: 25 Asian, five Black, one multiracial, one Pacific Islander, 29 White, two undisclosed, six Hispanic or Latinx, 56 non-Hispanic or Latinx, and one undisclosed. The participants had an average experience of 4.0 years (range = 0–14) playing musical instruments and/or receiving musical training. Three participants were excluded because their PSEs were not measurable, leaving a total of 60 participants in the analyzed data, equally distributed across the three between-subjects conditions, as described below. All participants self-reported having normal hearing and were screened via pure-tone audiometry to ensure audiometric thresholds no greater than 20 dB hearing level at octave frequencies between 250 and 8,000 Hz. Participants were recruited through introductory psychology courses. This study was approved by the University of Minnesota's institutional review board. All participants completed a consent form prior to the study and were awarded a digital gift card or extra course credit upon completion.

Stimuli

The stimuli used in this experiment were the same as those used in Experiment 1, with the exception that each stimulus was presented to only one ear at a time. In a given trial, all adaptors were presented to one ear, and the target was presented to either the same or different ear. All stimuli were scaled to a 70-dB sound pressure level and were delivered through Sennheiser H650 headphones (Wedemark, Germany) in a sound-attenuating booth.

Design and Procedure

This experiment adopted a mixed design, with ears (adaptor and target in same or different ears) and adaptor type (voice or instrument) as within-subjects factors and grouping method as a between-subjects factor. The three levels of grouping method were as follows. (a) Fixed ear: Within each block, all trials shared the same type of adaptor (i.e., voice or instrument) and were presented to a fixed combination of ears. For example, in a given block, all

adaptors were voice stimuli presented to the left ear, and all targets were presented to the right ear. (b) Mixed ear: Within each block, all trials shared the same type of adaptor and the same relationship (i.e., same or different ears) between the adaptor ear and the test ear, but whether the adaptors were presented to the left or the right ear was randomized on a trial-by-trial basis. For example, in a certain block, all adaptors were presented to the same ear as the targets, but whether they were presented to the left or the right ear would be randomly determined on each trial. (c) All random: both adaptor types and presentation ears were randomized on a trial-by-trial basis. These three conditions differed in the overall variability of stimuli presentation and adaptor types, allowing us to investigate whether the adaptation effects transfer across ears and whether they accumulate across trials.

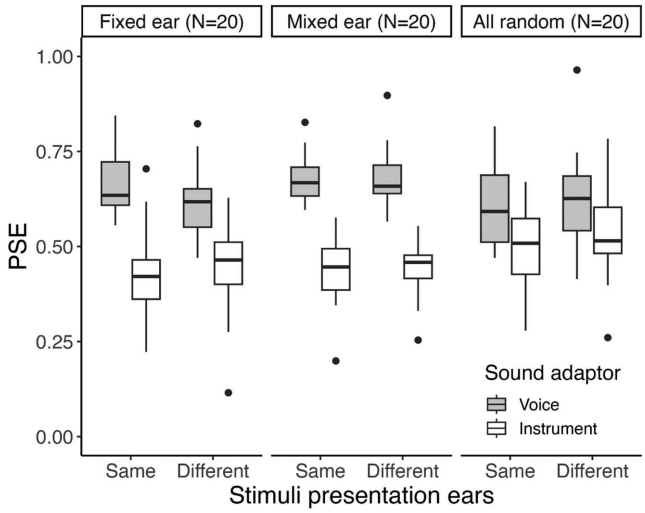
This experiment consisted of eight blocks of 63 trials. Across the blocks, each possible combination of adaptor type and ears had the same number of trials. With the exception of dichotic presentation and grouping methods, the experimental procedure was the same as Experiment 1. The experiment took each participant about 80 min to complete, including breaks.

Results

All missed trials (0.9% of all trials) were removed prior to data analysis. A repeated-measures ANOVA was performed in R v4.2.2 to examine the within-subjects effects of ears (same or different) and adaptor type (voice or instrument) and between-subjects effect of grouping method (fixed ear, mixed ear, or all random) on the PSE of targets (Figure 8). There was a significant main effect adaptor type, $F(1,57) = 238.68$, $p < .0001$, $\eta_p^2 = 0.81$, 90% CI [0.73, 0.85], where the PSE was higher following voice adaptors compared with instrumental adaptors, confirming significant overall adaptation effects. There was a significant interaction between adaptor type and ears, $F(1, 57) = 5.08$, $p = .028$, $\eta_p^2 = 0.08$, 90% CI [0.00, 0.21].

Figure 8

The Effects of Adaptor Type, Stimuli Presentation, and Grouping Methods on Point of Subjective Equality



Note. Larger PSE corresponds to a greater tendency to perceive ambiguous sounds as musical instruments. PSE = point of subjective equality.

Post hoc analyses with Bonferroni correction (p values adjusted for two comparisons) revealed significant adaptation effects in both the same ear and the different ear conditions ($p < .0001$ in both cases), but the effect was larger in the same ear condition than in the different ear condition. There was also a significant interaction between adaptor type and grouping method, $F(2, 57) = 13.46$, $p < .0001$, $\eta_p^2 = 0.32$, 90% CI [0.15, 0.46]. Post hoc analyses with Bonferroni correction (p values adjusted for three comparisons) revealed significant adaptation effects in all group types ($p < .0001$ in all cases); the effect was the smallest in the "all random" condition, while the "fixed ear" and the "mixed ear" conditions were more comparable. No other significant main effects or interactions were observed ($p > .08$ in all cases).

Discussion

Contrastive adaptation effects between voice and instrumental sounds were observed when the adaptors and the targets were presented to different ears, suggesting that the effects do not rely fully on peripheral mechanisms. This outcome might be expected, given that the gap between the last adaptor and the target was 700 ms—longer than most time constants associated with peripheral auditory-nerve adaptation (for a review, see Willmore & King, 2023). However, the adaptation effects were somewhat weaker when the adaptors and the targets were presented to different ears, an effect most apparent in the fixed-ear condition, suggesting some peripheral or spatial-filtering contributions to the effect.

The adaptation effects were larger when the trials were grouped by adaptor types (i.e., voice or instrument) than when randomly interleaved, suggesting that the effects can accumulate over time when consecutive trials share the same type of adaptor. When the trials were grouped by adaptor types, the effects were comparable whether the adaptors were presented to a fixed ear or randomly switched between the ears, suggesting that weakened effects in the "all random" condition were not due to the increased variability in the stimuli presentation ear. The homogeneity of adaptor types, rather than the homogeneity of the presentation ear between consecutive trials, seems crucial for the adaptation effects to accumulate across time.

Experiment 4

The results of Experiment 3 suggest that the adaptation effects between voice and instrumental sounds may accumulate over time when the trials with the same type of adaptors are grouped together. In Experiment 4, we investigated the time course of this accumulation—the rate at which it built up, as well as its persistence following the end of the adaptors. The buildup was measured by parametrically varying the length of blocks during which the type of adaptor remained fixed. To measure persistence, we followed the paradigm employed in a previous study by Schweinberger et al. (2008), in which voice gender adaptation effects were observed in a postadaptation block, where the adaptors were no longer present. We examined whether similar postadaptation effects exist following adaptation tasks between voice and instrumental sounds and whether the postadaptation effects were dependent on the accumulation of the adaptation effects across time.

Method

Participants

Sixty participants took part in this in-person experiment. Their ages ranged from 18 to 25 years (M = 20.0, SD = 1.6). Twenty identified as men, 39 as women, and one chose not to disclose gender. They identified their race/ethnicity as follows: 19 Asian, six Black, two multiracial, 30 White, three undisclosed, three Hispanic or Latinx, 55 non-Hispanic or Latinx, two undisclosed. The participants had an average experience of 4.1 years (range = 0–16) playing musical instruments and/or receiving musical training. One participant was excluded because their PSE was not measurable, leaving a total of 59 analyzed data sets. All participants self-reported having normal hearing and were screened via pure-tone audiometry to ensure audiometric thresholds no greater than 20-dB hearing level at octave frequencies between 250 and 8,000 Hz. Participants were recruited through introductory psychology courses. The study was approved by the University of Minnesota's institutional review board. All participants completed a consent form prior to the study and were awarded a digital gift card or extra course credit upon completion.

Stimuli

This experiment consisted of interleaved adaptation blocks and postadaptation blocks. In the adaptation blocks, the adaptors and targets were the same as in Experiment 3, but the trials were grouped differently, as described below. In the postadaptation blocks, the same set of targets as in Experiment 1 were presented randomly to either the left or right ear, without adaptors. Each trial in the postadaptation blocks started with a 1,000-ms fixation point, followed by a 600-ms prompt screen and then a target. The next trial started immediately after the response had been recorded. If the participant failed to respond within 2.8 s, a prompt "please respond sooner" would be displayed for 500 ms before the next trial commenced. All other details of the stimuli and presentation methods were the same as in Experiment 3.

Design and Procedure

Similar to Experiment 3, this experiment adopted a mixed design, with ears (i.e., same or different) and adaptor type (voice or instrument) as within-subject factors and homogeneity level as a between-subjects factor. Homogeneity level was defined as the number of consecutive trials that were guaranteed to share the same type of adaptors (voice or instrument) and were presented to fixed ears (e.g., the adaptors were always presented to the left ear, and the targets were always presented to the right ear). The three levels of homogeneity adopted in this experiment were 60, 30, and 1. Each of the conditions consisted of eight blocks, each block having 60 trials. Across the blocks, each possible combination of adaptor type and ears had the same number of trials. The "60" condition was largely the same as the "fixed ear" condition in Experiment 3, except that three trials were randomly selected and omitted from each block to make the total number of trials in each block more factorizable. Similarly, the "1" condition was equivalent to the "all random" condition in Experiment 3, where the adaptor type and presentation ear randomly varied on a trial-by-trial basis. The "30" condition was an intermediate condition where the trials within each half of a block shared the same adaptor type and were presented to fixed ears, but both adaptor type and ear varied randomly across different halves.

Each block of the adaptation task was followed by a short block (~2 min) of 63 postadaptation trials, where only the targets were presented. Between each adaptation block and its succeeding postadaptation block, a brief description of the task was displayed on the screen. Participants were instructed not to take a break between adaptation and postadaptation blocks and to press space to start the postadaptation block when they were ready. The entire experiment took each participant about 90 min to complete, including breaks.

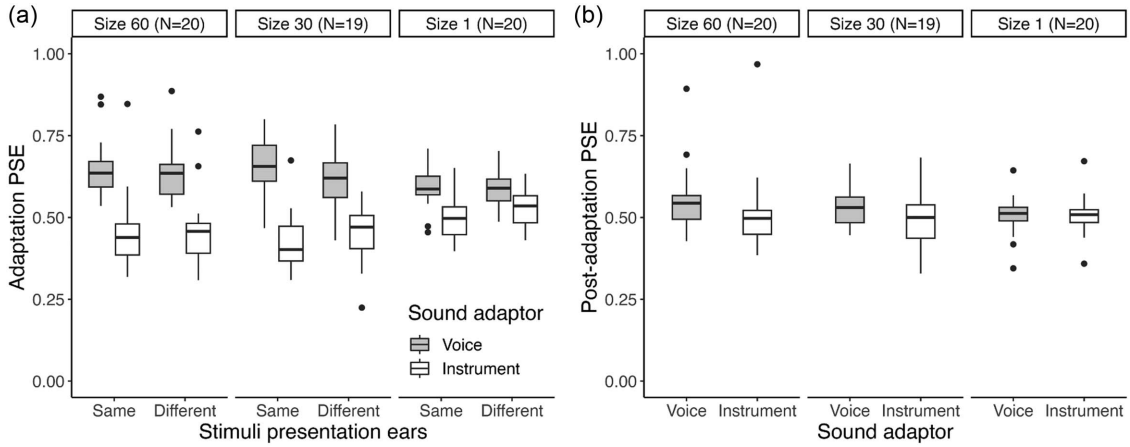
Results

All missed trials (0.7% of all trials) were removed prior to data analysis. A repeated-measures ANOVA was performed in R v4.2.2 on the PSE of targets in the adaptation blocks, with within-subjects factors of ears (same or different) and adaptor type (voice or instrument) and a between-subjects factor of homogeneity level (60, 30, or 1; Figure 9a). The results of Experiment 3 were largely replicated. There was a significant main effect adaptor type, $F(1,56) = 576.99$, $p < .0001$, $\eta_p^2 = 0.91$, 90% CI [0.88, 0.93], where the PSE was higher (implying a greater number of "instrument" responses) following voice adaptors compared with instrument adaptors. There was a significant interaction between adaptor type and ears, $F(1, 56) = 13.42$, $p = .0006$, $\eta_p^2 = 0.19$, 90% CI [0.06, 0.34]. Post hoc analyses with Bonferroni correction (p values adjusted for two comparisons) revealed significant adaptation effects in both the same ear and the different ear conditions ($p < .0001$ in both cases), but the effect was somewhat larger in the same ear condition than in the different ear conditions. There was also a significant interaction between adaptor type and homogeneity level, $F(2, 56) = 36.46$, $p < .0001$, $\eta_p^2 = 0.57$, 90% CI [0.42, 0.67]. Post hoc analyses with Bonferroni correction (p values adjusted for three comparisons) revealed significant adaptation effects in all levels of homogeneity ($p < .0001$ in all cases); the effect was the smallest in the "1" condition, while the "60" and "30" conditions were similar. No other significant main effects or interactions were observed ($p > .10$ in all cases).

A repeated-measures ANOVA was also performed in R v4.2.2 on the PSE of targets in the postadaptation blocks with within-subjects factors of adaptor type (voice or instrument) and between-subjects factor of homogeneity level (60, 30, or 1; Figure 9b). For the 30 and 1 levels of homogeneity, adaptor type was defined by the last adaptor within the block. The variable "ears" was not included in this analysis because the targets were randomly presented to the left and right ears regardless of adaptor type and morph level, and therefore the data may not be sufficient to get a reliable estimate of PSE for every possible combination of adaptor type and ears. There was a significant main effect adaptor type, $F(1,56) = 20.76$, $p < .0001$, $\eta_p^2 = 0.27$, 90% CI [0.12, 0.42], where the PSE was higher following voice adaptors compared with instrumental adaptors, suggesting that the adaptation effects can generally extend to the postadaptation blocks. There was also a significant interaction between adaptor type and homogeneity level, $F(2, 56) = 5.41$, $p = .0071$, $\eta_p^2 = 0.16$, 90% CI [0.03, 0.30]. Post hoc analysis with Bonferroni correction (p values adjusted for three comparisons) showed that the PSE of postadaptation targets was significantly

Figure 9

The Effects of Adaptor Type, Stimuli Presentation, and Adaptor Homogeneity Level on Point of Subjective Equality



Note. (a) Adaptation blocks; (b) postadaptation blocks. Larger PSE corresponds to a greater tendency to perceive ambiguous sounds as musical instruments. PSE = point of subjective equality.

higher for voice than for instrumental adaptors only under the "60" and "30" conditions (p = .0002 in both cases), but not the "1" condition (p = .965), suggesting that the postadaptation effects only existed when the adaptation effects were able to accumulate across trials.

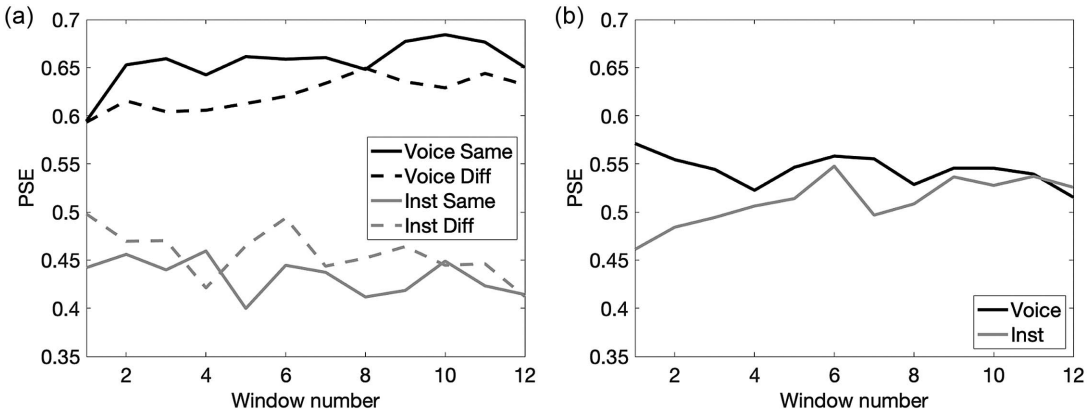
To further explore how adaptation effects build up across trials within a block at a finer time scale, we pooled the data (N = 40) from the "fixed ear" condition of Experiment 3 and the adaptation blocks in Experiment 4 "60" condition, where all trials in each block share the same adaptor type and presentation ears. For each combination of adaptor and ear conditions, PSE was calculated from every five consecutive trials pooled across all participants. As shown in Figure 10a, the difference between the voice and instrument adaptor conditions gradually increased across trials, especially in the first half of the block. Similarly, to explore the time course of postadaptation effects, we pooled the data (N = 39) from the postadaptation blocks in Experiment 4 "60" and "30" conditions. For each adaptor condition, PSE was calculated from every five consecutive trials pooled across all participants. As shown in Figure 10b, the difference between the voice and instrument adaptor conditions quickly decreased in the first 15–20 trials, corresponding to about 40–50 s.

Discussion

The results of Experiment 4 replicated those of Experiment 3 by showing larger adaptation effects when trials were grouped together based on adaptor types, suggesting that adaptation

Figure 10

Trajectories of Point of Subjective Equality Across Trials Within One Block, Estimated From Data Pooled Across Participants



Note. (a) Adaptation blocks; (b) postadaptation blocks. Data were plotted by adaptor type (voice vs. instrument) and presentation ears (same vs. different). Larger PSE corresponds to a greater tendency to perceive ambiguous sounds as musical instruments. Each window corresponds to five consecutive trials. PSE = point of subjective equality; Diff = different ears; Inst = instrument.

effects can accumulate across time. This result is consistent with previous observations on adaptive perception of auditory brightness (Siedenburg et al., 2021). Moreover, the adaptation effects were similar whether the trials in the entire block or just each half of the block were grouped by adaptor types and presentation ears, suggesting that the accumulation may saturate within 30 consecutive trials or less.

Contrastive shifts in PSEs were also observed in the postadaptation blocks, where the adaptors were no longer present, following the adaptation blocks where the adaptors were more homogeneous. Such effects decreased dramatically over the first 15–20 trials of the postadaptation blocks, suggesting limits to the persistence over time. When the adaptor types varied randomly on a trial-by-trial basis, the influence of the final adaptor did not extend into the postadaptation block, suggesting that some accumulation is required for the adaptation effects to persist.

It is also worth noting that the interstimulus interval (ISI) between adaptor and target has been shown to affect the degree of auditory context effects (Feng & Oxenham, 2015; Holt & Lotto, 2002). Although the ISI of the present study (700 ms) is comparable to previous studies on voice gender (Schweinberger et al., 2008) and speaker identity (Zäske et al., 2010) adaptation, early studies suggested that some auditory context effects become negligible at shorter ISIs (e.g., ~400 ms in Holt & Lotto, 2002). It is possible that the adaptation effects between higher level perceptual categories are less dependent on temporal proximity of stimuli than lower level spectral contrast effects, regardless of the duration of the adaptors. Alternatively, because the adaptors were repeated multiple times over the course of several seconds in each trial in the present study, as well as in previous studies on voice gender and speaker identity, it could be that the robustness of the effects under a longer ISI is attributable to the accumulation of the effects within each trial.

General Discussion

In four experiments, we investigated the contrastive adaptation effects between voice and musical instrumental sounds, focusing on the potentially ear- and modality-specific nature of the adaptation, as well as its time course. We discovered a contrastive shift in the PSE following repetitive exposure to voice and instrumental sound adaptors (Experiment 1), but not corresponding image adaptors (Experiment 2), suggesting the existence of modality-specific adaptation effects between voice and instrumental sounds. The effects were observed when the adaptors and the targets were played to either the same or different ears but were somewhat larger in the same-ear conditions (Experiment 3), suggesting that peripheral (ear-specific) or at least spatially selective mechanisms play a role. When consecutive trials share the same type of adaptor (i.e., voice or instrument), the adaptation effects were larger than when different types of adaptors were randomly interleaved (Experiments 3 and 4). Adaptation effects were observed in a separate 2-min postadaptation block where the adaptors were no longer present (Experiment 4), suggesting that the adaptation effects can accumulate and persist over substantial time periods.

Adaptation Is Ubiquitous in Auditory Perception

Over the past decades, adaptation effects have been observed in the perception of linguistic and nonlinguistic features of voice as well as nonvoice stimuli. Early studies focused on how acoustic features of sentences affect the perception of subsequent phoneme. For example, by manipulating the spectral features of the sentences, the perception induced by the same vowel utterance was shifted contrastively between /ɪ/ and /ɛ/ (Broadbent & Ladefoged, 1960; Watkins, 1991). Similar effects were observed when consonants were used as the context for vowels in "consonant–vowel–consonant" syllables and when these consonants were replaced with sine-wave tones that shared acoustic similarities with the consonants (Holt et al., 2000). Even when the target syllables were synthesized with sine waves representing only the first three formants, the frequency trajectory of preceding context was able to alter the perception of phonemes (/b/ vs. /d/) in a contrastive manner (Wang & Oxenham, 2014). Adaptation effects in speech perception have also been observed in cochlear-implant users (Feng & Oxenham, 2018; Winn et al., 2013) and to a lesser extent when vocoded speech was tested on listeners with normal hearing (Feng & Oxenham, 2020; Stilp, 2017).

It has been debated whether these effects are specific to linguistic features, but more recent adaptation studies on paralinguistic vocal features and nonvoice stimuli suggest that adaptation is likely a more general auditory mechanism. When categorizing voices along a continuum between genders (Schweinberger et al., 2008), speakers (Zäske et al., 2010), and affects (Bestelmeyer et al., 2010), repetitive presentation of one endpoint stimuli caused the categorization of the succeeding target to shift toward the opposite end of the continuum. Adaptation effects have also been observed in the perception of nonvoice sounds: Both string sounds filtered to emphasize the spectral features of horn or saxophone sounds (Frazier et al., 2019) and unfiltered horn or saxophone sounds (Lanning & Stilp, 2020) induced contrastive adaptation effects when average listeners were categorizing musical instrument sounds into horn and saxophone.

Although previous auditory adaptation studies covered a wide range of stimuli, to the best of our knowledge, the present study is the first to show that adaptation effects can occur between voices and instrument sounds. This finding provides further support for the ubiquity of auditory adaptation effects: Adaptation can occur when perceiving multiple features of meaningful harmonic sounds, both within the same sound source (e.g., different vowels pronounced by the same person) and between sound sources that are similar (e.g., two different people) or different (e.g., a person and a musical instrument) in nature. It seems likely, therefore, that auditory adaptation is a domain-general mechanism that assists in categorization, thereby enhancing the efficiency and effectiveness of information extraction, rather than a specialized mechanism for familiar stimuli such as the human voice.

No Evidence for Audiovisual Cross-Modality Adaptation in Voice–Instrument Contrasts

In the present study, adaptation effects between voice and musical instrument sounds were observed when sounds, but not images, were used as adaptors. This lack of audiovisual cross-modality adaptation effects is in line with previous research on voice gender (Schweinberger et al., 2008), but not speaker identity (Zäske et al., 2010). Schweinberger et al. (2008) did not observe voice gender adaptation effects using either image or muted video adaptors. In contrast, Zäske et al. (2010) observed speaker identity adaptation effects using muted video adapters, although smaller than when

sound adaptors were used. The lack of audiovisual cross-modality adaptation in our Experiment 2 could be due to the static nature of image adaptors: Both muted videos and sounds change and fluctuate across time, making images less of a visual counterpart than muted videos.

However, the lack of motion in images could not fully explain its inability to induce audiovisual cross-modality adaptation for all sound features because neither images nor muted videos of speakers pronouncing syllables were able to elicit voice gender adaptation effects. Another possible explanation is that the specificity of auditory imagery induced by visual stimuli differs across different sound features. While the visual presentation of a specific familiar person can give rise to the imagery of a specific speaker's vocal characteristics, the auditory imagery following the visual presentation of an unfamiliar person of a specific gender or a syllable annotation is likely much more general, making it more difficult to form a concrete auditory imagery comparable to the actual sound adaptors. Although it is possible that the visual presentation of musical instruments can generate the auditory imagery of specific sound characteristics, this would require a high level of familiarity with the timbres of the musical instruments used in this study. Future research could test the role of auditory imagery in cross-modal adaptation by comparing musicians' and nonmusicians' performance in audiovisual adaptation experiments where muted videos of musicians playing instruments are used as adaptors.

In addition, even if silent visual adaptors cannot induce adaptation effects in some sound features, audiovisual integration may still be able to moderate auditory adaptation effects. Previous audiovisual integration studies mostly investigated concurrent sounds and visual stimuli, and in some cases mismatch between modalities has been found to robustly bias the perceptual categorization of the auditory stimuli toward that of the visual stimuli (e.g., McGurk & MacDonald, 1976). Future studies could investigate whether auditory adaptation effects are different when adaptors are presented alone or with concurrent visual stimuli that can bias the perception of the adaptors.

Adaptation Effects Can Accumulate and Persist Across Time

We observed adaptation effects in separate postadaptation blocks, where the adaptors were no longer present, following the initial presentation of adaptors in adaptation blocks. This result is consistent with the previously observed postadaptation effects in voice gender adaptation experiments (Schweinberger et al., 2008). Taken together with previous research on long-term (up to 1 week) adaptation effects in face perception (e.g., Carbon & Ditye, 2011), it is possible that persistence across time is a common characteristic of adaptation effects, regardless of the specific modality or feature being tested. In addition, the current results suggest that the persistence of adaptation effects across time probably relies on their accumulation across time, which is in turn dependent on the homogeneity of adaptors among consecutive trials, as shown in Experiments 3 and 4. When more consecutive trials share the same type of adaptors, the adaptation effects are larger and can persist longer. This accumulation effect is likely to have a saturation state, where additional presentation of homogenous adaptors does not further increase the degree of the accumulated adaptation effects, because the adaptation effects were similar when the full block or only each half of the block shares the same type of adaptors in Experiment 4.

The accumulation of the adaptation effects was comparable whether the adaptors were presented to a fixed ear in an entire block or whether it changed randomly on a trial-by-trial basis in Experiment 3, as long as all trials in the block shared the same type of adaptors. Therefore, the accumulation effect seemed to be determined by the consistency in the type of the adaptors, regardless of the specific ears where the adaptors were presented, which suggests that the accumulation effect relies primarily on central mechanisms. It is still possible that the accumulation effect can reach the saturation state sooner when the adaptors are presented to a fixed ear than when randomly interleaved, which would suggest that peripheral mechanisms also contribute to the accumulation of adaptation effects.

Addressing a Common Challenge in Voice Perception Research

A central question of voice perception research is whether the voice is treated by the brain as a special type of stimulus and, if so, how the processing of voice is different from that of nonvoice auditory stimuli. To answer this question, some early studies compared brain activation patterns while listening (usually passively) to human voice and nonvoice stimuli and identified voice-sensitive brain regions (Belin et al., 2000, 2002) and responses (Charest et al., 2009; Levy et al., 2001). More recently, several studies have compared performance in psychoacoustic tasks when voice and nonvoice stimuli are used. Voice advantage effects have been observed in timbre recognition (Agus et al., 2012; Isnard et al., 2019) and melody memorization (Weiss et al., 2015); in contrast, voice disadvantage effects have been observed in sound features presumably less relevant to speech information processing, including intonation judgment (Hutchins et al., 2012), note naming (Gao & Oxenham, 2022; Vanzella & Schellenberg, 2010; Weiss et al., 2015), and fine-tuned pitch discrimination (Gao & Oxenham, 2022; Hutchins et al., 2012).

A common challenge faced by those voice perception studies is the appropriate selection of nonvoice control stimuli. Although a wide range of nonvoice stimuli have been tested, they are different from human voice in both acoustical properties and perceptual categorization, making it difficult to determine whether the voice-selective neural responses and behavioral effects are a result of low-level acoustic features of voice or perceptual categorization of a sound being voice. In an attempt to disentangle these two aspects, Agus et al. (2017) created "auditory chimeras" by combining the spectral content of voices with the temporal envelope of musical instrument tones or vice versa. Although some chimeras were more likely to be categorized as voice than nonvoice in a behavioral task, the functional magnetic resonance imaging results showed that only the original version of voice was able to induce voice-selective brain responses. These results suggest that voice-selective brain activation pattern is not a prerequisite of inducing a subjective percept of voice and that voice-selective neural responses may be more dependent on the acoustic feature of voice rather than perceptual categorization of voice. However, it is still unclear whether acoustic properties and perceptual categorization both play a role in voice-specific effects in behavioral tasks.

The current results on adaptation effects between voice and music instrument sounds could potentially be used to manipulate the perception of a sound being voice or nonvoice while keeping the low-level acoustic properties unchanged. For example, participants could be instructed to perform a pitch discrimination task between two sounds selected from a voice–nonvoice continuum, following prolonged exposure to either voice or nonvoice adaptors. In this way, adaptation effects between voice and nonvoice could be used as a tool to dissociate the roles of acoustic properties and perceptual categorization and therefore address a common challenge in previous voice perception research.

Limitations and Future Directions

The present study investigated adaptation effects along the voice–instrument continuum, the loci and time course of the effects, yet the acoustic basis of the adaptation is still unknown. Voice and musical instrument tones are acoustically different in multiple ways. While the mean F0 of the voice and instrument stimuli were adjusted to the same in the present study, the pitch contours were unmodified to keep the sounds naturalistic. Natural human voices were shown to have larger F0 variability than musical instrument sounds (Hutchins et al., 2012), and adaptation to the pitch contour of auditory context has been shown in both voice (Huang & Holt, 2011) and pure-tone-based pitch glides (Alais et al., 2015). Therefore, it is possible that differences in pitch contour contribute to the adaptation along the voice–instrument continuum. Alternatively, as vowels usually have more salient formants than nonvocal harmonic stimuli, spectral modulation could be another acoustic factor that drives the adaptation between voice and nonvoice. Future studies could use synthesized sounds as adaptors so that the contribution of multiple acoustic features can be tested separately in a more controlled manner.

Another limitation of the present study is that the selection of nonvoice was limited to musical instrument sounds, a familiar and ecologically valid type of stimuli for most listeners. We kept the speakers, vowels, and instruments of the adaptor and the target different in each trial to ensure the generalizability across voices and instruments. However, the generalizability to unfamiliar nonvoice sounds and the role of familiarity on the adaptation effects along the voice/instrument continuum remain unknown. Existing literature has produced mixed evidence for the effect of familiarity on context effects: While prior experience with accented speech facilitated audiovisual priming effects on the categorization of word versus nonword (Witteman et al., 2013), familiarity with speakers' voice failed to consistently augment speaker adaptation (Hatter et al., 2024). Future studies could include unfamiliar vocal and nonvocal stimuli, such as foreign vowels and synthesized complex tones, as well as recruit participants from a wide range of listening backgrounds, to understand the relationship between long-term perceptual experience and adaptation.

Finally, our attempt to dissociate peripheral from central mechanisms was limited to manipulating the presentation ear. To dissociate peripheral influences from spatial selectivity, other manipulations, such as using interaural time differences to induce perceptual spatial separation while maintaining the stimulation of both ears, would be necessary (e.g., Feng & Oxenham, 2018; Watkins, 1991). Similarly, we only tested one gap duration between the end of the last adaptor and the beginning of the target (700 ms). Testing different gaps may reveal an interaction between rapidly decaying and slower effects that may be due to more peripheral and central mechanisms, respectively.

Constraints on Generality

Our findings provide evidence of adaptation effects between human voice and musical instrument sounds in participants who are young adults and have normal hearing. The participants were recruited from students who were taking introductory psychology courses, without any requirements regarding native language or musical experience. Because the tasks did not require language or music skills beyond categorizing sounds into voice and instruments, we expect the results to generalize to young adults with normal hearing regardless of their level of education, native language, or musical experience. The results may, however, have limited generalizability to listeners who are hearing impaired or from a significantly older age group.

The stimuli used in this study were created from recorded vowels pronounced by female speakers and single tones played on musical instruments. Three different vowels and three instruments from different families (i.e., woodwind, brass, and string instruments) were selected to enhance the generalizability of the results. Although we did not test male voices in the present study, we expect the results to generalize to male voices because voice adaptation effects have been observed in previous studies where male voices were used as testing stimuli (Zäske et al., 2010). However, we do not have evidence that the findings will generalize to nonvoice sounds that are not musical instrument tones because it is difficult to find another category of nonvoice stimuli that has a high ecological validity comparable with musical instrument sounds. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

References

Agus, T. R., Paquette, S., Suied, C., Pressnitzer, D., & Belin, P. (2017). Voice selectivity in the temporal voice area despite matched low-level acoustic cues. *Scientific Reports*, *7*(1), Article 11526. https://doi.org/10.1038/s41598-017-11684-1

Agus, T. R., Suied, C., Thorpe, S. J., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *The Journal of the Acoustical Society of America*, *131*(5), 4124–4133. https://doi.org/10.1121/1.3701865

Alais, D., Orchard-Mills, E., & Van der Burg, E. (2015). Auditory frequency perception adapts rapidly to the immediate past. *Attention, Perception & Psychophysics*, *77*(3), 896–906. https://doi.org/10.3758/s13414-014-0812-2

Armington, J. C., & Biersdorf, W. R. (1958). Long-term light adaptation of the human electroretinogram. *Journal of Comparative and Physiological Psychology*, *51*(1), 1–5. https://doi.org/10.1037/h0044572

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research*, *13*(1), 17–26. https://doi.org/10.1016/S0926-6410(01)00084-2

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312. https://doi.org/10.1038/35002078

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., & Weenink, D. (2021). *Praat: Doing phonetics by computer (6.1.40)* [Computer software]. https://www.fon.hum.uva.nl/praat/

Broadbent, D. E., & Ladefoged, P. (1960). Vowel judgements and adaptation level. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, *151*(944), 384–399. https://doi.org/10.1098/rspb.1960.0005

Byrne, A. J., Stellmack, M. A., & Viemeister, N. F. (2013). The salience of enhanced components within inharmonic complexes. *The Journal of the Acoustical Society of America*, *134*(4), 2631–2634. https://doi.org/10.1121/1.4820897

Canévet, G., Scharf, B., & Botte, M.-C. (1985). Simple and induced loudness adaptation. *Audiology*, *24*(6), 430–436. https://doi.org/10.3109/00206098509078362

Carbon, C.-C., & Ditye, T. (2011). Sustained effects of adaptation on the perception of familiar faces. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 615–625. https://doi.org/10.1037/a0019949

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J.-P., & Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, *10*(1), Article 127. https://doi.org/10.1186/1471-2202-10-127

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Feng, L., & Oxenham, A. J. (2015). New perspectives on the measurement and time course of auditory enhancement. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1696–1708. https://doi.org/10.1037/xhp0000115

Feng, L., & Oxenham, A. J. (2018). Spectral contrast effects produced by competing speech contexts. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(9), 1447–1457. https://doi.org/10.1037/xhp0000546

Feng, L., & Oxenham, A. J. (2020). Spectral contrast effects reveal different acoustic cues for vowel recognition in cochlear-implant users. *Ear and Hearing*, *41*(4), 990–997. https://doi.org/10.1097/AUD.0000000000000820

Frazier, J. M., Assgari, A. A., & Stilp, C. E. (2019). Musical instrument categorization is highly sensitive to spectral properties of earlier sounds. *Attention, Perception & Psychophysics*, *81*(4), 1119–1126. https://doi.org/10.3758/s13414-019-01675-x

Gao, Z., & Oxenham, A. J. (2022). Voice disadvantage effects in absolute and relative pitch judgments. *The Journal of the Acoustical Society of America*, *151*(4), 2414–2428. https://doi.org/10.1121/10.0010123

Gao, Z., & Oxenham, A. J. (2023, October 11). *Contrastive adaptation effects along a voice/non-voice continuum*. PsyArXiv. https://osf.io/zjq4s/

Gibson, J. J., & Radner, M. (1937). Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. *Journal of Experimental Psychology*, *20*(5), 453–467. https://doi.org/10.1037/h0059826

Hatter, E. R., King, C. J., Shorey, A. E., & Stilp, C. E. (2024). Clearly, fame isn't everything: Talker familiarity does not augment talker adaptation. *Attention, Perception & Psychophysics*, *86*(3), 962–975. https://doi.org/10.3758/s13414-022-02615-y

Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research*, *167*(1–2), 156–169. https://doi.org/10.1016/S0378-5955(02)00383-0

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *The Journal of the Acoustical Society of America*, *108*(2), 710–722. https://doi.org/10.1121/1.429604

Huang, J., & Holt, L. L. (2011). Evidence for the central origin of lexical tone normalization (L). *The Journal of the Acoustical Society of America*, *129*(3), 1145–1148. https://doi.org/10.1121/1.3543994

Hutchins, S., Roquet, C., & Peretz, I. (2012). The vocal generosity effect: How bad can your singing be? *Music Perception*, *30*(2), 147–159. https://doi.org/10.1525/mp.2012.30.2.147

Isnard, V., Chastres, V., Viaud-Delmon, I., & Suied, C. (2019). The time course of auditory recognition measured with rapid sequences of short natural sounds. *Scientific Reports*, *9*(1), Article 8005. https://doi.org/10.1038/s41598-019-43126-5

Kawahara, H., & Matsui, H. (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *2003 IEEE international conference on acoustics, speech, and signal processing, 2003. Proceedings (ICASSP '03), 1* (pp. I-256–I-259). https://doi.org/10.1109/ICASSP.2003.1198766

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 *2008 IEEE international conference on acoustics, speech and signal processing* (pp. 3933–3936). https://doi.org/10.1109/ICASSP.2008.4518514

Lanning, J. M., & Stilp, C. (2020). Natural music context biases musical instrument categorization. *Attention, Perception & Psychophysics*, *82*(5), 2209–2214. https://doi.org/10.3758/s13414-020-01980-w

Levy, D. A., Granot, R., & Bentin, S. (2001). Processing specificity for human voice stimuli: Electrophysiological evidence. *Neuroreport*, *12*(12), 2653–2657. https://doi.org/10.1097/00001756-200108280-00013

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *The Journal of the Acoustical Society of America*, *102*(2), 1134–1140. https://doi.org/10.1121/1.419865

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Mueller, R., Utz, S., Carbon, C.-C., & Strobach, T. (2020). Face adaptation and face priming as tools for getting insights into the quality of face space. *Frontiers in Psychology*, *11*, Article 166. https://doi.org/10.3389/fpsyg.2020.00166

Pantle, A. (1974). Motion aftereffect magnitude as a measure of the spatio-temporal response properties of direction-sensitive analyzers. *Vision Research*, *14*(11), 1229–1236. https://doi.org/10.1016/0042-6989(74)90221-1

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

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

R Core Team. (2022). *R: A language and environment for statistical computing (4.2.2)* [Computer software]. https://www.R-project.org/

Schweinberger, S. R., Casper, C., Hauthal, N., Kaufmann, J. M., Kawahara, H., Kloth, N., Robertson, D. M. C., 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

Siedenburg, K., Barg, F. M., & Schepker, H. (2021). Adaptive auditory brightness perception. *Scientific Reports*, *11*(1), Article 21456. https://doi.org/10.1038/s41598-021-00707-7

Stilp, C. E. (2017). Acoustic context alters vowel categorization in perception of noise-vocoded speech. *Journal of the Association for Research in Otolaryngology: JARO*, *18*(3), 465–481. https://doi.org/10.1007/s10162-017-0615-y

Sussman, J. E. (1993). Focused attention during selective adaptation along a place of articulation continuum. *The Journal of the Acoustical Society of America*, *93*(1), 488–498. https://doi.org/10.1121/1.405629

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental*

Psychology A: Human Experimental Psychology, *43*(2), 161–204. https://doi.org/10.1080/14640749108400966

Vanzella, P., & Schellenberg, E. G. (2010). Absolute pitch: Effects of timbre on note-naming ability. *PLOS ONE*, *5*(11), Article e15449. https://doi.org/10.1371/journal.pone.0015449

Wang, N., & Oxenham, A. J. (2014). Spectral motion contrast as a speech context effect. *The Journal of the Acoustical Society of America*, *136*(3), 1237–1245. https://doi.org/10.1121/1.4892771

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *The Journal of the Acoustical Society of America*, *90*(6), 2942–2955. https://doi.org/10.1121/1.401769

Weiss, M. W., Vanzella, P., Schellenberg, E. G., & Trehub, S. E. (2015). Pianists exhibit enhanced memory for vocal melodies but not piano melodies. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *68*(5), 866–877. https://doi.org/10.1080/17470218.2015.1020818

Willmore, B. D. B., & King, A. J. (2023). Adaptation in auditory processing. *Physiological Reviews*, *103*(2), 1025–1058. https://doi.org/10.1152/physrev.00011.2022

Winn, M. B., Rhone, A. E., Chatterjee, M., & Idsardi, W. J. (2013). The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Frontiers in Psychology*, *4*(824), Article 824. https://doi.org/10.3389/fpsyg.2013.00824

Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception & Psychophysics*, *75*(3), 537–556. https://doi.org/10.3758/s13414-012-0404-y

Zäske, R., Fritz, C., & Schweinberger, S. R. (2013). Spatial inattention abolishes voice adaptation. *Attention, Perception & Psychophysics*, *75*(3), 603–613. https://doi.org/10.3758/s13414-012-0420-y

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

Received October 22, 2023
Revision received May 9, 2024
Accepted August 6, 2024 ■