Attention, Perception, & Psychophysics (2023) 85:2488–2501
https://doi.org/10.3758/s13414-023-02722-4

Talker adaptation or "talker" adaptation? Musical instrument variability impedes pitch perception

Anya E. Shorey[1] · Caleb J. King[1] · Rachel M. Theodore[2,3] · Christian E. Stilp[1]

Accepted: 26 April 2023 / Published online: 31 May 2023

Abstract
Listeners show perceptual benefits (faster and/or more accurate responses) when perceiving speech spoken by a single talker versus multiple talkers, known as talker adaptation. While near-exclusively studied in speech and with talkers, some aspects of talker adaptation might reflect domain-general processes. Music, like speech, is a sound class replete with acoustic variation, such as a multitude of pitch and instrument possibilities. Thus, it was hypothesized that perceptual benefits from structure in the acoustic signal (i.e., hearing the same sound source on every trial) are not specific to speech but rather a general auditory response. Forty nonmusician participants completed a simple musical task that mirrored talker adaptation paradigms. Low- or high-pitched notes were presented in single- and mixed-instrument blocks. Reflecting both music research on pitch and timbre interdependence and mirroring traditional "talker" adaptation paradigms, listeners were faster to make their pitch judgments when presented with a single instrument timbre relative to when the timbre was selected from one of four instruments from trial to trial. A second experiment ruled out the possibility that participants were responding faster to the specific instrument chosen as the single-instrument timbre. Consistent with general theoretical approaches to perception, perceptual benefits from signal structure are not limited to speech.

Keywords Talker adaptation · Speech perception · Music perception · Musical instruments · Pitch

Introduction

The acoustic environment is full of variability. The types of sounds that we hear, including music and speech, are highly variable. There is also substantial variability within sound types, like the array of musical instruments playing a variety of notes to create a song. To navigate the world around us, we must manage this constant acoustic variability. Speech is the most canonical example as it is replete with acoustic variability both within and across talkers. For example, one talker's production of a certain sound might overlap with how another talker produces a different speech sound (Hillenbrand et al., 1995; Peterson & Barney, 1952). Listeners must overcome this variability when listening to different talkers, especially in succession.

When hearing the same talker, listeners adapt to that talker's speech. This process is referred to as talker adaptation (sometimes called talker normalization). In talker adaptation paradigms, listeners often make categorization judgments such as vowel quality (e.g., /i/ as in "beet" versus /u/ as in "boot") when the sounds are spoken by either a single talker or one of several talkers presented in random orders. Importantly, listening to speech from a single talker provides perceptual benefits (faster and/or more accurate responses) over listening to speech from multiple talkers. This has been demonstrated in a wide range of speech perception tasks, including word identification (Choi et al., 2018; Mullennix et al., 1989; Stilp & Theodore, 2020), word list recall (Goldinger et al., 1991; Martin et al., 1989), digit list recall (Bressler et al., 2014), vowel monitoring (Barreda, 2012; Magnuson & Nusbaum, 2007), voice classification (Mullennix & Pisoni, 1990), phoneme categorization (Assmann

✉ Anya E. Shorey
anya.shorey@louisville.edu

[1] Department of Psychological and Brain Sciences, University of Louisville, 317 Life Sciences Building, Louisville, KY 40272, USA

[2] Department of Speech, Language, and Hearing Sciences, University of Connecticut, 2 Alethia Drive, Unit 1085, Storrs, CT 06269-1085, USA

[3] Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut, 337 Mansfield Road, Unit 1272, Storrs, CT 06269-1272, USA

et al., 1982; Rand, 1971), lexical tone categorization (Zhang & Chen, 2016), and vowel categorization shaped by acoustic context effects (Assgari & Stilp, 2015).

Several theoretical approaches offer explanations as to why adapting to one talker's speech yields perceptual benefits relative to multiple talkers' speech. The first three approaches discussed here focus on speech-specific explanations. First, the *episodic approach* (Goldinger, 1996, 1998) posits listeners retain lexical examples of every speaker they have ever heard. When a listener encounters a new speaker, they must compare the signal to other stored templates (i.e., "exemplars"). For familiar talkers, the listener matches the newly encountered token to a previously stored speech exemplar; for unfamiliar talkers, the listener matches the new token to either the closest exemplar or a summary of all previously heard exemplars. Hearing the same talker might continuously facilitate those comparisons (single-talker condition) and hearing different talkers (multiple-talker condition) might require switching between different sets of exemplars, thus challenging perception. Second, pursuant to the *active control approach* (Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007), when a new talker is heard, an active control mechanism is engaged to test and update hypotheses in response to that variability. There is then an increased cognitive cost resulting in increased response time and/or decreased accuracy for hearing multiple talkers compared with a single talker stemming from variation in the available alternative interpretations of the input. This active, attentionally guided system is posed as an alternative to a passive, deterministic system that utilizes invariant mappings between input and output. Third, according to the *Bayesian belief updating approach* (Kleinschmidt & Jaeger, 2015), listeners track and update expected distributions of input statistics, which may be updated in a context-dependent manner (e.g., for individual talkers). Here, optimal perception may require inferring the correct distribution for the current talker and/or for the current sound(s), and then updating prior beliefs to integrate the present input. It might be easier for listeners to retrieve, update, and maintain a single distribution for a single talker than it is to retrieve multiple distributions for multiple talkers. While Kleinschmidt and Jaeger (2015) state that the problem of overcoming signal variability is not specific to speech perception, Bayesian belief updating requires extensive perceptual experience to populate these distributional representations of inputs. Listeners hear speech far more than any other sound, potentially limiting its generalizability to different input domains.

Other approaches address talker adaptation in a more domain-general nature. *Efficient coding* (Attneave, 1954) broadly states that perceptual systems have evolved to exploit structure and predictability in the environment. Applied to talker adaptation, the efficient coding approach (Stilp & Theodore, 2020) puts forth that when there is structure in the input (e.g., single-talker cases), perceptual processing is efficient; when there is less structure (e.g., multiple-talker cases), processing is less efficient, as evidenced by longer response times. Finally, the *streaming approach* (Choi & Perrachione, 2019) suggests listeners deploy attention to a coherent stream from a single source (e.g., a single talker speaking) which facilitates perception. When a new talker is heard, this forms a new stream. Switching between multiple streams (as when there are multiple talkers speaking in succession) is resource intensive, which slows perception.

These five accounts make the same prediction for talker adaptation tasks: superior performance in single-talker cases compared with multiple-talker cases. However, they differ considerably in their ability to generalize beyond speech perception. Both streaming (Bregman, 1990) and efficient coding (Attneave, 1954) are well supported in nonspeech (or even nonauditory) domains. Bayesian belief updating addresses a domain-general problem but utilizes an architecture whose generalizability beyond speech (to domains with which listeners have comparatively less experience) is unclear. Conversely, episodic and active control theories are exclusively developed to account for human speech perception (mappings of acoustic cues to phonemes, lexical access, and distinguishing between alternative linguistic utterances, respectively). It remains an open question whether some aspects of talker adaptation may reflect more general auditory principles as suggested by streaming and efficient coding accounts.

Music, like speech, is a common sound in our environment. Western music contains structure in the form of meter, harmony, and repetition, contrasted with variability such as deviant note choices, changing timbres, and rhythmic variations. The need to adapt to consistency of the acoustic input is thus not limited to speech, but a possibility within music as well. In music perception, adapting to consistency in musical sounds improves performance relative to when inconsistency (i.e., variability) is present. Music perception research has long recognized the bidirectional interference of pitch and timbre (Krumhansl & Iverson, 1992; Melara & Marks, 1990; Pitt, 1994). For example, Krumhansl and Iverson (1992, Experiment 1) presented listeners with low- and high-pitched notes in isolation presented with either a trumpet or piano timbre. Participants were required to identify the pitch when timbre was held constant, identify the timbre when pitch was held constant, or to identify one stimulus dimension while the other dimension was also varying. Listeners could not ignore task-irrelevant information such that changes in timbre impaired pitch judgments and changes in pitch impaired timbre judgments (Krumhansl & Iverson, 1992; Melara & Marks, 1990; Pitt, 1994; Van Hedger et al., 2015). Listeners must overcome this acoustic variability to interpret and appreciate the music to which they are listening. Thus, the consistency and variability within music (e.g., varying pitches and instrumental timbres) can be conceived as analogous to the variability within

speech (e.g., varying speech sounds and talkers). The open question is the extent to which the mechanisms that facilitate overcoming variability in both domains are similar.

The present study modified a talker adaptation experimental paradigm using instruments instead of talkers and pitches instead of speech sounds. A low–high pitch judgment task was created to parallel the word choice selection in the speech paradigm. On each trial, participants categorized a sound as low or high in pitch that was played by either a single instrument (analogous to the single talker condition) or one of four instruments (like the multiple talker condition) that varied from trial to trial. Using a musical task accessible to musicians and nonmusicians alike offers a way to test generalizability of mechanisms underlying adaptation. Consistent with previous research on the interdependence between pitch and timbre reviewed above (Krumhansl & Iverson, 1992; Melara & Marks, 1990; Pitt, 1994; Van Hedger et al., 2015), we predicted participants would be faster making their low–high pitch judgments for a single timbre than when timbre is varying. If the predicted results are observed, this would point toward domain-general theoretical approaches to talker adaptation (streaming and efficient coding) rather than speech-specific accounts (Bayesian belief updating, episodic, and active control). Since the two pitches tested (D4 and F#4 in musical terms) were highly discriminable on both acoustic (see Wier et al., 1977) and musical grounds (Zarate et al., 2012, 2013), we also predicted that accuracy would not differ as a function of block (with both blocks at ceiling; accuracy is also typically at ceiling in traditional talker adaptation paradigms, e.g., Choi et al., 2018; Stilp & Theodore, 2020).

Experiment 1

Method

Participants

The final sample included 40 participants (three male, 36 female, one other) who were undergraduates participating in exchange for course credit. They were at least 18 years of age (M = 20.23 years, 95% CI [18.87, 21.59]) and were required to have self-reported healthy hearing. Participants had, on average, 3.53 (95% CI [2.25, 4.80]) years of formal musical training. An additional 30 participants completed the study but were not included in the final sample for failing to meet inclusion criteria as outlined below.

The present study was patterned after a recent talker adaptation study that compared reaction time across a single-talker block and a highly variable multiple-talker block (Stilp & Theodore, 2020). Results from that study were entered into a power simulation using the package simr (Green & Macleod, 2016) in R (R Core Team, 2021), from which a sample size of 40 participants yielded >99% power. Based upon these results, the present sample size should be sufficient to detect adaptation of a similar magnitude across blocks. This study was approved by the Institutional Review Board at the University of Louisville. Participants provided electronic informed consent at the beginning of the study.

Stimuli

Stimuli were musical instrument notes selected from the McGill University Musical Samples Database (Opolko & Wapnick, 1989). Selected instruments were required to have a constant pitch throughout the duration of the note (i.e., no vibrato). Instruments were chosen to span a wide range of timbres. Recordings of plucked violin (practice); alto saxophone (single-instrument block); and clarinet, French horn, marimba, and piano (mixed-instrument block) playing the notes D4 (294 Hz) and F#4 (370 Hz) were selected. To mirror how accuracy performance is often at ceiling in talker adaptation paradigms (e.g., Choi et al., 2018; Stilp & Theodore, 2020), the interval of a major third (four semitones; 400 cents) was selected because it comfortably exceeds the interval discrimination threshold for nonmusicians (≈ 100–125 cents; Zarate et al., 2012, 2013).

The first 1000 ms of each note was extracted at zero crossings from the original instrument recording in Praat (Boersma & Weenick, 2021). The durations of the plucked violin notes were less than 1000 ms; thus, these sounds were used for practice to ensure all instrument notes in the main task had equal 1000 ms duration. In MATLAB, each note was ramped with a 2-ms linear offset and set to a constant root-mean-squared amplitude.

Procedure

Participants completed the experiment online on a personal computer. The entire experiment took approximately 15 minutes to complete. After providing informed consent, participants completed a six-trial headphone screen (Woods et al., 2017). On each trial, listeners heard three tones and reported which was the quietest. This task is designed to be easy while wearing stereo headphones but difficult using external speakers due to destructive interference between tones heard in open air. All participants completed the Woods et al. (2017) headphone screen; 24 out of the 40 final participants passed the screen (defined as at least five out of six trials correct within two attempts). The pattern of pitch categorization accuracy and response time (RT) collapsed across blocks did not differ between participants who passed the screener (accuracy: M = 91.6%, 95% CI [87.7%, 95.5%]; RT: M = 1,018 ms, 95% CI [907, 1,128]) relative to participants who passed all the other checks that were germane

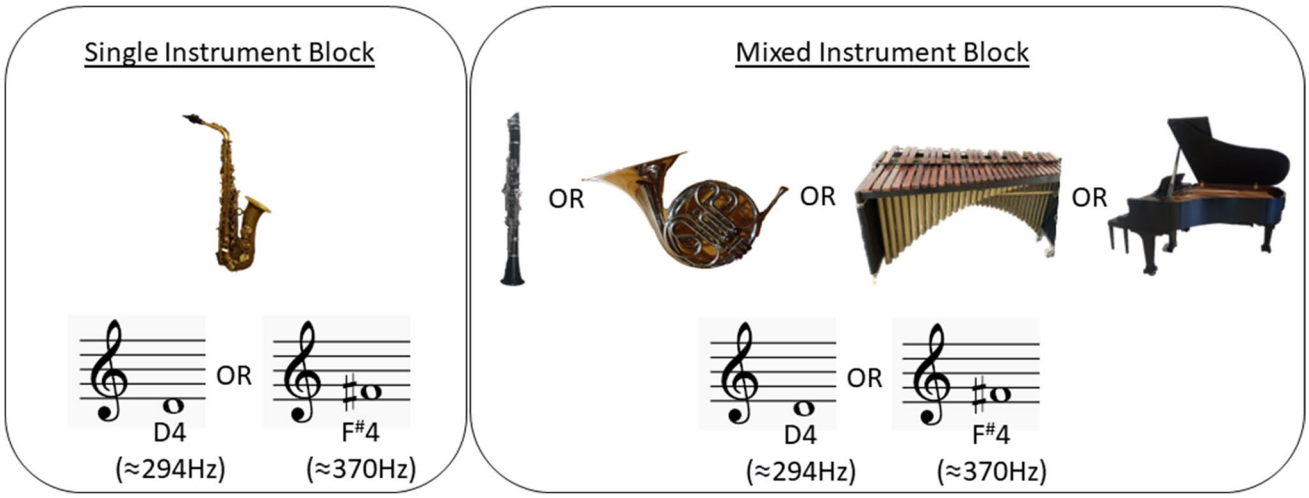

Fig. 1 Diagram of trial structure in single- and mixed-instrument blocks. On each trial, participants heard an instrument play either a low note (D4) or a high note (F#4). In the single-instrument block, all notes were played by an alto saxophone. In the mixed-instrument block, notes were played by either a clarinet, French horn, marimba, or piano as shown from left to right. Participants reported whether they heard the low or high note on each trial

to this task regardless of headphone status (accuracy: M = 89.5%, 95% CI [86.4%, 92.7%]; RT: M = 1,045 ms, 95 % CI [961, 1,130]), so all participants at this step were retained.

Participants then heard the verbal label of "low" followed by the note D4 and the verbal label "high" followed by the note F#4, both played on plucked violin. Participants could listen to the labels and notes up to five times. Next, listeners practiced categorizing the plucked violin sounds as low or high with feedback (10 trials at each pitch height, in random orders). Participants were required to achieve 90% accuracy on practice within three attempts to be included in the final sample; four individuals did not meet this criterion and so were excluded.

Following practice, participants made the same low–high pitch judgments in a single-instrument block and in a multiple-instrument block (Fig. 1). On each trial of the single-instrument block, listeners categorized notes played by an alto saxophone as low or high. Each note was presented 40 times in random orders (2 notes × 40 repetitions = 80 total trials). On each trial of the multiple-instrument block, listeners categorized notes played by one of four instruments (clarinet, French horn, marimba, piano) as low or high. Each note as played by each instrument was presented ten times in random orders (4 instruments × 2 notes × 10 repetitions = 80 total trials). All participants completed both blocks, and the presentation order of the blocks was counterbalanced across participants. To facilitate comparison to results from talker adaptation paradigms, participants were required to maintain 90% accuracy in the single-instrument block to be included in the final sample (as listeners in talker adaptation experiments using this paradigm routinely approach ceiling levels of performance in the single-talker condition, if not all conditions). As a result, an additional 25 were excluded for failing to meet this criterion.[1]

Finally, all participants answered general demographic questions and were asked if they had any musical experience. Participants who responded "yes" to having musical experience answered additional questions such as their number of years of musical instruction, performing experience, and the instrument(s) they play. One participant was removed from the sample for reporting that they had diagnosed hearing loss in this questionnaire, resulting in the final sample of 40 participants.

Results

Accuracy

Mean accuracy was 97.2% (95% CI [96.4%, 98.1%]) for the single-instrument block and 81.8% (95% CI [76.5%, 87.1%]) for the mixed-instrument block, d = .93 (Fig. 2a).

The binary outcome variable, accuracy (1 = correct, 0 = incorrect), was assessed using a generalized linear mixed-effects model using lme4 (Bates et al., 2015) in R (R Core

[1] Since participants for Experiment 1 were recruited without regard for their musical training, an alternative analysis was conducted including the participants who did not meet the 90% accuracy criterion in the single block. Participants were still more accurate and responded faster in the single instrument block (M acc = 77.6%, 95% CI [69.9%, 85.3%]; M RT = 909 ms, 95% CI [833, 985]) than the mixed block (M acc = 71.7%, [67.1%, 76.4%]; M RT = 1,161, 95% CI [1,067, 1,254]). Mixed effects models using the same architecture as described in the main text revealed the differences across the blocks were significant (accuracy model: $\hat{\beta}$ = −1.25, 95% CI [−2.12, −0.42], Z = −3.17; RT model: $\hat{\beta}$ = .28, 95% CI [0.18, 0.37], t = 6.08). Thus, the same pattern of results was observed with or without the single-instrument block performance criterion.

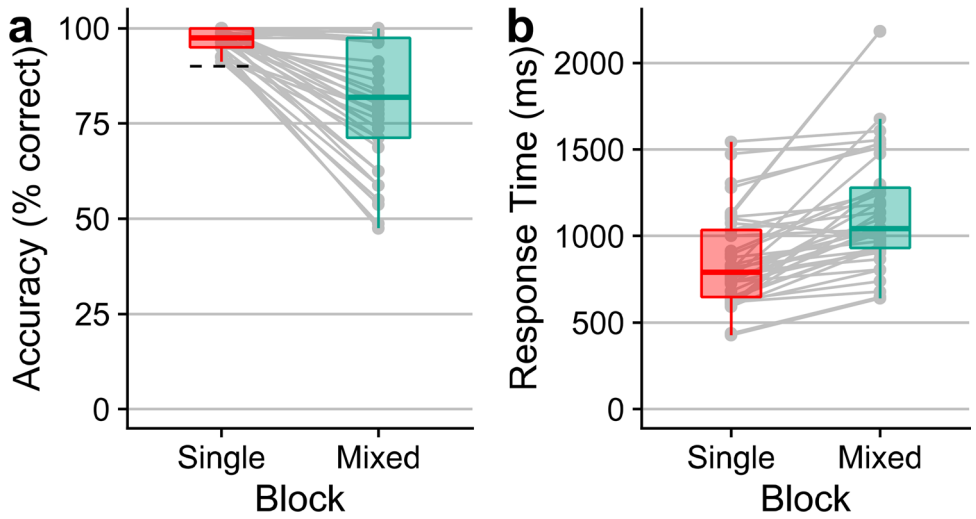

Fig. 2 Individual and aggregated results for pitch categorization. **a** Each participant is represented as an individual dot with lines connecting their accuracy between blocks. The boxplots show the aggregate accuracy for pitch judgments within each block. The dashed line beneath the single-instrument block boxplot represents the 90% minimum correct accuracy criterion to be included in the final sample. **b** Participants are again represented as dots with lines connecting their response times (for correct trials only) between blocks. Boxplots display aggregate response times within each block. (Color figure online)

Team, 2021) with the binomial logit linking function. The model included a fixed effect of block which was contrast-coded (single block as −0.5, mixed block as 0.5). The model included random intercepts by subject and by instrument, and random slopes by subject for block, reflecting the maximal random effects structure (Barr et al., 2013). The model demonstrates that participants were more accurate in the single-instrument block than the mixed-instrument block ($\widehat{\beta} = -1.59$, 95% CI [−2.36, −0.80], $Z = -4.52$). The estimated marginal means for accuracy were 98.0% (95% CI [96.5%, 98.9%]) for the single-instrument block and 90.9% (95% CI [84.2%, 95.0%]) for the mixed-instrument block, as calculated using the emmeans package (Lenth, 2019).

Response time

In talker adaptation studies, it is common practice to analyze response times only for correct trials that are within three standard deviations of a participant's mean response time (e.g., Choi et al., 2018; Stilp & Theodore, 2020). Accordingly, prior to analyses of response time, 670 incorrect trials (10.5% of all trials) were removed, and an additional 92 trials (0.8%) were removed for being outliers relative to a participant's mean response time. Response time was positively skewed (skewness = 2.81), so it was transformed using the natural logarithm prior to analysis (skewness of log-transformed response times = 0.26). For ease of interpretation, untransformed response times are reported here. Mean response time was 852 ms (95% CI [768, 935]) for the single-instrument block and 1,144 ms (95% CI [1030, 1,258]) for the mixed-instrument block, $d = .94$ (Fig. 2b).

Trial-level logarithmically transformed response times were submitted to a linear mixed effects model using lme4 (Bates et al., 2015) in R (R Core Team, 2021). The model included a fixed effect of block, which was sum-coded (single block as −0.5, mixed block as 0.5). The final random effects structure consisted of random intercepts by subject and by instrument, and random slopes by subject for each block, again corresponding to the maximal random effects structure (Barr et al., 2013). Participants were faster in the single-instrument block than the mixed-instrument block ($\widehat{\beta} = 0.29$, 95% CI [0.18, 0.40]), $t(7.56) = 5.43$, $p < .001$. The estimated marginal means for reaction time were 787 ms (95% CI [701, 884]) for the single-instrument block and 1050 ms (95% CI [957, 1152]) for the mixed-instrument block, as estimated by the emmeans package (Lenth, 2019).

Exploratory analyses of musical training

Talker adaptation studies test adults who are native speakers of the language and are thus experts at categorizing the presented speech sounds. This task was created to mirror talker adaptation studies, but participants were largely nonexperts in the present study. Here, individuals were recruited to participate in this experiment irrespective of their musical training, but they did vary in their musical backgrounds (years of formal musical instruction: $M = 3.53$ years, 95% CI [2.25, 4.80]), with the caveat that only 15 of the 40 participants had five or more years of formal musical instruction. Exploratory analyses were conducted to assess potential relationships between musical training and task performance. Unsurprisingly, years of formal music training was positively skewed (skewness = 1.00), so

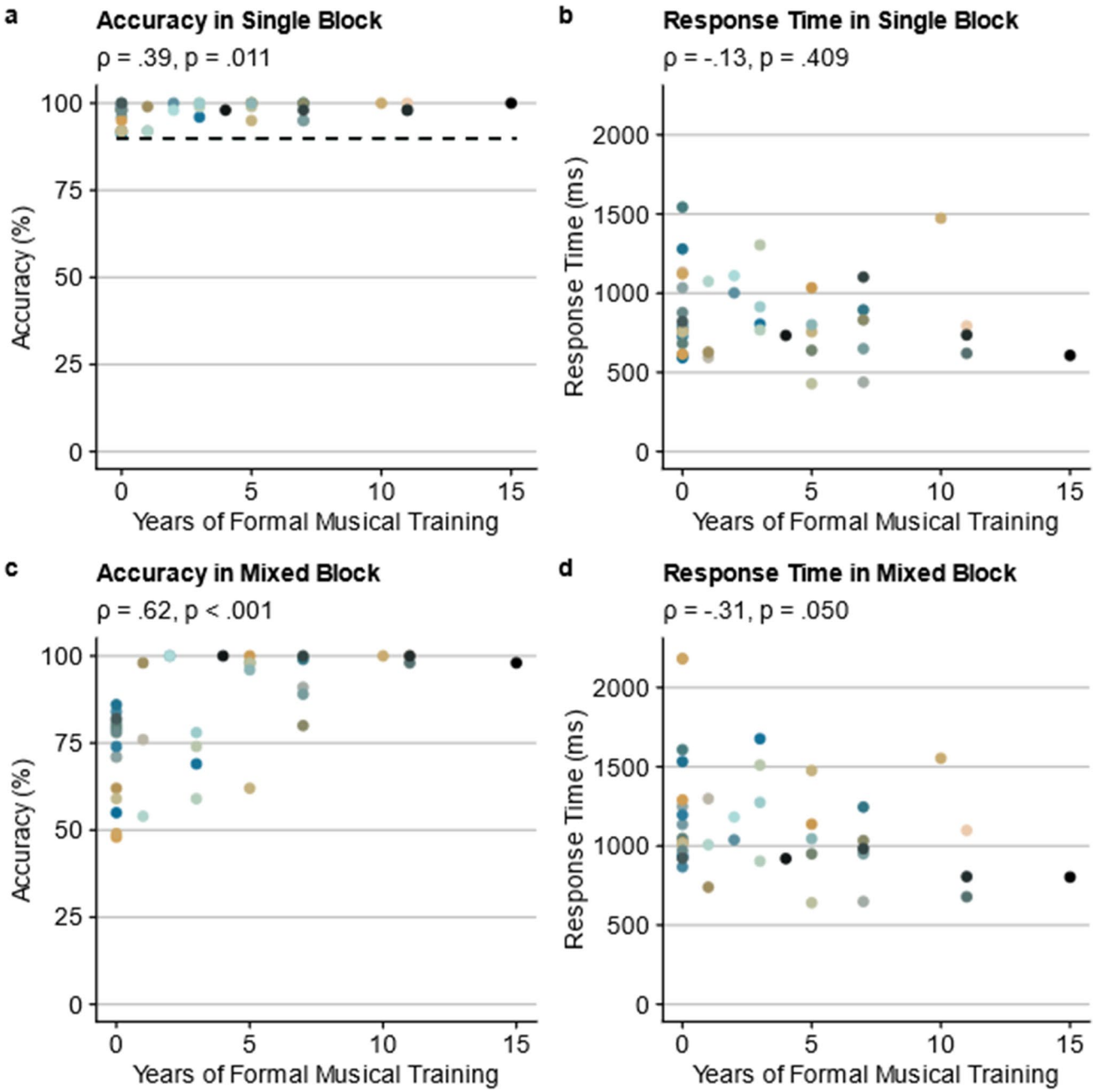

Fig. 3 Scatterplots depicting the relationship between years of formal musical training and task performance. Each color represents a unique participant, with the same participant represented by the same color across panels (color online). The dashed line in panel (**a**) represents the 90% inclusion criterion to be included in the final sample as described in the main text. (Color figure online)

Spearman's rho was used for correlation analyses with aspects of performance in this experiment (Fig. 3). Participants with more years of formal music training exhibited higher accuracy in the single ($\rho = .39$, $p = .011$) and mixed block ($\rho = .62$, $p < .001$), and trended toward having faster response times ($\rho = -.31$, $p = .050$) in the mixed-instrument block. Musical training was not correlated with response time in the single-instrument block ($\rho = -.13$, $p = .409$).

Discussion

Experiment 1 showed that listeners were faster and more accurate to categorize low and high pitches when played by a single instrument (alto saxophone) than when the instrument was changing (random orders of clarinet, French horn, marimba, and piano). These findings parallel results using speech in talker adaptation paradigms, in which listeners are faster and more accurate to categorize speech sounds when spoken by a single

talker than when the talker varies. These parallels promote domain-general approaches to adapting to consistency in the environment, while speech-specific approaches to talker adaptation cannot as parsimoniously accommodate these results.

These results are also reminiscent of the pitch and timbre interdependence studies. For example, Krumhansl and Iverson (1992) found that listeners were faster to make pitch judgments for one instrument than for two. In the present study, additional variability was present in the mixed-instrument block, in which timbre randomly varied from trial-to-trial among four instruments. To examine timbre-specific effects more closely, we used a mixed effects model with fixed effects of instrument and random intercepts by participant to predict response times for the mixed-instrument block only. Responses were fastest to marimba (M = 1,058 ms, 95% CI [1,022, 1,093]) relative to clarinet (M = 1,147 ms, 95% CI [1,105, 1,190]), horn (M = 1,147 ms, 95% CI [1,107, 1,187]), and piano (M = 1,109 ms, 95% CI [1,065, 1,153]; all $\widehat{\beta}$ s > 0.05, ts > 3.18, ps < .001) and were faster to piano than horn ($\widehat{\beta}$ = .04, $t(2{,}510.39) = 2.30$, p = .022). Marimba and piano are both percussion instruments which have relatively sharp attack onsets, so listeners reasonably responded to pitches produced by these timbres the fastest, although the mean response time for the mixed-instrument block was still slower the single-instrument block in which the presented timbre was not percussive.

Experiment 1 was designed by randomly selecting one instrument to serve in the single-instrument condition; participants were not expected to respond preternaturally faster to the alto saxophone than to any other instrument. An alternative explanation for the data in Experiment 1 could be that participants respond faster to saxophone due to some quality of its timbre rather than because it was presented as the single, consistent instrument. To test this possibility, in Experiment 2, each of the five instruments (saxophone plus clarinet, horn, marimba, and piano) was presented in its own single-instrument block where the trial-by-trial instrument consistency was the same as in the alto saxophone single-instrument block of Experiment 1. We hypothesized that participants would not systematically respond faster or more accurately to alto saxophone sounds than to the other instrument sounds.

Experiment 2

Method

Participants

The final sample included 30 participants who were undergraduates participating in exchange for course credit. They were at least 18 years of age and had self-reported healthy hearing. An additional 19 participants completed the study but were not included in the final sample for failing to meet inclusion criteria as outlined below. None participated in Experiment 1. This study was approved by the Institutional Review Board at the University of Louisville. Participants provided electronic informed consent at the beginning of the study.

Stimuli

Stimuli were recordings of plucked violin (practice), alto saxophone, clarinet, French horn, marimba, or piano playing the notes D4 (294 Hz) and F#4 (370 Hz). All details of stimuli were the same as those reported in Experiment 1.

Procedure

Participants completed the experiment online on a personal computer. The entire experiment took approximately 15 minutes to complete. After providing informed consent, participants completed the same six-trial headphone screen (Woods et al., 2017) followed by a practice block of labelling low and high violin notes as detailed in Experiment 1. As before, due to an unexpectedly high rate of failure to meet the headphone screen performance criterion (at least five out of six trials correct within two attempts), exclusion was determined by other performance checks that were specific to this task. Three individuals did not achieve 90% accuracy within three attempts of the practice block and so were excluded from analyses.

Following practice, participants made the same low–high pitch judgments in five blocks each containing only one instrument (alto saxophone, clarinet, French horn, marimba, or piano). On each trial of a block, listeners categorized the note played by an instrument as low or high. Each note was presented 40 times in random orders (2 notes × 40 repetitions = 80 trials per block). All participants completed all five blocks, and the presentation order of the blocks was randomized across participants. Participants were required to average 90% accuracy across all blocks to be included in the final sample; 16 were excluded for failing to meet this criterion. Thus, the final sample included 30 participants.

Results

Accuracy

All trials were included in analyses. Mean accuracy was 95.6% (95% CI [94.3%, 97.0%]) for the alto saxophone, 96.2% (95% CI [94.6%, 97.7%]) for the clarinet, 96.0% (95% CI [94.3%, 97.7%]) for the French horn, 96.8% (95% CI [95.4%, 98.2%]) for the marimba, and 94.5% (95% CI [91.2%, 97.9%]) for the piano (Fig. 4a).

The binary outcome variable, accuracy (1 = correct, 0 = incorrect), was assessed using a generalized linear

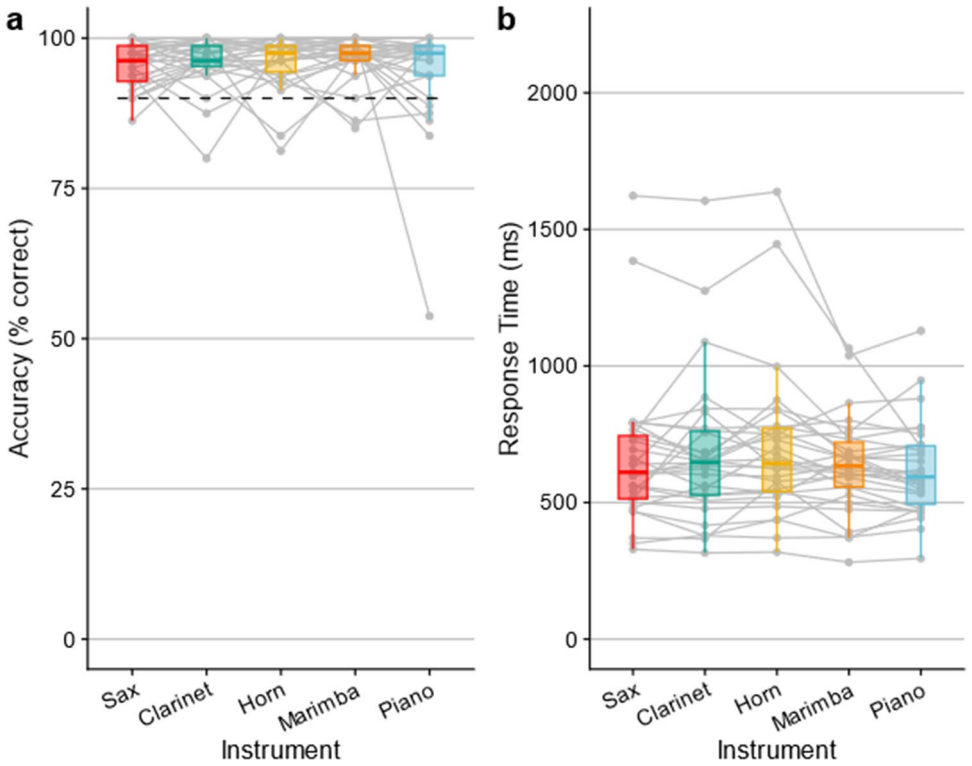

Fig. 4 Accuracy and response time results by instrument. Alto saxophone is displayed as "Sax" and French horn is displayed as "Horn." **a** Each participant is represented as an individual dot with lines connecting their accuracy across blocks. The boxplots show the aggregate accuracy distribution within each block. The dashed line represents the 90% minimum correct accuracy criterion to be included in the final sample. **b** Participants are again represented as dots with lines connecting their response time across blocks. Boxplots display aggregate response time distributions within each block. (Color figure online)

mixed-effects model using lme4 (Bates et al., 2015) in R (R Core Team, 2021) with the binomial logit linking function. The model included a fixed effect of instrument which was treatment-coded with the alto saxophone as the reference level, with random slopes for instrument and random intercepts by subject. No accuracy comparisons across instruments were significant (all $\widehat{\beta}$ s < 0.46, Zs < 1.81, ps $> .071$).

Response time

As in Experiment 1, only correct trials that were within three standard deviations of a participant's mean response time were analyzed. Accordingly, prior to analysis, 498 incorrect trials (4.2% of all trials) were removed, and an additional 185 trials (0.8%) were removed for being outliers relative to a participant's mean response time. The response time variable was positively skewed (skewness = 3.79), so it was transformed using the natural logarithm prior to analysis (skewness of log-transformed response times = 0.44). For ease of interpretation, untransformed response times are reported here. Mean response time was 662 ms (95% CI [563, 762]) for the alto saxophone, 677 ms (95% CI [576, 778]) for the clarinet, 700 ms (95% CI [597, 804]) for the French horn, 637 ms (95% CI [572, 702]) for the marimba, and 621 ms (95% CI [557, 685]) for the piano (Fig. 4b).

Trial-level logarithmically transformed response times were submitted to a linear mixed effects model using lme4 (Bates et al., 2015) in R (R Core Team, 2021). The model included a fixed effect of instrument, which was treatment-coded with the alto saxophone as the reference level. The random effect structure included random slopes for instrument and random intercepts for participant. Compared with the alto saxophone, participants were significantly slower for the French horn ($\widehat{\beta} = 0.06$, 95% CI [0.001, 0.10]), $t(29.03) = 2.38$, $p = .02$. No other response time comparisons were significant.

Discussion

Experiment 2 was conducted to address a potential confound in Experiment 1 to determine whether faster responses to alto saxophones were instrument-specific or condition-specific (i.e., because it was the instrument presented in the single-instrument block). In Experiment 2, each of the five instruments tested in Experiment 1 was presented in its own single-instrument block. Mean response times were faster

for saxophone than for horn, although the mean response time for French horn when presented in a single block of Experiment 2 (M = 700 ms) was still markedly faster than the mean response time for the French horn when presented in the mixed-instrument block of Experiment 1 (M = 1,147 ms). This pattern of results also held true for the clarinet (Experiment 1 mean RT = 1,147 ms, Experiment 2 mean RT = 677 ms), marimba (Experiment 1 mean RT = 1,058 ms, Experiment 2 mean RT = 637 ms), and piano (Experiment 1 mean RT = 1,109 ms, Experiment 2 mean RT = 621 ms). Responses to saxophone were not faster than the remaining instruments. In Experiment 1, four instruments made up the mixed block (only one of which is French horn); Experiment 2 suggests that this is not because participants respond faster to the saxophone than to instruments in the mixed block but are rather capitalizing on the structured nature of making pitch judgments when the instrument is consistent. Together, these points discount the possibility that the results from Experiment 1 were due to participants responding faster to saxophone sounds in general, but instead they were responding more quickly to structured input (the single-instrument block) relative to unstructured input (the mixed-instrument block).

Additionally, in Experiment 1 responses were faster to percussive instruments with sharp attacks (piano and marimba) relative to nonpercussive instruments. In Experiment 2, the effect of faster responses to sharp attacks is gone when each instrument is presented individually. Thus, at least for the present data, it seems that the response time benefit for sharper attacks is only relative to longer attacks as in the mixed-instrument block, not when the piano and marimba are in blocks by themselves. Thus, the conclusion presented in Heald et al. (2017, para. 1) that "the process of auditory recognition cannot be divorced from the short-term context in which the auditory object is presented" also rings true across our Experiments 1 and 2.

Regrettably, information about the musicianship of participants in Experiment 2 was not collected. Therefore, it is unclear whether musicianship influenced the data pattern as it did for Experiment 1. Since Experiments 1 and 2 were drawn from the same pool of participants (undergraduate students in psychology courses at the same university), one might infer similar distributions of musical training as in Experiment 1 due to random sampling. After removing the outliers, the range in RT data is much smaller in Experiment 2 (≈500 ms) than in Experiment 1 (>1,500 ms), possibly due to an unknown difference in musical training across the samples. However, it is inadvisable to statistically compare the RTs across Experiments 1 and 2, because although the primary task was the same, the context within which the task was situated was not. In Experiment 1 (for which participants' musical training background was collected), training was only correlated with response time in the mixed-instrument block. Because Experiment 2 instruments were presented only in single-instrument blocks, we might conclude that the similar response times across blocks are due to markedly reducing the timbral variability within each block, and not because our sample hypothetically consists of highly trained musicians who are better at this task than the participants in Experiment 1. The question of the role of musical training in talker or music adaptation task performance remains an open one. Future work specifically recruiting musicians and nonmusicians will illuminate if musical training might attenuate the costs associated with timbre variability in pitch perception, and/or decrease reaction times in both single and mixed-instrument conditions.

General discussion

Many speech studies have demonstrated that speech sound perception is faster and/or more accurate when listening to a single talker across trials than when the talker changes from trial to trial. There are several accounts for this "talker adaptation" effect, some of which are specific to speech (episodic, active control, Bayesian belief updating) and some of which are general to auditory perception (streaming and efficient coding). Music perception research has shown that there is a perceptual interdependence between pitch and timbre, and that variability in one dimension influences perception of the other. The present study investigated music perception to assess if "talker" adaptation can extend beyond speech. Participants made low–high pitch judgments for notes played either by the same instrument or by multiple instruments presented in random orders. Consistent with talker adaptation paradigms and our primary hypothesis, participants were faster to make pitch judgments in the single-instrument block compared with the mixed-instrument block. Contrary to our prediction, participants were also less accurate in the mixed-instrument block relative to the single-instrument block. This is not problematic, as the direction of this difference is consistent with the increase in reaction times in the mixed-instrument block. Also, there are reports of lower accuracy in mixed-talker blocks of some speech perception experiments examining talker adaptation (e.g., Assmann et al., 1982; Creelman, 1957; Martin et al., 1989; Mullennix et al., 1989; Nusbaum & Morin, 1992; Sommers et al., 1994).

While the present study replicated the direction of the established effect in talker adaptation experiments (faster responses to a single, consistent source than variable sources), the magnitude of the effect appears to differ by domain. Adapting to a single instrument produced a larger perceptual benefit (i.e., a decrease in reaction time relative to the mixed-instrument block) than in talker adaptation studies using the same paradigm. On average, participants responded 292 ms faster (Cohen's d = 0.94) in

the single-instrument block than in the mixed-instrument block of Experiment 1. In two recent studies that used this paradigm (Choi et al., 2018; Stilp & Theodore, 2020), participants categorized speech targets 62–141 ms faster (*d*s = 0.34–0.68) in single-talker blocks than in mixed-talker blocks. Differences in effect magnitudes cannot be attributed to differences in stimulus durations, as stimuli in the present study (1,000 ms) all had longer durations than the monosyllabic words tested by Choi et al., 2018 (Experiment 1, all durations = 300 ms) and in Stilp and Theodore (2020; mean duration = 625 ± 81 ms). Previous research has established perceptual interdependencies between pitch and timbre in music (Krumhansl & Iverson, 1992) and between speech content and talker identity in speech studies (Mullennix & Pisoni, 1990). The present results of a larger effect size for the music task relative to previous speech studies suggests that separating pitch and timbre in music might be more difficult than separating speech sounds and talker identity. However, two points temper this conclusion. First, this is generalization from the first experiment of its kind (converting a talker adaptation paradigm to measure pitch height judgments); further replications and extensions examining matched durations and difficulty would be beneficial. Second, these are between-subjects comparisons at present; future studies should test the same listener sample in both speech adaptation and music adaptation paradigms to test this relationship explicitly.

Not all acoustic variability proves consequential for a given perceptual task. For example, in speech, Sommers et al. (1994) reported that variability in talker or in speaking rate, but not amplitude, affected spoken word recognition (see also Bradlow et al., 1999; Nygaard et al., 1995). Listeners rely less on the spectral properties of a preceding context for categorizing subsequent target vowels (smaller spectral contrast effects) when the mean fundamental frequency of the context sentence varied from trial to trial relative to a more consistent fundamental frequency (Assgari et al., 2019). However, similar trial-to-trial variability for characteristics that would be consistent within a given talker (resonances of the vocal tract; mean F1, mean F3) did not affect perception of vowels (Mills et al., 2022). In music, Van Hedger et al. (2015) demonstrated similar effects in which listeners with absolute pitch were slower to recognize a designated pitch class when there was trial-to-trial instrument variability or octave variability, but not when there was amplitude variability (compared with when a given dimension did not vary; see also Krumhansl & Iverson, 1992; Melara & Marks, 1990; Pitt, 1994). Furthermore, in naturalistic speech or music listening, listeners must continuously evaluate which cues are important for perception based on an evolving window of context, and they can exploit any structure that is present in one dimension even if there is variability in another (Heald et al., 2017). Thus, not all types of variability are expected to induce processing costs; one possible determining factor in whether a cost is incurred is whether the varying dimension is diagnostic for perception of the target. Additionally, when processing costs do arise from variability in the signal, not all costs or types of variability are equally detrimental to perception. As previously discussed, the effect size is much larger here than in speech studies, suggesting that the processing cost associated with changing timbres might be more expensive than that of changing talkers. Together, not all variability results in challenges to perception, and variability which does influence perception does so to different degrees. Continuing work in these types of adaptation studies needs to explicitly situate experiments with these broader points in mind.

Musical experience could promote resiliency to acoustic variability across musical instruments. While the relationship between musicianship and task performance in this experiment could only be examined post hoc, the participants with the most musical training were the most accurate and trended toward being faster in the mixed-instrument block (Fig. 3). This is consistent with the "musician advantage" literature in which musicians consistently outperform nonmusicians in musical pitch-related tasks (e.g., Madsen et al., 2017; Micheyl et al., 2006; Schön et al., 2004; Spiegel & Watson, 1984; Tervaniemi et al., 2005). While musical training appears to be related to increased accuracy and decreased reaction time in the mixed-instrument block, the participants had little musical training overall (M = 3.53 years). Participants with more formal musical training likely had more experience with (at least some of) the stimuli being tested as well as making pitch judgments. However, even our most musical participants have far less experience perceiving music than they do with perceiving speech. Taken together, this suggests that perceptual experience might protect against processing costs associated with less structure in the stimuli (as seen for the higher-performing participants in the mixed-instrument block). Additional work is necessary to delineate between how musical training experience contributes to this perceptual resiliency to stimulus variability. Future studies should explicitly define musicianship and recruit listeners with a wide range of musical experience (from nonmusicians to highly trained musicians) to investigate the role of increasing experience in music in these adaptation paradigms.

One limitation of this study is the unexpected difficulty of the task. The major third interval was chosen because it far exceeds the just noticeable difference for discrimination of pitch intervals (Zarate et al., 2012, 2013). However, this study employed a categorization task, which is more difficult than a discrimination task. Because of this difference, a 90% correct performance criterion was implemented in the single-instrument block of Experiment 1 to ensure participants were able to reliably make the pitch judgments. This was also done to match how accuracy is typically at ceiling

in talker adaptation paradigms (e.g., Choi et al., 2018; Stilp & Theodore, 2020). Although categorizing pitches within a major third interval was more difficult than originally anticipated, most participants still performed far above chance levels. However, choosing a larger musical interval might not necessarily decrease task difficulty, as (nonmusical) participants might categorize pitches in the wrong octave (Shepard, 1964). The difficulty of the task underscores that in traditional talker adaptation paradigms, listeners are experts at the task (categorizing speech sounds of their native language) while in these experiments, listeners are nonexperts at the task (generally nonmusicians categorizing pitches in music). The intersections between perceptual expertise and overcoming acoustic variability as discussed above, as well as how difficult the listener finds the task, are rich areas for future investigations.

In talker adaptation paradigms, participants are assumed to be expert speech perceivers who exploit existing well-defined categories for speech sounds. The present study does not presume that participants have extensive music experience or existing categories for "low pitch" and "high pitch" (or at least, not as well-defined as those categories they have for speech sounds). Yet, the same pattern of response times was observed across the two domains, suggesting that adaptation might not depend on having long-established categories in place. Active control (Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007) and episodic approaches (Goldinger, 1996, 1998) to talker adaptation are specific to speech, in which listeners are expert perceivers; Bayesian belief updating also draws heavily on a listener building up extensive perceptual experience with the sounds they are hearing (Kleinschmidt & Jaeger, 2015). Conversely, the streaming (Choi & Perrachione, 2019) and efficient coding approaches (Stilp & Theodore, 2020) do not require such experience, and thus most readily explain benefits from adapting to structure in nonspeech sounds. The benefits of structure for auditory perception apply beyond *talker* adaptation and are not dependent on having categories as thoroughly defined as those for speech sounds.

One of the most enduring debates surrounding speech perception is the extent to which it is "special" compared with other perceptual abilities (e.g., Diehl et al., 2004; Liberman, 1996; Liberman et al., 1967; Schouten, 1980). Although several perceptual behaviors have been initially reported as indicating specialized speech processing, this view is challenged given that some of these behaviors were subsequently replicated with nonspeech stimuli. These behaviors include: categorical perception (speech: Liberman et al., 1967; Studdert-Kennedy et al., 1970; nonspeech: Locke & Kellar, 1973; Miller et al., 1976), duplex perception (speech: Mattingly et al., 1971; Rand, 1974; nonspeech: Fowler & Rosenblum, 1990), trading relations between correlated cues (speech: Repp, 1982; nonspeech: Parker et al., 1986), and compensation for preceding context on shorter (speech: Mann, 1980; Miller & Liberman, 1979; nonspeech: Pisoni et al., 1983; Lotto & Kluender, 1998) and longer timescales (speech: Ladefoged & Broadbent, 1957; nonspeech: Stilp et al., 2010). Some of these behaviors have been demonstrated in listeners who have far less experience perceiving speech than adults, ranging from human infants (e.g., Eimas et al., 1971; Fowler et al., 1990; Miller & Eimas, 1983; Werker & Tees, 1984) to nonhuman animals (e.g., Dooling et al., 1988; Kuhl & Miller, 1975; Kluender et al., 1987; Sinnott et al., 1976). In the present study, parallel patterns of results across the present music perception task and previously published similar speech perception tasks endorse reconsideration of domain-specific (i.e., speech-specific) theoretical accounts to talker adaptation (Heald & Nusbaum, 2014; Kleinschmidt & Jaeger, 2015; Liberman, 1996; Liberman et al., 1967; Magnuson & Nusbaum, 2007) in favor of domain-general accounts of capitalizing on consistency in sensory inputs rooted in first principles of how perceptual systems operate most broadly (Attneave, 1954; Barlow, 1961; Bregman, 1990; Diehl et al., 2004). Since listeners must overcome acoustic variability when perceiving speech, music, and other classes of natural sounds, it is perhaps unsurprising that domain-general approaches proved most useful in explaining the nonspeech behavior observed in the current work.

Conclusion

Variability is pervasive in the acoustic environment. Previous studies report that structure in speech yields perceptual benefits (i.e., faster and/or more accurate responses to a single talker compared with multiple talkers), while work in music perception has repeatedly demonstrated that listeners' judgments about pitch are influenced by the timbre of the instrument playing the note, and vice versa. As such, the perceptual benefits for making judgments about a single, consistent source were predicted to extend beyond speech. Here, listeners categorized pitches as low or high when produced by a single instrument or by one of four instruments that varied from trial to trial. As predicted, listeners were faster in the single-instrument condition than the multiple-instrument condition, and were also more accurate for the single instrument (Experiment 1), and the specific timbre of the single instrument was not responsible for these results (Experiment 2). Thus, at least some aspects of "talker" adaptation appear to be a general response to structure in the acoustic environment. The efficient coding and streaming accounts of "talker" adaptation most readily accommodate the benefits of structure in less familiar nonspeech stimuli. Future work should investigate the role of perceptual experience in resiliency to stimulus variability as well as continue to delineate between approaches to adaptation more broadly.

Open practices statement

This experiment was not preregistered. Stimuli, de-identified data, and data analysis scripts for this experiment are available (https://osf.io/6gwxv/).

Acknowledgments The authors thank Lauren Girouard-Hallam, Raina Isaacs, Vitor Neves Guimaraes, and Carolyn Mervis for feedback on an earlier version of this manuscript, and Aidan Shorey, Micki Shorey, and Ralph Shorey for providing instrument photos for Fig. 1. The authors declare no financial support nor any conflict of interests pertaining to this manuscript.

References

Assgari, A. A., & Stilp, C. E. (2015). Talker information influences spectral contrast effects in speech categorization. *The Journal of the Acoustical Society of America, 138*(5), 3023–3032. https://doi.org/10.1121/1.4934559

Assgari, A. A., Theodore, R. M., & Stilp, C. E. (2019). Variability in talkers' fundamental frequencies shapes context effects in speech perception. *The Journal of the Acoustical Society of America, 145*(3), 1443–1454. https://doi.org/10.1121/1.5093638

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America, 71*(4), 975–989. https://doi.org/10.1121/1.387579

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183–193. https://doi.org/10.1037/h0054663

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication, 1*(01), 217–233.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barreda, S. (2012). Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis. *Journal of the Acoustical Society of America, 132*(5), 3453–3464.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 2021. https://doi.org/10.18637/jss.v067.i01

Boersma, P., & Weenick, D. (2021). Praat: Doing phonetics by computer (Version 6.1.50) [Computer program]. http://www.praat.org

Bradlow, A. R., Nygaard, L. C., & Pisani, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61*(2), 206–219.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Bressler, S., Masud, S., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research, 78*(3), 349–360. https://doi.org/10.1007/s00426-014-0555-7

Choi, J. Y., Hu, E. R., & Perrachione, T. K. (2018). Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention, Perception, & Psychophysics, 80*(3), 784–797. https://doi.org/10.3758/s13414-017-1395-5

Choi, J. Y., & Perrachione, T. K. (2019). Time and information in perceptual adaptation to speech. *Cognition, 192*(June), 103982. https://doi.org/10.1016/j.cognition.2019.05.019

Creelman, C. D. (1957). Case of the unknown talker. *The Journal of the Acoustical Society of America, 29*(5), 655–655. https://doi.org/10.1121/1.1909003

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology, 55*, 149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028

Dooling, R. J., Okanoya, K., & Brown, S. D. (1989). Speech perception by budgerigars (Melopsittacus undulatus): The voiced-voiceless distinction. *Perception & Psychophysics, 46*(1), 65–71.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*(3968), 303–306.

Fowler, C. A., & Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance, 16*(4), 742–754. https://doi.org/10.1037/0096-1523.16.4.742

Fowler, C. A., Best, C. T., & Mcroberts, G. W. (1990). Young infants' perception of liquid coarticulatory influences on following stop consonants. *Perception & Psychophysics, 48*, 559–570.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 22*(5), 1166–1183. https://doi.org/10.1037/0278-7393.22.5.1166

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(1), 152–162.

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8*(MAR), 1–15. https://doi.org/10.3389/fnsys.2014.00035

Heald, S. L. M., Van Hedger, S. C., & Nusbaum, H. C. (2017). Perceptual plasticity for auditory object recognition. *Frontiers in Psychology, 8*(MAY), 781. https://doi.org/10.3389/fpsyg.2017.00781

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustic Society of America, 97*(5), 3099–3111.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203. https://doi.org/10.1037/a0038695

Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science, 237*(4819), 1195–1197.

Krumhansl, C. L., & Iverson, P. (1992). Perceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 739–751.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science, 190*(4209), 69–72.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America, 29*(1), 98–104.

Lenth, R. (2019). *emmeans: Estimated marginal means, aka least-squares means* (R package Version 1.6.3). https://cran.r-project.org/package=emmeans

Liberman, A. M. (1996). *Speech: A special code*. MIT Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431–461.

Locke, S., & Kellar, L. (1973). Categorical perception in a nonlinguistic mode. *Cortex, 9*(4), 355–369.

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics, 60*(4), 602–619.

Madsen, S. M. K., Whiteford, K. L., & Oxenham, A. J. (2017). Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds. *Scientific Reports, 7*(1), 1–9. https://doi.org/10.1038/s41598-017-12937-9

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*(2), 391–409. https://doi.org/10.1037/0096-1523.33.2.391

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28*(5), 407–412.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(4), 676–684. https://doi.org/10.1037/0278-7393.17.1.152

Mattingly, I. G., Liberman, A. M., Syrdal, A. K., & Halwes, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology, 2*(2), 131–157. https://doi.org/10.1016/0010-0285(71)90006-5

Melara, R. D., & Marks, L. E. (1990). Processes underlying dimensional interactions: Correspondences between linguistic and nonlinguistic dimensions. *Memory & Cognition, 18*(5), 477–495.

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research, 219*(1/2), 36–47. https://doi.org/10.1016/j.heares.2006.05.004

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics, 25*, 457–465.

Miller, J. L., & Eimas, P. D. (1983). Studies on the categorization of speech by infants. *Cognition, 13*(2), 135–165.

Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., & Dooling, R. J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America, 60*(2), 410–417. https://doi.org/10.1121/1.381097

Mills, H. E., Shorey, A. E., Theodore, R. M., & Stilp, C. E. (2022). Context effects in perception of vowels differentiated by F1 are not influenced by variability in talkers' mean F1 or F3. *The Journal of the Acoustical Society of America, 152*(1), 55–66. https://doi.org/10.1121/10.0011920

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*(4), 379–390. https://doi.org/10.3758/BF03210878

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*(1), 365–378. https://doi.org/10.1037/0278-7393.15.4.676

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 113–134). IOS Press.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics, 57*(7), 989–1001. https://doi.org/10.3758/BF03205458

Opolko, F., & Wapnick, J. (1989). *McGill University Master Samples user's manual*. McGill University.

Parker, E. M., Diehl, R. L., & Kluender, K. R. (1986). Trading relations in speech and nonspeech. *Perception & Psychophysics, 39*(2), 129–142. https://doi.org/10.3758/BF03211495

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*(2), 175–184.

Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics, 34*, 314–322.

Pitt, M. A. (1994). Perception of pitch and timbre by musically trained and untrained listeners. *Journal of Experimental Psychology: Human Perception and Performance, 20*(5), 976–986. https://doi.org/10.1037/0096-1523.20.5.976

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rand, T. C. (1971). Vocal tract size normalization in the perception of stop consonants. *Journal of the Acoustical Society of America, 50*, 139.

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America, 55*(3), 678–680. https://doi.org/10.1121/1.1914584

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92*(1), 81–110. https://doi.org/10.1037/0033-2909.92.1.81

Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology, 41*(3), 341–349. https://doi.org/10.1111/1469-8986.00172.x

Schouten, M. E. H. (1980). The case against a speech mode of perception. *Acta Psychologica, 44*(1), 71–98. https://doi.org/10.1016/0001-6918(80)90077-3

Shepard, R. N. (1964). Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America, 36*(12), 2346–2353. https://doi.org/10.1121/1.1919362

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America, 96*(3), 1314–1324. https://doi.org/10.1121/1.411453

Spiegel, M. F., & Watson, C. S. (1984). Performance on frequency-discrimination tasks by musicians and nonmusicians. *Journal of the Acoustical Society of America, 76*(6), 1690–1695. https://doi.org/10.1121/1.391605

Sinnott, J. M., Beecher, M. D., Moody, D. B., & Stebbins, W. C. (1976). Speech sound discrimination by monkeys and humans. *The Journal of the Acoustical Society of America, 60*(3), 687–695.

Stilp, C. E., & Theodore, R. M. (2020). Talker normalization is mediated by structured indexical information. *Attention, Perception, & Psychophysics, 82*(5), 2237–2243. https://doi.org/10.3758/s13414-020-01971-x

Stilp, C. E., Alexander, J. M., Kiefte, M., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, & Psychophysics, 72*(2), 470–480.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., & Cooper, F. S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review, 77*(3), 234–249. https://doi.org/10.1037/h0029078

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs nonmusicians: An event-related potential and behavioral study. *Experimental Brain Research, 161*(1), 1–10. https://doi.org/10.1007/s00221-004-2044-5

Van Hedger, S. C., Heald, S. L. M., & Nusbaum, H. C. (2015). The effects of acoustic variability on absolute pitch categorization: Evidence of contextual tuning. *The Journal of the Acoustical*

Society of America, 138(1), 436–446. https://doi.org/10.1121/1.4922952

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*(1), 49–63.

Wier, C. C., Jesteadt, W., & Green, D. M. (1977). Frequency discrimination as a function of frequency and sensation level. *The Journal of the Acoustical Society of America, 61*(1), 178–184.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Zarate, J. M., Ritson, C. R., & Poeppel, D. (2012). Pitch-interval discrimination and musical expertise: Is the semitone a perceptual boundary? *The Journal of the Acoustical Society of America, 132*(2), 984–993. https://doi.org/10.1121/1.4733535

Zarate, J. M., Ritson, C. R., & Poeppel, D. (2013). The effect of instrumental timbre on interval discrimination. *PLOS ONE, 8*(9), e75410. https://doi.org/10.1371/journal.pone.0075410

Zhang, C., & Chen, S. (2016). Toward an integrative model of talker normalization. *Journal of Experimental Psychology: Human Perception and Performance, 42*(8), 1252–1268.