Short-term Recognition of Timbre Sequences: Music Training, Pitch Variability, and Timbral Similarity

Kai Siedenburg
McGill University, Montreal, Canada & Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

Stephen McAdams
McGill University, Montreal, Canada

The goal of the current study was to explore outstanding questions in the field of timbre perception and cognition—specifically, whether memory for timbre is better in trained musicians or in nonmusicians, whether short-term timbre recognition is invariant to pitch differences, and whether timbre dissimilarity influences timbre recognition performance. Four experiments examined short-term recognition of musical timbre using a serial recognition task in which listeners indicated whether the orders of the timbres of two subsequently presented sound sequences were identical or not. Experiment 1 revealed significant effects of sequence length on recognition accuracy and an interaction of music training and pitch variability: musicians performed better for variable-pitch sequences, but did not differ from nonmusicians with constant-pitch sequences. Experiment 2 yielded a significant effect of pitch variability for musicians when pitch patterns varied between standard and comparison sequences. Experiment 3 highlighted the impact of the timbral dissimilarity of swapped sounds and indicated a recency effect in timbre recognition. Experiment 4 confirmed the importance of the dissimilarity of the swap, but did not yield any pertinent role of timbral heterogeneity of the sequence. Further analyses confirmed the strong correlation of the timbral dissimilarity of swapped sounds with response behavior, accounting for around 90% of the variance in response choices across all four experiments. These results extend findings regarding the impact of music training and pitch variability from the literature on timbre perception to the domain of short-term memory and demonstrate the mnemonic importance of timbre similarity relations among sounds in sequences. The role of the factors of music training, pitch variability, and timbral similarity in music listening is discussed.

Received September 4, 2017, accepted April 8, 2018.

Key words: musical timbre, auditory short-term memory, serial recognition, pitch variability, similarity

The 20th century has witnessed a flourishing of interest in the manipulation of timbre by means of music composition and music technology. In fact, there are a variety of musical styles for which sequences of timbres act as the primary conveyors of musical information. Apart from abundant examples in popular and non-Western music (Nattiez, 2007), a popular example from 20th-century art music is the so-called *Klangfarbenmelodie* ("timbre melody"), featuring timbral configurations that are sculpted over time (Erickson, 1975). Around the beginning of the last century, the composer Arnold Schoenberg famously conjectured, "Tone-color melodies (Klangfarbenmelodien)! How acute the senses that would be able to perceive them! How high the development of spirit that could find pleasure in such subtle things! In such a domain, who dares ask for theory!" (Schoenberg, 1911/1978, p. 422). Schoenberg's statement is prophetic in the sense that despite a long history of research on timbre—at least dating back to von Helmholtz (1877/1954)—we only have a coarse understanding of the cognitive processing of timbral structures in musical contexts until today. Most empirical work has focused on the "royal couple" of music theory; that is, pitch and duration. Timbral structures, omnipresent and decisive in most contemporary (popular or art) music, are hardly captured by the theoretical network spun by mainstream music cognition research.

Timbre is here understood as an umbrella term (Siedenburg & McAdams, 2017a) that primarily concerns the bundle of perceptual attributes that lends tones a sense of "color" or "shape" and identity (Handel, 1995). It encompasses continuous perceptual qualities of sounds such as brightness, sharpness of attack, spectrotemporal irregularity, roughness, and noisiness in addition to auditory features specific to certain instruments. The perceptual structure of timbre has been modeled by multidimensional scaling (MDS) of pairwise dissimilarity judgments, yielding spatial configurations of timbres (cf. McAdams, 2013, for a review).

Music Perception, volume 36, issue 1, pp. 24–39, issn 0730-7829, electronic issn 1533-8312. © 2018 by the regents of the university of california all rights reserved. please direct all requests for permission to photocopy or reproduce article content through the university of california press's Reprints and Permissions web page, http://www.ucpress.edu/journals.php?p=reprints. DOI: https://doi.org/10.1525/mp.2018.36.1.24

McAdams, Winsberg, Donnadieu, De Soete, and Krimphoff (1995) found spectral, temporal, and, to a lesser extent, spectrotemporal properties of tones to be the major acoustic correlates of the resulting timbre space (for a recent review, see Siedenburg, Fujinaga, & McAdams, 2016).

McAdams and Goodchild (2017) have formulated a taxonomy of timbral contrasts that occur frequently in the orchestral repertoire. Important contrast types include antiphonal alternation of instrument groups ("call and response"), timbral echoing in which repeated musical phrases appear with different orchestrations, and timbral shifts in which musical materials are reiterated with varying orchestrations. In order to apprehend any of these contrast types, listeners must track timbral changes over time. Here we report on a series of four experiments that probe a cognitive process foundational for timbre's function in musical contexts: the capacity to recognize and match timbre sequences from short-term memory (STM). Formally, we define STM as the cognitive faculty responsible for the retention of sensory and categorical information over spans in the range of seconds (Jonides et al., 2008). In music listening, short-term sequence recognition is at the foundation of the cognitive sequencing of musical timbre, including the parsing and integration of musical events into phrase structures and eventually the experience of musical form (McAdams, 1989). Outside the lab, however, timbral contrast does not occur in isolation, but mostly covaries with other parameters, such as pitch or loudness. A central goal of the current study thus is to clarify whether timbre pattern recognition is robust to concurrent variation in pitch. Given that timbre recognition may seem to be a specialist domain (at least in Schoenberg's eyes one century ago), we were also interested in whether performance would differ across groups of trained musicians and nonmusicians who do not have any experience in playing or analyzing music. Finally, we also wished to study the role of timbre dissimilarity relations in STM in order to investigate how timbre's perceptual topology affects timbre sequence recognition.

TIMBRE RECOGNITION IN THE LITERATURE

Empirical research on short-term memory for timbre is a rather recent endeavor, and selected studies have started to address basic issues with regards to the function and structure of short-term memory for timbre. Investigating the domain-specificity of memory for timbre, Schulze and Tillmann (2013) compared the recognition of sequences with five and six sounds differing in timbre or pitch with that of sequences of words. They used a serial recognition task that required listeners to assess whether two subsequently presented sequences of sounds were in the same order or not. They observed that, contrary to words and pitches, there was no effect of length for sequences of sounds differing in timbre. The authors interpreted this finding as an indication that timbre is stored via sensory representations, which in contrast to words and pitches may not engage motor-based rehearsal mechanisms. Siedenburg, Mativetsky, and McAdams (2016) explored auditory and verbal STM in a case study of a North Indian style of drumming that incorporates vocalizations of drum sounds (a type of timbre solfège for drumming). They observed strong effects of sequence structure on serial recognition accuracy, but could not find a principled advantage for verbal compared to instrumental sounds. The only advantage for verbal sounds occurred for sequential structures that participants were familiar with.

Using electrophysiology to study STM for timbre, Nolden et al. (2013) recorded EEG during a serial recognition task. In a control condition without memory load, participants ignored the standard sequence and merely judged a property of the last tone of the comparison sequence. Differences in event-related potentials (ERP) between control and memory conditions were found during the retention interval, and the higher the memory load, the stronger was the ERP negativity. Similar findings have been reported by Alunni-Menichini et al. (2014), who demonstrated that the same ERP component robustly indexes STM capacity. These results indicate that the retention of timbre requires an active, attention-dependent form of STM.

THE ROLE OF CONCURRENT PITCH VARIABILITY

Concurrent variability in pitch adds another source of complexity to studying STM for timbre. Most studies on the perceptual interaction of pitch and timbre processing are based on pairwise discrimination with only short retention times below 1 s. Providing a groundwork for many later studies on interactions of auditory dimensions, Melara and Marks (1990) used speeded classification of stimuli varying in pitch and timbre with either independent or correlated changes along the two dimensions. Participants were asked to discriminate stimuli only along one dimension. Reaction times were slower when changes in attended and unattended attributes were independent, but faster when both dimensions were correlated. This was interpreted as evidence for integral processing of the two auditory attributes, conceptualized as a cross-talk between "higher-level channels" responsible for the computation of the perceptual attributes pitch and timbre. These findings were replicated for nonmusicians and musicians (Krumhansl & Iverson, 1992; Pitt, 1994), and recently by Caruso and Balaban (2014), showing that the greater a concurrent change in pitch, the harder it was to correctly discriminate timbre. Further extending this line of work, Allen and Oxenham (2014) measured difference limens for musicians and nonmusicians using stimuli with concurrent random variations along the nonattended dimension. Ensuring that the experimental units of timbre and pitch were of the same perceptual magnitude, they found symmetric mutual interference of pitch and timbre in the discrimination task. Musicians yielded higher discrimination overall, but there was no interaction of musicianship and auditory parameter (pitch/timbre) that would have pointed to a structural difference in the processing of these two attributes.

More specifically probing STM, Starr and Pitt (1997) used an interpolated tone paradigm (cf., Deutsch, 1970) that required participants to match a standard and a comparison stimulus, separated by a 5-s interval with intervening distractor tones. Their first experiment demonstrated an effect of timbre similarity without marked differences in performance between musicians and nonmusicians. Using a mixed group of participants, they further tested whether pitch variability in the interference tones affected timbre matching, which turned out not to be the case. Thus, the reliable perceptual interaction of timbre and pitch in discrimination tasks (as reviewed above) appeared to vanish in the domain of STM. Based on this rather incongruent set of findings, we conclude that further research is required to investigate how STM for timbre is affected by pitch variability, and whether other factors such as music training could play a role in this context.

MUSIC TRAINING

Musicians are well known to have superior memory for pitch (see, e.g., Schulze, Zysset, Mueller, Friederici, & Koelsch, 2011), and therefore musicians might be less likely to confuse variation in pitch and timbre in memory tasks. More generally, this issue relates to an open question in timbre research, namely whether music training affects timbre processing. So far, no systematic differences between musicians and nonmusicians have been found in experiments on the perception of timbral dissimilarity (Alluri & Toiviainen, 2012; Kendall, Carterette, & Hajda, 1999; Lakatos, 2000; McAdams et al., 1995). As mentioned above, Starr and Pitt (1997) also did not find differences between groups. On the other hand, Chartrand and Belin (2006) reported that musicians possess superior discrimination abilities for vocal and instrumental timbres. Specifically investigating short-term recognition, Siedenburg and McAdams (2017b) recently demonstrated that musicians more accurately retained the timbre of sounds across short retention times (2 and 6 s) compared to nonmusicians, independent of whether familiar or unfamiliar sounds were presented. Additional evidence for enhanced timbre processing in musicians is accrued by several neurophysiological studies (Pantev, Roberts, Schulz, Engelien, & Ross, 2001; Shahin, Roberts, Chau, Trainor, & Miller, 2008; Strait, Chan, Ashley, & Kraus, 2012). The impact of music training on timbre cognition and STM for timbre thus remains an open question.

SIMILARITY

Similarity effects are ubiquitous in verbal and visual STM (e.g., Baddeley, 2012; Sekuler & Kahana, 2007): despite being perceptually discriminable, similar items are more frequently confused in memory compared to dissimilar ones. When it comes to memory for timbre, however, only a few studies have taken into account the role of similarity. Starr and Pitt (1997) had participants match synthesized harmonic complexes differing in brightness in the presence of interfering tones in the retention interval. They observed that both musicians and nonmusicians performed with greater accuracy when the timbre of the distractor tones was dissimilar to the target timbre, an effect that was robust over distractors with varying pitch. Golubock and Janata (2013) tested recognition performance of isolated tones for a set of synthesized electronic sounds (varing along the dimensions of spectral centroid, attack time, and spectral variability over time) along with a perceptually more diverse set of sounds from a commercial synthesizer. They observed a significantly greater working memory capacity for the latter set, which suggests that acoustic diversity (proportional to the pairwise perceptual similarity between items in a test set) enhances recognition, which could point to an underlying similarity effect.

Studying commonalities of auditory and visual STM, Visscher, Kaplan, Kahana, and Sekuler (2007) used amplitude-modulated sinusoid complexes in audition and Gabor patches in vision in an item-recognition task: lists (i.e., sequences) of items followed by probe items were presented and participants indicated whether the probe was part of the previous list or not (“old” vs. “new”). The obtained recognition accuracy exhibited similar patterns across domains and salient effects of probe-to-list similarity as well as sequence homogeneity were observed. Specifically, it was shown that on “new” trials, an increase in heterogeneity yields a decrease of “new” responses independent of probe-list similarity (cf., Kahana & Sekuler, 2002; Viswanathan, Perl, Visscher, Kahana, & Sekuler, 2010). These findings were confirmed and reinterpreted as an adaptive shift of participants' response criteria by Nosofsky and Kantner (2006): the more homogeneous a list is, the more likely a participant is to respond "old." On the other hand, testing timbre recognition specifically with an item-recognition task, Siedenburg and McAdams (2017b) observed a positive correlation of response choices with the average (summed) timbre dissimilarities of the probe to the list, but no correlation of response behavior with sequence homogeneity was found. To our knowledge, no study has yet tested effects of homogeneity (or its inverse, heterogeneity) in serial recognition. More generally, although the literature suggests that timbral similarity could well play a role in serial recognition of timbre, it is as yet unclear how this aspect would manifest itself in a sequential context.

THE PRESENT STUDY

Whereas many studies that use pairwise discrimination have found interactive processing and interference, a study using a task that more strongly tapped into STM found non-congruent results (Starr & Pitt, 1997). Therefore, the first central goal of this study was to investigate the robustness of timbre-sequence recognition to interference by concurrent variability in pitch. Experiment 1 included a factor of pitch variability that compared timbre recognition on sequences with constant and variable-pitch. Because a factor of sequence length would allow us to draw a connection to other recent timbre memory studies that used serial recognition tasks (Nolden et al., 2013; Schulze & Tillmann, 2013), we tested sequences with four to six items. In order to better understand the role of music training in timbre perception and cognition, we tested a group of musicians and nonmusicians. Experiment 2 followed up on the impact of pitch variability and tested a group of musicians on a subset of trials from Experiment 1 in conjunction with a more complex structuring of pitch patterns. Regarding the variable of timbre similarity, Starr and Pitt (1997) found similarity-based interference in an interpolated tone task, and Siedenburg and McAdams (2017b) underlined the importance of list-to-probe similarity in an item-recognition task. However, it is unclear whether similarity among items plays a role in serial recognition of easily discriminable timbres when only the order of items may change. Therefore, Experiment 3 examined the effect of varying the timbral similarity of the swapped pair of sounds. Finally, Experiment 4 was conceived to explore the impact of an additional facet of similarity, namely the timbral heterogeneity within sequences.

Experiment 1: Group, Length, and Pitch Variability

The experiment used a serial recognition task to test groups of musicians and nonmusicians on the within-subject factors of sequence length (4, 5, 6 tones), and concurrent pitch variability (constant vs. variable pitch). We expected lower performance for longer sequences, as well as for sequences with variable pitch. We also expected musicians to generally outperform nonmusicians.

METHOD

The research reported in this manuscript was carried out according to the principles expressed in the Declaration of Helsinki, and the Research Ethics Board II of McGill University has reviewed and certified this study for ethical compliance (certificate #67-0905).

Participants. Sixty listeners participated in the experiment. These consisted of 30 music students, recruited from a mailing list of the Schulich School of Music at McGill University, and 30 nonmusicians (required not to have more than a year of experience in playing a musical instrument, nor any musical instruction after elementary school) who were recruited via web-based, classified advertisements. Musicians had an average age of 23 years ($SD = 4.4$, range: 19–33), included 19 male participants, and featured an average of 14 years of instrumental training ($SD = 5.2$) and 4 years ($SD = 2.8$) of formal music-theoretical instruction or ear training. Nonmusicians were on average 25 years old ($SD = 7.2$, range: 19–50), included 25 female participants, had an average of 0.3 years ($SD = 0.66$) of instrumental instruction and an average of 1.1 ($SD = 1.74$) years of music instruction in elementary school and no further formal music training from there on. In this and the following experiments, participants reported normal hearing, which was confirmed by conducting a standard pure-tone audiogram measured right before the main experiment (ISO 398-8, 2004; Martin & Champlin, 2000). They were required to have hearing thresholds of 20 dB HL or better, assessed at octave spaced frequencies from 125 to 8000 Hz.

Stimuli. The same stimuli were used as in McAdams et al. (1995) based on FM-synthesized sounds (Yamaha DX7, Yamaha Corp., Shizuoka, Japan) created by Wessel, Bristow, and Settel (1987), to some extent emulating instruments from the classical orchestra. All sounds were synthesized at pitch E♭4 (fundamental frequency of 311 Hz) and had been perceptually normalized with regards to loudness and duration in the original study. We used this particular set of sounds because it not only had been perceptually normalized in pitch and loudness, but also allowed us to make use of its extant timbre dissimilarity data. These had been collected through pairwise timbre dissimilarity judgments of sounds on a 1–9 rating scale, collected from 98 participants. We were thus able to construct timbre sequences with varying degrees of inter-item similarity (see Experiments 3–4). From the 18 sounds, we selected a diverse subset of eight sounds, featuring sounds from four instruments with impulsive excitation (plucked, struck) and four with continuous excitation (blown, bowed). The selected instruments were electronic emulations of the bassoon, clarinet, guitar, harpsichord, horn, piano, trumpet, and vibraphone.

The original sounds contained subtle hiss noise, which was removed by using a state-of-the-art audio noise removal algorithm (Siedenburg & Dörfler, 2013) implemented in MATLAB version R2013a (The MathWorks, Inc., Natick, MA). In order to construct sequences with variable pitch height, and given that the original synthesizer was no longer available to us, we created transposed versions of plus/minus one whole tone using the audio-editing software AudioSculpt (IRCAM, Paris, France). As confirmed informally by several expert listeners, no audible artifacts were introduced by noise removal or transposition. Sounds were then cropped to a duration of 500 ms using a linear fade out from 480–500 ms. Note that similarity relations could have been affected by this truncation in minor ways. Further note that timbre dissimilarity relations and timbral identity can be assumed to remain stable for pitch transpositions that are below one octave (Handel & Erickson, 2001; Marozeau, de Cheveigné, McAdams, & Winsberg, 2003; Steele & Williams, 2006). More specifically, Zacharakis, Pastiadis, and Reiss (2015, p. 411) showed that digital pitch shifting by two semitones has negligible effects on timbre dissimilarity, as indicated by almost coinciding positions of original and transposed tones in MDS-based timbre spaces.

Between any two sounds in a sequence, there was a silent interval of 20 ms. Sequences contained 4, 5, or 6 items with 520-ms interonset interval (IOI), similar to the design used by Schulze and Tillmann (2013). Tone sequences had constant or varying pitch. There were 180 sequences in total with 30 sequences per length × pitch condition. Figure 1 shows a sketch of the two different pitch conditions used in Experiments 1 and 2.

Comparison sequences followed the standard after a silent interval of 3 s and were generated either by using the identical sequence or by swapping the last and third-to-last items. One might suspect that the fixed position of swaps (which subjects were not aware of) could lead to an optimization of listening strategy. This aspect is further addressed by varying the position of swap in Experiment 3. In each condition, 50% of the comparison sequences were identical, and 50% were different from the standard sequence.

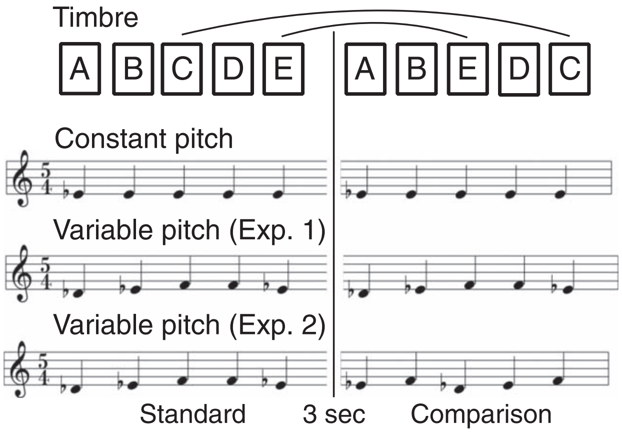

FIGURE 1. Schematic of pitch variability in Experiments 1 and 2 for an exemplary 5-item sequence. Although pitch sequences are identical for standard and comparison in Experiment 1, they differ in Experiment 2. Experiments 3 and 4 only used constant-pitch sequences.

Pitches were drawn from the set D♭4, E♭4, and F4, and pitch patterns were created by interleaving two random permutations of that set and truncating according to the length of 4, 5, or 6 items. For instance, given the permutations (D♭4, E♭4, F4) and (F4, E♭4, D♭4) and a five-item sequence, this would yield the pitches (D♭4, F4, E♭4, E♭4, F4). This ensured that any third-to-last and last tone would have different pitches (i.e., that the same two pitches would not occur in the positions at which the swapped timbres were located). For any given trial, the same pitch pattern was used for standard and comparison sequences.

Apparatus. All four experiments took place in a double-walled sound-isolation chamber (IAC Acoustics, Bronx, NY). Stimuli were presented on Sennheiser HD280Pro headphones (Sennheiser Electronics GmBH, Wedemark, Germany) using a Macintosh computer with digital-to-analog conversion on a Grace Design m904 (Grace Design, Boulder, CO) monitor system. The output level was 67 dB SPL on average (range: 58–75 dB) as measured with a Brüel & Kjær Type 2205 sound-level meter (A-weighting) with a Brüel & Kjær Type 4153 artificial ear to which the headphones were coupled (Brüel & Kjær, Nærum, Denmark). The experimental interface and data collection were realized with the audio software Max/MSP (Cycling 74, San Francisco, CA).

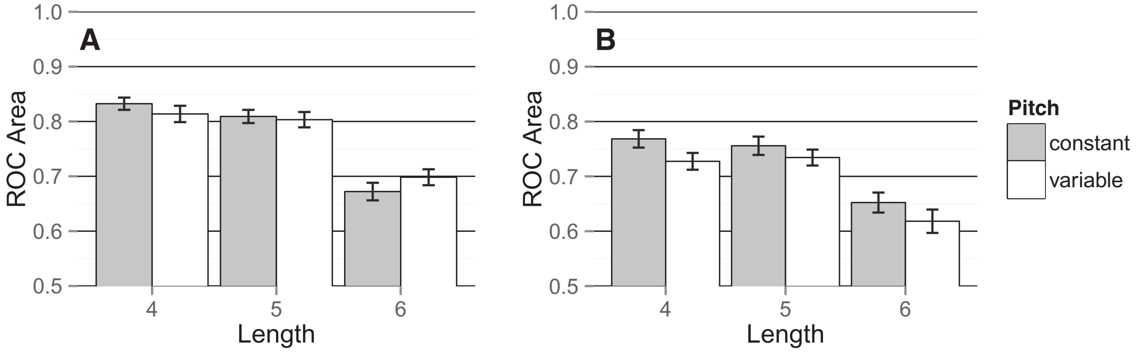

FIGURE 2. Experiment 1: Recognition performance for the factors of sequence length and pitch variability for groups of musicians (A) and nonmusicians (B). Error bars represent (within-subject) 95% confidence intervals.

Procedure. There was one between-subjects factor of music training and two within-subject factors: pitch variability and sequence length. Trials were split into two blocks corresponding to the constant-pitch and variable-pitch conditions. The order of the presentation of the blocks was counterbalanced across participants. The order of sequences within blocks was fully randomized. After having completed the audiogram, participants read through the experimental instructions and completed a set of six training trials, not part of the main experiment. They were instructed that if there was a pitch change during the sequence, only the order of the sounds (timbres) might or might not change, but not the order of the pitches and that they could ignore pitch. The first three training trials were from the constant-pitch condition, in order to ensure that participants understood that they should focus on timbre. The latter three training trials were from the variable-pitch condition. Feedback on response correctness was provided for training trials, and potential questions could be clarified with the experimenter. Participants could listen to the sequences as often as they wished.

In the main experiment, participants listened to the standard sequence, followed by 3 s of silence and the comparison sequence, and then gave their response by clicking on the appropriate button (*same/different*). They subsequently provided an assessment of their level of confidence. Participants could then proceed to the next trial. In contrast to the training, no feedback was provided and no repetition of the stimulus was possible. After finishing the first experimental block, which took around 20 min, they were asked to take a break for 5 min. Having finished both blocks, participants filled out a questionnaire concerning their musical background. Overall, the experiment lasted around 50 min for which participants received a compensation of $10 CAD.

Data analysis. As the dependent measure of recognition performance of all experiments, we calculated the area under the empirical receiver-operating characteristic (ROC) curves for each participant in each condition. Empirical ROCs were computed by combining the same/different responses and confidence ratings (Macmillan & Creelman, 2005, Chapter 3). This measure is monotonically related to the sensitivity index d' but does not assume any specific distribution of the sensory observations. It ranges between .5 (chance level) and 1 (perfect score) and equals the proportion correct in a four-interval same/different (4IAX, or *dual-pair comparison*) task (Micheyl & Dai, 2008). Here and in all following analyses, no violations of sphericity were observed (Mauchly's test). We report original p values in post hoc comparisons, but compare them against the Bonferroni-corrected α-level to account for multiple comparisons.

RESULTS

Grand averages for musicians and nonmusicians were $M = .77$ and $M = .71$, respectively. Mean scores were monotonically decreasing with sequence length ($M = .78, .77, .66$). Musicians' average scores were almost identical across pitch conditions (both $M = .77$ after rounding), but nonmusicians' performance was higher for constant-pitch ($M = .72$) compared to variable-pitch sequences ($M = .69$). Figure 2 depicts all condition means.

Employing a mixed Group (2) × Pitch Variability (2) × Length (3) ANOVA yielded an effect of experimental group, $F(1, 58) = 6.38, p = .01, \eta_p^2 = .10$. There was no main effect of pitch variability, $F(1, 58) = 2.69, p = .11$, but a strong effect of sequence length, $F(2, 116) = 81.22$ $p < .001, \eta_p^2 = .58$. Post hoc comparisons revealed that performance was not significantly different for sequences of length 4 and 5, $t(59) = 0.99, p = .32$, but scores were higher for lengths 4 and 5 compared to 6,

$t(59) > 9.72$, $p < .001$. There was a marginally significant interaction between group and pitch variability, $F(1, 58) = 3.05$, $p = .09$, $\eta_p^2 = .05$. It was due to significantly lower accuracy of nonmusicians compared to musicians in the variable-pitch condition, two-sample $t(58) = 2.96$, $p = .004 < \alpha_{crit} = .0125$, but no significant differences between groups in the constant-pitch condition, two-sample $t(58) = 1.69$, $p = .095$. There were no within-subject differences across pitch variability for musicians, paired $t(29) = .08$, $p > .10$, but marginal differences for the group of nonmusicians, paired $t(29) = 2.22$, $p = .03 > \alpha_{crit} = .0125$.

DISCUSSION

The experiment revealed that multiple factors play into timbre sequence recognition. Performance decreased with sequence length, in line with results from Nolden et al. (2013), but directly contrasting the null result of Schulze and Tillmann (2013). Whereas in the latter study no differences in recognition accuracy between timbre sequences of length 5 and 6 were observed, the current experiment obtained a strong effect.

The current results further touch on the role of music training and its relation to concurrent pitch variability in timbre sequence recognition. Because performance differed only marginally between musicians and nonmusicians in the constant-pitch condition, the significant main effect of music training was at least partially driven by the relatively strong decrease of nonmusicians' performance in the variable-pitch condition. This observation suggests that the musician advantage in timbre recognition might be not very marked as far as timbre is concerned in isolation. However, as soon as concurrent pitch variability enters the picture—requiring listeners to disentangle pitch and timbre in short-term memory—musicians exhibit clearly better memory fidelity than nonmusicians.

In this context, it should be acknowledged that our set of three pitches (D♭4, E♭4, F4) was very small (although it comprises some of the most frequently occurring musical intervals, see e.g., Huron, 2006, Chapter 5). Drawing from larger pitch sets may have created stronger effects, also for musicians. Given the size of the pitch set, the question yet remains whether there are situations in which musicians show impaired timbre recognition due to interference by pitch. For instance, if pitch patterns changed across standard and comparison sequences, would musicians' timbre matching accuracy still be unaffected? This question was considered next, using an experiment specifically designed for musicians with an even harder variable-pitch condition than Experiment 1.

Experiment 2: Length and Pitch Variability

To further explore the effect of concurrent variation in pitch on memory for timbre sequences in musicians, we presented distinct pitch patterns for standard and comparison sequences and compared this condition to a constant-pitch baseline. As in Experiment 1, sequences of lengths 4, 5, and 6 were presented. With the more complex type of pitch variability, we now expected musicians to show lower performance in the variable pitch condition.

METHOD

Participants. Twenty-two normal-hearing musicians were recruited over mailing lists of the Schulich School of Music at McGill University. None of them had participated in the previous experiment. Based on pilot data, we presumed that a smaller number of participants (reduced by around one fourth compared to Experiment 1) would suffice for this experiment. The group had a mean age of 26 years (*SD* = 7.6, range: 19–51), included 8 females, and featured an average of 18 years (*SD* = 7.8) of instrumental training and 5 years (*SD* = 3.4) of formal music-theoretical instruction or ear training.

Stimuli. There were 60 constant-pitch sequences (E♭4) and 60 variable-pitch sequences, both resulting from 20 sequences at each of the three lengths 4, 5, and 6. These 20 sequences per length condition were obtained from a subset of ten sequences from Experiment 1 that were here presented once with identical and once with non-identical comparison sequences with a swap of the last and third-to-last items. The variable-pitch condition was constructed as follows: As in Experiment 1, the pitch set D♭, E♭, F was used and two random permutations of these three pitches were interleaved. Contrary to Experiment 1, we now used different pitch patterns for the standard and comparison sequences. We did not allow pairs of standard and comparison sequences $P_1^S, \ldots, P_L^S$ and $P_1^C, \ldots, P_L^C$, to have pitch progressions that paralleled the potential swap of last and third-to-last timbres, i.e., we discarded pairs for which both $P_{L\text{-}2}^S = P_L^C$ and $P_L^S = P_{L\text{-}2}^C$. Finally, in order to enhance the contrast between standard and comparison, we selected pairs of pitch sequences that had a fairly high edit distance (or "Levenshtein Distance," LD), which measures the minimum number of single-item edits (insertion, deletion, substitution) needed to transform one sequence into another. To transform "123" into "321," for instance, one requires at least two replacements, yielding an LD of two. We selected standard-comparison pairs of six items, whose LD equaled five (n being the maximum LD for sequences of length n), before truncating pitch templates to the appropriate length of four to six items.

Procedure. The experiment featured the within-subject factors of length (4, 5, 6) and pitch variability (constant, variable). Each condition contained 20 sequences with 50% identical and 50% nonidentical trials, yielding 120 trials overall. The two levels of the pitch factor were presented in two blocks and their order was counterbalanced across participants. Otherwise, the procedure was identical to Experiment 1.

RESULTS

Overall performance was greater for constant-pitch ($M = .85$) compared to variable-pitch sequences ($M = .78$) and decreased monotonically with length ($M = .85, .83, .77$). Moreover, the difference across pitch conditions was most pronounced for five-tone sequences. Figure 3 displays all condition means.

A within-subject Pitch Variability (2) × Length (3) ANOVA confirmed significant main effects of pitch, $F(1, 21) = 12.75$, $p = .002$, $\eta_p^2 = .38$, and length, $F(2, 42) = 12.78$, $p < .001$, $\eta_p^2 = .38$. Post hoc tests showed that the effect of length was due to significant differences of length 6 from lengths 4 and 5, paired $t(21) > 3.48$ $p < .002$, whereas the difference between lengths 4 and 5 did not reach significance, $t(21) = 1.47$, $p = .15$.

The effect of pitch was qualified by an interaction between pitch and length, $F(2, 42) = 4.72$, $p = .014$, $\eta_p^2 = .18$, which was due to only insignificant differences across pitch condition for 4- and 6-item sequences, paired $t(21) < 2.02$, $p > .056$, but significant differences for sequences of length 5, $t(21) = 5.43$, $p < .001$.

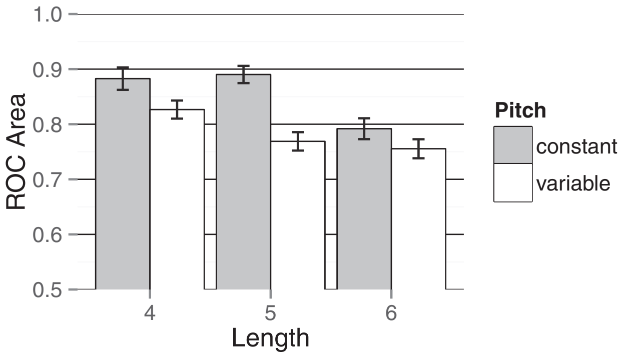

FIGURE 3. Experiment 2: Performance of musician participants for the factors sequence length and pitch variability. Error bars: 95% CI.

DISCUSSION

Testing a group of musicians, the experiment replicated the main effect of length from Experiment 1. The order of the timbres of six sounds was significantly harder to match compared to four or five sounds. We further observed a significant main effect of pitch variability by using a variable condition with distinct pitch patterns for standard and comparison sequences. Here the accuracy of the musician participants was lower for sequences with variable pitch compared to the constant-pitch baseline.

The interaction between pitch and length complicates the interpretation of the effect of pitch, because the effect appears to be partially driven by differences across pitch conditions for five-tone sequences. Because the construction of sequences was identical across lengths, there is no straight-forward explanation for this circumstance. Nonetheless, this observation indicates that constant pitch sequences with five tones appear to be particularly sensitive test items, because the addition of another tone or the addition of pitch variability yielded drastic drops in performance.

Comparing the constant-pitch condition of Experiment 2 to the corresponding trials in Experiment 1, musicians in Experiment 1 tended to score lower (M = .80) compared to the musicians in Experiment 2 ($M = .85$), $t(50) = -1.98$, $p = .054$. We did not find any biographical factors related to music training that accounted for this trend. This difference in performance across experiments implies that the main effect of music training in Experiment 1 would have been much stronger with the population of musicians from Experiment 2. More generally, these findings suggest that well-trained musicians—music students/professionals with a mean age of 26 years and a mean of 18 years of instrumental instruction—are not immune to cross-channel interference by pitch in STM for timbre, if pitch patterns vary across the timbre patterns to be matched (i.e., if pitch patterns evolve over time). Note that in the simpler scenario of Experiment 1 with pitch patterns fixed across standard and comparison sequences, there was only interference for nonmusicians, but not for musicians. We thus conclude that both musical expertise and the complexity of the concurrent pitch pattern affect listeners' memory fidelity for timbre sequences.

Experiment 3: Similarity and Position

The third experiment was designed to specifically explore the potential role of timbre similarity relations between the tones that were swapped. We expected better accuracy for swaps with more dissimilar pairs. The length of sequences was held fixed with four tones. We further included a factor that varied the serial position of the swap, which allowed us to observe whether similarity effects could be specific to certain serial positions and which would additionally provide some context for the swap at one fixed position in the previous two experiments. In contrast to the factor of pitch variability in Experiment 1, we did not assume the dissimilarity factor to differentially affect musicians compared to nonmusicans. Therefore, only groups of musicians were tested in Experiments 3 and 4.

METHOD

Participants. Twenty-two normal-hearing musicians (9 female) with an average age of 24 years ($SD = 4.3$, range: 19–36) participated. They had received an average of 15 years ($SD = 4.9$) of instrumental training and 4 years ($SD = 2.9$) of formal music-theoretical instruction including ear training. None of the listeners had participated in either of the two previous experiments.

Stimuli. This experiment used the same tones as before, held at constant pitch, and concatenated to sequences of length 4. Four different swap positions were employed: 1 & 2, 2 & 3, 3 & 4, 2 & 4.

For the swap, we used each of the $(7 \times 8)/2 = 28$ possible pairs, given our set of eight sounds. The timbral dissimilarity of swap (TDS) of the pairs was given by the raw mean dissimilarity ratings (on a scale from 1–9) taken from McAdams et al. (1995). The TDS factor partitioned the full range of TDS as obtained from the 28 pairings described above into a lower and upper half, such that the factor's first level comprised the 14 swaps of low TDS, and the second comprised the 14 pairs with high TDS values.

The remaining two items per sequence were chosen randomly without replacement from the resulting set of sounds. A set of sequences was constructed by utilizing each pair of items A-B in both orders (e.g., C-A-B-D and C-B-A-D), yielding 2 (order) × 28 (pairs) × 4 (position of swap) = 224 sequences in total. In every condition, half of the sequences were presented with identical and half with nonidentical comparison sequences.

Procedure. This experiment featured a within-subject design with factors of swap position (4 levels) and timbral dissimilarity of swap (TDS, 2 levels). The 224 trials were presented in fully randomized order, partitioned into two blocks of around 22 min duration each. Otherwise, the procedure was identical to the previous experiments.

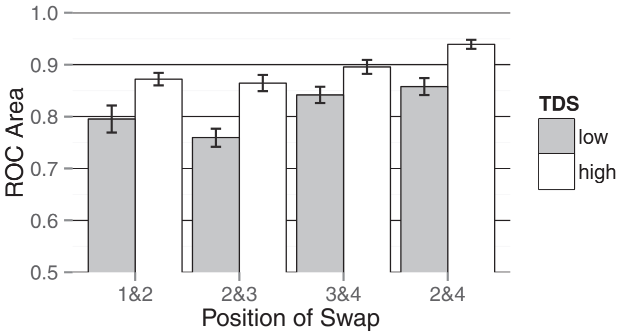

FIGURE 4. Experiment 3: Recognition performance for the main effects of swap position and timbral dissimilarity of swap (TDS). Error bars: 95% CI.

RESULTS

Performance was worse for low TDS (M = .81) compared to high TDS (M = .89), in accordance with the hypothesized role of dissimilarity. The four different positions of swaps (1 & 2, 2 & 3, 3 & 4, and 2 & 4) yielded averages of $M = .83$, .81, .87, and .90, respectively. This means that performance was particularly good for swaps that included the last serial position. Figure 4 shows all condition means.

A within-subject TDS (2) × Position (4) ANOVA revealed main effects of TDS, $F(1, 21) = 45.87$ $p < .001$, $\eta_p^2 = .69$, and of Position, $F(3, 63) = 10.59$ $p < .001$, $\eta_p^2 = .34$.

There was no interaction between TDS and Position, F (3, 63) < 1. Post hoc comparisons showed that the effect of position was due to significantly better performance for positions including the last tone (i.e., 2 & 4 vs. 1 & 2, 3 & 4 vs. 2 & 3, and 2 & 4 vs. 2 & 3), $t(21) < -3.3$, $p < .001$, but no significant differences otherwise, $|t(21)| > 2.80$, $p > .01$ ($\alpha_{crit} = .0083$).

DISCUSSION

This experiment demonstrated the important role of the timbral dissimilarity of swapped items (TDS). As a memory similarity effect, this finding may be considered to parallel the phonological similarity effect for verbal material in serial recognition (Nimmo & Roodenrys, 2005): highly dissimilar swaps yield better recognition performance than similar ones. Similarity factors have similarly been proven to be crucial for nonverbal auditory, as well as visual, short-term memory (e.g., Kahana & Sekuler, 2002; Visscher et al., 2007). In order to satisfactorily describe response choices, however, these studies were also required to consider an additional property of the presented memory sequences, namely their heterogeneity (i.e., the average

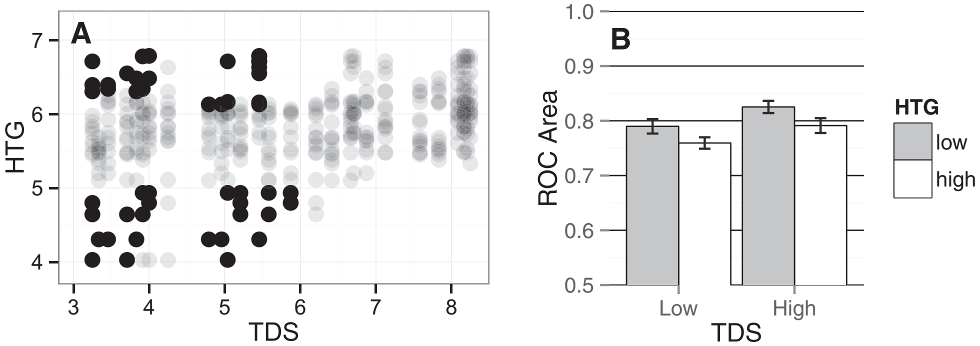

FIGURE 5. A) Timbral dissimilarity of swap (TDS) and heterogeneity (HTG) for the 12 sequences per condition selected for Experiment 4 (black dots) and all other possible sequences composed (without replacement) with four out of the eight sounds (gray dots). B) Recognition performance as a function of timbral dissimilarity of swap (TDS) and heterogeneity (HTG) averaged across all 26 participants. Error bars: 95% CI.

dissimilarity between all items of a sequence). In a final experiment, we therefore explored the potential role of heterogeneity in our current timbre sequence recognition scenario. This also allowed us to reassess the robustness of the effect of TDS.

The experiment further demonstrated a main effect of serial position of swap, which was due to increased performance for sequences with swaps occurring at final positions (2 & 4, 3 & 4), indexing a recency effect in serial recognition of timbre. We note that the 2 & 4 condition in this experiment was equivalent to the swaps from Experiments 1 and 2, thus providing some perspective on these previous choices. From a formal standpoint, it could be argued that the constancy of the position of the swap in Experiments 1 and 2 could have allowed participants to optimize their strategy by only focusing on the two relevant positions (once they would have noticed this circumstance over the course of the experiment). Yet, a comparison with the data from Experiment 3, where swaps were equally distributed, renders this stance implausible. In fact, there were no significant differences between scores from the 2 & 4 condition in Experiment 3 ($M = .90$) and the corresponding condition with constant pitch and sequence length 4 from Experiment 2 ($M = .88$), two-sample $t(42) = -0.75$, $p = .46$. Mean scores of the corresponding condition from Experiment 1 ($M = .83$) were even significantly smaller compared to the mean of the swap position 2 & 4 in Experiment 3, $t(50) = 2.85$, $p = .006$, which speaks against any marked optimization of response strategies for the constant positions of swaps in Experiments 1 and 2. This stance is further supported by the strong effect of length in both experiments, which is equally unlikely in conjunction with participants' attention distributed selectively to only one or two serial positions.

Experiment 4: Similarity and Heterogeneity

This experiment studied the role of sequence heterogeneity (HTG) and timbral dissimilarity of swapped items (TDS) in a 2 × 2 factorial design. We expected higher scores for high TDS trials and lower scores for high HTG trials (Visscher et al., 2007).

METHOD

Participants. Twenty-six normal-hearing musicians (11 female) with an average age of 23 years ($SD = 3.1$, range: 19–28) participated. They had received an average of 14 ($SD = 3.8$) years of instrumental training and 4 ($SD = 3.4$) years of formal music-theoretical instruction including ear training. None of the participants had participated in any of the previous experiments. As in Experiments 2 and 3, we ran this experiment with 22 participants initially. Because we observed that there were several chance performers (see Results below), we decided to add a few more participants (ca. 25%) in order to further stabilize the experimental results.

Stimuli. The interest in this experiment was to independently manipulate the two factors of timbral dissimilarity of swap (TDS) and timbral heterogeneity (HTG), each with two levels. Because both variables are correlated (e.g., an increase in TDS implies an increase in HTG), sequences had to be selected carefully in order to guarantee an independent factor design. Figure 5A graphs TDS and HTG values of all possible four-item sequences based on the eight sounds in use. It also shows the 12 sequences per condition that were selected for the current experiment. Low TDS sequences ranged between 3.2 and 4.0 dissimilarity units, high TDS sequences between 4.8 and 5.9. Low HTG sequences ranged between 4.0 and 4.9 units, high HTG from 6.1 to 6.8. None of the TDS and HTG distributions from the sub-conditions (e.g., TDS-low x HTG-low) differed significantly on their corresponding factors (i.e., TDS did not differ for TDS-low x HTG-low as compared to TDS-low x HTG-high), as indicated by two-sample *t*-tests, all $p > .45$. We used the swap position 2 & 3, and the pair to be swapped occurred in both orders as part of standard sequences (i.e., ABCD and ACBD). All sequences were presented both with identical and nonidentical comparison sequences. This yielded 12 (sequences) × 2 (low and high TDS) × 2 (low and high HTG) × 2 (order of pair) × 2 (same/different) = 192 trials overall.

Procedure and design. The order of trials was fully randomized in two experimental blocks of around 20 min duration each. Otherwise, the procedure was identical to that of Experiments 1–3.

RESULTS

Both factors generated rather small differences in performance. Performance was worse for low TDS ($M = .77$) compared to high TDS ($M = .81$), as expected. Scores were higher in the low HTG condition ($M = .81$) compared to the high HTG condition ($M = .78$), respectively. Figure 5B depicts the results.

A repeated-measures TDS (2) × HTG (2) ANOVA indicated that scores were affected by TDS, $F(1, 25) = 8.64$, $p < .007$, $\eta_p{}^2 = .26$, but only marginally by HTG, $F(1, 25) = 4.08$, $p = .05$, $\eta_p{}^2 = .14$, with no interaction, $F(1, 25) < 1$.

Note that an analysis of the dependent variable of proportion of "same" responses (as considered in the original homogeneity studies, cf., Kahana & Sekuler, 2002) did not yield different results for the HTG factor: there was a significant effect of TDS, $F(2, 50) = 77.12$, $p < .0001, \eta_p{}^2 = .75$, but neither an effect of HTG, $F(1, 25) = 0.10$, $p = .75$, nor an interaction, $F(2, 50) = 1.78$, $p = .17$. Closer inspection of the data suggested that the trend of HTG was likely driven by only a few participants: the overall average performance of participants was distributed bimodally, with mean scores of $M = .64$ ($SD = .06$) for a low-performance group of nine participants and $M = .88$ ($SD = .05$) in a high-performance group, without overlap of distribution. These groups differed significantly as indicated by a two-sample *t*-test, $t(24) = 10.94$, $p < .0001$. We did not find biographical factors that accounted for this gap. Specifically, three participants from the low-performance group scored at around chance level in the high HTG condition. After removal of these three participants, the trend of HTG vanished, $F(1, 22) = 0.90$, $p = .35$, $\eta_p{}^2 = .04$, but the effect of TDS remained stable, $F(1, 22) = 13.86$, $p = .001$, $\eta_p{}^2 = .39$.

DISCUSSION

This experiment confirmed the role of the timbral dissimilarity of the swap (TDS) as an essential variable allowing participants to distinguish between identical and nonidentical trials. At the same time, the experiment casts doubt on the relevance of sequence heterogeneity (HTG) in the current paradigm. HTG only affected the performance of three out of 26 participants, who generated the overall marginally significant effect of HTG. After removal of these participants from the analysis, however, HTG failed to affect performance significantly. The next section analyses the role of TDS and HTG in determining response choice in more detail.

Response Choice and Timbre Dissimilarity

In this last analysis, we reconsider the role of the timbral distance of swap (TDS) and heterogeneity (HTG) across all four experiments. The factorial designs of Experiments 3 and 4 partitioned a wide range of continuously distributed TDS and HTG values into two levels. Beyond such coarse distinctions, however, there could also be more fine-grained effects, which we attempt to explore here. Instead of analyzing recognition accuracy via the ROC area that requires an accumulation of responses from several identical and non-identical trials, we now consider response choice; that is, the proportion of "same" responses as the dependent variable of interest. As in other studies of recognition memory (e.g., Visscher et al., 2007), this allows us to analyze the relation of response choices and similarity on a per-trial basis.

Specifically, we only considered four-item sequences at constant pitch for Experiments 1 and 2, and used all trials from Experiments 3 and 4. In order to increase the signal-to-noise ratio of the choice data with regards to the factor of interest, i.e., similarity, we averaged data over all identical TDS × HTG conditions, such that there are no two data points that have identical (TDS, HTG) pairs (models for Experiments 1–4 contained 29, 19, 72, 84 points, respectively).

Table 1 shows the estimated regression coefficients for all four experiments. All experiments achieve a good fit with roughly 90% explained variance in the choice data. However, only in Experiment 3 is there a significant contribution from HTG, and in Experiment 1 there only is a marginally significant impact. In Experiments 2 and 4, however, HTG does not make a significant contribution to the regression. At the same time, absolute values of standardized β coefficients are roughly an order of magnitude lower for HTG compared to TDS, again reflecting its significantly inferior predictive power. Figure 6 shows the corresponding scatter plots of response

TABLE 1. ***Multiple Linear Regression Results for Experiments 1–4 with Timbral Dissimilarity of Swap (TDS) and Heterogeneity (HTG) as Independent Variables***

	Variable	*B*	*SE B*	β	*p*
Experiment 1	Intcpt.	.486	.108	1.86	< .0001
(R^2 = .92)	TDS	−.079	.005	−.953	< .0001
$n = 60$	HTG	.037	.018	.010	.058
Experiment 2	Intcpt.	.140	.468	0.425	.769
(R^2 =.90)	TDS	−.092	.008	−.940	< .0001
$n = 22$	HTG	.092	.080	.091	.262
Experiment 3	Intcpt.	.431	.117	1.413	< .0001
(R^2 =.90)	TDS	−.095	.004	−.958	< .0001
$n = 22$	HTG	.055	.021	.100	.009
Experiment 4	Intcpt.	.729	.058	3.011	< .0001
(R^2 =.89)	TDS	−.101	.004	−.940	< .0001
$n = 26$	HTG	.005	.009	.020	.585

Note: For all four experiments, response choice probability acts as dependent variable. Leftmost column: proportion of variance explained and number of participants.

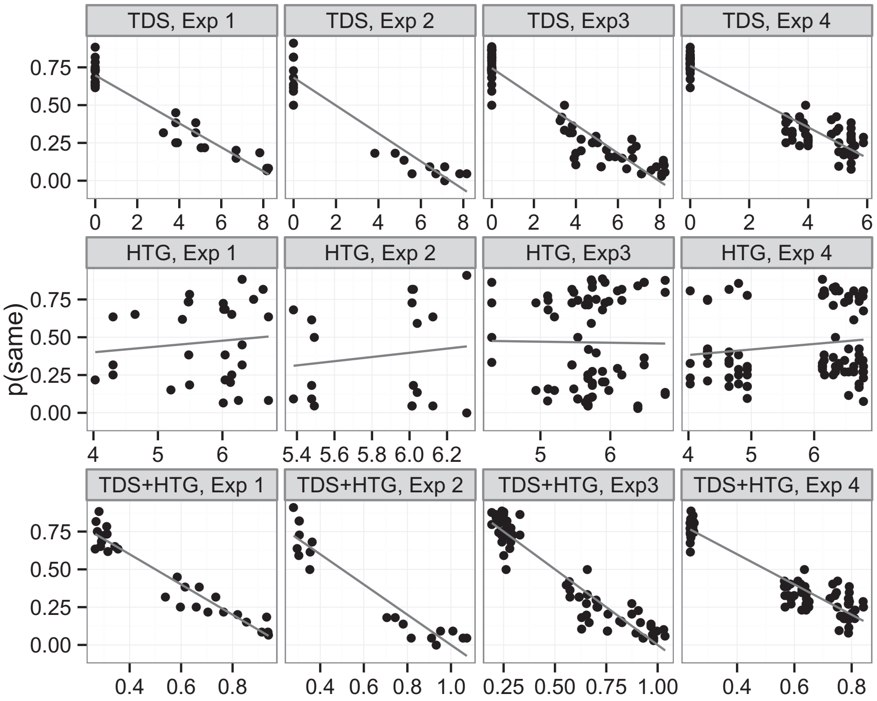

FIGURE 6. Experiments 1–4: Timbral dissimilarity of swap (TDS, top row), heterogeneity (HTG, middle row), and their linear combination (TDS+HTG, bottom row) as predictors of response choice (probability of "same").

choice and both variables (top and middle row) as well as their best linear combination (bottom row). Again, it is hard to visually recognize any consistent trends based on HTG.

Even more important is the fact that stepwise multiple regression did not enter HTG into the model for any of the data from the four experiments (the deviations of the resulting parameter estimates are only marginal as compared to those listed in Table 1 and are thus omitted for the sake of brevity). The R^2 value in these univariate models is .89 for Experiments 2, 3, and 4, and .91 for Experiment 1, i.e., not more than 1% below the proportion of variances explained by the full models as listed in Table 1. A parsimonious account of response behavior thus does not require HTG.

General Discussion

The current series of experiments provides a comprehensive picture of short-term recognition of musical timbre sequences. Four experiments explored how the cognitive sequencing of timbre is affected by music training (Experiment 1), concurrent pitch variability (Experiments 1 & 2), and similarity (Experiments 3 & 4). It was shown that musicians exhibited slightly better performance, but mainly outperformed nonmusicians for concurrent variability in pitch. It was further shown that the timbral dissimilarity of the swapped sounds predicted a large majority of response choices across all four experiments.

The results regarding the differences across the groups of musicians and nonmusicians contribute to a diverse set of findings in the literature. In timbre dissimilarity, no consistent differences in the ratings of musicians and nonmusicians have been found (Lakatos, 2000; McAdams et al., 1995). Using an interpolated tone task, Starr and Pitt (1997) neither observed an effect of interference of pitch nor an effect of music training in STM for timbre. This stands in contrast to discrimination tasks (Chartrand & Belin, 2006) and enhanced neurophysiological responses of musicians (Pantev et al., 2001; Shahin et al., 2008; Strait et al., 2012). Furthermore, a recent study observed a clear advantage for musicians in an item recognition task (Siedenburg & McAdams, 2017b) with recorded sounds from acoustic instruments. The present data do not show reliable differences between musicians and nonmusicians when pitch is constant. Differences between musicians' and nonmusicians' timbre processing may thus be subtle and highly task-dependent. A role could also be played by the stimuli used, which were quite different from the timbres (mostly originating from acoustic instruments) that musicians typically deal with on a daily basis. The demonstration of impaired timbre recognition for variable-pitch sequences in Experiment 2 adds another aspect to this issue, because it shows that musicians' memory fidelity can also be impaired by the presence of concurrent pitch patterns that evolve over time.

The presented similarity effects extend results from Starr and Pitt (1997) and Visscher et al. (2007) to serial recognition. In short, our observations imply that similarity relations play an integral role in STM for timbre. A correlation analysis of trial-wise data from all experiments showed that TDS, the timbral dissimilarity of swapped items, is a good predictor of response choice in serial recognition. This single variable predicted the greatest portion of variance of response choices throughout all four experiments. To the best of our knowledge, this is the first study on auditory STM that has introduced a parsimonious and parametric notion of similarity for serial recognition.

Moreover, we tested the homogeneity-computation hypothesis (Viswanathan et al., 2010), which suggests that short-term recognition memory for multiple stimuli is affected by the items' mutual similarity. In the current data, we could not observe any strong effect of homo/heterogeneity. There could be a multitude of factors that explain why studies starting with Kahana and Sekuler (2002) have demonstrated strong effects of homogeneity in visual and auditory STM, and why this does not extend to the current scenario (cf., Nosofsky, Little, Donkin, & Fific, 2011). Among these factors could be differences in the employed tasks, because serial recognition probes memory for serial order and not item identity as is the case in item recognition. It has been argued that both types of memory signals rely on different mnemonic mechanisms (Henson, Hartley, Burgess, Hitch, & Flude, 2003), which could constrain the utility of sequence heterogeneity to the item identity case. Yet, given the insignificant role of HTG in a recent experiment using an item identity task (Siedenburg & McAdams, 2017b), task differences are unlikely to be the only factor at play. Another aspect concerns the confusability of items that are used in typical heterogeneity studies, such as Gabor patches in vision or moving ripples in audition (cf. Visscher et al., 2007). Here, we used clearly distinguishable sounds with timbres that varied on multiple perceptual dimensions (as opposed to a single one). This distinctiveness, which would set apart sounds, might be a case in which homogeneity does not become relevant.

The TDS variable also enables us to conceptualize the matching process for timbre sequences. There are two potential strategies that can be distinguished. On the one hand, there may be a strategy to match complete, integrated sequences as *Gestalts*. On the other hand, participants could also match sequences using an item-by-item strategy. In the current experiments, response choices could be well predicted by the summed item-by-item dissimilarities (which are all zero apart from the two items that were swapped, i.e., yielding TDS). However, this variable could equally be a strong correlate of any integrated, sequence-wise distance measure, and we cannot therefore distinguish between these two hypotheses at present. Listening strategies could have depended on characteristics of the listeners (e.g., music training) as well as the experimental situation (e.g., the presence or absence of pitch variability).

In particular, a solely item-wise strategy cannot explain the effect of pitch variability for nonmusicians

(Experiment 1), because the pitch templates were constant across standard and comparison sequences, and therefore there was no item-wise difference in pitch that could have hampered the computation of timbral difference. For nonmusicians, it thus seems plausible to posit a sequence-wise discrimination process with cross-channel interference from pitch to timbre (Melara & Marks, 1990). In the variable-pitch condition of Experiment 1, the (match) result of the pitch-sequence discrimination then impairs the discrimination in the target attribute (timbre). To the contrary, the results obtained for musicians may be better explained by item-wise strategies, which, particularly in the variable-pitch condition, may have better allowed timbre to be isolated from pitch on a local level (what Jones & Boltz, 1989, called "analytic attending"). Musicians' performance may not have differed between constant and variable-pitch conditions in Experiment 1, because with an item-wise strategy, the lack of item-wise pitch differences did not impair the computation of timbral difference. In Experiment 2, the item-wise differences in pitch could have interfered with item-wise matching of timbre (because here pitch patterns differed across standard and comparison sequences). Needless to say, further experimentation is required to develop and scrutinize this hypothesis about the matching process's dependency on experimental scenario and musical expertise.

Overall, our findings resonate with what Sekuler and Kahana (2007) have dubbed the "stimulus-oriented approach to memory" that emphasizes the interrelatedness of sensory representation and short-term recognition. As the authors note, "But when memory models fail to link their stimulus representations to measures of perceptual similarity, they needlessly limit their ability to account for a variety of important phenomena" (p. 305). Together with the current results, considerations such as these imply that it would be hazardous to neglect similarity relations in future studies on short-term recognition of timbre, even if the employed stimuli are easily discriminable. On the contrary, if perceptual similarity is part of the experimental design, similar effects emerge across domains as diverse as musical timbre, musical pitch (Williamson, Baddeley, & Hitch, 2010), words and non-words (Nimmo & Roodenrys, 2005), and even auditory ripple noise and visual Gabor patches (Visscher et al., 2007).

Regarding the three variables of interest, pitch variability, music training, and perceptual similarity, our results can thus be seen as natural extensions of findings from work on perceptual processing. This characterization favors the view of short-term memory not as a dedicated neural system (Baddeley, 2003), but as an active, top-down-type of trace maintenance that is based on sensory recruitment, i.e., the dedication of attention to sensory representations (D'Esposito & Postle, 2015). Expressed in other words, our results point towards a sensory-cognitive continuum (cf. Collins, Tillmann, Barrett, Delbé, & Janata, 2014), in which the faculty of memory for timbre naturally grows on the basis and the properties of sensory representation, rather than being one of many separate "cognitive shoe-boxes" for the retention of modality-specific information.

It is curious to note that our results regarding pitch variability constitute an interesting analogy to practices in 20th century music composition. In fact, many composers who wished to draw their listeners' attention towards timbral structures often drastically reduced concurrent complexity in pitch. Classical examples include Schoenberg's archetypal Klangfarbenmelodie, *Five pieces for orchestra*, Op. 16, No. 3 ("Farben"), as well as works by Giacinto Scelsi, Tristan Murail's *Mémoire/Erosion*, and many others (cf. Erickson, 1975; Murail, 2005). One can also observe that the focus on sound qualities in popular music has led to pitch structures that, at times, almost constitute a diminutive feature (see, e.g., Osborn, 2016, for examples from the rock group *Radiohead*). In this study, we touched on potential cognitive undercurrents of this facet of composition practice: we observed that when complex sound structures exhibit variability in pitch, nonmusicians and musicians (albeit to a lesser extent) have difficulties tracking timbral changes over time. In this sense, the musical discourse is limited by cognitive constraints, because listeners are not able to recognize arbitrarily subtle or complex configurations of auditory attributes. The discussed role of timbral dissimilarity further indicates that timbral possibilities can be (and likely are) used strategically by composers and music producers, following the basic principle that perceptually similar timbral changes are difficult to track in memory. Expressed in other words, timbral diversity may be a good predictor of how well sounds' timbral structures can be recognized by listeners and ultimately contribute to the experience of musical form.

Author Note

The authors wish to thank the anonymous reviewers for insightful remarks on earlier versions of this manuscript. This work was supported by a Carl von Ossietzky Young Researchers' Fellowship from the

University of Oldenburg awarded to Kai Siedenburg, and a grant from the Canadian Natural Sciences and Engineering Research Council (RGPIN 2015-05280) and a Canada Research Chair (950-223484) awarded to Stephen McAdams.

Correspondence concerning this article should be addressed to Kai Siedenburg, Department of Medical Physics and Acoustics, Carl von Ossietzky University of Oldenburg, Küpkersweg 74, 26129 Oldenburg, Germany. E-mail: kai.siedenburg@uni-oldenburg.de

References

Allen, E. J., & Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *Journal of the Acoustical Society of America*, *135*(3), 1371–1379.

Alluri, V., & Toiviainen, P. (2012). Effect of enculturation on the semantic and acoustic correlates of polyphonic timbre. *Music Perception*, *29*, 297–310.

Alunni-Menichini, K., Guimond, S., Bermudez, P., Nolden, S., Lefebvre, C., & Jolicoeur, P. (2014). Saturation of auditory short-term memory causes a plateau in the sustained anterior negativity event-related potential. *Brain Research*, *1592*, 55–64.

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839.

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29.

Caruso, V. C., & Balaban, E. (2014). Pitch and timbre interfere when both are parametrically varied. *PLoS ONE*, *9*(1), e87065.

Chartrand, J.-P., & Belin, P. (2006). Superior voice timbre processing in musicians. *Neuroscience Letters*, *405*(3), 164–167.

Collins, T., Tillmann, B., Barrett, F. S., Delbé, C., & Janata, P. (2014). A combined model of sensory and cognitive representations underlying tonal expectations in music: From audio signals to behavior. *Psychological Review*, *121*(1), 33–65.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*(28), 1–28.

Deutsch, D. (1970). Tones and numbers: Specificity of interference in immediate memory. *Science*, *168*(3939), 1604–1605.

Erickson, R. (1975). *Sound structure in music*. Berkeley, CA: University of California Press.

Golubock, J. L., & Janata, P. (2013). Keeping timbre in mind: Working memory for complex sounds that can't be verbalized. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 399–412.

Handel, S. (1995). Timbre perception and auditory object identification. In B. C. Moore (Ed.), *Hearing* (Vol. 2, pp. 425–461). San Diego, CA: Academic Press.

Handel, S., & Erickson, M. L. (2001). A rule of thumb: The bandwidth for timbre invariance is one octave. *Music Perception*, *19*, 121–126.

Helmholtz, H. L. F. von (1954). *On the sensations of tone* (A. J. Ellis, Trans.). New York: Dover. (Original work published 1877)

Henson, R., Hartley, T., Burgess, N., Hitch, G., & Flude, B. (2003). Selective interference with verbal short-term memory for serial order information: A new paradigm and tests of a timing-signal hypothesis. *Quarterly Journal of Experimental Psychology: Section A*, *56*(8), 1307–1334.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review*, *96*(3), 459–491.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, *59*, 193–224.

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research*, *42*(18), 2177–2192.

Kendall, R. A., Carterette, E. C., & Hajda, J. M. (1999). Perceptual and acoustical features of natural and synthetic orchestral instrument tones. *Music Perception*, *16*, 327–363.

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

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception and Psychophysics*, *62*(7), 1426–1439.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Marozeau, J., de Cheveigné, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America*, *114*(5), 2946–2957.

Martin, F. N., & Champlin, C. A. (2000). Reconsidering the limits of normal hearing. *Journal of the American Academy of Audiology*, *11*(2), 64–66.

McAdams, S. (1989). Psychological constraints on form-bearing dimensions in music. *Contemporary Music Review*, *4*, 181–198.

McAdams, S. (2013). Musical timbre perception. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 35–67). San Diego, CA: Academic Press.

McAdams, S., & Goodchild, M. (2017). Musical structure: Sound and timbre. In R. Ashley & R. Timmers (Eds.), *The Routledge companion to music cognition* (pp. 129–139). New York: Routledge.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*(3), 177–192.

Melara, R. D., & Marks, L. E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception and Psychophysics*, *48*(2), 169–178.

Micheyl, C., & Dai, H. (2008). A general area theorem for the same-different paradigm. *Attention, Perception, and Psychophysics*, *70*(5), 761–764.

Murail, T. (2005). The revolution of complex sounds. *Contemporary Music Review*, *24*(2/3), 121–135.

Nattiez, J.-J. (2007). Le timbre est-il un paramètre secondaire? [Is timbre a secondary parameter?]. *Cahiers de la Société Québécoise de Recherche en Musique*, *9*(1–2), 13–24.

Nimmo, L. M., & Roodenrys, S. (2005). The phonological similarity effect in serial recognition. *Memory*, *13*(7), 773–784.

Nolden, S., Bermudez, P., Alunni-Menichini, K., Lefebvre, C., Grimault, S., & Jolicoeur, P. (2013). Electrophysiological correlates of the retention of tones differing in timbre in auditory short-term memory. *Neuropsychologia*, *51*(13), 2740–2746.

Nosofsky, R. M., & Kantner, J. (2006). Exemplar similarity, study list homogeneity, and short-term perceptual recognition. *Memory and Cognition*, *34*(1), 112–124.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, *118*(2), 280–315.

Osborn, B. (2016). *Everything in its right place: Analyzing radiohead*. Oxford, UK: Oxford University Press.

Pantev, C., Roberts, L. E., Schulz, M., Engelien, A., & Ross, B. (2001). Timbre-specific enhancement of auditory cortical representations in musicians. *Neuro Report*, *12*(1), 169–174.

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.

Schoenberg, A. (1978). *Theory of harmony [Harmonielehre]* (R. E Carter, Trans.). Berkeley, CA: University of California Press. (Original work published 1911)

Schulze, K., & Tillmann, B. (2013). Working memory for pitch, timbre, and words. *Memory*, *21*(3), 377–395.

Schulze, K., Zysset, S., Mueller, K., Friederici, A. D., & Koelsch, S. (2011). Neuroarchitecture of verbal and tonal working memory in nonmusicians and musicians. *Human Brain Mapping*, *32*(5), 771–783.

Sekuler, R., & Kahana, M. J. (2007). A stimulus-oriented approach to memory. *Current Directions in Psychological Science*, *16*(6), 305–310.

Shahin, A. J., Roberts, L. E., Chau, W., Trainor, L. J., & Miller, L. M. (2008). Music training leads to the development of timbre-specific gamma band activity. *Neuroimage*, *41*(1), 113–122.

Siedenburg, K., & Dörfler, M. (2013). Persistent time-frequency shrinkage for audio denoising. *Journal of the Audio Engineering Society (AES)*, *61*(1/2), 29–38.

Siedenburg, K., Fujinaga, I., & McAdams, S. (2016). A comparison of approaches to timbre descriptors in music information retrieval and music psychology. *Journal of New Music Research*, *45*(1), 27–41.

Siedenburg, K., Mativetsky, S., & McAdams, S. (2016). Auditory and verbal memory in north indian tabla drumming. *Psychomusicology: Music, Mind, and Brain*, *26*(4), 327–336.

Siedenburg, K., & McAdams, S. (2017a). Four distinctions for the auditory "wastebasket" of timbre. *Frontiers in Psychology*, *8*, 1747.

Siedenburg, K., & McAdams, S. (2017b). The role of long-term familiarity and attentional maintenance in auditory short-term memory for timbre. *Memory*, *25* (4), 550–564. DOI: 10.1080/09658211.2016.1197945

Starr, G. E., & Pitt, M. A. (1997). Interference effects in short-term memory for timbre. *The Journal of the Acoustical Society of America*, *102*(1), 486–494.

Steele, K. M., & Williams, A. K. (2006). Is the bandwidth for timbre invariance only one octave? *Music Perception*, *23*, 215–220.

Strait, D. L., Chan, K., Ashley, R., & Kraus, N. (2012). Specialization among the specialized: Auditory brainstem function is tuned in to timbre. *Cortex*, *48*(3), 360–362.

Visscher, K. M., Kaplan, E., Kahana, M. J., & Sekuler, R. (2007). Auditory short-term memory behaves like visual short-term memory. *PLoS Biology*, *5*(3), e56.

Viswanathan, S., Perl, D. R., Visscher, K. M., Kahana, M. J., & Sekuler, R. (2010). Homogeneity computation: How interitem similarity in visual short-term memory alters recognition. *Psychonomic Bulletin and Review*, *17*(1), 59–65.

Wessel, D. L., Bristow, D., & Settel, Z. (1987). Control of phrasing and articulation in synthesis. In J. Beauchamp (Ed.), *Proceedings of the International Computer Music Conference* (pp. 108–116). San Francisco, CA: Computer Music Association.

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory and Cognition*, *38*(2), 163–175.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2015). An interlanguage unification of musical timbre. *Music Perception*, *32*, 394–412.