MEMORY, 2016
http://dx.doi.org/10.1080/09658211.2016.1197945

Routledge
Taylor & Francis Group

The role of long-term familiarity and attentional maintenance in short-term memory for timbre

Kai Siedenburg[a,b] and Stephen McAdams[a]

[a]Schulich School of Music, McGill University, Montreal, QC, Canada; [b]Department of Medical Physics and Acoustics, University of Oldenburg, Oldenburg, Germany

ABSTRACT
We study short-term recognition of timbre using familiar recorded tones from acoustic instruments and unfamiliar transformed tones that do not readily evoke sound-source categories. Participants indicated whether the timbre of a probe sound matched with one of three previously presented sounds (item recognition). In Exp. 1, musicians better recognised familiar acoustic compared to unfamiliar synthetic sounds, and this advantage was particularly large in the medial serial position. There was a strong correlation between correct rejection rate and the mean perceptual dissimilarity of the probe to the tones from the sequence. Exp. 2 compared musicians' and non-musicians' performance with concurrent articulatory suppression, visual interference, and with a silent control condition. Both suppression tasks disrupted performance by a similar margin, regardless of musical training of participants or type of sounds. Our results suggest that familiarity with sound source categories and attention play important roles in short-term memory for timbre, which rules out accounts solely based on sensory persistence.

ARTICLE HISTORY
Received 9 December 2015
Accepted 30 May 2016

KEYWORDS
Short-term memory; stimulus familiarity; timbre; levels of processing; attentionalmaintenance

Timbre refers to the auditory attributes that lend sounds a sense of "colour" and "shape" in time and enable the inference of sound sources. The percept emerges from acoustic cues such as the spectral envelope distribution, attack sharpness, spectrotemporal variation or modulation, roughness, and noisiness, in addition to features that may be idiosyncratic to certain instruments (McAdams, 2013; Siedenburg, Fujinaga, & McAdams, 2016a). Although timbre is a major component of audition, many facets of its mnemonic processing have only started to be explored. A question of particular concern for the present study is whether short-term timbre recognition is facilitated by long-term familiarity with sounds produced by acoustic instruments. If that was the case, it would cast doubt upon accounts that portray short-term memory (STM) for timbre as based on a "one-size-fits-all" principle of sensory persistence. A natural follow-up question would then be whether a memory advantage of familiar over unfamiliar sounds is due to differences in encoding or maintenance strategies. For instance, a simple explanation could state that prior knowledge of instrument categories support verbal labelling and in turn give rise to maintenance via verbal rehearsal. In order to approach these questions, we here compared the recognition of timbres from familiar musical instruments that afford verbal labelling with recognition of timbres from unfamiliar transformed sounds, the underlying sound sources of which were obscured by means of digital signal processing. Questioning the contribution of prior knowledge of instrument categories to short-term recognition not only opens a novel window into the mechanisms involved in memory for timbre, but may also inform emerging theories of non-verbal sensory memory in general (e.g., Jolicoeur, Levebre, & Martinez-Trujillo, 2015). In the following, we briefly provide a general background on STM, before outlining relevant experimental results on timbre with a specific focus on the role of instrument categories and potential maintenance strategies.

STM and lexicality

STM (here not specifically differentiated from *working* memory) is usually regarded as a memory system of limited capacity that decays within seconds. Although STM is often conceived as separate from long-term memory (LTM), there is good reason to assume strong interrelations between STM and LTM (see e.g., Jonides et al., 2008). A corresponding experimental cornerstone in the verbal domain, constituting a major portion of STM research in general, is the *lexicality effect*: STM for item identity is generally better for words than for closely matched *pseudo-words* (Thorn, Frankish, and Gathercole,

CONTACT Kai Siedenburg ✉ kai.siedenburg@uni-oldenburg.de Schulich School of Music, McGill University, 555 Sherbrooke Street West, Montreal, QC, Canada

2008). The latter are defined as vocables that respect phonotactic constraints of a language but are meaningless, that is, not part of the dictionary. Similar effects have also been shown for variables such as word frequency and imaginability (Thorn et al., 2008). Whether caused by greater activation strength, facilitated rehearsal, or more robust memory retrieval (cf., Macken, Taylor, & Jones, 2014; Thorn, Gathercole, & Frankish, 2002), these effects underline the importance of long-term knowledge and familiarity in verbal short-term remembering.

Another important characteristic trait of verbal STM is its reliance on active maintenance of the memory trace. Words may be rehearsed via recall and subvocal articulation, as described by the *phonological loop* component in Baddeley's influential model of working memory (Baddeley, 2012). In effect, verbal STM partly functions via a translation of auditory sensory representations into rehearseable sensorimotor codes (Schulze and Koelsch, 2012). A process called *attentional refreshing* has been proposed as an alternative form of active trace maintenance (Camos, Lagner, & Barrouillet, 2009). Refreshing emerges through the reactivation of a target's mental representation by means of attentional focusing (Cowan, 1988; Johnson, 1992). The target briefly reenters conscious awareness, whereby its representation is kept in an active state. The process has been shown to be independent of subvocalisation-based rehearsal (Camos et al., 2009) and is preferentially employed in verbal working memory tasks with low concurrent processing load (Camos, Mora, & Oberauer, 2011).

However, these findings on the contribution of prior knowledge and maintenance processes to short-term remembering all emerged for verbal STM, and it is currently unclear whether similar phenomena could be of relevance for timbre. The goal of this study was to explore this question.

Sound source categories and timbre familiarity

In contrast to verbal memory and perhaps also musical pitch structures, the cognitive processing of timbre has not been studied extensively. In fact, timbre has traditionally been treated as a primarily sensory phenomenon that resides "in the moment" and is not subject to long-term familiarisation. Neurophysiological studies on timbre processing have started to provide evidence for the contrary position. Pantev, Roberts, Schulz, Engelien, and Ross (2001) observed that professional trumpet players and violinists exhibited stronger N1 event-related potential components to sounds from their own instrument, indexing stronger pre-attentive processes related to stimulus detection. Shahin et al. (2008) showed that gamma-band (25–100 Hz) oscillations in EEG-recordings can be enhanced by a year of piano training in children. The same gamma signal differentiated adult musicians from non-musicians in their non-attentive response to different musical timbres. Further research showed that learning not only affects cortical activity, but even modulates "low-level" processing: Strait, Chan, Ashley, and Kraus (2012) demonstrated that brainstem recordings of pianists more closely correlated with the amplitude envelopes of the original piano sounds, compared to recordings of musicians who did not have extensive experience with the piano, but there was no difference between groups for sounds from the tuba and bassoon. This result indicates that there may be instrument-specific neural adaptation that affects the perceptual processing of certain classes of instrumental sounds. It remains unclear, however, whether these effects index conscious perceptual experience and whether they play into STM. The present study investigated the effect of prior knowledge of instrument categories on STM fidelity as indexed by a behavioural short-term recognition task.

Coming back to verbal lexicality may yield an instructive analogy. In the simplest terms, many words reference things or activities in the world. Timbre has similar properties, in the sense that familiar timbres from acoustic instruments can be perceived as referents to sound sources (e.g., a violin) and the cause or activity that set them into vibration (e.g., plucking), likely by virtue of learned, long-term associations (McAdams, 1993). Comparing STM for unfamiliar tones with hidden underlying source/causes to familiar tones from acoustic instruments may therefore create a scenario that is analogous to experiments that give rise to the verbal lexicality effect.

A particular challenge lies in the selection of unfamiliar sounds (perhaps corresponding to "pseudo-words"). A simple idea would be to use abstract digitally synthesised sounds, created by additive synthesis of sinusoidal components. One problem of such an approach is that the overall acoustic complexity (or variability) of a stimulus set appears to affect STM. Golubock and Janata (2013) observed severe capacity limits of STM for the timbre of tones created by additive synthesis, but less so when a more variable set of tones, selected from commercial synthesisers, was used. Therefore, a desirable property of unfamiliar stimuli would be that they feature a similar degree of acoustic complexity compared to natural recordings. Here we used a specifically tailored signal transformation, based on a purposeful mismatch of the quickly-varying temporal fine structure and the more slowly varying spectro-temporal envelope of acoustic signals (e.g., Smith, Delgutte, & Oxenham, 2002). The resulting "hybrid" sounds featured similar overall acoustic properties compared to the original recordings, but were hard to identify and were rated as perceptually less familiar. One might suspect potential differences in memory performance for such "referential" (familiar) and "non-referential" (unfamiliar) timbres to emerge from encoding, where familiar timbres may be assumed to more strongly activate semantic LTM representations than unfamiliar timbres, affording a level-of-processing phenomenon (Craik & Lockhart, 1972). At the same time, differences in memory maintenance strategies may be involved, a topic that researchers have only started to explore for timbre.

Maintenance of timbre

Three basic scenarios for the maintenance of timbre in STM may be distinguished. First, timbre recognition may be a passive process (i.e., maintenance would in fact not play a strong role) such that the retention of timbre primarily relies on the persistence of the sensory memory trace (McKeown, Mills, & Mercer, 2011; Schulze & Tillmann, 2013). Second, participants may attach labels to timbres (e.g., "piano–violin–harp") and subsequently rehearse the verbal labels. This would constitute a verbal surrogate of auditory memory (cf., Schulze, Vargha-Khadem, & Mishkin, 2012). Third, listeners may allocate attention to the auditory memory trace, and "mentally replay" timbres, akin to what has been described as attentional refreshing above.

A few studies have started to probe these hypotheses. McKeown et al. (2011) had participants discriminate small changes in spectral aspects of timbre and showed that sensitivity was above chance even for extended retention intervals of 5–30 s. Notably, this effect was robust to an articulatory suppression task in which participants were required to read aloud during the retention time. These results were interpreted as evidence for a type of sensory persistence that is "neither transient nor verbally coded nor attentionally maintained." (p. 1202). Nonetheless, they alsoemphasised that there may be various other forms of memory for timbre. Schulze and Tillmann (2013) compared the serial recognition of timbres, pitches, and words in various experimental variants, using sampled acoustic instrument tones and recorded verbalisations. They found that the retention of timbre, contrary to that of pitches and words, did not suffer from concurrent articulatory suppression. On the basis of these results, they suggested that working memory for timbre is structured differently than working memory for words or pitches and is unlikely to be facilitated by verbal labelling and rehearsal.

Other studies have underlined the necessity of attentional maintenance. Nolden et al. (2013) recorded electroencephalographic signals during a serial order recognition task with synthesised timbres differing in spectral envelope. In a control condition, participants received the same stimuli but were asked to ignore the standard and to judge a property of the last tone of the comparison sequence. Significant differences in event-related potentials (ERP) were found during the retention interval; the higher the memory load, the stronger the ERP negativity. These findings cohere with those of Alunni-Menichini et al. (2014), demonstrating that the same ERP component robustly indexes STM capacity, providing evidence for an attention-dependent form of STM. Most recently, Soemer and Saito (2015) observed that short-term item recognition of timbre was only inconsistently disrupted by articulatory suppression, but was more strongly impaired by a concurrent auditory imagery task. This was interpreted as evidence that memory for timbre can be an active process that deteriorates when attentional resources are removed.

Importantly, research has already provided evidence for the feasibility of imagery for timbre. Halpern, Zatorre, Bouffard, and Johnson (2004) let musicians rate perceived dissimilarity of subsequently presented pairs of timbres while recording brain activity with functional magnetic resonance imaging. The same procedure was repeated in a condition in which the auditory stimuli were to be actively imagined. Both conditions featured activity in auditory cortex with a right-sided asymmetry, and behavioural ratings from the two conditions correlated significantly. Results such as these speak for the potential accuracy of auditory imagery for timbre: sensory representations activated by timbre perception may at times resemble those activated by imagery (also see Crowder, 1989; Pitt & Crowder, 1992, for earlier behavioural results). Overall, the reviewed findings suggest that attentional refreshing, quite similar to imagery in its active and reconstructive nature, may be a viable candidate mechanism for the active maintenance of timbre.

The present study

For exploring the role of long-term familiarity and sound source categories, as well as the interconnected role of maintenance strategies in STM for timbre, we compared the recognition of familiar acoustic musical instrument sounds and their unfamiliar digital transformations. Exp. 1 tested effects of timbre familiarity and list-probe delay, as well as effects of serial position and list-probe dissimilarity. In order to more thoroughly disentangle the role of active maintenance strategies, Exp. 2 used a subset of trials from Exp. 1 and exposed participants to articulatory suppression, a visual distractor task, and a silent control condition. Because effects of familiarity may be less pronounced for non-musicians who can be assumed to be less exposed to orchestral instrument sounds, Exp. 2 compared the performance of musician and non-musician participants.

Experiment 1: Material and delay

We studied the effect of long-term timbre familiarity and delay interval on musicians' short-term item recognition performance. Because we expected the timbral memory traces of unfamiliar transformations to be more transient, we hypothesised that a potential familiarity advantage would even be greater at 6 s compared to 2 s of delay.

Methods

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

Thirty musicians (22 female) participated in the experiment for monetary compensation. They were recruited from a mailing list of the Schulich School of Music at McGill University and had an average age of 21 years (SD = 3.7, range: 18–29). They had 10 years (SD = 3.8) of instruction on at least one musical instrument and had received 5 years (SD = 3.6) of formal music-theoretical training. Participants reported normal hearing, which was confirmed in a standard pure-tone audiogram measured before the main experiment (ISO 398-8, 2004; Martin & Champlin, 2000) and had hearing thresholds of 20 dB HL or better for octave-spaced frequencies from 125 Hz to 8000 Hz.

Stimuli

Recorded and transformed sounds. A material factor contained two conditions with different types of sounds: familiar acoustic recordings, and unfamiliar synthetic transformations. The first set consisted of 14 recordings of single tones from common musical instruments, all played at mezzo-forte without vibrato. Piano and harpsichord samples were taken from Logic Professional 7 (Apple Computer, Cupertino, CA), and all others were drawn from the Vienna Symphonic Library (http://vsl.co.at, last accessed April 12, 2014); see for a complete list. The audio sampling rate was 44.1 kHz with 16-bit amplitude resolution. Sounds had a fundamental frequency of 311 Hz (E♭4), and only the left channel of the stereo sound file was used. According to VSL, the samples were played as 8th notes at 120 beats per minute, that is, of 250 ms “musical duration”. Nonetheless, actual durations were all slightly longer than 500 ms. We therefore applied barely noticeable fade-outs of 20 ms duration (raised cosines) in order to obtain a uniform stimulus duration of 500 ms.

A set of 70 unfamiliar sounds was generated digitally in order to obscure associations with an underlying source while retaining a similar degree of “acoustic complexity”. However, an important piece of the problem is to define what the latter notion actually means. Digitally synthesised tones usually vary on a small number of dimensions, whereas natural sounds vary in manifold ways. Here, we utilised a perspective that has proven to be of relevance in a variety of studies in hearing science and signal processing, namely the distinction between the quickly varying temporal fine structure and the more slowly varying temporal envelope of acoustic signals (e.g., Moore, 2015). Each novel sound was derived by superimposing the spectrotemporal envelope of one sound onto the temporal fine structure of another. We thereby generated unfamiliar “chimæric” tones that have similar physical properties compared to the original set of recorded acoustic tones (also see Agus et al., 2012; Smith et al., 2002). More details on the sound synthesis and familiarity and dissimilarity ratings can be found in the appendix.

Using the 14 recorded acoustic tones and the 70 resulting transformations, 15 musicians rated perceived familiarity on an analog-categorical scale (1-highly unfamiliar, 5-highly familiar) (Weber, 1991) and identified sounds by selecting one out of eight options (including six instrument names and the labels “unidentifiable” and “identifiable, but not in the list”). The 14 transformations that received the smallest mean familiarity ratings were selected for the main experiment. Mean familiarity of the 14 original recordings ($M = 4.2$, range: 3.1–4.8) was significantly higher than that of the 14 selected transformations ($M = 2.0$, range: 1.6–2.4) as indicated by an independent-samples t-test, $t(26) = 15.5, p < .001$. The median proportions of “unidentifiable” ratings selected for the 14 recordings ($Mdn = 0$, $M = 0.04$, SD = 0.06) and the 14 selected transformations (Mdn = .53, $M = 0.52$, SD = 0.11) were significantly different (Wilcoxon signed-rank test, $Z = -4.5$, $p < .001$). Pearson correlations between the proportion of “unidentifiable” votes per stimulus and mean familiarity ratings were strong and negatively associated, $r(82) = -.88$, $p < .001$. Table A.1 lists the stimuli used for the current memory experiments.

Perceived loudness was matched on the basis of six expert listeners’ adjustments. Subsequently, 24 musicians rated pairwise dissimilarity for both sets of sounds on an analog-categorical scale (1 – identical, 9 – very dissimilar).

Memory sequences. We used an item-recognition task for the main experiment. Every trial featured a “study list”, that is, a sequence of three distinct sounds of 500 ms duration each, which were concatenated with an inter-stimulus interval of 100 ms. The list was followed by a delay of 2 or 6 s before a probe tone was presented.

Fourteen study lists were generated by drawing sounds (nos. 1–14) randomly without replacement under the constraint that every tone occurred equally often (i.e., 3 times) in the 14 lists. Note that the underlying list structure was identical for both material conditions (i.e., recordings and transformations); only the individual sounds that represented the numbering scheme differed. Per material condition, every list was paired with two matching and two non-matching probes. Matching probes were taken from all three serial positions, such that there were overall 8, 10, and 10 probes from the first, second, and third serial positions, respectively. New probes were selected among the remaining $14-3=11$ sounds from the set of recordings or transformations such that for every list there was one probe that was dissimilar (i.e., with a list-probe dissimilarity above the median) and another that was similar (i.e., a below-median dissimilarity). The fact that timbre dissimilarity relations are different between recordings and transformations required us to use differently numbered non-matching probes in the two material conditions. Figure 1 illustrates this graphically.

List-probe dissimilarity has been proven to be important in various short-term item recognition studies (see e.g., Visscher, Kaplan, Kahana, & Sekuler, 2007). In our case, the resulting distribution of dissimilarities did not differ

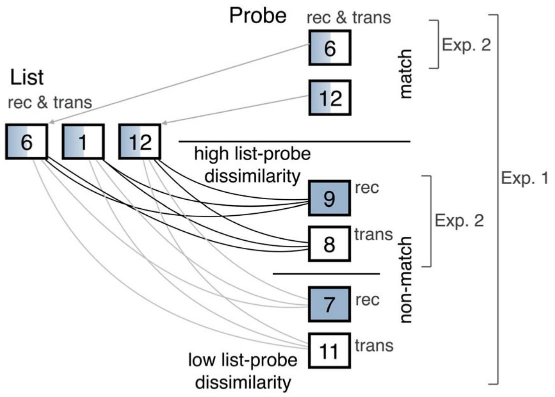

Figure 1. Illustration of the construction of list-probe sequences. Digits refer to individual sounds (#1–14), blue boxes to recordings (rec), white boxes to transformations (trans), half blue/half white boxes to numbers that are instantiated by both materials. Per list, there were two matching probes, equally selected from all three serial positions across the different trials (see Table 3). Non-matching probes were selected such that both materials' lists had one probe with high and another with low list-probe dissimilarity. (The distribution of dissimilarities did not differ across material.) Exp. 2 only used a subset of the trials.

between recording ($M = 5.2$, $\mathrm{SD} = 1.09$) and transformation trials ($M = 5.2, \mathrm{SD} = 0.90$), neither in terms of means, $t(54) < 1$ (two-sample t-test), nor in terms of shape, $D = 0.18$, $p = .72$ (two-sample Kolmogorov–Smirnov test). The complete list of memory sequences is given in (appendix). Overall, there were 14 × 2 (new probes: low and high dissim.) × 2 (old probes) × 2 (delay: 2 s and 6 s) = 112 trials per material condition.

Presentation and apparatus. The average presentation level after loudness-normalisation was 66 dB SPL (range = 58–71 dB SPL) 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). Experiments took place in a double-walled sound-isolation chamber (Industrial Acoustics Company, Bronx, NY). Stimuli were presented on Sennheiser HD280Pro headphones (Sennheiser Electronics GmbH, Wedemark, Germany), using a Macintosh Pro 5 computer (Apple Computer, Cupertino, CA) with digital-to-analog conversion on a Grace Design m904 (Grace Digital Audio, San Diego, CA) monitor system. The experimental interface and data collection were conducted with the audio software Max/MSP (Cycling 74, San Francisco, CA).

Procedure and design. In the item recognition task, participants were asked to respond to the question "Did the final sound exactly match any previous sound from the sequence?" by pressing a button on a response box corresponding to "Yes" or "No". If participants responded "Yes", they were asked to indicate the serial position of the match by pressing the corresponding number on the computer keyboard. We only consider the data from the first binary task for the current analyses.

Trials were presented in four blocks, with two containing recordings and two containing transformations. They were interleaved (e.g., rec, trans, rec, trans) with order counterbalanced across subjects. Within each material condition, the order of trials was fully randomised. Every block required around 15 min to complete, and participants took a mandatory break of 5 min between blocks. In order to get used to the recognition task, participants received four example trials using the recordings for which correct responses were provided. After completion of the experiment, participants filled out a questionnaire concerning biographical information and reactions to the experiment itself.

Data analysis. We measured sensitivity with d′ scores and response bias with the criterion location c, as provided by the Yes/No model (Ch. 1–2 Macmillan and Creelman, 2005). Hits were defined as a correctly recognised match trial (i.e., "old"), false alarms as incorrectly identified non-match trials (reporting "old" to new probes). The sensitivity d′ thus indicates how well participants discriminate between old and new trials. The criterion c describes whether participants are biased toward responding "non-match" ($c > 0$) or "match" ($c < 0$). We did not consider individual responses that were faster than 200 ms or slower than 4000 ms (less than 5% of overall responses). We did not analyse response times in the full factorial designs because instead of reflecting memory fidelity, response times may have been confounded by the factors of delay in Exp. 1 and suppression in Exp. 2. The following set of analyses considers the variables of material, delay, serial position, and list-probe dissimilarity, as well as potential effects of online familiarisation. ANOVAs are conducted for the dependent variables of (i) sensitivity and (ii) bias as a function of material and delay. The factor of position could not be included in this analysis, because it is only defined on match trials, whereas the signal detection theoretic variables require match and non-match trials to be combined. We thus computed another ANOVA for an analysis of (iii) hit rate as a function of material, delay, and position. For non-match trials, we analysed (iv) correlations between list-probe dissimilarities and correct-rejection rates. In order to assess potential effects of online familiarisation, we finally computed two ANOVAs on (v) sensitivity and (vi) bias as a function of experimental block[1] (1st vs. 2nd) and material. Because multiple null hypothesis tests (such as the five ANOVAs just mentioned) inflate experiment-wise Type I error rates, we used the adjusted significance level of $\alpha = .01$ for the main analyses.[2]

Results

Sensitivity. A repeated-measures ANOVA on d′ scores yielded effects of material, $F(1, 29) = 11.1$, $p = .002$, $\eta_p^2 = .276$, and delay, $F(1, 29) = 30.3$, $p < .001$, $\eta_p^2 = .511$, but no significant interaction. Performance is worse for transformed sounds compared to recordings and at a 6-s delay compared to a 2-s delay. So both familiarity and delay affect recognition (Figure 2 (a)).

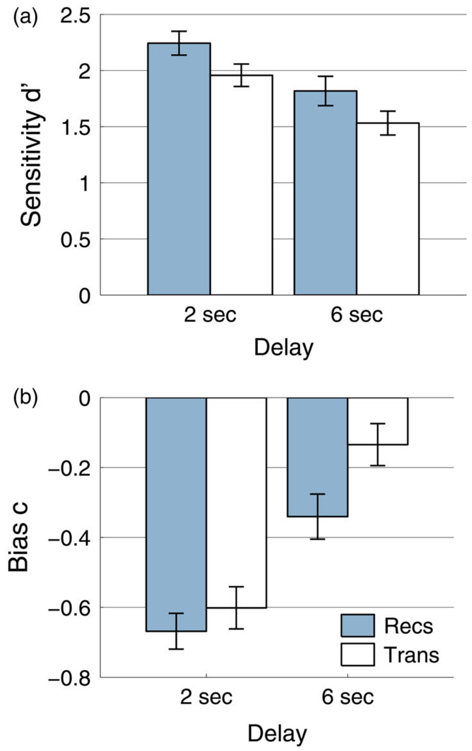

Figure 2. Exp. 1: d' scores (a), and response biases (b) for recordings (recs) and transformations (trans). Error bars depict standard errors of the mean.

Bias. The criterion location c was significantly affected by material, $F(1, 29) = 12.3$, $p = .002$, $\eta_p^2 = .297$, and delay, $F(1, 29) = 100$, $p < .001$, $\eta_p^2 = .776$, but the interaction of both factors failed to reach significance ($\alpha = .01$), $F(1, 29) = 4.66$, $p = .039$, $\eta_p^2 = .139$. The bias toward responding "match" was greater at the shorter delay and was greater for recorded than for transformed sounds (Figure 2 (b)). In other words, participants gravitated towards providing "non-match" responses for the longer delay and the unfamiliar transformed sounds.

Serial position. Considering effects of serial position, a repeated-measures ANOVA on hit rates with the factors position, material, and delay yielded an effect of position, $F(2, 58) = 13.4$, $p < .001$, $\eta_p^2 = .316$, and of material, $F(1, 29) = 17.9$, $p < .001$, $\eta_p^2 = .382$, as well as a significant interaction between the two, $F(2, 58) = 12.6$, $p < .001$, $\eta_p^2 = .304$. The main effect of position stemmed from significantly lower performance in the second position compared to the first and third positions, paired $t(29) > 2.9$, $p < .007$, but only a marginal difference between first and third positions, $t(29) = -2.3$, $p = .028$ (n = 3 comparisons, Bonferroni-corrected $\alpha_{\text{crit}} = .0167$). The interaction of position and material was due to higher hit rate for recordings in the second position, paired $t(29) = 5.2$, $p < .001$ (see Figure 3), but no differences between recordings and transformations in the other two serial positions, $p > .040$ ($n = 3$ comparisons, Bonferroni-corrected $\alpha_{\text{crit}} = .0167$).

There was also an effect of delay, $F(1, 29) = 52.4$, $p < .001$, $\eta_p^2 = .644$, and an interaction of delay and position, $F(2, 58) = 4.2$, $p = .002$, $\eta_p^2 = .127$ (see Table 1 for the raw accuracy data). The latter was due to the fact that in addition to the main effect of position (featuring lowest performance in the second serial position) hit rates were particularly low in this serial position with 6 s of delay ($M = .75$, $\text{SD} = .20$, compared to $M = .90$, $\text{SD} = .12$ for 2 s), as confirmed by *post-hoc* contrasts, $t(325) = 5.14$, $t(325) = 5.14$, $p < .001$.

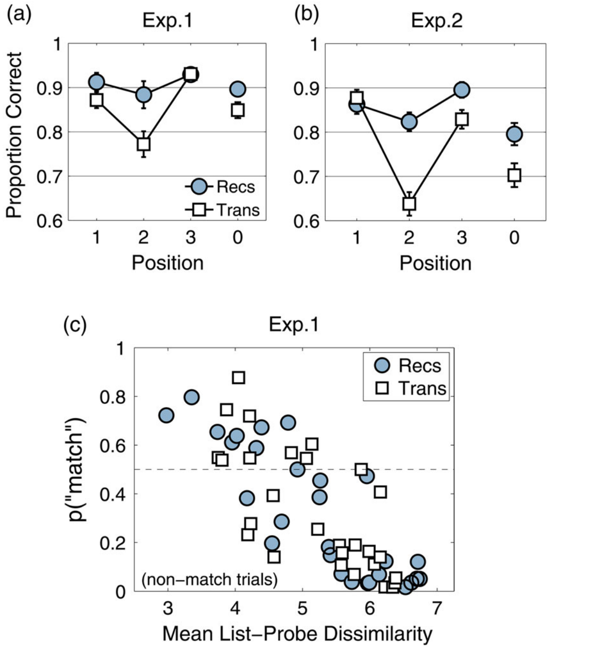

Figure 3. Proportion of correct responses as a function of serial position and material conditions in Exp. 1 (a) and Exp. 2 (b) for recordings (recs) and transformations (trans). Position "0" refers to the non-match trials with high list-probe dissimilarity (see Table 1 for the low dissimilarity condition in Exp. 1, not displayed here in order to retain a reasonable resolution). Error bars show standard error of the mean. (c) List-probe dissimilarity and response choice for non-match trials of Exp. 1.

Dissimilarity and non-match trials. Figure 3(c) shows the strong association between response choice and the dissimilarity between list and probe timbres with significant correlations for recordings, $r(26) = .85, p < .001$, and transformations, $r(26) = .72, p < .001$. Note that we did not observe a significant correlation between the timbral heterogeneity of the list items and correct rejection rate or hit rate, $r(27) < .40$, $p > .12$.

The figure also demonstrates that responses are strongly biased, because trials with the lowest dissimilarity ratings received correct-rejection rates of less than 50%. This bias of participants to preferentially select "match" responses for low list-probe dissimilarities warrants the usage of the signal-detection-theory measures for the analysis of global variables involving both match and non-match trials. With the unbiased d' measure, performance on the lower half of list-probe dissimilarities ranged above chance with $M = 1.4$ ($\text{SD} = 0.65$), and $M = 1.3$ ($\text{SD} = 0.48$) for recordings and transformations, respectively. For the other half of trials with high dissimilarities, sensitivity was at $M = 2.9$ ($\text{SD} = 0.69$) and $M = 2.3$ ($\text{SD} = 0.60$) for the two respective material conditions.

Table 1. Proportion of correct responses for familiar recordings (recs) and unfamiliar transformations (trans) and all other within-subjects conditions across musician participants in Exp. 1 and musicians and non-musicians in Exp. 2.

		Experiment 1 ($n = 30$)				Experiment 2 ($n = 48$)					
		Familiar (recs)		Unfamiliar (trans)		Familiar (recs)			Unfamiliar (trans)		
		2 s	6 s	2 s	6 s	Sil	Vis	Count	Sil	Vis	Count
*Position (HI)	1	.96	.86	.93	.82	.88	.88	.83	.91	.86	.86
	2	.94	.83	.87	.67	.83	.80	.84	.63	.60	.69
	3	.97	.89	.99	.88	.94	.87	.88	.86	.81	.82
Dissimilarity (CR)	High	.91	.89	.84	.86	.83	.80	.76	.76	.70	.64
	Low	.41	.50	.43	.57	–			–		

Note: The values correspond to hit rates for position conditions and to correct rejection rates for dissimilarity conditions.

The accuracy data presented in Table 1 and Figure 3, panel (a), show that the advantage in sensitivity of recordings over transformations was not only due to match trials (as might be concluded from the presentation of the position effects above), but also due to non-match trials: Recordings yielded significantly higher correct rejection rates for high dissimilarity trials compared to transformations, $t(29) = 2.9, p = .007$. Due to the strong response bias mentioned above, however, correct rejection rates of both recordings and transformations did not differ from chance in the low dissimilarity condition, $t(29) < 1.3, p > .19$. Overall, this suggests that the main effect of material on d′ scores originated from both match and non-match trials.

Familiarisation. Finally, we addressed potential effects of online familiarisation in a dedicated repeated-measures ANOVA with the factors of material and experimental block. If there was online familiarisation with the initially unfamiliar transformations, one would expect an interaction between the two variables. Besides the main effect of material on d′ scores already analysed above, there was neither an effect of block, $F(1, 29) = 0.4$, $p = .53$, nor an interaction, $F(1, 29) = 1.2$, $p = .27$. The criterion location c was affected by material (analysed above), but not significantly affected by experimental block, $F(1, 29) = 0.17$, $p = .68$, and the interaction of material and block failed to reach significance ($\alpha = .01$), $F(1, 29) = 5.4$, $p = .026, \eta_p^2 = .158$.

Discussion

We compared short-term item recognition of musicians for a set of familiar orchestral tones and a set of unfamiliar synthetic tones. The stimulus sets were closely matched in terms of physical properties such as spectrotemporal envelope profiles, and so were the resulting sets of memory lists and probes, which were almost identical in structure and did not differ with regards to list-probe similarity (cf., Visscher et al., 2007). The main effect of material on sensitivity was coherent with our hypotheses. Familiar timbres that musicians can associate with well-known instrument categories are overall better recognised than are unfamiliar timbres.

The current data featured a detrimental effect of delay on hit rates (see Table 1), but not on correct rejections. This means that listeners are less likely to provide a "match" response when the retention time increases (which may seem intuitive), but delay does not affect correct rejections. The same pattern was observed by Golubock and Janata (2013). We had expected an even larger difference in sensitivity across material conditions at 6 s of list-probe delay where we thought the multiple affordances for encoding and maintenance of familiar timbres would lead to more robust recognition. This was not the case, although there was a tendency for an interaction effect on response bias: participants judged more transformation trials than recording trials as "new", and this was particularly so for the longer delay condition. That is, the interplay of material and retention time only tended to weakly affect response behaviour, rather than affecting memory fidelity as such. Memory representations of familiar and unfamiliar timbres can be concluded to possess similar decay over time and thus might be subject to similar maintenance processes.

Considering the serial position data, there was not only a main effect of position on hit rate, but transformations were even less well recognised when they were presented in the least salient medial position of the sequence. It is important to note, however, that the main effect of familiarity was not only due to the superior performance for that position, but also due to non-match trials (i.e., correct rejections), at least for the high dissimilarity condition (which was not corrupted by the marked response bias). We will postpone further interpretation of this pattern of results until the discussion of Exp. 2.

There was a strong correlation between correct rejections and dissimilarity: the more dissimilar the probe was to the elements of the list, the more likely it was to be recognised as new. Note that we did not find any significant effect of list homogeneity (pairwise similarity of a study list) on correct rejections or on hit rates. This contrasts with the findings from Visscher et al. (2007), who demonstrated that an increase in list heterogeneity leads to a decrease in correct rejection rates in a short-term item recognition task involving auditory moving ripple stimuli.

By differentiating between list-probe dissimilarity and list homogeneity, one can also refine an interpretation of results from Golubock and Janata (2013). Here the authors observed an increase in memory capacity across two experiments, which they interpreted to be caused by a larger acoustic variability of the set of sounds used in the second experiment. Our current results suggest that the increase in memory capacity may more particularly be due to the overall larger list-probe dissimilarities prevailing in their more variable second set.

An intricate question is whether the initially unfamiliar transformations become more familiar over the course of the experiment. The intriguing repetition priming results of Agus, Thorpe, and Pressnitzer (2010) showed that after only a few exposures, participants implicitly learned features of white noise clips, which led to enhanced processing fluency in the detection of clip repetitions. Other STM studies (e.g., Golubock & Janata, 2013; Soemer & Saito, 2015) have selected large numbers of supposedly unfamiliar stimuli by relying on the subjective familiarity judgments of the authors alone, as well as audio-descriptor-based models of timbre dissimilarity (which have only been perceptually validated to a limited extent). We chose a “closed set” design that repeats items, because we wanted to thoroughly control the items’ perceptual familiarity and identifiability as well as perceptual dissimilarities between target list items and probe items on the basis of experimental data (as reported in the stimulus section above). Because the number of pairwise dissimilarity ratings grows quadratically with set size, we thus needed to settle on two relatively small sets of tones. Every sound, whether as part of a sequence or as probe, appeared on average around 32 times over the course of the entire experiment. In that sense, the current design may conflate the aspects of familiarity and source identification, which theoretically may have different dynamics: The transformed sounds do not readily evoke sound source categories, and this is unlikely to change with repeated listening (because there aren’t any). On the contrary, it could be that listeners became progressively more familiar with the transformations, supporting processing fluency.

Our data, however, do not feature significant effects of online familiarisation, as would have been indicated via a material×block interaction for sensitivity or bias. Although there was a tendency of an interaction effect for the latter variable, participants did not manage to adapt their strategy for the transformed sounds in a way that optimised sensitivity. For that reason, we conclude that the current data are not substantially affected by an online gain in processing fluency.

Turning towards the underpinnings of the observed effect of material, a simple, maintenance-based explanation could posit that musicians verbally labelled recordings but not transformations and subsequently rehearsed verbal labels. Exp. 2 set out to test this hypothesis and to further clarify the role of active maintenance in timbre recognition.

Experiment 2: Material, suppression, and group

In order to assess the contribution of maintenance to the observed familiarity effect, Exp. 2 compared a silent delay condition with a classic articulatory suppression task that required participants to count aloud during the retention interval, which should impair their ability to verbally label and rehearse timbres. If this was the driving force behind the observed material effect, the advantage of familiar over unfamiliar timbres should vanish (or at least be reduced) under articulatory suppression. A perhaps more obscure hypothesis would be that the material effect is due to participants’ reliance on visual associations, which could again be more readily available in the case of familiar acoustic tones. In order to control for this possibility, we also included an attention-demanding visual suppression condition.

Whereas the suppression factor was primarily included in order to test the contribution of active maintenance strategies to the observed familiarity effect, it further served as a useful tool for differentiating between the different maintenance strategies themselves. In fact, articulatory suppression not only requires verbal resources, but also generates interference with the auditory memory trace. Therefore, it can be assumed to have a detrimental impact on all three discussed candidate mechanisms for the retention of timbre: sensory decay, labelling, and refreshing (although the magnitude of such an effect would likely differ across mechanisms). The visual suppression condition can be assumed to be more specific in this case, because it leaves the auditory trace fully intact while withdrawing attentional resources. Consequently, if visual suppression impaired performance, this would indicate that maintenance of timbre requires attentional resources (which cannot be strictly inferred from the contrast of articulatory suppression and the control condition alone).

The experiment further compared musicians with a group of non-musicians, which we assumed to be less experienced and less familiar with orchestral instrument sounds. Accordingly, we expected a diminished advantage of recordings over transformations for non-musicians (as expressed in a material×group interaction).

In sum, the experiment contained a between-subjects factor of musical training, and besides the within-subject factor of material, a novel within-subject suppression factor with the conditions of articulatory suppression, visual suppression, and a silent control condition.

Methods

Participants. Forty-eight listeners participated in the experiment for monetary compensation. A group of 24 musicians (13 female) was recruited from a mailing list of the Schulich School of Music at McGill University. They had mean ages of 23 years (SD = 4.2, range: 18–34), had received 15 years (SD = 4.5) of instruction on at least one musical instrument (including the voice) and had received 6 years (SD = 4.3) of formal music-theoretical instruction. None of them had participated in Exp. 1. The group of 24 non-musicians was recruited via classified advertisements on a McGill University webpage. They had a mean age of 28 years (median: 23.5, SD = 11.6, range: 19-67), 0.4 years (SD = 0.91) of instruction on a musical instrument, and no formal music-theoretical or instrumental training beyond elementary school. Normal hearing was confirmed as in Exp. 1.

Stimuli

Memory sequences. We used the memory lists from Exp. 1 in conjunction with the group of non-match probes that possessed high list-probe dissimilarity, plus one of the two subsets of old probes (see Table A.2). This yielded 14 × 2 (match, non-match) = 28 trials per material condition. Every trial was presented in each of the three suppression conditions. As in the second delay condition from Exp. 1, lists and probes were separated by 6 s.

Suppression conditions. There was a silent condition, a visual distractor task, and an articulatory suppression condition. In the visual task, a sequence of 4 × 4 grids of filled black and white squares appeared on the screen, similar to the method used by Pechmann and Mohr (1992) and Schendel and Palmer (2007). Participants were asked to indicate, using the same yes/no buttons on the response box, whether there was a direct repetition of a grid in the sequence or not. The visual sequence appeared 100 ms after the offset of the study list and contained 6 grids, each of which was presented for 600 ms. The grids were created randomly such that 5 of the 16 squares were always filled (Pechmann & Mohr, 1992). The grids occupied a 10 × 10 cm area on the computer screen. In 50% of the visual suppression trials, there was a direct repetition of a visual grid, distributed across the serial positions of the visual sequence. After the end of the visual sequence, subjects had at least 2300 ms to respond to the visual task and to prepare for the auditory task. One second before the onset of the probe stimuli, the screen into which the grids were embedded disappeared, signalling participants to get ready to respond to the probe. Figure 4 illustrates the task demands of the three suppression conditions.

In the articulatory suppression task, a screen appeared 100 ms after offset of the study list. It asked participants to count aloud into a microphone, starting at one. The screen disappeared 1 s before the onset of the probe, which indicated to participants to stop counting and prepare for the auditory task.

Presentation and apparatus. Presentation and apparatus were identical to those in Exp. 1.

Procedure and design

Participants completed the audiogram and read through the experimental instructions. They were then introduced to the basic item recognition task that was used in all three suppression conditions. For that purpose, two example trials without suppression were presented for which the correct responses were provided on the experimental interface. Each suppression condition was then presented block-wise and was preceded by six training trials that familiarised participants with the current task. During training, participants could clarify questions with the experimenter. All training trials used sounds from the recordings.

In sum, we considered one between-subjects factor (musicians, non-musicians) and two global within-

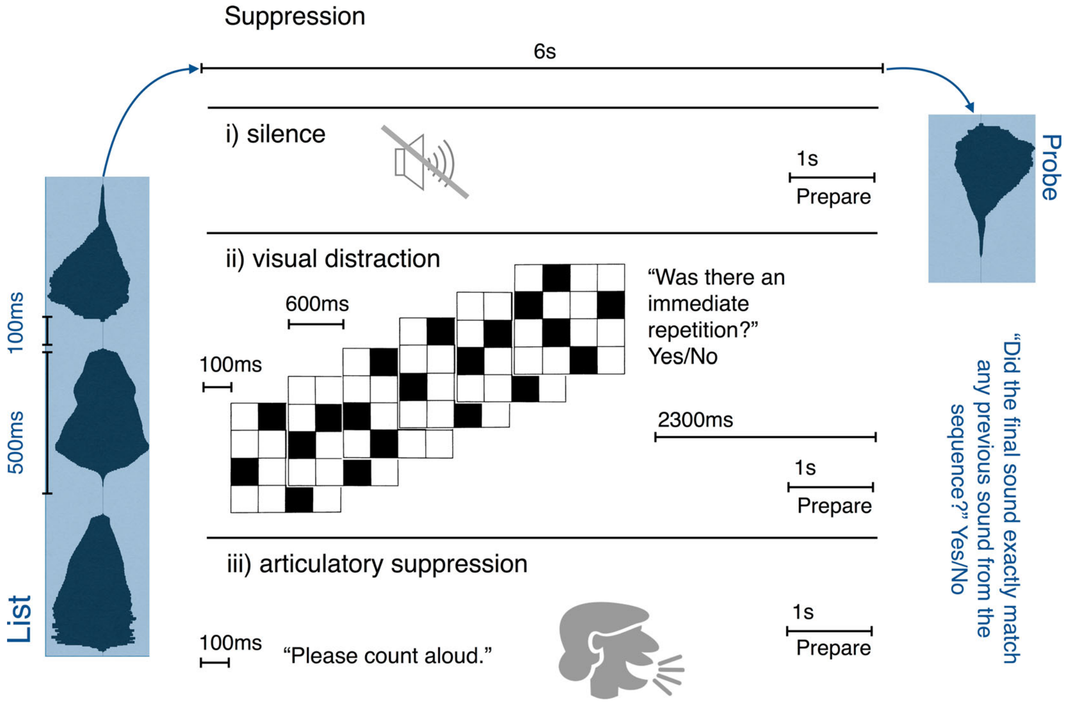
Suppression
6s
i) silence
1s
Prepare
ii) visual distraction
600ms
100ms
"Was there an immediate repetition?" Yes/No
2300ms
1s
Prepare
iii) articulatory suppression
100ms
"Please count aloud."
1s
Prepare
List
500ms
100ms
Probe
"Did the final sound exactly match any previous sound from the sequence?" Yes/No

Figure 4. Sketch of the three different suppression conditions in Exp. 2.

subject factors, suppression (silence, visual, articulatory) and material (recordings vs. transformations). The serial position factor was nested within the subset of matching probes. The six possible orders of presenting the three suppression blocks were counterbalanced across participants (i.e., participants 1 & 7, 2 & 8, etc. received the same order of suppression blocks). The material condition was presented block-wise and was nested within the suppression conditions, with order counterbalanced orthogonally to the suppression factor (i.e., participants 1 & 3, 2 & 4, etc. received the same succession of material conditions). A questionnaire was administered after the experiment.

Data analysis. To ensure visual distraction, only trials with correct responses to the visual task were taken into account (on average 93%, SD = 6). In the articulatory suppression interval, participants' vocalisations were recorded in order to verify aurally that they counted aloud in all test trials of the articulatory suppression condition. ANOVAs were computed for the variables of (i) sensitivity and(ii) bias as a function of suppression, material, and musical training, (iii) hit rate as a function of these latter three independent variables and serial position. The robustness of effects found in analysis (iii) was confirmed by a cross-experiment ANOVA with the variables of material, position, and experiment (iv). For non-match trials, we considered (v) correlations between list-probe similarities and correct-rejection rate. Otherwise, the data analysis was identical to that in Exp. 1.

Results

Sensitivity. A mixed ANOVA indicated that all three factors of group, material, and suppression affected memory fidelity significantly. Figure 5 shows the corresponding d′ scores. Musicians had higher sensitivity than non-musicians, $F(1,46) = 25.6$, $p < .001$, $\eta_p^2 = .357$, and recordings were easier to recognise than transformations, $F(1,46) = 65.0$, $p < .001$, $\eta_p^2 = .586$. There was a main effect of suppression, $F(2,92) = 13.8$, $p < .001$, $\eta_p^2 = .231$, because the silence condition was easier than both the visual condition, paired $t(47) = 4.01$, $p < .001$, and the articulatory suppression condition, paired $t(47) = 4.96$, $p < .001$. However, there was no difference between visual and articulatory suppression conditions, paired $t(47) = -0.88$, $p = .383$. None of the interactions were significant.

Bias. Response bias was not affected by material, $F(1,46) < 1$, but was by group, $F(1,46) = 16.4$, $p < .001$, $\eta_p^2 = .262$, and weakly by suppression condition, $F(2,92) = 4.68$, $p < .001$, $\eta_p^2 = .092$ (Figure 6). The latter effect arose through a significant difference between the silence and counting condition, paired $t(47) = 3.19$, $p = .008$, but no differences otherwise, $p = .078$ ($n = 3$ comparisons, Bonferroni- corrected $\alpha_{\text{crit}} = .0167$).

Serial position. Regarding effects of serial position, a mixed ANOVA on hit rates did not yield main effects of group, $F(1,46) < 1$, or suppression, $F(2,92) = 1.99$, $p = .142$, but did reveal significant effects of serial position, $F(2,92) = 44.6$, $p < .001$, $\eta_p^2 = .492$, and material,

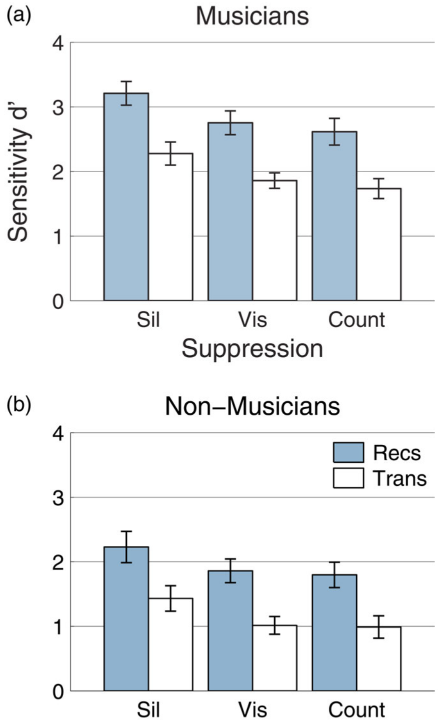

Figure 5. Exp. 2: d′ scores for musicians (a) and non-musicians (b) for recordings (recs) and transformations (trans) in the suppression conditions of silence (Sil), visual suppression (Vis), and articulatory suppression (Count).

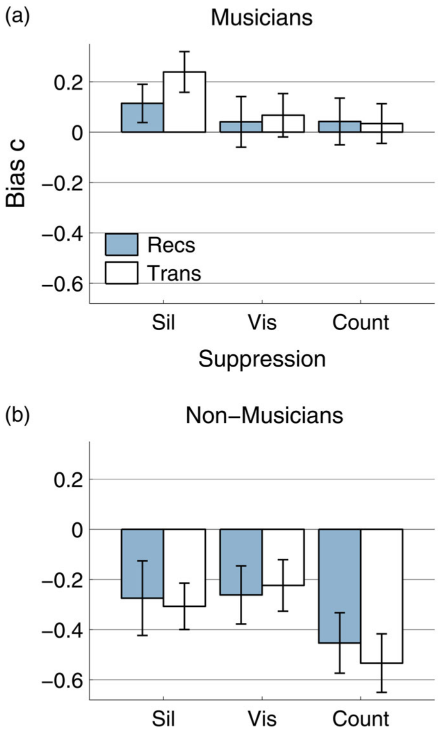

Figure 6. Exp. 2: Response bias for recordings (recs) and transformations (trans) as measured with response criterion location c for musicians (a) and non-musicians (b) in the suppression conditions of silence (Sil), visual suppression (Vis), and articulatory suppression (Count).

$\eta_p^2 = .410$, $p < .001$, $\eta_p^2 = .410$. The effect of position was due to inferior performance in the second position compared to the first and third, paired $t(47) > 8.0$, $p < .001$, but no differences between first and third position, $t < 1$. There was a strong interaction of position and material, $F(2, 92) = 29.1$, $p < .001$, $\eta_p^2 = .387$, that was due to no differences between recordings and transformations in the first serial position, paired $t(47) = -.687$, $p = .49$, but significant differences in the medial, $t(47) = 8.1$, $p < .001$, and last position, $t(47) = 3.7$, $p < .001$. There was no other significant interaction. Figure 3(b) displays the corresponding hit rates

Serial position across experiments. The robustness of the position-related effects was confirmed by a *post-hoc*, cross-experiment ANOVA on hit rate as a function of serial position, material, and experiment, using the subset of musicians from Exp. 2 in the silent suppression condition, and musicians from Exp. 1 in the 6-s-delay condition. There were significant main effects of material, $F(1, 52) = 43.5$, $p < .001$, $\eta_p^2 = .45$, and position, $F(2, 104) = 31.7$, $p < .001$, $\eta_p^2 = .38$, as well as an interaction of material and position, $F(2, 104) = 23.3$, $p < .001$, $\eta_p^2 = .31$. This interaction arose through significant differences between recordings and transformations in the medial position, paired $t(53) = 7.7$, $p < .001$, but no other significant differences, $p > .06$. Furthermore, there was no significant main effect of experiment, $F(1, 52) < 1$. Although the comparison of panels (a) and (b) in Figure 3 may suggest a differential effect of position in Exps. 1 and 2 (i.e., an experiment×position interaction), and even differential interactions of material and position across experiments (i.e., a three-way interaction), both two- and three-way interactions failed to fulfill the strict significance level of $\alpha = .01$ and, more importantly, had comparatively small effect sizes, $F(2, 104) < 3.7$, $p > .028$, $\eta_p^2 < .066$.

Dissimilarity and non-match trials. Correct rejection rates were not significantly correlated with list-probe dissimilarities for recording trials in any of the three suppression conditions, $r(13) < .511$, $p > .011$, nor was this the case for transformations, $r(13) < .30$, $p > .29$. The lack of a correlation in Exp. 2 may have been due to its smaller range of dissimilarities (rec: 5.4–6.7, trans: 5.5–6.4) compared to Exp. 1 (rec: 2.8–6.7, trans: 3.7–6.4), where significant correlations were obtained for both groups of sounds.

Correct rejection rates were higher for recordings compared to transformations, paired $t(47) = 5.6$, $p < .001$ (see Figure 3, panel (b)). Similarly to Exp. 1, this finding once again suggests that the effect of material on sensitivity was due to better performance on recordings in both match and non-match trials.

Discussion

Exp. 2 reproduced the advantage of familiar recordings over unfamiliar transformations with regards to recognition sensitivity. Notably, the effect persisted throughout the articulatory and visual suppression conditions. This suggests that the locus of the effect is due to differences in encoding rather than to maintenance mechanisms. Notably, this result rules out the oversimplified hypothesis that the familiarity effect is solely based on verbal labelling and subsequent rehearsal.

The interaction of serial position and material from Exp. 1 was replicated, see Figure 3 (panels (a) and (b)). A cross-experiment ANOVA further confirmed that this effect was robust across experiments, even though Exp. 1 presented a larger set of stimuli than Exp. 2, and both experiments featured different contextual variables, such as delay in Exp. 1 and suppression in Exp. 2. The position×material interaction on the set of match trials appears to suggest that unfamiliar matching probes are particularly difficult to recognise when they are in the medial serial position, and the cross-experiment analysis implies that there is no robust difference from familiar probes in the first and last serial positions. Nonetheless, the material effect was not limited to the medial position of match trials, but also occurred on non-match trials in Exps. 1 and 2.[3] For that reason, we interpret the current results as a familiarity-based mnemonic advantage and not as a mere difficulty in "parsing" unfamiliar sounds in the medial positions of match trials.

There remains the question of why no effect of material became apparent in the first and last serial positions. Note that in the last serial position of Exp. 1 (for which there was no significant difference between material conditions), 21 out of 30 participants obtained hit rates that were greater than or equal to 90% for both material conditions. In the last serial position of Exp. 2 (the position which did not yield significant differences between recordings and transformations), 18 out of 48 participants scored higher than 90% in both conditions. For both material conditions, scores of many listeners thus ranged close to a ceiling level in these non-medial positions, which may well have blurred any differences in observed memory fidelity across material conditions.

Regarding the between-subjects factor of musical training, we saw that musicians featured higher sensitivity and less bias than non-musicians. Note that this is not due to a different approach to the speed-accuracy trade-off, as musicians were also faster overall with a grand average response time of $M = 1358$ ms ($SD = 306$) compared to $M = 1710$ ms ($SDSD = 337$) for non-musicians, independent-samples $t(46) = -3.8$, $p < .001$.

Contrary to our hypotheses, sensitivity was not affected by an interaction of material and group. This may be surprising at first glance, because one can assume that musicians are more familiar with orchestral instrument sounds (Douglas, 2015), and therefore the difference in their encoding and maintenance of familiar acoustic and unfamiliar synthetic sounds should be particularly large. Nonetheless, considering unfamiliar sounds as a neutral baseline across groups may have been a flawed assumption because musicians possess better auditory skills (Kraus & Chandrasekaran, 2010; Patel, 2012) and may be generally more experienced in memorising and categorising sounds, even if novel.

The main effect of suppression was due to reduced performance in both suppression tasks relative to the control condition, and the advantage of recordings persisted throughout all suppression conditions. We obtained a significant decrease of sensitivity through articulatory suppression, contrary to the lack of effects in (McKeown et al., 2011; Schulze & Tillmann, 2013) and what was more ambiguous in the results from Soemer and Saito (2015) where only performance on lists with two items was reduced. The fact that the material effect persisted under articulatory suppression speaks against verbal labelling as a dominant maintenance strategy, because even performance on unfamiliar sounds (unlikely to be labelled) was impaired. It seems more likely that the detrimental effect of articulatory suppression was due to interference with the auditory trace. Note that passive sensory decay as a retention mechanism is ruled out by the detrimental effect of visual suppression, leaving the auditory sensory trace intact. Attentional refreshing, to the contrary, can be assumed to be moderately disrupted by both types of suppression because the visual distractor task reduces attentional resources which refreshing is based on, and articulatory suppression interferes with the very auditory trace to be refreshed (beyond its minor attentional requirements). Attentional refreshing therefore seems to be best supported by the current results.

The finding that articulatory suppression significantly impaired timbre recognition (Exp. 2) is novel and does not cohere with a number of studies (McKeown et al., 2011; Schulze & Tillmann, 2013; Soemer & Saito, 2015). Discerning potential differences with previous studies, it should be first noted that McKeown et al. (2011) used a drastically different experimental scenario. Their task was to discriminate subtle changes in spectral envelope. They tested three participants (two of which were co-authors), and participants underwent daily training for from one to two months with a test phase that lasted for around 10h over 20 days. It thus seems hard to exclude the possibility that their finding – reading aloud does not impair timbre discrimination over long retention intervals – reflects rather specific training effects. In one of their experiments, Schulze and Tillmann (2013) did not find effects of articulatory suppression in a backward serial recognition task, requiring subjects to match the order of a mentally reversed timbre sequence to a comparison. Given the stark differences of item and backward sequence recognition tasks, it is hard to draw direct comparisons to the current results because task demands likely affect which strategies are used preferentially (e.g., Camos et al., 2009, 2011). In an experimental design that was relatively close to the current study, Soemer and Saito (2015) only observed a detrimental effect of articulatory suppression in a 2-item list condition (always presented first), but not for lists of length 3 or 4. These results are difficult to reconcile with the current data and may require further empirical study.

Conclusion

Musicians and non-musicians better recognised timbres from acoustic instruments compared to timbres from digital transformations. Across material conditions, stimuli were otherwise matched in terms of spectrotemporal envelope properties, temporal fine structure, loudness, and list-probe dissimilarities. We interpret these findings as evidence that familiarity with sound source categories plays a salient role in short-term timbre recognition, an effect that arose independently of musical training. Furthermore, sensitivity for both familiar and unfamiliar sounds was equally impaired by articulatory and visual dual tasks, which rules out the hypothesis that the familiarity effect is due to differential maintenance strategies, such as simple verbal labelling of familiar sounds.

In effect, these findings point toward a more robust form of encoding for timbral properties of familiar acoustic instruments. Prior knowledge of instrument categories for familiar acoustic instrument sounds helps to associate sounds with auditory knowledge schemes. In other words, familiar instrument sounds not only activate auditory sensory representations, but to some extent also semantic, visual, and even sensorimotor networks, which could act as representational anchors for the associated auditory sensory traces. Consequentially, familiar timbres possess more affordances for "deep" encoding. As noted by Craik,

> Deep processing can be carried out on any type of material: the general principle is that the new information is related conceptually to relevant pre-existing schematic knowledge. Thus familiar odors, pictures, melodies and actions are all well remembered if relating to existing bases of meaning at the time of encoding. On the other hand, stimuli that lack an appropriate schematic knowledge base […], are extremely difficult to remember. (Craik, 2007, p. 131)

Although level-of-processing effects (Craik & Lockhart, 1972) have traditionally been sought in the domain of LTM, Rose, Buchsbaum, and Craik (2014) have recently shown that there can be effects of encoding depth (shallow vs. deep, i.e., based on orthographic/phonemic vs. semantic perceptual analysis) on working memory when participants use attentional refreshing. Regarding the nature of timbre maintenance itself, results from Exp. 2 provide support for attentional refreshing as an important maintenance strategy in STM for timbre. Refreshing relies on domain-general attention as well as the fidelity of an item's representation. Should the integrity of either component be disrupted, such as by removal of attention (as in the visual task) or by interference with the auditory trace and reduction of attentional resources (as in articulatory suppression), the process may be assumed to become prone to errors. This is coherent with the pattern of results of Exp. 2, in which both articulatory and visual suppression impaired recognition performance.

The generality of refreshing is supported by the fact that the suppression effects occurred regardless of whether familiar recordings or chimæric transformations were presented. For these reasons, the current results resonate with previous studies (Nolden et al., 2013; Soemer & Saito, 2015) in that they portray attention as a major factor of short-term timbre recognition.

By and large, our results suggest that timbre (re)cognition is a multifaceted and active process. It therefore not only functions on the basis of the persistence of sensory features, but evolves through the interplay of attention, different representational formats (i.e., sensory and sound-source-specific information), and LTM. The more a timbre affords multilayered and deep encoding, the more robust becomes its recognition. STM for timbre should then be seen not as a mere "echo" in the mind of a listener, but rather as a flexible "workspace" that revolves around auditory sensory representations and trades with a plurality of other mental currencies.

Notes

1. We did not analyse material, delay, and block conjointly because in our randomisation scheme, each subject was presented with a varying number of trials for a given block×material×delay condition, which would have rendered the calculation of signal detection theoretic measures problematic (Macmillan & Creelman, 2005, pp. 8–9).
2. See for instance the statistical guidelines of the *Psychonomic Bulletin & Review* for corresponding recommendations: http://www.springer.com/psychology/cognitive+psychology/journal/13423
3. The low dissimilarity trials of Exp. 1 were an exception to this, because proportion correct scores were at chance, and thus blurred any distinction between material conditions.

Acknowledgments

We thank Rachel Na Eun Kim and Shrinkhala Dawadi for running experimental participants, and Bennett K. Smith for assistance with programming the experiment. We also would like to thank Barbara Tillmann, Ilja Frissen, and the two anonymous reviewers for valuable comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by 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.

References

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

Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: Insights from noise. *Neuron*, *66*, 610–618.

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

Camos, V., Lagner, P., & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language*, *61*(3), 457–469.

Camos, V., Mora, G., & Oberauer, K. (2011). Adaptive choice between articulatory rehearsal and attentional refreshing in verbal working memory. *Memory & Cognition*, *39*(2), 231–244.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163–191.

Craik, F. I. (2007). Encoding: A cognitive perspective. In H. L. Roediger III, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 129–136). Oxford: Oxford Univ Press.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Crowder, R. G. (1989). Imagery for musical timbre. *Journal of Experimental Psychology: Human Perception and Performance*, *15* (3), 472–478.

Douglas, C. (2015). *Perceived affect of musical instrument sounds* (Unpublished Master's thesis). McGill University.

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.

Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, *42*(9), 1281–1292.

Johnson, M. K. (1992). MEM: Mechanisms of recollection. *Journal of Cognitive Neuroscience*, *4*(3), 268–280.

Jolicoeur, P., Levebre, C., & Martinez-Trujillo, J. (2015). *Mechanisms of sensory working memory/ attention and perfomance XXV*. London: Academic Press.

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.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, *11*(8), 599–605.

Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx), 10–15 September* (pp. 237–244). Bordeaux, France.

Macken, B., Taylor, J. C., & Jones, D. M. (2014). Language and short-term memory: The role of perceptual-motor affordance. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *40*(5), 1257–1270.

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

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. (1993). Recognition of sound sources and events. In S. McAdams & E. Bigand, (Eds.), *Thinking in sound: The cognitive psychology of human audition* (pp. 146–198). Oxford: Oxford University Press.

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

McKeown, D., Mills, R., & Mercer, T. (2011). Comparisons of complex sounds across extended retention intervals survives reading aloud. *Perception*, *40*(10), 1193–1205.

Moore, B. C. (2015). *Auditory processing of temporal fine structure: Effects of age and hearing loss*. Singapore: World Scientific.

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.

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.

Patel, A. D. (2012). The OPERA hypothesis: Assumptions and clarifications. *Annals of the New York Academy of Sciences*, *1252*(1), 124–128.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., & Allerhand, M. (1992). Complex sounds and auditory images. *Auditory Physiology and Perception*, *83*, 429–446.

Pechmann, T., & Mohr, G. (1992). Interference in memory for tonal pitch: Implications for a working-memory model. *Memory & Cognition*, *20*(3), 314–320.

Pitt, M. A., & Crowder, R. G. (1992). The role of spectral and dynamic cues in imagery for musical timbre. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 728–738.

Rose, N. S., Buchsbaum, B. R., & Craik, F. I. (2014). Short-term retention of a single word relies on retrieval from long-term memory when both rehearsal and refreshing are disrupted. *Memory & Cognition*, *42*(5), 689–700.

Schendel, Z. A., & Palmer, C. (2007). Suppression effects on musical and verbal memory. *Memory & Cognition*, *35*(4), 640–650.

Schulze, K., & Koelsch, S. (2012). Working memory for speech and music. *Annals of the New York Academy of Sciences*, *1252*(1), 229–236.

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

Schulze, K., Vargha-Khadem, F., & Mishkin, M. (2012). Test of a motor theory of long-term auditory memory. *Proceedings of the National Academy of Sciences*, *109*(18), 7121–7125.

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., Fujinaga, I., & McAdams, S. (2016a). 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., Jones-Mollerup, K., & McAdams, S. (2016b). Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds. *Frontiers in Psychology*, *6*(1977). doi:10.3389/fpsyg.2015.01977

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90.

Soemer, A., & Saito, S. (2015). *Maintenance of auditory-nonverbal information in working memory*. Psychonomic Bulletin & Review Published online. doi:10.3758/s13423-015-0854-z

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.

Thorn, A. S., Frankish, C. R., & Gathercole, S. E. (2008). The influence of long-term knowledge on short-term memory: Evidence for multiple mechanisms. In A. S. Thorn & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 198–219). New York, NY: Psychology Press.

Thorn, A. S. C., Gathercole, S. E., & Frankish, C. R. (2002). Language familiarity effects in short-term memory: The role of output delay and long-term knowledge. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(4), 1363–1383.

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, 0662–0672.

Weber, R. (1991). The continuous loudness judgement of temporally variable sounds with an “analog” category procedure. In *5th Oldenburg Symposium on Psychological Acoustics* (pp. 267–289). Oldenburg: BIS.

Appendix. Transformation and selection of sounds

Sound synthesis. We used MATLAB version R2013a (The MathWorks, Inc., Natick, MA) and a linear 24-band Gammatone-filterbank decomposition (Patterson et al., 1992) as implemented in the MIRtoolbox (Lartillot & Toiviainen, 2007). Transformations were derived by mismatching the temporal fine structure (TFS) of the filterbands of one signal with the bands' spectrotemporal envelope (ENV) of another. Additionally, we included “filterbank scrambled” (FBS) sounds as input signals for the transformation process. These were generated in the following way: We (i) decomposed the acoustic sounds into 24 Gammatone-filterbands, (ii) randomly selected 4 sounds from the 14, (iii) allocated their filterbands such that each of the four sounds contributed to the new sound with six different bands, and (iv) added all 24 distinct bands. Six such sounds were selected, denoted as FBS 1–6 below. Among these, FBS 1 & 2 possessed a slow attack, FBS 3 & 4 a sharp attack, and FBS 5 & 6 attacks in between the two extremes. See Siedenburg, Jones-Mollerup, and McAdams (2016b) for more details on the transformation process.

Familiarity and identification judgments. Among the around 400 resulting transformations, we selected 70 to be rated in a dedicated experiment on perceptual familiarity and other variables. The selection favoured timbres that seemed unfamiliar to the experimenters, but did not contain too much narrowband noise (an artifact that was introduced in some transformations by boosting the amplitude of filterbands with low energy). All sounds were normalised in peak amplitude. An experiment assessed perceptual familiarity and source identification of the resulting 70 transformed tones and 14 original recorded acoustic tones. Fifteen musicians participated. In every trial of the experiment, a single stimulus from the 84 tones was presented to participants. They were asked to choose an identifier from a list of eight possible options. The list consisted of six musical instrument names. For recorded timbres, it contained the correct label and five randomly chosen

Table A.1. List of tones used in Exps. 1 and 2 with mean familiarity ratings.

	Set 1 (Recordings)		Set 2 (Transformations)		
#	Instrument	Famil.	TFS	ENV	Famil.
1	Bass Clarinet	4.3	Bass Clarinet	FBS2	1.6
2	Bassoon	3.1	Bassoon	Harpsichord	1.9
3	Flute	4.1	FBS1	Violoncello	1.8
4	Harpsichord	4.5	FBS2	Violoncello	2.1
5	Horn	4.2	FBS3	FBS2	2.1
6	Harp	4.1	FBS6	Trumpet	1.9
7	Marimba	4.6	Flute	FBS1	2.1
8	Piano	4.3	Harp	FBS3	1.7
9	Trumpet	4.8	Harpsichord	FBS4	2.3
10	Violoncello	4.7	Horn	FBS6	2.0
11	Violonc. Pizz.	4.5	Marimba	Harpsichord	2.0
12	Vibraphone	4.3	Trumpet	FBS5	2.3
13	Violin	3.4	Violin	Piano	2.4
14	Violin Pizz.	4.4	Violoncello	Vibraphone	2.0

Notes: TFS, temporal fine structure; ENV, envelope; FBS, filterbank scrambling (see text).

Table A.2. List of memory sequences.

			Probes					
			recs & trans match		recs non-match		trans	
Lists recs & trans			A	B	A	B	A	B
11	12	6	11	12	1	7	8	1
11	4	3	11	4	13	6	2	10
10	7	4	10	7	5	14	12	11
2	1	9	2	1	11	5	8	12
5	14	13	14	13	1	9	8	12
1	5	11	5	11	7	2	13	4
2	6	8	6	8	13	5	4	13
8	14	2	14	2	1	6	11	12
10	13	3	13	3	14	1	8	14
9	4	7	7	9	12	5	5	3
10	13	7	7	10	5	14	4	2
5	3	9	9	5	11	2	4	10
8	12	14	14	8	2	7	1	2
6	1	12	12	6	9	7	8	11

Notes: Digits 1–14 refer to the materials of recordings (recs) and transformations (trans) as provided in Table A.1. Lists and matching probes rely on the same numbering structure for both materials. Non-matching probes are selected differently across materials, in order to obtain a similar distribution of list-probe dissimilarities across material conditions. Non-match probes in the A columns feature high list-probe dissimilarity, and the B columns contain low-dissimilarity probes. Exp. 1 uses all listed trials. In Exp. 2, only the probes listed in the columns A are used.

labels from the remaining set. For transformations, it involved the two labels of the instruments that had been involved with their TFS or ENV, plus four labels chosen randomly from the remaining set. For instance, if a transformation was derived from a piano's TFS and a violin's ENV, then both instrument names, piano and violin, would be part of the list. The list further contained the two options "unidentifiable" and "identifiable but not contained in list". If the participant selected the latter option, a dialogue box appeared prompting them to enter an appropriate identifier in the text box on the screen. They could then continue, whereupon they heard the sound a second time and were presented with two analog-categorical scales on which they had to rate familiarity (1 – highly unfamiliar, 5 – highly familiar) and artificiality (1 – very natural, 5 – very artificial) (Weber, 1991). The 14 transformations that received the smallest mean familiarity ratings were selected for use in the main experiment (see Table A.1).

Dissimilarity ratings. Subsequently, six expert musician listeners equalised the perceived loudness of familiar recordings and unfamiliar transformations against a reference sound (marimba) by adjusting the amplitude of the test sound until it matched the loudness of the reference sound. The levels were then set to the median of the loudness adjustments.

In order to be able to control for perceptual similarity among timbres, 24 musicians rated pairwise dissimilarity for both sets of sounds. Sets were presented separately, and the order of sets was counterbalanced across participants. The 105 pairs of stimuli (14 identical, 91 non-identical) were presented at a 300-ms inter-stimulus-interval and participants provided dissimilarity ratings on an analog-categorical scale (1 – identical, 9 – very dissimilar). The order of stimulus presentation (AB vs. BA) was counterbalanced across participants. See Siedenburg et al. (2016b) for more details on individual sounds and their familiarity and dissimilarity relations.