Semantic Crosstalk in Timbre Perception

Zachary Wallmark

Abstract
Many adjectives for musical timbre reflect cross-modal correspondence, particularly with vision and touch (e.g., “dark–bright,” “smooth–rough”). Although multisensory integration between visual/tactile processing and hearing has been demonstrated for pitch and loudness, timbre is not well understood as a locus of cross-modal mappings. Are people consistent in these semantic associations? Do cross-modal terms reflect dimensional interactions in timbre processing? Here I designed two experiments to investigate crosstalk between timbre semantics and perception through the use of Stroop-type speeded classification. Experiment 1 found that incongruent pairings of instrument timbres and written names caused significant Stroop-type interference relative to congruent pairs, indicating bidirectional crosstalk between semantic and auditory modalities. Pre-Experiment 2 asked participants to rate natural and synthesized timbres on semantic differential scales capturing luminance (brightness) and texture (roughness) associations, finding substantial consistency for a number of timbres. Acoustic correlates of these associations were also assessed, indicating an important role for high-frequency energy in the intensity of cross-modal ratings. Experiment 2 used timbre adjectives and sound stimuli validated in the previous experiment in two variants of a semantic-auditory Stroop-type task. Results of linear mixed-effects modeling of reaction time and accuracy showed slight interference in semantic processing when adjectives were paired with cross-modally incongruent instrument timbres (e.g., the word “smooth” with a “rough” timbre). Taken together, I conclude by suggesting that semantic crosstalk in timbre processing may be partially automatic and could reflect weak synesthetic congruency between interconnected sensory domains.

Keywords
Automaticity, cross-modal correspondences, semantics, speeded classification, Stroop, timbre

Submission date: 2 August 2018; Acceptance date: 5 April 2019

Musical timbre is commonly described using adjectives borrowed from non-auditory senses—think of a “smooth” saxophone, “dark” voice, or “brilliant” violin. These surprisingly consistent associations between qualities of sound and other sensory modalities, particularly vision and touch, have been documented across varying historical periods, languages, and cultural contexts (for reviews, see Wallmark & Kendall, in press; Saitis & Weinzierl, 2019). Cross-modal metaphors were essential to Jean-Jacques Rousseau’s (1772) first formal definition of timbre as “[a sound’s] harshness or softness, its dullness or brightness.” To talk about timbre is often to synesthetically evoke other congruent domains of human sensory experience. While the ubiquity of cross-modal correspondence in timbre semantics has long been noted, the cognitive mechanisms undergirding these common descriptive practices remain poorly understood. In this article, I examine the interrelationship between semantic and timbre perception by exploring the degree to which this common lexicon may reflect automatic associations. To do so, I use a Stroop-style speeded classification task to evaluate “semantic crosstalk” (Melara & Marks, 1990) between auditory perception and common cross-modal adjectives.

Cross-modal correspondence refers to the perceptual and/or semantic interactions between two or more sensory modalities. Correspondences between vision and hearing are pervasive and well documented: writers since antiquity have remarked upon them, and psychologists have been...
studying them since the late 19th century (for historical review, see Marks, 1975). Contemporary researchers often investigate these interactions using a speeded classification paradigm (e.g., see Garner, 1974; Garner & Felfoldy, 1970). In this type of design, participants react times and error rates are collected for classifications of two concurrently presented cross-modal stimuli that vary both orthogonally (i.e., in attended sensory dimension) and in degree of correlation (i.e., whether the stimuli seem to match or mismatch). This paradigm is related to the classic color-word task developed by Stroop (1935), which demonstrated that incongruent task-irrelevant color words interfere with the classification of font colors, but not vice versa. This task has been extensively validated and broadened to explore the automaticity of many other cross-modal and semantic relationships (for reviews, see MacLeod, 1991; Marks, 2004; Spence, 2011). It has also been used in a small handful of studies pertaining to automaticity in music perception, including pitch recognition among possessors of absolute pitch (Akiva-Kabiri & Henik, 2012), notated pitches and note names among trained musicians (Grégoire, Perruchet, & Poulin-Charronnat, 2013; Zakay & Glicksohn, 1985), and cognitive interference in response to musical consonance/dissonance (Masataka & Perlovsky, 2004) and loudness have been shown to map onto pitch height and loudness are commonly associated with visual brightness, as operationalized in the form of dim/bright light, colors, and shapes (Klapetek, Ngo, & Spence, 2012; Marks, 1987; Marks, Hammeal, & Bornstein, 1987; Martino & Marks, 2000; Parise & Pavani, 2011). Third, loudness and pitch depth are commonly associated with physical size: loud, low sounds provide a better perceptual correlation with large objects than small ones (Eitan, Schupak, Gotler, & Marks, 2014; Gallace & Spence, 2006; Parise & Spence, 2012; Plazak & McAdams, 2017). Additionally, though less consistently, pitch height and loudness have been shown to map onto touch, both in tactile sensation (Peeva, Baird, Izmirlı, & Blevins, 2004) and metaphorically (Eitan & Rothschild, 2010): higher/louder sounds are generally associated with rougher textures and with the adjectives “rough,” “sharp,” and “hard.”

As evinced above, scholars of multisensory perception commonly speak of sound as possessing only two perceptual dimensions—pitch and loudness (Marks, 2004; Spence, 2011). Often neglected in this discourse is the multidimensional auditory parameter of timbre, which refers to the character or quality of sound, “what it sounds like” covarying with pitch and loudness (Handel, 1995; McAdams & Goodchild, 2017). In many ways, this is a surprising omission; since the semantic norms of musical timbre are so thoroughly infused with cross-modal metaphor, timbre seems tailor-made for study using a multisensory paradigm. Indeed, a few studies have reported correspondences between timbral attributes and visual texture and shape. Giannakis (2006) found that aspects of visual texture cross-modally mapped onto timbral attributes: texture contrast related to high-frequency energy, coarseness to spectral spread, and pattern periodicity to auditory roughness. Parise and Pavani (2011) asked participants to vocalize the vowel /a/ after being primed with visual stimuli of varying shapes, luminance, and sizes and reported an association between shape and spectral energy distribution (triangles were associated with increased energy in higher-frequency partials compared with dodecagons).

Evidence for cross-modality in timbre semantics comes from a number of sources, including perceptual studies (Alluri & Toiviainen, 2010; von Bismarck, 1974; Eitan & Rothschild, 2010; Kendall & Carterette, 1993a, 1993b; Lichte, 1941; Pratt & Doak, 1976), ethnographic fieldwork and interviews (Fales, 2002; Nykänen & Johansson, 2003; Porcello, 2004; Reymore & Huron, 2018), and online discourse analyses (Fales, 2018; Ferrer & Eerola, 2011). In a comparative perceptual study between English and Greek speakers, Zacharakis, Pastiadias, and Reiss (2014, 2015) reported that both language groups tended to structure timbre terminology according to three primary semantic categories: luminance (e.g., “bright, dim”), texture (“smooth, rough”), and mass (“heavy, light”). Similarly, in an analysis of timbre description in a corpus of orchestration treatises, Wallmark (2019) found that around 20% of all terms consisted of cross-modal metaphors. In addition to pitch and loudness, then, cross-modality appears intrinsic to how we perceive, conceptualize, and communicate about musical timbre.

How automatic are these cross-modal associations between timbre and the other senses? Although there is robust evidence for cross-modality in the descriptive conventions of timbre, the mechanisms governing these routines have not been adequately theorized. One open question is whether common cross-modal adjectives for timbre represent a cognitively mediated, top-down processing stage in which auditory sensations are—arduously at times—translated into lexical categories (Glaser & Glaser, 1989) or whether they suggest interactions at an earlier, perceptual stage of encoding (Melara & Marks, 1990; Spence & Deroy, 2013). In the case of cross-modal correspondences involving pitch and loudness, it appears likely that multiple mechanisms are at work: all interrelations are semantically mediated through linguistic and cultural filters, but audio-visual correspondences have also been documented in preverbal infants as young as 20–30 days old (Lewkowicz & Turkewitz, 1980; Walker et al., 2010).
Accordingly, Spence (2011) suggests that while correspondences are semantically congruent (i.e., matching in identity or meaning, thus sensitive to statistical learning, linguistic mediation, and strategic input), some may also be synesthetically congruent (i.e., structurally resembling each other irrespective of semantics, and present early in human development). These “low-level” synesthetic interactions in turn influence semantic conventions downstream, as iconic modes of representation transition to symbolic (Bruner, 1964). Do cross-modal timbre adjectives reflect a synesthetic level of correspondence?

In the context of timbre, it also seems relevant to draw a distinction between categorical and continuous modes of lexical activation. Here, “categorical” refers to sound source identification, while “continuous” refers to source description, often in the form of adjectival scales. These two processes overlap and interact in complex ways (Siedenburg, Jones-Mollerup, & McAdams, 2016) but are nonetheless useful to disambiguate. Categorical timbre semantics has been approached from an information processing orientation, in which sensation, perception, and representation are organized into a multi-stage, hierarchical encoding process (see McAdams, 1993); it has also been treated through the lens of ecological psychology as a direct pick-up of source mechanics and identity (Handel, 1995). By contrast, the processing of continuous lexical responses has received less attention. On the face of it, it would seem to constitute a second-order level of lexical activation that involves additional cognitive load. Yet, if valid, the ecological perspective might complicate this account, implying that identification and meaning are mutually entwined and roughly synchronous (Clarke, 2005; Gaver, 1993; Gibson, 1966). For example, corroborating this view, recent neuro-imaging data suggest that perception of aversive timbres commonly described with negative texture metaphors (“harsh,” “rough”) activates regions of the secondary somatosensory cortex also responsible for processing tactile sensation (Wallmark, Iacoboni, Deblieck, & Kendall, 2018; see also Schützmann, Caetano, Hlushchuk, Jousmäki, & Hari, 2006).

Given the above, we might anticipate that cross-modal interference would be more pronounced in cases of stimulus conflict involving categorical classification (e.g., hearing a piano but seeing a guitar). But if cross-modal timbre terms in some instances reflect quick and automatic multisensory associations, it is plausible that continuous mismatches may likewise elicit interference as cross-modal adjectives clash with the semantic implications of incongruent timbres (e.g., the word “gentle” presented alongside a highly distorted electric guitar). Relatedly, we could predict that congruent timbre–semantic associations should exhibit “redundancy gain” (Melara & O’Brien, 1987) in the form of faster reaction times and lower error relative to incongruent matches and would again likely be more pronounced for categorical identification. To date, however, this hypothesis has never been explicitly tested.

**Study Aim**

The present study investigates semantic crosstalk, a specific kind of cross-modal correspondence located within the lexical access stage of sensory processing (Melara & Marks, 1990; Pomerantz, Pristach, & Carson, 1989). Semantic crosstalk in this context refers to the interaction of timbre processing with categorical/continuous modes of semantic identification and description. The central goal of this study was to test whether semantic crosstalk occurs in timbre perception—that is, whether mismatches between sensory inputs will result in Stroop-style interference using a speeded classification paradigm. Here is my logic: if descriptive words for timbre represent a more strategic, deliberative process—something we apply to our perceptions somewhat haphazardly, and fairly late in appraisal—then we should not expect to see any semantic crosstalk. That is, the processing of cross-modal adjectives and timbre should be independent and unrelated. If reaction time is slowed and/or error elevated by incongruent matches of words and timbres, however, perhaps timbre terms exhibit some degree of automaticity in processing when coupled with congruent or incongruent timbres.

One could posit weak (H1) and strong (H2) versions of this hypothesis corresponding to categorical and continuous semantic modes. In H1, interference is expected to be observed when instrument names and sounds disagree. This premise is analogous to cross-modal studies using naturalistic sounds and images/names (e.g., seeing an elephant but hearing a sheep): authors report that congruent matches facilitate classification (Chen & Spence, 2010, 2011; Yuval-Greenberg & Deouell, 2009), with quick and effortless sound-source elisions (Dell’acqua & Job, 1998; Glaser & Glaser, 1989). This result would suggest a failure of selective attention as participants struggle to reconcile competing semantic and sonic frames. Moreover, consistent with the multisensory perception literature (e.g., see Stroop, 1935; Donohue, Appelbaum, Park, Roberts, & Woldorff, 2013), words will likely have a greater effect on timbre perception than vice versa.

Correspondingly, H2 predicts that Stroop-style interference will be observable between mismatching cross-modal timbre adjectives and timbres. Since semantic and timbral dimensions are integral in the categorical context but possibly separable here (Garner & Felfoldy, 1970), it is likely that semantic crosstalk in incongruent adjective/timbre relationships will be substantially weaker than in instrument name/timbre pairings. This result would indicate that semantic and timbre processing might be imbricated at a fairly automatic level.

Additionally, the study explores whether this hypothesized crosstalk is associated with musical training. Does extensive exposure to certain instrumental timbres and timbre-descriptive lexicons strengthen automaticity of cross-modal associations? Musical training has been shown to play a role in some perceptual evaluations of timbre
(Chartrand & Belin, 2006; McAdams, Douglas, & Vempala, 2017; Siedenburg & McAdams, 2018; but see Filipic, Tillmann, & Bigand, 2010): we may expect, then, that trained musicians will possess a more finely tuned timbre–semantic repertoire than nonmusicians. On the other hand, if cross-modal correspondence reflects some degree of synesthetic congruency, then musical background would plausibly play an indeterminate role. Of interest in this regard, too, is the possible role of conventional sound-source associations in the perceptual immediacy of descriptive conventions. To attempt to control for these possible associations, this study used both natural acoustic instruments and synthetic signals as stimuli (e.g., see Grey & Moorer, 1977; McAdams, Beauchamp, & Meneguzzi, 1999; Golubock & Janata, 2013).

Two experiments were designed to explore the association between timbre perception and semantic representation. Experiment 1 used a modified audio-visual/semantic Stroop task to test the effect of categorical word stimuli (instrument names) on instrument timbre perception and, vice versa, the effect of timbre on word classification (H1). Reaction time and error were measured in attend-sound and attend-word tasks, and stimuli were paired in congruent and incongruent conditions (along with a semantically neutral control) in order to measure the “total Stroop effect” (Brown, Gore, & Pearson, 1998) associated with mismatch between timbral and semantic frames.

To select stimuli for Experiment 2, Pre-Experiment 2 tested a large set of natural instrument and synthesizer signals using bipolar semantic differential scales (Osgood, Suci, & Tannenbaum, 1957) consisting of common luminance and texture terms for timbre, “dark–bright” and “smooth–rough” (Zacharakis, Pastiadis, & Reiss, 2014, 2015). Additionally, acoustic data from this stimuli set were computationally extracted in order to explore which spectrotemporal characteristics best predicted semantic ratings.

Experiment 2 extended the earlier paradigm to continuous timbre semantics in order to test for Stroop-type interference on the speeded identification of four common cross-modal descriptive adjectives when presented with a smaller subset of validated signals as exemplars of these adjectives (e.g., the “darkest” and “brightest” sounds). Two variants of this experiment were conducted: Experiment 2a used a large subset of signals (64) and a semantically neutral control signal, while Experiment 2b used a smaller subset of repeating signals (12) and a no-sound baseline. Reaction times and error rates were analyzed in an attend-word task to determine whether crosstalk is present in the perception of timbre terms.

**Experiment 1: Speeded Classification of Instruments**

**Participants**

Forty-six undergraduates (26 female, 20 male) were recruited to participate ($M$ age = 21.35, $SD$ = 3.68).

**Table 1.** Median reaction time (ms), variance (IQR), and error rate by task and congruency.

| Congruency | Attend-sound | Attend-word | Overall |
|------------|--------------|-------------|---------|
| Baseline   | Med 519 116 0.02 | Med 562 180 0.05 | Med 553 150 0.04 |
| Congruent  | Med 542 184 0.02 | Med 592 242 0.03 | Med 567 191 0.03 |
| Incongruent| Med 738 349 0.08 | Med 640 318 0.09 | Med 670 341 0.08 |

Note. IQR: interquartile range; Med: median.

Musical training ranged from 0 to over 10 years of formal instruction on an instrument or voice (assessed using the Goldsmiths Musical Sophistication Index; see Müllensiefen, Gingras, Musil, & Stewart, 2014). Participants with 3 years or less of musical training were referred to here as “nonmusicians,” while those with 4 or more years of formal training on an instrument (which included a number of music majors) were considered “musicians” (see Supplementary Materials, SM Table 1 for distribution of musical training among participants). All participants reported normal hearing and normal or corrected-to-normal vision. None of the participants took part in any of the other experiments reported in this study. Students received extra course credit for their contribution. All experiments were approved by the Southern Methodist University (SMU) Institutional Review Board.

**Materials**

Sound stimuli consisted of three instrumental signals taken from the McGill University Master Samples (MUMS) library (Opolko & Wapnick, 1987): trumpet, clarinet, and violin. Signals were 1.5 s in duration, with a 200 ms linear fade-out applied to the end. Following Eerola, Ferrer, and Alluri (2012), the pitch was D#4 (311 Hz), which was selected due to maximum overlap between different natural instruments’ ranges as well as proximity to “average pitch.” The sampling rate was 44.1 kHz. Since perceived loudness can never really be equal for every listener (Hajda, Kendall, Carterette, & Harshberger, 1997), loudness was equalized manually by the author. Signals were drawn from a larger stimuli set used in Pre-Experiment 2, as discussed later (see Supplementary Materials for stimuli.)

**Procedure**

All experiments were conducted in a quiet room using iMac computers, and stimuli were presented at a subjectively determined comfortable volume level through Bose SoundTrue headphones. Experiment 1 was administered using DirectRT software (Jarvis, 2016a). A beige fixation cross appeared for 2 s in the center of a black screen (27-inch display, 2,560 × 1,440-pixel resolution; Times New Roman font, 24-point size). Participants were instructed to focus their vision on this cross. After 2 s, it was replaced
by one of three word cues: (1) TRUMPET, (2) CLARINET, or (3) a neutral baseline condition consisting of the letters XXXX, which is a common stimulus in the Stroop literature used to determine reaction time in response to visual information that is not linguistically meaningful (see MacLeod, 1991). All word cues were presented in the same color/font/size as the fixation cross. (Visual cues are hereafter referred to in capital letters; auditory cues in lowercase.) Simultaneous to the presentation of the visual cue, participants heard one of the three sound stimuli (trumpet, clarinet, or violin).

Participants took part in two orthogonally varied tasks: (1) attend-sound, in which they were asked to identify as quickly and accurately as possible the instrument they heard through the headphones while disregarding the word displayed on the screen, and (2) attend-word, in which they were asked to quickly and accurately identify the word they saw on the screen while ignoring the sound. The procedure was identical for both tasks save for the difference in response, and the order of the two tasks was randomized between participants. Each of the three response types were assigned to different keys (left-arrow, down-arrow, right-arrow). To familiarize themselves with the response key assignment and the stimuli, prior to beginning each task subjects completed brief practice sessions consisting of six word/sound combinations (two trumpet, two clarinet, and two violin, paired with XXXX control). Since participants in Stroop-type tests can be sensitive to training effects (see MacLeod, 1991), practice was kept to a minimum. Participants were instructed not to “cheat” by closing their eyes or looking away from the fixation cross and visual cues during the attend-sound task.

Each task consisted of three congruency conditions varying in degree of correlation: (1) congruent relationship between word and auditory stimuli (TRUMPET-trumpet and CLARINET-clarinet), (2) incongruent relationship between word and auditory stimuli (TRUMPET-clarinet and CLARINET-trumpet), and (3) a baseline condition consisting of the simultaneous presentation of XXXX and the violin signal. Each condition appeared 10 times in a random order (30 total conditions presented for each of the two trials = 60 total per participant, 2,760 total responses collected). The experiment took approximately 10 minutes.

**Results**

All data preparation and analyses were conducted using R Version 3.4.4 (R Core Team, 2004–2016). (See Supplemental Materials for data and R scripts.) Table 1 summarizes reaction time (RT) and accuracy data collapsed between the different tasks and congruency conditions to produce median RTs, interquartile ranges, and error rates. Only correct RT responses to the congruent and incongruent conditions were included in the subsequent RT analysis. Additionally, since excessively short or long RTs typically indicate inattention in multisensory psycholinguistic tasks, following Whelan (2008), an outlier threshold of < 100 ms and > 2000 ms was applied, resulting in the exclusion of 36 responses (all over 2,000 ms). Finally, data were scanned for participants with a total error rate worse than chance (≥ 33%), although none of the participants met this threshold. (For more on analysis methods, see Lachaud & Renaud, 2011).

To test the effect of task (two levels: attend-word vs. attend-sound), congruency (two levels: congruent vs. incongruent), and musical training (two levels: musician vs. nonmusician) on log-transformed RTs, a linear mixed-effects model was computed. Since incorrect responses were trimmed for the analyses presented in this article, leading to incomplete datasets, the linear mixed model (LMM) approach offers more flexibility than traditional repeated-measures analysis of variance (ANOVA), which typically requires listwise deletion or median imputation in the event of missing data. To account for repeated measurements, participant variability was modeled as a random effect. All mixed-effects models were created with the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015), and significance levels of fixed main effects and interactions were calculated using Type II Wald chi-square tests in the car package (Fox & Weisberg, 2010). Following the suggestion of Grégoire, Perruchet, and Poulin-Charronat (2013), the baseline condition was excluded from analysis since it did not shed light on the “total Stroop effect” (Brown et al., 1998) between congruent and incongruent pairings of words and sounds.

The model explained close to half of the variance in reaction time ($R^2 = .41, p < .0001$) with significant main fixed effects of task and congruency (Table 2). For the latter, incongruent RTs were significantly longer than congruent ($M = 640$ ms vs. $592$ ms, $t = -3.58, p < .0001$). Moreover, task and musical training exhibited a significant interaction: the attend-word task was faster, irrespective of musical training, but the difference between tasks among nonmusicians was not as pronounced. No main fixed effect of musical training was observed.

As shown in Figure 1, results indicate that overall participants were 48 ms quicker in the congruent condition

| Table 2. Results of linear mixed-effect models exploring fixed effects of musical training, task, and congruency on log-transformed RTs and error rate in Experiment 1. |
|-----------------|-----------|-----------------|-----------|-----------------|-----------|
|                 | log-RT ($R^2 = .41$) | Error ($R^2 = .13$) |
|                 | df | Wald $\chi^2$ | $p$ | df | Wald $\chi^2$ | $p$ |
|-----------------|-----------|-----------------|-----------|-----------------|-----------|
| Musical training | I            | 1.51            | .22 | I            | 1.16            | .28 |
| Task            | 1          | 276.19          | <.0001 | 1          | 0.58            | .45 |
| Congruency      | I          | 37.24           | <.0001 | I          | 28.7            | <.0001 |
| Training * task | I          | 11.57           | .001 | I          | 0.84            | .36 |

*Note. RT: reaction time. $p$-values <.05 indicated in bold. Non-significant interactions not included.*
was the interaction between task and musical training. In the RT analysis, task was not significant in this model, nor were the responses to congruent pairs more error-prone than responses to congruent pairs. Unlike responses to incongruent pairs both took longer and were substantially more error-prone than attend-word. The rate for incongruent pairs was 9.13\% (t = 1.6, p < .0001); however, echoing the RT results, a significant main fixed effect of congruency was found. Congruent word-sound pairs produced an average error rate of 3\% while the rate for incongruent pairs was 9\% (t = 1.6, p < .0001); that is, responses to incongruent pairs both took longer and were more error-prone than responses to congruent pairs. Unlike the RT analysis, task was not significant in this model, nor was the interaction between task and musical training.

### Summary of Experiment 1

This experiment investigated categorical timbre perception through the speeded classification of musical instrument names and tones. Verifying H1, participants suffered from significant Stroop interference in auditory classification (attend-sound) when names conflicted with timbres; likewise, though more subtly, in the word classification task (attend-word), interference obtained when timbral stimuli were negatively correlated with written instrument names. Bidirectional semantic crosstalk suggests that in both trials participants exhibited a failure of selective attention. This result is consistent with the literature on cross-modal relations between environmental sounds and semantic classification. For example, it has been proposed that images, words, and sounds indicating specific referents (e.g., a dog’s bark and a picture of a dog) can directly access the same underlying semantic category (Chen & Spence, 2010, 2011; Yuval-Greenberg & Deouell, 2009). Similarly, categorical musical timbre perception is premised to a large degree on the stability of referential sound-source associations. It is therefore intuitive that stimulus conflict between written instrument names and tones would lead to decisional processing delays and response error, since source identification is one of the most fundamental properties of timbre perception.

Do common descriptive adjectives for timbre—and more specifically, cross-modal adjectives (Zacharakis et al., 2014, 2015)—elicit a similar effect? In Experiment 2, I address this question using two variants of a Stroop-type speeded word identification task.

### Pre-Experiment 2: Stimuli Selection—Cross-Modal Ratings of Natural and Synthesized Timbres and Their Acoustic Correlates

#### Methods

A pre-experiment was conducted to select appropriate natural instrument and synthesizer signals for use as stimuli in Experiment 2 (for complete details, see Supplementary Materials). Participants (N = 29) evaluated a total of 93 signals—50 natural, 43 synthesized—that were selected to represent a diverse and ecologically valid range of timbres common in western classical and popular music. As in Experiment 1, all signals were 1.5 s in duration (with 200 ms fade-out), pitch D#4, sampled at 44.1 kHz, and manually equalized for loudness. Natural signals were selected from MUMS library (Opolko & Wapnick, 1987) and synthesized signals were generated using Apple GarageBand (version 10.1.6) software instrument presets.

Participants were instructed to listen to the 1.5 s tones on the computer and rate them on 7-point bipolar semantic scales measuring intensity of cross-modal associations (Osgood et al., 1957; Zacharakis et al., 2014): dark–bright (luminance) and smooth–rough (texture), with 1 corresponding on the luminance scale to “very dark” and the texture scale to “very smooth” and 7 corresponding on the luminance scale to “very bright” and on the texture scale to “very rough”. Participants were advised to use the full extent of the scale in their answers. In order to familiarize them with the stimuli, a random subset of 10 signals was played before beginning each of the two sections (natural
and synthesized = 20 signals presented); additionally, participants practiced using the horizontal rating scale on three 1.5 s test signals (sine, square, and sawtooth waves, all D#4 and equalized for loudness).

The pre-experiment was presented using MediaLab software (Jarvis, 2016b). In order to control for differences in perceptual attributes within this heterogeneous set of signals, natural and synthesized stimuli were presented separately, and the order of the two trials was randomized. Within each trial, the two verbal scales were also presented separately in a randomized order. Stimuli were likewise randomized, with a single rating judgment for each stimulus. Each participant thus evaluated a total of 186 signals: 50 natural stimuli × 2 modalities (luminance and texture), and 43 synthesized stimuli × 2 modalities. The complete experiment took approximately 20 minutes.

**Stimuli Selection**

The internal consistency of the natural instrument and synthesizer luminance/texture verbal scales was acceptable to very good (natural-luminance: $M_{\text{Cronbach's } \alpha} = .75$; natural-texture: $M_{\text{Cronbach's } \alpha} = .89$; synthesizer-luminance: $M_{\text{Cronbach's } \alpha} = .9$; synthesizer-texture: $M_{\text{Cronbach's } \alpha} = .84$). The two modality scales were strongly correlated, $r(91) = .84$, $p < .0001$, suggesting that perhaps participants were responding to a latent magnitude or intensity dimension underlying both sensory modalities (Smith & Sera, 1992). Ratings data were subjected to linear mixed-effects modeling in order to quantify the strength of cross-modal associations among individual stimuli (see SM Table 2).

To determine the most consistently extreme timbres (“darkest,” “brightest,” “smoothest,” and “roughest”) for use in Experiment 2, mean stimuli ratings were sorted in ascending order (see SM Figure 1). The lowest and highest timbres for each block and modality were selected as exemplars of the perceptual scales; in borderline cases, $SD$ was used as a tie-breaker, with preference given to timbres with the lower variance. For complete details, see Supplementary Materials.

**Acoustic Analysis**

In addition to validation and stimuli selection, an exploratory aim of this pre-experiment was to determine the acoustical determinants of semantic ratings. To explore this question, 14 common acoustic descriptors were computationally extracted using MIRtoolbox 1.6.1 (Lartillot & Toiviainen, 2007; see SM Table 3 for complete list) in MATLAB (Release, 2016a; The MathWorks, Inc.). A Principal Components Regression (PCR) was performed predicting luminance and texture ratings from acoustic descriptor data using the pls package in R (Mevik, Wehrens, & Liland, 2016). PCR uses orthogonally transformed (thus uncorrelated) principal components as predictors in a least-squares linear regression: for this reason, it is ideally suited for models in which there are numerous related predictors, such as acoustic descriptors. Because of the high correlation between luminance and texture scales ($r = .84$), the two modalities were collapsed to a single index of visuo-tactile intensity (i.e., the perceived brightness and roughness of each timbre).

PCR on scaled acoustic data revealed three principal components (Eigenvalues $> 1$) that together explained 76% of variance in visuo-tactile intensity ratings. Table 3 displays factor loadings for the final cross-validated model, along with estimates of regression coefficients and significance levels for each acoustic descriptor. PC1 (50%) was associated with a number of indices of high-frequency energy, including spectral centroid. PC2 (13%) was

| Acoustic descriptor | PCs and factor loadings |
|--------------------|-------------------------|
|                    | 1 (50%) | 2 (13%) | 3 (12%) | b   | SE  | t     | p     |
| Low-freq. spectral flux (< 800 Hz) | —— | —— | —— | —— | —— | —— | —— |
| High-freq. spectral flux (> 800 Hz) | —— | —— | —— | —— | —— | —— | —— |
| Zero-cross rate | .27 | —— | —— | 12 | .04 | 2.89 | <.01 |
| Rolloff | .35 | —— | —— | .09 | .01 | 5.86 | <.0001 |
| Brightness | .33 | —— | —— | .09 | .02 | 3.61 | <.001 |
| Spectral centroid | .36 | —— | —— | .07 | .02 | 3.89 | <.001 |
| Spectral spread | .24 | —— | —— | —— | —— | —— | —— |
| Spectral skewness | —— | —— | —— | —— | —— | —— | —— |
| Spectral flatness | .28 | —— | —— | .39 | —— | —— | —— |
| Spectral entropy | .34 | —— | —— | .1 | .02 | 6.02 | <.0001 |
| Spectral irregularity | —— | —— | —— | —— | —— | —— | —— |
| Roughness | .29 | —— | —— | .11 | .03 | 3.36 | .001 |
| Inharmonicity | —— | —— | —— | —— | —— | —— | —— |
| Attack time (log) | —— | —— | —— | .31 | —— | —— | —— |

Note. Loadings < .2 omitted; loadings > .3 are shown in bold (or where none meet this criterion, highest loading in bold). $T$ and $p$-values estimated using jackknife resampling technique; variables with $p$-values $>.05$ are omitted.
associated with decreased spectral fluctuation in lower frequency components (below 800 Hz), plus decreasing inharmonicity and spectral irregularity. Finally, PC3 (12%) was related to increasing spectral spread, flatness, and length of attack, and inversely related to high-frequency spectral flux (above 800 Hz) and roughness. The prevalence of high frequency energy in predicting intensity of cross-modal semantic response (“brighter” and “rougher” timbres) resonates with much of the psychoacoustics literature in linking perceived timbral “brightness,” “sharpness,” and/or “nasality” to spectral center of mass (e.g., Beauchamp, 1982; Schubert & Wolfe, 2006). Full details of this acoustic analysis can be found in the Supplementary Materials.

Experiment 2a: Preliminary Stroop Test of Cross-Modal Timbre Semantics

Participants

Fifty-three undergraduates were recruited for this experiment (36 female, 17 male; \( M_{\text{age}} = 20.18, SD = 2.52 \)). Participants self-reported number of years of formal musical training from 0 to over 10 years (see Supplementary Materials, SM Table 1): “musicians” \( N = 35 \) had 4 or more years of formal training on an instrument/voice, and “nonmusicians” \( N = 17 \) had 3 or fewer years of training. None of the participants were involved in the other experiments reported in this study. All participants reported normal hearing and normal or corrected-to-normal vision and received extra course credit for their participation.

Stimuli

Stimuli consisted of 64 natural and synthesized timbres validated in Pre-Experiment 2 as being most consistently and strongly associated with the two cross-modal adjective scales. Eight stimuli were selected from the complete natural and synthesizer stimuli sets as exemplars of each end of the luminance (“dark,” “bright”) and texture (“smooth,” “rough”) scales, as shown in SM Table 4. Many of the exemplar timbres overlapped between the two modalities: for example, 6 out of 8 natural “bright” timbres were also considered the most “rough,” and 7 of 8 synthesized “dark” timbres were also considered the most “smooth.” Of the 64 total stimuli, 42 were unique. Interestingly, none of the timbres at the extremes of one scale cross-mapped to the opposite end of the other scale: that is, no “dark” timbres were also perceived as “rough,” and no “bright” timbres were perceived as “smooth.” This was true for both natural and synthesized stimuli, again suggesting a structural relationship between the luminance and texture scales based on magnitude of intensity (Smith & Sera, 1992).

Additionally, the grand piano sample from MUMS and the “mysterious synth lead” preset from GarageBand were used as auditory control stimuli for the two trials (1.5 s, D#4, loudness equalized manually). These signals were rated as neutral on luminance and texture scales.

Procedure

The experiment was designed and administered using DirectRT software (Jarvis, 2016a). A beige fixation cross appeared for 2 s in the center of a black screen (see Experiment 1 Procedure section for additional details). After 2 s, the cross was replaced by one of two visual cues: (1) a cross-modal adjective (DARK, BRIGHT; SMOOTH, ROUGH) or (2) a semantically neutral control condition consisting of the letters XXXX, all in the same color/font/size as the fixation cross. To account for the difference between semantic and timbral domains in this task, which presumably adds to cognitive load compared with the fairly straightforward instrument classification task in Experiment 1, a stimulus onset asynchrony (SOA) of 200 ms was applied to the signals. SOA was used to adjust for the speed and automaticity of reading, as demonstrated in many studies (Donohue et al., 2013; Molholm, Ritter, Javitt, & Foxe, 2004; Yuval-Greenberg & Deouell, 2009). For example, Chen and Spence (2011) suggested that a 200–350 ms head start of auditory stimuli may be required to access associated semantic information in time for the presentation of visual stimuli.

Participants were instructed to respond as quickly and accurately as possible by pressing the arrow key assigned to the word that appeared on the screen: DARK/SMOOTH, BRIGHT/ROUGH, or the XXXX control. In contrast to Experiment 1, the entire test was “attend-word”: participants responded to word cues as quickly and accurately as possible while disregarding task-irrelevant sounds. To familiarize themselves with the response key assignment and the stimuli range, prior to beginning each trial participants completed brief practice sessions consisting of eight randomized word/sound combinations of the experimental stimuli.

Natural and synthesized stimuli were presented in separate randomly ordered trials. Modalities (luminance and

| Condition | Luminance Med | Luminance IQR | Luminance Error | Texture Med | Texture IQR | Texture Error |
|-----------|---------------|---------------|----------------|-------------|-------------|---------------|
| Natural   | 508           | 114           | .01            | 506         | 116         | .02           |
| Congruent | 504           | 167           | .05            | 508         | 186         | .06           |
| Incongruent | 519         | 175           | .08            | 508         | 172         | .05           |
| Synthesized | 514         | 112           | .03            | 510         | 102         | .03           |
| Congruent | 513           | 190           | .07            | 509         | 156         | .03           |
| Incongruent | 513         | 168           | .07            | 515         | 150         | .04           |

Note. IQR: interquartile range; Med: median.
texture) were likewise presented separately in a randomized order. In each trial/modality, participants were presented with 48 word-timbre pairs in three conditions (16 randomized pairs each): (1) congruent—e.g., the word BRIGHT with a “bright” timbre; (2) incongruent—e.g., BRIGHT with a “dark” timbre; and (3) control, a neutral XXXX word cue with control timbre. Participants responded to a total of 192 word-timbre pairs (3 conditions \( \times \) 16 pairs = 48/modalit; 48 \( \times \) 2 modalities = 96/trial; 96 \( \times \) 2 trial = 192). The experiment took approximately 30 minutes.

**Results**

Only correct RT responses were included in the RT analysis and an outlier threshold of < 100 and > 2,000 ms was applied, resulting in the trimming of 53 responses (out of 9,696 total). Due to technical difficulties, 8 participants only completed one of the two trial blocks. No participants’ error rates exceeded chance (33%). Table 4 summarizes the data.

Log-transformed RTs were analyzed using a linear mixed-effects model that included three within-subject variables—stimuli block (two levels: natural, synthesized), modality (two levels: luminance, texture), and congruency condition (three levels: congruent, incongruent, and control)—and one between-subject variable, musical training (musician vs. nonmusician), with participants modeled as a random effect. In contrast to Experiment 1, the XXXX-neutral signal control condition was included in this analysis in order to simultaneously eliminate variance associated with both linguistic meaning and cross-modal timbre associations (i.e., to explore pairings that are neither semantically mismatched nor matched). If there were not some degree of association between the validated signals and these cross-modal adjectives, we should expect roughly similar RT/error responses between the three congruency conditions.

The model explained one-third of total variance in RT (\( R^2 = .33, p < .0001 \)). As shown in Table 5, main fixed effects of training and modality failed to reach significance. Block and congruency, however, significantly affected reaction times: natural instruments elicited quicker correct RTs than synthesizers, and congruent pairings were an average of 5 ms faster than incongruent (509 vs. 514 ms). However, while a significant difference between the control and the two experimental conditions was found (for both, \( p < .0001 \) in post-hoc comparison, corrected using False Discovery Rate [FDR]; see Benjamini & Hochberg, 1995), the difference between congruent and incongruent failed to reach significance \( (t = -1.43, p = .07) \). More importantly, though, two-way interactions were observed between block*congruency, training*congruency, and training*block. No three- or four-way interactions were significant.

As shown in Figure 2, the interaction of block and congruency was driven by the incongruent condition. Incongruent pairings of adjectives and natural timbres slowed reaction time by an average of 8 ms relative to congruent and control pairs, while the synthesized stimuli produced interference of just a couple milliseconds. Next, musical expertise appeared to be associated with faster RT in the control condition (20 ms faster than nonmusicians), which is perhaps the result of increased perceptual and encoding acuity of trained musicians in responding to the repeated control timbres (Landry & Champoux, 2017; Shahin, Roberts, Chau, Trainor, & Miller, 2008). Finally, musical training was associated with slightly faster RTs in both stimuli blocks, though the difference is more pronounced for the natural instruments. To be clear, the effects for all significant interactions in this experiment were very small, although they were reliably observed.

Lastly, a logistic LMM predicting error from the independent variables yielded an explanatorily weak but statistically significant model (\( R^2 = .05, p < .0001 \)) (Table 5).

**Table 5.** Results of linear mixed-effect models exploring fixed effects of musical training, adjective modality, block (stimuli type), and congruency on log-transformed RTs and error rate in Experiment 2a.

|             | log-RT (\( R^2 = .33 \)) | Error (\( R^2 = .05 \)) |
|-------------|--------------------------|------------------------|
|             | Wald \( \chi^2 \) df p   | Wald \( \chi^2 \) df p   |
| Musical training | 1 0 .98 1 1.45 .23          |
| Modality     | 1 0.29 .59 1 8.55 .003       |
| Block (stimuli type) | 1 9.89 .002 1 0.76 .38          |
| Congruency   | 2 139.33 <.0001 2 33.31 <.0001 |
| Block * congruency | 2 8.99 .01 2 7.53 .02           |
| Training * congruency | 2 45.59 <.0001 2 2.27 .32     |
| Training * block | 2 4.67 .03 2 1.7 .19            |
| Modality * block | 1 0.4 .82 1 3.9 .05             |

Note. \( p \)-values < .05 indicated in **bold.** Non-significant interactions not included.
Results largely conformed to the RT analysis. Of particular interest was the main fixed effect of congruency on accuracy. All three conditions were significantly different from one another: unsurprisingly, control (M error = 2%) was the most accurate compared with the others (p < .0001), and congruent pairs (5% error) elicited slightly higher accuracy than incongruent (6% error) (both p < .0001, FDR-corrected).

Summary of Experiment 2a

Sixty-four natural instrument and synthesizer signals rated as extreme on cross-modal adjective scales (Pre-Experiment 2) were paired with these same adjectives in a Stroop task, with a cross-modally neutral signal and XXXX-type control. Results suggest that incongruent matches with task-irrelevant natural instrument stimuli produced an 8 ms total Stroop effect relative to congruent matches. Furthermore, interactions involving training indicate that musical expertise was associated with faster RTs to natural timbres and the control condition (both stimuli trials) compared with nonmusicians. Analysis of error rate revealed a slight accuracy advantage of congruent pairs compared with incongruent.

Small Stroop effects in Experiment 2a might be explained by limitations with study design (see also Supplementary Materials, Limitations section). A relatively large number of signals were used, and, despite validation using a rating procedure in the pre-experiment, total Stroop effects at the level of individual stimuli varied widely (see SM Figures 2 and 3). In the follow-up Experiment 2b, then, a smaller subset of repeating signals was employed, along with a larger sample size necessary to detect subtle effects.

Experiment 2b: Stroop Test of Cross-Modal Timbre Semantics

Participants

Participants were 110 undergraduates (66 female, 41 male; M age = 19.87, SD = 2.83) with diverse musical backgrounds (Müllensiefen et al., 2014). As in the other experiments, participants with 3 years or less of musical training were classified as “nonmusicians” (N = 62), while those with 4 or more years of formal training on an instrument were considered “musicians” (N = 47; see SM Table 1 for distribution). All participants reported normal hearing and normal or corrected-to-normal vision, and nobody took part in any of the other experiments reported in this study. Students received extra course credit for their contribution. All experiments were approved by the SMU IRB.

Materials

Sound stimuli consisted of 12 natural instrument and synthesizer signals (three per modality/block) validated as most consistently and strongly associated with the two cross-modal scales (visual: dark–bright, tactile: smooth–rough), as shown in Table 6. Validation was achieved in two ways: first, participants in the pre-experiment rated these stimuli consistently at the extremes of the cross-modal scales. Second, they were found in Experiment 2a to elicit the strongest total Stroop effects (see Supplementary Materials for complete details of the Experiment 2a exploratory analysis of individual stimuli). All signals were 1.5 s long, D♯4, of sample rate 44.1 kHz, and manually equalized for loudness. Additionally, for this experiment, a no-sound XXXX baseline condition was included in order to check response time and accuracy when neither semantic nor timbral stimuli were given.

Procedure

The experimental procedure was identical to Experiment 2a except for the new stimuli and baseline condition. Participants were instructed to respond as quickly and accurately as possible by pressing the arrow key assigned to the word that appeared on the screen. Key allocation was counterbalanced. Prior to beginning each block and modality participants completed brief practice sessions consisting of six randomized word/sound combinations of the experimental stimuli.

Table 6. Experiment 2b stimuli.

| Luminance | Texture |
|-----------|---------|
| Dark      | Bright  | Smooth     | Rough     |
| Natural   |         |            |           |
| Archlute  | Alto shawm | Archlute   | Bass viol |
| Tenor crumhorn | Bass viol | Oboe      | Cornetto  |
| Tenor sax (growl) | Tenor viol | Vibraphone (hard mlt.) | English horn |
| Synthesized |           | Heavy sub bass |           |
| Infinity pad | Big pulse waves | Infinity pad | Paper-thin lead |
| FM piano  | Chip tune lead | Paper-thin lead | Paper-thin lead |
| Soft square lead | Paper-thin lead | Soft square lead | Soft plucky lead |

Note. Sounds that were validated as extreme in both modalities are listed in italics. Natural instrument signals taken from the McGill University Master Samples (MUMS) library; synthesized signals derived from GarageBand software instrument library.
Natural and synthesized stimuli were presented in separate randomly ordered blocks. Modalities (luminance and texture) were likewise presented separately in a randomized order. For each congruency condition per block and modality, participants were presented with 3 unique word-sound pairs repeated once (in a random order) to serve as a measure of within-subject consistency. Thus 36 total word-sound pairs were divided into the three congruency conditions (12 pairs each): (1) congruent, (2) incongruent, and (3) control, a neutral XXXX word cue presented with no accompanying sound. In total, participants responded to 144 word-sound pairs (3 congruency conditions × 12 pairs = 36/modality; 36 × 2 modalities = 72/block; 72 × 2 blocks = 144). The experiment took approximately 20 minutes.

### Results

Conforming to the exclusion criteria of the other experiments (only accurate responses above/below a critical RT threshold) resulted in the trimming of 51 responses out of 14,781 total, and data from five participants were removed due to an error rate higher than chance (33%), which suggested random guessing or confusion with the task. Table 7 lists RT and accuracy data collapsed between the different blocks (natural vs. synthesized stimuli) and congruency conditions. Following the suggestion of Grégoire et al. (2013), since the XXXX no-sound baseline was not relevant to total Stroop effects and functioned here as a manipulation check, it was not included in the analysis. Significantly longer RTs for baseline relative to congruent and incongruent conditions was largely to be expected and likely reflect the preparation enhancement hypothesis (Nickerson, 1973), which posits that auditory stimuli, particularly presented with SOA, can act as a “warning signal” prompting participants to respond.

Log-transformed RTs were analyzed using a linear mixed model, as itemized in Table 8. Model fit was poor to moderate ($R^2 = .18$). No significant main fixed effects were found for any of the independent variables, including congruency. An interaction between block and musical training revealed that musicians were faster for both blocks, with an even greater advance with the natural stimuli compared with nonmusicians.

Test–retest reliability analysis on log-RTs between each repeated stimulus revealed moderate within-subject consistency, $r(4,940) = .24, p < .0001$, and the four experimental variables did not appear to greatly affect consistency ($r$ in range $2–3, p < .0001$). This suggests that while participants performed reliably the second time they heard each signal, the magnitude of this association was not particularly strong.

Finally, a logistic LMM was computed to test the effect of the independent variables on error rate. A significant main fixed effect of congruency was found, indicating that error rates were systematically higher for incongruent pairs compared with congruent (respectively, 6.6% vs. 5.5%, $p = .03$). Thus, although incongruent word-sound pairs did not lead to significant processing delays compared with congruent pairs, accuracy was negatively impacted by mismatched stimuli, as shown in Figure 3. Furthermore,
reinforcing the RT analysis, musical training interacted with stimuli block, with musicians outperforming nonmusicians on the natural acoustic instrument signals. The error model performed relatively weakly, however ($R^2 = .10$, $p < .0001$), indicating that most of the total variance in error was driven by factors not captured in this experiment.

**Summary of Experiment 2b**

This variant of the timbre Stroop task paired a small set of repeating stimuli with adjectives varying in degree of cross-modal congruency. Semantic congruency exhibited no discernible impact on processing speed. Analysis of error rates, however, revealed a statistically significant (though small) total Stroop effect of 1.1% difference between congruency conditions. Additionally, test–retest reliability indicated that participants were modestly consistent in their responses to repeating stimuli.

**Discussion**

The present study examined semantic crosstalk in timbre perception by way of speeded classification. The investigation confirmed the first hypothesis (H1) and partially confirmed aspects of the second (H2). First, *instrument name classification was impeded by task-irrelevant incongruent timbres and timbre classification was slowed by task-irrelevant incongruent instrument names* (H1). This bidirectional processing interference was largely to be expected, though to date has not been formally demonstrated. Second, *task-irrelevant timbral stimuli modulated processing of cross-modal adjectives, with mismatched dimensions eliciting certain mild Stroop-type impairments relative to congruent matches* (H2). As anticipated, this effect was much more muted than in H1. To be clear, moreover, results were inconsistent between Experiments 2a and 2b: the magnitude of Stroop-type interference was small in both variants, and model fit was moderate to poor. Clearly, more evidence is needed to elucidate the links between timbre perception and cross-modal descriptive conventions. Nevertheless, certain subtle effects of timbre congruency on semantic response—primarily involving response accuracy, less so speed—may indicate weak crosstalk in timbre perception. I interpret these results and discuss their implications below.

This study demonstrates for the first time that certain common cross-modal adjectives for musical timbre may reflect automatic associations between auditory and cross-modal semantic domains. The extent to which cross-modal interactions are automatic or top-down remains an open question in multisensory integration, perhaps the big question of current research (for review, see Spence & Deroy, 2013). Automaticity is tricky to characterize, but most researchers tend to agree on a number of defining attributes. According to Moors and De Houwer (2006), the most important are the goal-independence criterion and the non-conscious criterion. Goal-independence means that a process cannot be susceptible to intentional control (e.g., trying to choose congruent stimuli pairs faster than incongruent); relatedly, non-conscious implies that a process occurs independent of awareness and deliberation. The results of this study conform to both of these criteria: task-irrelevant auditory/semantic dimensions systematically (though subtly) affected participants’ performance, even when they were not attending to them. This suggests that timbre processing can bleed into semantics, both in categorical source identification and cross-modal adjectival description. (Additionally, Experiment 1 found that semantics influences timbre perception.)

To be clear, it is unlikely that interactions are *entirely* automatic; rather, both automatic and top-down, strategic processes are probably at work in semantic responses to timbre. It is plausible to theorize that certain cross-modal semantic congruencies—for example, the connection between brightness and strength in high-frequency portions of the spectrum, as revealed in the pre-experiment acoustic analysis—may be based in part in synesthetic correspondence. However, early perceptual processing may influence semantic conventions downstream, as reflected in multiple documented languages (Wallmark & Kendall, in press). High frequencies have been shown to correspond with visual brightness among preverbal infants (Walker et al., 2010): if spectral center of mass is perceptually related to pitch height and taps a similar “brightness” schema, which seems probable, then perhaps common luminance and texture terms may reflect multisensory integration, or “weak synesthesia” (Marks, 1975), that has in turn historically calcified into symbolic conventions.

An important caveat regarding this possibility must be acknowledged, however. Since this study examined written cross-modal timbre adjectives (and instrument names), not sensory experience itself—the difference between seeing a bright light and reading the word “bright”—these results do not constitute direct evidence for genuine multisensory integration. Rather, results of Experiment 2 suggest that automatic crosstalk may occur at the *semantic level only*—and very faintly, at that—although it is indeed suggestive and deserving of further study that the adjectives under examination are themselves cross-modal metaphors. These effects should therefore be interpreted in light of the *semantic coding hypothesis* (Martino & Marks, 2000), which appears to reconcile synesthetic and semantic accounts. “According to the semantic coding hypothesis,” the authors state (p. 751), “the locus of these cross-modal interactions is post-perceptual, occurring after stimulus information is coded or recoded into an abstract semantic representation that captures the synesthetic correspondence common to dimensions of both modalities.” The elaborate dance of perceptual and cognitive levels ensures that continuous timbre semantics will always be less stable and consistent than categorical identification, and more
malleable to the influence of other factors in the environment, as well as the individual background of the listener.

Dovetailing on the above, although both semantic and synesthetic correspondences are operative in timbre–semantic processing, as suggested by the semantic coding hypothesis, they are not necessarily coterminal. Differences between individual stimuli in Pre-Experiment 2 ratings (a highly semantically mediated task) and Experiment 2a—in addition to only moderately consistent responses to repeated stimuli in Experiment 2b—suggest that the same timbres and adjectives can elicit conflicting associative patterns depending on context and dominant mode of congruency, semantic or synesthetic (Spence, 2011). They also might reflect differences in stimulus-response compatibility (McClain, 1983). To illustrate: English horn was deemed extremely “bright” in the pre-experiment ratings task, but RT and error analysis in Experiment 2a suggest that “dark” offers a more congruent fit (see Supplementary Materials). That is, conscious deliberation diverged from implicit association. Moreover, certain timbre–semantic relationships are probably influenced by valence transfer (Weinreich & Gollwitzer, 2016): for example, the association between aversive growled saxophone and “dark” likely relates more to negative valence than vision, since the growl timbre is actually quite “bright” in terms of spectral distribution (see also Wallmark et al., 2018). This suggests that, at least in certain instances, synesthetic congruency is modulated by cultural context. Indeed, synesthetic and semantic correspondences interact in complex and unpredictable ways with valence, arousal, and potency in the expansive arena of human cultural differences. In future research, it is necessary to disambiguate and contextualize these processing levels with greater sensitivity to these intricacies.

How much of semantic crosstalk is accountable in the acoustic signal itself? Analysis of acoustic descriptors in Pre-Experiment 2 indicated that strength of high-frequency components explained half of all variance in visuo-tactile intensity ratings (“bright” and “rough”; see Table 4). Higher fundamental frequencies have been shown to correspond with visual brightness (Klapetek et al., 2012; Marks, 1987; Marks et al., 1987; Martino & Marks, 2000; Parise & Pavani, 2011): this result suggests something similar in regard to energy distribution within the spectrum (though, as noted, using a semantic instead of visual task), and is consonant with previous psychoacoustic findings (Beauchamp, 1982; Schubert & Wolfe, 2006). However, it is also conceivable that acoustical attributes are secondary to instrument-specific associations with the timbre lexicon that are acquired through musical training and culture more generally, and possibly somewhat arbitrary. In contrast to the automatic model previously proposed, can the timbre–semantic connections operationalized in this study be explained primarily by associative learning?

Although musical training did not produce a main fixed effect on RT or error in any of the Experiment 2 models, it was involved in a number of statistically significant interactions. For example, semantic interference was found only for the natural instrument signals in Experiment 2a, and musicians outperformed nonmusicians on these stimuli. This may seem to confirm research supporting the positive effects of musical training on the encoding of timbral properties (Chartrand & Belin, 2006; Siedenburg & McAdams, 2018). Why did synthesizers in Experiment 2a fail to elicit the same magnitude of effect as natural instruments? It is possible that this asymmetry was due to familiarity: any effects of timbre on continuous semantics may, in this account, simply represent a second-order categorical schema of learned adjectival associations. The trumpet is commonly referred to as “bright,” the horn is “dark”; synthesizer timbres, by contrast, are not generally accompanied by a parallel lexicon of roughly agreed-upon descriptive terms (Golubock & Janata, 2013; though see Wallmark, Frank, & Nghiem, in review). It follows that perhaps prevalence of timbre–semantic correlations in music discourse conditions cross-modal associations by dint of repeated exposure.

Though this would likely be true for certain instruments in certain registers, I am dubious of the strong version of this explanation, for three reasons. First, a number of the cross-modal exemplar timbres validated in Pre-Experiment 2a are diametrically opposed to what we might assume from the common discourse (see SM Table 4). For example, oboe was deemed a very “dark” timbre, and English horn the “brightest” of the original set. By way of comparison, in a representative sample of orchestration treatises, Wallmark (2019) found that the most frequent descriptors of the oboe were “nasal,” “thin,” “penetrating,” and related adjectives—“dark” did not make the top ten—and the English horn was most commonly described as “melancholic,” “rich,” and “dark.” Since D#4 is a low register (“dark”) note on the oboe, this inconsistency may be the result of tessitura. Second, most natural signals in Experiment 2 were fairly obscure historical instruments unlikely to be associated with specific learned adjectives (e.g., crumhorn, shawm, cornetto), even among trained musicians. Finally, and most crucially, the time course for semantic crosstalk observed in this study is arguably too short for consistent lexical access of this sort to take place, even if there were strong pre-established links. In short, though associative learning may account for some of this processing interference, it is unlikely to have played a decisive role.

The modest effects documented here do not represent true cross-modal correspondences, but rather the semantic correlates of cross-modality. In the future, it would be interesting to investigate the association between timbral qualities and visual/tactile perception more directly using speeded classification or a priming paradigm. Additionally, sensory scales or icon ratings may provide a useful non-linguistic index of cross-modal timbre associations.
(Murari, Rodà, Canazza, De Poli, & Da Pos, 2015; Murari, Schubert, Rodà, Da Pos, & De Poli, 2018). The automaticity hypothesis might also be fruitfully explored using physiological methods: for example, Bien, ten Oever, Goebel, and Sack (2012) used TMS, EEG, and behavioral measures together to demonstrate pitch-size automaticity at the level of 250 ms in intraparietal regions of the cortex. Convergent methods and paradigms promise to offer greater analytical precision in the investigation of the timbre–language connection.

**Conclusion**

This study offers a novel approach for the investigation of semantic automatisms in timbre perception. The findings presented here suggest that timbre perception and its categorical and continuous semantic correlates, while dissociable in certain aspects, may be loosely coupled when we reconcile sensory and semantic frames of reference. Semantic processing is mildly impaired in the presence of mismatched task-irrelevant timbres; in the case of instrument identification, moreover, impairment of auditory classification was observed when sounds were paired with mismatched task-irrelevant words. In the stimuli selection pre-experiment, participants made systematic cross-modal semantic judgments for a large set of musical timbres, and intensity of visuo-tactile response was predicted to a significant degree by strength of high-frequency energy in the spectra.

In the final analysis, why does it matter that cognitive processing is fleetingly compromised when timbre and semantics are mismatched? According to the semantic coding hypothesis (Martino & Marks, 2000), evidence of crosstalk, while certainly closely related to linguistic conventions that are culturally and historically situated, may also reflect some level of correspondence between auditory, visual, and tactile modalities. The fact that impairments in decision-making result from incongruency of sound and semantics may indicate that the two mechanisms are to some degree coupled at a fairly early processing stage. Seen from this perspective, the cross-modal terms that are so ubiquitous in talking about timbre may reflect structural resemblances between qualities of sound and both vision and touch.

A number of implications follow from this conclusion. For multisensory integration researchers, the present study demonstrates the value of adding timbre to the repertoire of audio-visual correspondences. Semantic crosstalk has long demonstrated the pervasiveness of multisensory interactions and acoustical correlates of polyphonic timbre. The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Acknowledgments**

I am grateful to Roger Kendall, W. Jay Dowling, and Stephen McAdams for feedback at various stages in this project, and for the Analysis, Creation and Teaching of Orchestration (ACTOR) partnership, which has provided intellectual nourishment for this project. I also wish to acknowledge Ian Cross, Emily Payne, and three peer reviewers for their invaluable feedback during the revision process. Benjamin Tabak for his assistance with non-music-major recruitment; and my team of SMU MuSci Lab research assistants, who helped to gather most of the data presented in this study: Jay Appaji, Grace Kuang, Caitlyn Etter, Lea Hobbie, Mary Lena Bleile, Jordan Tenpas, Jessie Henderson, and Camille Van Dorpe. Preliminary findings from this study were presented at the Timbre 2018: Timbre is a Many-Splendored Thing conference at McGill University and appear in the conference proceedings.

**Declaration of conflicting interests**

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author received no financial support for the research, authorship, and/or publication of this article.

**ORCID iD**

Zachary Wallmark is now affiliated with University of Oregon, School of Music and Dance, Eugene, OR, USA.

**Peer Review**

Kai Siedenburg, University of Oldenburg, Department of Medical Physics and Acoustics.
Zohar Eitan, Tel Aviv University, School of Music.
Antonio Rodà, University of Padova, Department of Information Engineering.

**Supplemental Material**

Supplemental materials are available online with the article.

**References**

Akiva-Kabiri, L., & Henik, A. (2012). A unique asymmetrical Stroop effect in absolute pitch possessors. *Experimental Psychology*, 59, 272–278. doi:10.1027/1618-3169/a000153
Alluri, V., & Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception*, 27, 223–241.
Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. doi:10.18637/jss.v067.i01
Beauchamp, J. (1982). Synthesis by spectral amplitude and “brightness” matching of analyzed musical instrument tones. The Journal of Audio Engineering Society, 30, 396–406.

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. Perception & Psychophysics, 57, 1151–1162.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, 57, 289–300.

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. Journal of Experimental Psychology, 87, 241–247. doi:10.1037/h0030524

Bien, N., ten Oever, S., Goebel, R., & Sack, A. T. (2012). The sound of size: Crossmodal binding in pitch-size synesthesia: A combined TMS, EEG and psychophysiology study. NeuroImage, 59, 663–672. doi:10.1016/j.neuroimage.2011.06.095

Brown, T. L., Gore, C. L., & Pearson, T. (1998). Visual half-field Stroop effects with spatial separation of words and color targets. Brain and Language, 63, 122–142. doi:10.1006/brln.1997.1940

Bruner, J. S. (1964). The course of cognitive growth. American Psychologist, 19, 1–15. doi:10.1037/h0044160

Chartrand, J.-P., & Belin, P. (2006). Superior voice timbre processing in musicians. Neuron, 405, 154–167.

Chen, Y.-C., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. Perception & Psychophysics, 68, 1191–1203.

Chen, Y.-C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. Journal of Experimental Psychology: Human Perception and Performance, 37, 1554–1568. doi:10.1037/a0024329

Chen, Y.-C., & Spence, C. (2012). The paradox of timbre. In Wallmark (Eds.), Digital multimedia perception and emerging from social and acoustic descriptions of music. Music Perception: An R companion to applied regression (2nd ed.). Thousand Oaks, CA: Sage.

Clarke, E. F. (2005). Ways of listening: An ecological approach to the perception of musical meaning. Oxford, England: Oxford University Press.

Cox, A. (2016). Music and embodied cognition: Listening, moving, feeling, and thinking. Bloomington, IN: Indiana University Press.

Dell’acqua, R., & Job, R. (1998). Is object recognition automatic? Psychonomic Bulletin & Review, 5, 496–503. doi:10.3758/BF03208828

Donohue, S. E., Appelbaum, L. G., Park, C. J., Roberts, K. C., & Woldorff, M. G. (2013). Cross-modal stimulus conflict: The behavioral effects of stimulus input timing in a visual-auditory Stroop task. PloS ONE, 8, e62802. doi:10.1371/journal.pone.0062802

Eerola, T., Ferrer, R., & Alluri, V. (2012). Timbre and affect dimensions: Evidence from affect and similarity ratings and acoustic correlates of isolated instrument sounds. Music Perception, 30, 49–70.

Eitan, Z., & Rothschild, I. (2010). How music touches: Musical parameters and listeners’ audio-tactile metaphorical mappings. Psychology of Music, 39, 449–467.

Eitan, Z., Schupak, A., Gotler, A., & Marks, L. E. (2014). Lower pitch is larger, yet falling pitches shrink. Experimental Psychology, 61, 273–284. doi:10.1027/1618-3169/a000246

Eitan, Z., & Timmers, R. (2010). Beethoven’s last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. Cognition, 114, 405–422. doi:10.1016/j.cognition.2009.10.013

Fales, C. (2002). The paradox of timbre. Ethnomusicology, 46, 56–95.

Fales, C. (2018). Hearing timbre: Implicit perceptual learning among early Bay Area ravers. In R. Fink, M. Latour, & Z. Wallmark (Eds.), The relentless pursuit of tone: Timbre in popular music (pp. 21–42). New York, NY: Oxford University Press.

Ferrr, R., & Eerola, T. (2011). Semantic structures of timbre emerging from social and acoustic descriptions of music. Journal of Audio, Speech, and Music Processing, 11, 1–16.

Filipic, S., Tillmann, B., & Bigand, E. (2010). Judging familiarity and emotion from very brief musical excerpts. Psychonomic Bulletin & Review, 17, 335–341. doi:10.3758/prb.17.3.335

Fox, J., & Weisberg, H. S. (2010). An R companion to applied regression (2nd ed.). Thousand Oaks, CA: Sage.

Gallace, A., & Spence, C. (2006). Multisensory interactions in the speeded classification of visual size. Perception & Psychophysics, 68, 1191–1203.

Garner, W. R. (1974). The processing of information and structure. Oxford, England: Lawrence Erlbaum.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. Cognitive Psychology, 1, 225–241. doi:10.1016/0010-0285(70)90016-2

Gaver, W. W. (1993). What is the world do we hear?: An ecological approach to auditory event perception. Ecological Psychology, 5, 1–29.

Giannakis, K. (2006). Perceptual-based visualization of auditory information using visual texture. In G. Ghinea & S. Y. Chen (Eds.), Digital multimedia perception and design (pp. 152–169). London, England: IGI Global.

Gibson, J. J. (1966). The senses considered as perceptual systems. Boston, MA: Houghton Mifflin.

Glaser, W. R., & Glaser, M. O. (1989). Context effects in Stroop-task. Journal of Experimental Psychology: General, 118, 13–42. doi:10.1037/0096-3445.118.1.13

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, 399–412. doi:10.1037/a0029720

Grégoire, L., Perruchet, P., & Poulin-Charronnat, B. (2013). The musical Stroop effect: Opening a new avenue to research on automatisms. Experimental Psychology, 60, 269–278. doi:10.1027/1618-3169/a000197
Grey, J. M., & Moorer, J. A. (1977). Perceptual evaluation of synthesized musical instrument tones. *Journal of the Acoustical Society of America*, 62, 454–462.

Hailstone, J. C., Omar, R., Henley, S. M. D., Frost, C., Kenward, M. G., & Warren, J. D. (2009). It’s not what you play, it’s how you play it: Timbre affects perception of emotion in music. *The Quarterly Journal of Experimental Psychology*, 62, 2141–2155.

Hajda, J., Kendall, R. A., Carterette, E. C., & Harshberger, M. L. (1997). Methodological issues in timbre research. In I. Deliège & J. Sloboda (Eds.), *Perception and cognition of music* (pp. 253–306). Hove, England: Psychology Press.

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

Jarvis, B. (2016a). *DirectRT* [Computer software]. New York, NY: Empirisoft.

Jarvis, B. (2016b). *MediaLab* [Computer software]. New York, NY: Empirisoft.

Kendall, R. A., & Carterette, E. C. (1993a). Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck’s adjectives. *Music Perception*, 10, 445–467.

Kendall, R. A., & Carterette, E. C. (1993b). Verbal attributes of simultaneous wind instrument timbres: II. Adjectives induced from Piston’s Orchestration. *Music Perception*, 10, 469–501.

Klapetek, A., Ngo, M. K., & Spence, C. (2012). Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Attention, Perception, & Psychophysics*, 74, 1154–1167. doi:10.3758/s13414-012-0317-9

Lachaud, C. M., & Renaud, O. (2011). A tutorial for analyzing human reaction times: How to filter data, manage missing values, and choose a statistical model. *Applied Psycholinguistics*, 32, 389–416.

Landry, S. P., & Champoux, F. (2017). Musicians react faster and are better multisensory integrators. *Brain and Cognition*, 111, 156–162. doi:10.1016/j.bandc.2016.12.001

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

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory–visual intensity matching. *Developmental Psychology*, 16, 597–607. doi:10.1037/0012-1649.16.6.597

Lichte, W. H. (1941). Attributes of complex tones. *Journal of Experimental Psychology*, 28, 455–480.

Lidji, P., Kolinsky, R., Lochy, A., & Morais, J. (2007). Spatial associations for musical stimuli: A piano in the head? *Journal of Experimental Psychology: Human Perception and Performance*, 33, 1189–1207. doi:10.1037/0096-1523.33.5.1189

Luce, R., & MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109, 163–203.

Marks, L. E. (1975). On colored-hearing synesthesia: Cross-model translations of sensory dimensions. *Psychological Bulletin*, 82, 303–331.

Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 384–394. doi:10.1037/0096-1523.13.3.384

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 85–105). Cambridge, MA: MIT Press.

Marks, L. E., Hammele, R. J., & Bornstein, M. H. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development*, 52, 1–102.

Martino, G., & Marks, L. E. (2000). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, 28, 903–923. doi:10.1068/p2866

Masataka, N., & Perlovsky, L. (2013). Cognitive interference can be mitigated by consonant music and facilitated by dissonant music. *Scientific Reports*, 3, 2028. doi:10.1038/srep02028

Meadow, 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, England: Clarendon.

McAdams, S., Beauchamp, J. W., & Meneguzzi, S. (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrottemporal parameters. *The Journal of the Acoustical Society of America*, 105, 882–897.

McAdams, S., Douglas, C., & Vempala, N. N. (2017). Perception and modeling of affective qualities of musical instrument sounds across pitch registers. *Frontiers in Psychology*, 8, 153. doi:10.3389/fpsyg.2017.00153

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

McClain, L. (1983). Stimulus–response compatibility affects auditory Stroop interference. *Perception & Psychophysics*, 33, 266–270. doi:10.3758/BF03202864

Melara, R. D., & Marks, L. E. (1990). Processes underlying dimensional interactions: Correspondences between linguistic and nonlinguistic dimensions. *Memory & Cognition*, 18, 477–495. doi:10.3758/BF03198481

Melara, R. D., & O’Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, 116, 323–336. doi:10.1037/0096-3445.116.4.323

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2016). pls: Partial least squares and principal component regression (Version R package version 2.6-0). Retrieved from https://CRAN.R-project.org/package=pls

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex*, 14, 452–465.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, 132, 297–326. doi:10.1037/0033-2909.132.2.297
Wallmark, Z., & Kendall, R. A. (in press). Describing sound: The cognitive linguistics of timbre. In E. I. Dolan & A. Rehding (Eds.), Handbook of timbre. New York, NY: Oxford University Press. Prepublished June 12, 2018. doi: 10.1093/oxfordh/9780190637224.013.14

Weinreich, A., & Gollwitzer, A. (2016). Automaticity and affective responses in valence transfer: Insights from the crossmodal auditory-visual paradigm. Psychology of Music, 44, 1304–1317. doi:10.1177/0305735615626519

Whelan, R. (2008). Effective analysis of reaction time data. The Psychological Record, 58, 475–482. doi:10.1007/BF03395630

Woolridge, J. M. (2001). Econometric analysis of cross section and panel data. Cambridge, MA: MIT Press.

Yuval-Greenberg, S., & Deouell, L. Y. (2009). The dog’s meow: Asymmetrical interaction in cross-modal object recognition. Experimental Brain Research, 193, 603–614. doi:10.1007/s00221-008-1664-6

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2014). An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. Music Perception, 31, 339–358.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2015). An interlanguage unification of musical timbre: Bridging semantic, perceptual, and acoustic dimensions. Music Perception, 32, 394–412.

Zakay, D., & Glicksohn, J. (1985). Stimulus congruity and S-R compatibility as determinants of interference in a Stroop-like task. Canadian Journal of Psychology/Revue Canadienne de Psychologie, 39, 414–423. doi:10.1037/h0080069

Zbikowski, L. M. (2002). Conceptualizing music: Cognitive structure, theory, and analysis. New York, NY: Oxford University Press.