Semantic crosstalk in timbre perception

Zachary Wallmark

Supplementary Materials

SM Table 1. Number of participants in each musical training category (by experiment)

Years of musical training (non/musician = NM/M)	Experiment 1 (*N* = 46)	Experiment 2a (*N* = 53)	Experiment 2b (*N* = 110)
0 (NM)	9	11	29
1 (NM)	1	4	12
2 (NM)	5	1	9
3 (NM)	1	1	12
4–5 (M)	5	2	10
6–9 (M)	7	10	18
10 + (M)	15	23	19

Note: Three participants in Experiment 1 did not provide musical training information. Lifetime years of formal instrument training (including voice) was assessed using the Goldsmiths Musical Sophistication Index (Müllensiefen, Gingras, Musil, & Stewart, 2014).

Pre-Experiment 2: Stimuli selection: Cross-modal ratings of natural and synthesizer timbres and their acoustic correlates

Participants

Twenty-nine participants were recruited from the SMU community (15 females, 14 males), 18 of whom were music majors. Ages of participants ranged from 18 to 26 (age *M* = 20.03, *SD* = 1.91), with self-reported formal musical training from 0 to 14 years (*M* = 6.51, *SD* = 3.48). All participants reported normal hearing. None of them were involved in the other experiments. Students received extra course credit for their participation.

Stimuli

The original set of stimuli consisted of 93 signals—50 natural instrument and 43 synthesizers—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 1.5s (with 200ms fade-out), D#4, and equalized for loudness. (A complete list of stimuli can be found in Supplemental Materials).

Natural instrument stimuli were selected from the McGill University Master Samples (MUMS) collection (Opolko & Wapnick, 1987). The 50 samples were chosen to represent a broad range of instrumental timbres, from common orchestral instruments to jazz and historical instruments, in addition to auxiliary playing techniques. MUMS has a long history of use in similar timbre perception studies (for review, see Eerola & Ferrer, 2008).

Synthesizer stimuli were 43 software instrument pre-sets in the Apple GarageBand (version 10.1.6) music production program. GarageBand is a ubiquitous free software application, thus offering easy reproducibility. Additionally, since GarageBand software instruments are intended for general use, the timbral palette of the synthesizer library is naturalistic and figures prominently in a range of contemporary popular music genres. All software instruments were in the default mode. Three inclusion criteria were considered: (1) a clear and unambiguous fundamental frequency, (2) no prominent temporal variability in the steady-state portion of the signal, and (3) a fairly rapid attack time. Stimuli were selected from the Bass, Bell, Brass, Classics, Lead, Pad, and Strings collections, and were recorded directly into GarageBand at a 44.1 kHz sampling rate using an M-Audio Keystation 49 controller keyboard. Loudness was equalized manually and matched with the natural stimuli.

Procedure

Participants were instructed to listen to the 1.5s tones on the computer and rate them on 7-point bipolar semantic differential scales measuring intensity of cross-modal associations (Osgood, Suci, & Tannenbaum, 1957; Zacharakis, Pastiadis, & Reiss, 2014): *dark* to *bright* (luminance) and *smooth* to *rough* (texture), with 1 corresponding on the luminance scale to "very dark" (texture: "very smooth"), 4 to a neutral condition, and 7 to "very bright" (texture: "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 for them before beginning each of the two sections (natural and synthesized = 20 signals presented); additionally, participants practiced using the horizontal rating scale on three 1.5s test signals (sine, square, and sawtooth waves, all D#4 and equalized for loudness).

The 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 synthesizer stimuli were presented separately, and the order of the two trials was randomized (see Susini, Lemaitre, & McAdams, 2012). 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 x 2 conditions (visual and tactile), and 43 synthesized stimuli x 2 conditions. The complete experiment took approximately 20 minutes.

Stimuli selection, part 1: Results

The internal consistency of the natural and synthesizer luminance/texture verbal scales was acceptable to very good (natural-luminance: *M* Cronbach's α = .75; natural-texture: *M* α = .89;

synthesizer-luminance: *M* α = .9; synthesizer-texture: *M* α = .84). To determine whether cross-modal ratings varied reliably according to scale modality, stimulus, and musical training, separate linear mixed-effects models were computed for the two blocks (natural and synthesizer), with one between-subject (two levels: musician vs. non-musician) and two within-subject fixed effects (modality: luminance vs. texture; stimuli: individual signals), and participant variability modeled as a random effect.

The natural timbre model accounted for 48% of variance, $R^2 = .48$, $p < .0001$. Using Wald chi-squared tests (Type II), statistically significant main effects were observed for all variables except musical training, and interactions were likewise significant, as listed in SM Table 2 (interactions involving stimuli were omitted from the table; see note below). Similarly, the synthesizer model (also 48% variance accounted for) revealed significant main fixed effects of modality and stimuli, as well as interactions. This tells us that musical expertise alone, as indexed by the musician vs. non-musician grouping variable, did not much affect cross-modal ratings.

SM Table 2. Results of mixed-effects models for natural instrument and synthesizer blocks

	Natural (R^2 = .48)			Synthesizer (R^2 = .48)		
	df	Wald χ^2	*p*	*df*	Wald χ^2	*p*
Musical training	1	0.82	.37	1	0.02	.88
Modality	1	86.39	**<.0001**	1	88.49	**<.0001**
Stimuli	49	1713	**<.0001**	42	1356	**<.0001**
Training*modality	1	29.2	**<.0001**	1	9.15	**.002**

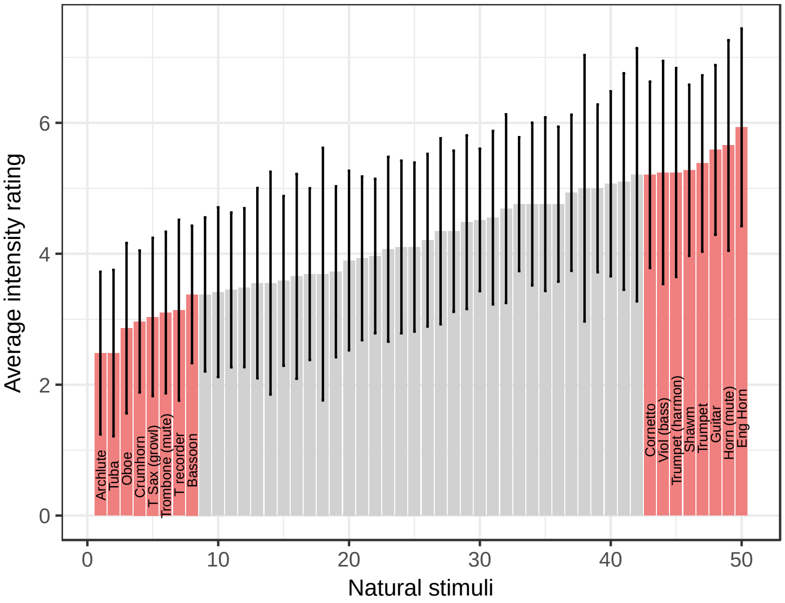
Interactions involving stimuli have been omitted. P-values < .05 indicated in **bold**.

Effects associated with modality indicate that the semantic differential scale was not perceptually equivalent between luminance and texture modalities: in general, the luminance scale tended toward significantly more intense responses (natural $M = 4.36$, $SD = 1.67$; synthesizer $M = 4.53$, $SD = 1.72$) compared to the texture (natural $M = 3.89$, $SD = 1.77$; synthesizer $M = 4.04$, $SD = 1.71$). However, the two scales were nonetheless 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).

Interactions between musical training and modality indicate the same pattern for both blocks: the difference between luminance and texture responses was narrower for non-musicians, suggesting that perceptual asymmetry between scales was more intense for musicians. Finally, due to the unwieldy number of stimuli levels (50 and 43, respectively), interactions involving this variable were difficult to interpret and somewhat meaningless for our purposes. Suffice it to say that systematic differences were observed between a number of the individual timbres in each of the blocks, indicating that certain signals were perceived as more or less "bright" and "rough" than others. In other words, ratings were not randomly distributed among the stimuli: participants tended to agree in their cross-modal responses to some timbres, particularly those at the extremes of the bipolar adjective scales.

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. The lowest and highest eight timbres for each trial and modality were selected as exemplars of the perceptual scales for the stimuli selection part 2 pilot testing; in borderline cases, SD was used as a tie-breaker, with preference given to timbres with the lower variance. For example, SM

Fig. 1 shows exemplars of "dark" (left) and "bright" (right) timbres from the 50 natural stimuli. (See "Stimuli" section of SM Experiment 2a below for more details.)



SM Figure 1: Luminance ratings for natural instrument stimuli in ascending order, with eight "darkest" (left) and "brightest" (right) timbres in red. *Error bars:* SD.

Acoustic data analysis

In addition to validation and stimuli selection, an important purpose of this pre-experiment was to determine the acoustical determinants of semantic ratings. What timbral parameters were consistently correlated with cross-modal impressions? To explore this question, common acoustic descriptors were computationally extracted using MIRtoolbox 1.6.1 (Lartillot & Toiviainen, 2007) in MATLAB (Release 2016a; The MathWorks, Inc.). Twenty-three total spectral and temporal features of the signals were originally assessed, as outlined in SM Table 3. Values consisted of an average taken over all frames of the 1.5s signals. Additionally, I extracted data pertaining to the fluctuation of energy within ten octave-scaled subdivisions of the spectrum, as first described by Alluri and Toiviainen (2010): these “sub-band flux” regions provide an index of spectrotemporal change throughout the frequency content of a signal, and have been shown to be relevant to timbre semantics (Alluri & Toiviainen, 2012, 2010; Wallmark, Frank, & Nghiem, submitted).

SM Table 3: Acoustic descriptors

Descriptor	Definition
Sub-band flux (10)	Spectrotemporal fluctuation within 10 frequency bands
Zero-cross rate	Number of signal changes per unit of time
Rolloff	Frequency threshold below which 95% of energy is contained
Brightness	Proportion of total spectral energy above 1500 Hz
Centroid	Center of spectral energy distribution
Spread	Standard deviation of spectral energy
Skewness	Asymmetry of spectrum
Flatness	Wiener entropy of signal
Kurtosis	Flatness of spectrum around mean
Entropy	Shannon entropy of signal
Irregularity	Degree of variation between successive spectral peaks over time
Roughness	Sensory dissonance averaged through time
Inharmonicity	Frequency deviation of partials from ideal harmonic series
Attack time (log)	Duration of attack phase

An initial correlation analysis was performed on raw acoustic data in order to screen for redundant parameters. Although a number of parameters were strongly correlated—for example, centroid and roughness, $\rho(91) = .8$, $p < .0001$—only a few variables exhibited close to perfect correlations, including some sub-band flux correlations between proximal bands ($\rho(91) > .9$), in addition to kurtosis and spectral skewness, ρ(91) = .96, *p* < .0001. (Spearman's ρ is reported here due to non-normality of the distribution of acoustic means.) For this reason, kurtosis was trimmed and the ten original sub-band flux variables were reduced to two: low-frequency (0–800Hz) and high-frequency (800Hz–22kHz). Following this initial screening, 14 acoustic descriptors for all 93 stimuli were tested for normality of distribution (Kolmogorov-Smirnov); variables that did not conform were transformed using an inverse-normal procedure into normally distributed Z scores (Templeton, 2011).

In order to assess latent acoustical patterns to determine which descriptors best predicted cross-modal semantic judgments, I next performed a Principal Components Regression (PCR) on visual and tactile dependent variables using the pls package in R (Mevik, Wehrens, & Liland, 2016). Data were first assessed for collinearity. Variance Inflation Factors were high (VIFs > 5) for many acoustic variables, indicating likely multicollinearity. 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. Prior to the analysis all data were scaled by subtracting the variable mean and dividing by its standard deviation. Models indicated a close resemblance between the luminance and texture scales: because of the high correlation between responses (*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). The initial PCR was thus performed to model the effect of 14 acoustic descriptors on intensity ratings. Leave-one-out cross-validation (LOO) was achieved using the Root Mean Square error prediction (RMSEP) rate to assess 10 random samples.

To select the appropriate number of components for the model, RMSEP was plotted for all 14 original PCs. Three components (Eigenvalues > 1) were found to best minimize model error, together explaining a total of 76% of variance in visuo-tactile intensity ratings. Table 4 in the main article displays factor loadings for the final cross-validated model, along with estimates of regression coefficients and significance levels for each acoustic descriptor.

PC1 explains half of visuo-tactile intensity in the model, and roughly conforms to the first component found in the acoustic PCA. It is associated with a range of features indexing high-frequency energy, including brightness, centroid, and rolloff, while showing a moderate negative association with skewness. This result replicates many previous psychoacoustic findings linking perceived timbral "brightness" to strength in high-frequency components of the spectrum (Beauchamp, 1982; Parise & Spence, 2012; Wessel, 1979). Corroborating the perceptual importance of high-frequencies, PC2 (13%) is associated with decreasing fluctuation in the low frequency range (including fundamental frequency, difference tones below the fundamental, and up to the second harmonic), as well as decreasing irregularity and inharmonicity of the spectrum. PC3 (12%) is related to increasing spread, flatness, and length of attack, and inversely related to high-frequency spectral flux and roughness.

Summary of Pre-Experiment 2

Pre-Experiment 2 investigated associations between a large set of natural instrument and synthesizer signals and common cross-modal adjectives using semantic differential ratings. Although participants scored the timbres as significantly more "bright" than "rough," luminance and texture scales were strongly correlated, indicating a structural resemblance between these modalities based on intensity or magnitude (the valence dimension of the respective scales were ordered inversely). Musical training alone did not appear to affect ratings; however, greater polarity between modalities was found for the musicians. Additionally, results indicate that some signals were consistently rated at the extremes of the bipolar adjective scales. These particular stimuli were taken to be exemplars of timbral "darkness," "brightness," etc., and selected for use in the Experiment 2 pilot study.

Finally, acoustic descriptors were extracted from the 93 signals and subjected to a PCR to predict intensity of visuo-tactile ratings. The model converged on three components explaining about three-quarters of the variance in intensity. Results confirm the long-established connection between spectral centroid and timbral "brightness."

Experiment 2a: Preliminary Stroop test of cross-modal timbre semantics

SM Table 4: Experiment 2a stimuli

Natural			
Luminance		Texture	
"Dark"	"Bright"	"Smooth"	"Rough"
archlute	alto shawm	*archlute*	alto sax (growl)
tenor baroque recorder	*bass viol*	celesta	*bass viol*
bassoon	*cornetto*	horn	*cornetto*
tenor crumhorn	*English horn*	oboe d'amore	*English horn*
oboe	*classical guitar*	*oboe*	*classical guitar*
tenor sax (growl)	horn (mute)	recorder (renaissance)	*tenor viol*
trombone (mute)	*tenor viol*	*tuba*	*trumpet (harmon mute)*
tuba	*trumpet (harmon mute)*	vibraphone (hard mallet)	viola (pizz.)

Synthesizer

antarctic sun	*big pulse waves*	*deep sub bass*	*big pulse waves*
deep sub bass	*bright synth brass*	evolving currents	*bright synth brass*
FM piano	*chip tune lead*	*FM piano*	bright synth strings
heavy sub bass	*icy synth lead*	*heavy sub bass*	*chip tune lead*
infinity pad	*monster bass*	*infinity pad*	*icy synth lead*
soft square lead	*paper-thin lead*	*soft square lead*	*monster bass*
starlight vox	percussive square lead	*starlight vox*	*paper-thin lead*
synth e-bass	soft saw lead	*synth e-bass*	short plucky lead

Timbres that were validated as extreme in both modalities are listed in *italics*. Natural timbres taken from MUMS library; synthesizer timbres derived from GarageBand software instrument library.

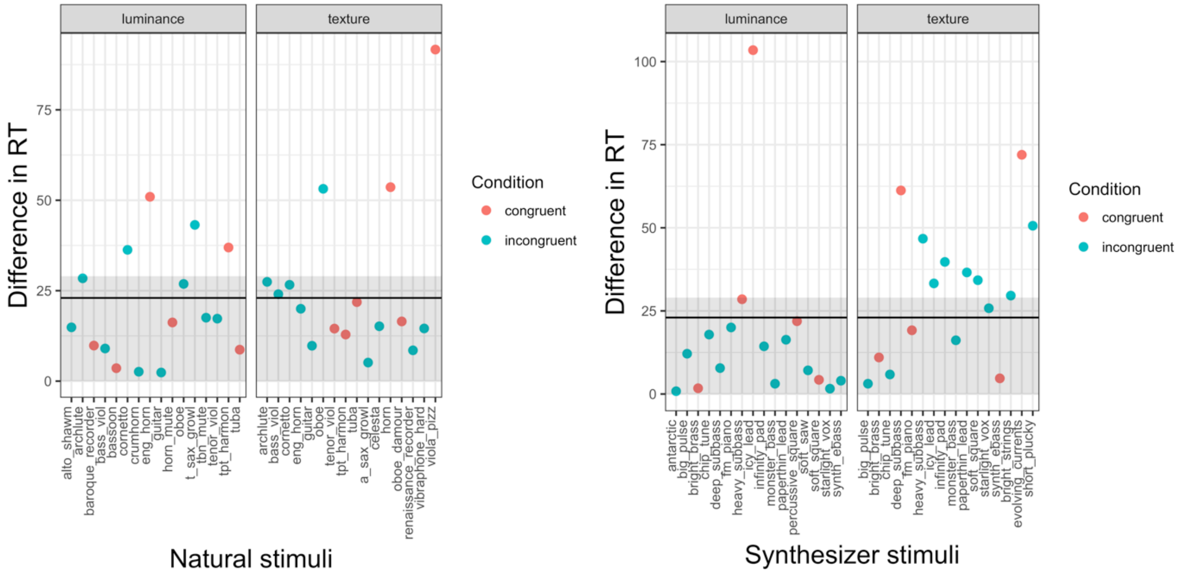
Exploratory analysis of individual stimuli in Exp. 2a

Congruent and incongruent stimuli groupings in Experiment 2a were selected on the basis of results from Pre-Experiment 2. This RT analysis, which collapsed individual stimuli by congruency condition, thus relied upon the *a priori* assumption that results obtained through scale ratings would be generalizable to this Stroop-style task. Scale rating procedures are more deliberative and cognitively mediated than RT tasks: it is possible, for example, that not all timbres considered “bright” when asked to consciously reflect on the matter would necessarily activate a lexical “brightness” schema in speeded response. We might therefore ask: Did any specific stimuli perform reliably better than others?

As a second stage of the Experiment 2a analysis, then, I sought post hoc to evaluate differences in total Stroop effect as a function of individual stimuli. Total Stroop effects (TSE) refer to the absolute difference between congruent and incongruent conditions (Brown, Gore, &

Pearson, 1998). To do so, differences in RT and error rates between the incongruent and congruent conditions for all stimuli were calculated along with the direction of the difference (i.e., towards congruent pairs or incongruent), then averaged across participants. Of the 64 stimuli, 42 demonstrated TSE in the incongruent direction, $\chi^2(1) = 10.56$, $p = .001$: that is, when viewed dichotomously, RTs were longer when paired with incongruent compared to congruent timbres.

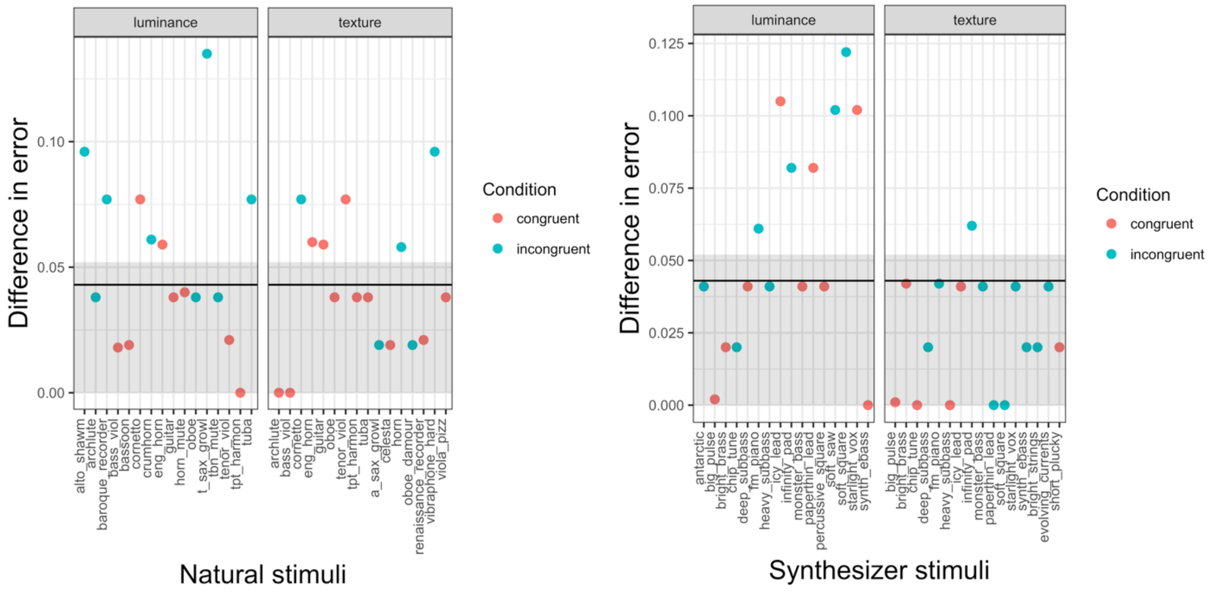
SM Figure 2 plots TSE for each stimuli. Direction of the TSE RT difference in SM Fig. 2 is mapped onto color: stimuli that produced TSEs in the *congruent* direction (i.e., that slowed RT when word/timbre pairs were congruent) are shown in red; stimuli that produced TSEs in the expected, *incongruent* direction are blue. The mean TSE (23 ms) is indicated with the horizontal reference line, and the shaded area represents the 97.5% confidence interval of this mean [CI: 0, 29]: stimuli that fall above this upper bound can thus be considered extreme compared to the average TSE; thus, blue points above the gray area indicate stimuli that produced strong interference on RT in that condition. As indicated, among natural stimuli a couple of timbres—most notably the viola pizzicato—produced the opposite effect from what we might infer from the pre-experiment: that is, the *congruent* condition slowed viola pizzicato RTs (by 92 ms) relative to the incongruent pairings. Similarly, the most extreme value among synthesized timbres fell in the congruent direction (icy lead synth, 103 ms). These results might suggest miscategorization of these particular stimuli, at least in respect to the difference between semantic ratings and speeded word response. In the case of the pizzicato, it might also represent an acoustical outlier (this attack profile differed quite a bit from most of the others in its rapid rise time).



SM Figure 2: Total Stroop effects (TSE) for RT between congruency conditions. Gray area denotes 97.5% CI; horizontal reference line is the mean TSE. Red points show stimuli that produced a TSE in the *congruent* direction (i.e., RT was longer when timbre-word pair was congruent); blue points show stimuli that produced a TSE in the hypothesized, *incongruent* direction (i.e., RT was longer when timbre-word pair was incongruent). The y-axis denotes absolute difference in RT (ms).

The best performers in TSE among natural instruments in the incongruent direction were the cornetto and tenor sax growl in the luminance modality (36 and 43 ms, respectively), and the oboe for the texture task (53 ms). Among synthesized luminance timbres, all but one fell within the 97.5% CI; texture terms, however, revealed seven timbres associated with extreme TSEs. This cluster of stimuli may indicate that incongruent pairings of synthesized timbres and texture terms (e.g., “rough” synth with the word SMOOTH) were uniquely susceptible to cross-modal processing interference, although when pooled with the other stimuli in this condition the total effect was nonsignificant.

Differences between stimuli error rates were also calculated as above. There was no difference between the number of congruent (34) versus incongruent (30) TSEs, $\chi^2(1) = 0.25$, $p = .62$. The mean TSE on error was 4% with 97.5% CI [0, 5.2%]. In the natural trial, a handful of stimuli produced substantially higher error rates in the incongruent condition; for example, tenor sax growl (14%), shawm (10%), and vibraphone (10%).



SM Figure 3: Total Stroop effects (TSE) for error rate between congruency conditions. Gray area denotes 97.5% CI; horizontal reference line is the mean TSE. Red points show stimuli that produced a TSE on error in the *congruent* direction (i.e., error was higher when timbre-word pair was congruent); blue points show stimuli that produced a TSE on error in the hypothesized, *incongruent* direction (i.e., error was higher when timbre-word pair was incongruent). The y-axis denotes absolute difference in error rate (percentage).

Putting RT and error TSEs together, we can see that tenor sax growl (luminance) and infinity pad (texture) exhibited both increased RT *and* error rate in incongruent pairings; conversely, English horn (luminance) and icy lead synth (luminance) gave more extreme RT and errors in the congruent condition. These consistencies indicate that, for example, the saxophone growl slowed RT by an average of 43 ms and drove up error a full 14% higher in response to the word BRIGHT than when paired with DARK. Moreover, while English horn was considered a "bright" timbre in Experiment 2, it interfered with speeded identification of the word BRIGHT and led to higher error rates. This would seem to indicate that, contrary to expectations, English horn timbre may have jarred with the word cue BRIGHT, where DARK provided a more consistently seamless association. Although the stimuli were too variant and inconsistent to product robust Stroop-style interference when collapsed into the *a priori* conditions—recall that congruent/incongruent pairings were just shy of significance in fixed main effects, though natural instrument RTs took significantly longer in the incongruent condition—a handful of individual timbres exhibited cross-modal interactions generally consistent with the hypothesis that incongruent word-timbre pairings would result in cognitive interference.

In sum, exploratory analyses of individual stimuli TSEs revealed that, despite significant interference during the incongruent condition in the natural instrument trial, not all pre-validated stimuli performed as expected in this task: more RT TSEs were found in the incongruent direction, as hypothesized, and some stimuli produced both extreme RT and error TSEs in the incongruent position. Other stimuli, however, exhibited the opposite effects.

The three signals per modality and stimuli block and produced the most pronounced incongruent RT and error TSEs in this Experiment 2a were selected for inclusion as stimuli in Experiment 2b, as reported in the main article.

Limitations

The control condition in Experiments 1 and 2a may have been ineffective in establishing a baseline for semantic crosstalk. In their study of the "musical Stroop effect," Grégoire, Perruchet, and Poulin-Charronnat (2013) reported that inclusion of a control condition in a speeded note-naming task among musicians reduced statistical power while offering no appreciable benefits in situations where the main research question hinged on the absolute differences between congruent versus incongruent conditions. Would eliminating the control change the results? Rerunning the RT analysis of Experiment 2a with the control removed led to a model with comparable predictive power to the original analysis ($R^2 = .21$, $p < .0001$) and a significant difference between congruent and incongruent conditions, Wald $\chi^2(1) = 4.8$, $p = .03$. This counterfactual result suggests that perhaps in future studies using a similar paradigm, only the two main experimental conditions should be included.

Auditory signals were presented with a 200ms stimulus onset asynchrony (SOA) to attempt to control for the differential in processing speed between modalities (Chen & Spence, 2011; Donohue et al., 2013). This may have facilitated the coupling of the two stimuli such that early presentation of the control timbre effectively divulged the identity of the XXXX word cue once the association was learned. If this were the case, we should theoretically expect to observe an ordering effect over the course of each trial. Conceivably this would be especially true for the expert musician sample: to be sure, a very weak though statistically significant negative correlation was found in this group between (randomized) presentation order and RT in the control condition, possibly indicating an extremely faint learning effect, $r(5758) = –.03$, $p = .01$. In contrast, non-music-majors exhibited no such correlation, $r(3934) = –.007$, $p = .66$. Given this negligible effect, it is doubtful that the significant differences in RT between control and congruent/incongruent conditions could be accounted for by associative learning alone. In further studies, it would be instructive to systematically vary SOA times to observe the effect on Stroop interference (Donohue et al., 2013).

Finally, additional design limitations must be noted in conclusion:

- Loudness levels were subjectively determined by participants, meaning that loudness was not matched between participants
- Loudness of stimuli were equalized manually
- No counterbalancing of pre-experiment scales was carried out
- No counterbalancing of key allocation in Experiments 1 and 2a
- The XXXX-type control may not be semantically neutral if sound symbolism (*a la* bouba/kiki effect) is taken into account

References

Alluri, V., & Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception*, *27*(3), 223–241.

Alluri, V., & Toiviainen, P. (2012). Effect of enculturation on the semantic and acoustic correlates of polyphonic timbre. *Music Perception*, *29*(3), 297–310.

Beauchamp, J. (1982). Synthesis by spectral amplitude and “brightness” matching of analyzed musical instrument tones. *The Journal of Audio Engineering Society*, *30*, 396–406.

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*(1), 122–142. https://doi.org/10.1006/brln.1997.1940

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*(5), 1554–1568. https://doi.org/10.1037/a0024329

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*(4), e62802. https://doi.org/10.1371/journal.pone.0062802

Eerola, T., & Ferrer, R. (2008). Instrument library (MUMS) revised. *Music Perception*, *25*(3), 253–255.

Grégoire, L., Perruchet, P., & Poulin-Charronnat, B. (2013). The musical Stroop effect: Opening a new avenue to research on automatisms. *Experimental Psychology*, *60*(4), 269–278. https://doi.org/10.1027/1618-3169/a000197

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

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

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLOS ONE*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Opolko, F., & Wapnick, J. (1987). *McGill University master samples*. Montreal, Quebec, Canada: McGill University.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: A study using the implicit association test. *Experimental Brain Research, 220*(3), 319–333. https://doi.org/10.1007/s00221-012-3140-6

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology, 24*(1), 99–142. https://doi.org/10.1016/0010-0285(92)90004-L

Susini, P., Lemaitre, G., & McAdams, S. (2012). Psychological measurement for sound description and evaluation. In B. Berlund, G. B. Rossi, J. T. Townsend, & L. R. Pendrill (Eds.), *Measurement with persons: Theory, methods, and implementation areas* (pp. 227–253). New York and London: Psychology Press.

Templeton, G. F. (2011). A two-step approach for transforming continuous variables to normal: Implications and recommendations for IS research. *Communications of the Association for Information Systems, 28,* 41–58.

Wallmark, Z., Frank, R. J., & Nghiem, L. (submitted). Creating novel tones from adjectives: An exploratory study using FM synthesis.

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer Music Journal, 3*(2), 45–52.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2014). An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception, 31*(4), 339–358.