JASA ARTICLE

Brightness perception for musical instrument sounds: Relation to timbre dissimilarity and source-cause categories

Charalampos Saitis[1,a)] and Kai Siedenburg[2]
[1]Audio Communication Group, TU Berlin, Einsteinufer 17c, D-10587 Berlin, Germany
[2]Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Oldenburg 26129, Germany

ABSTRACT:
Timbre dissimilarity of orchestral sounds is well-known to be multidimensional, with attack time and spectral centroid representing its two most robust acoustical correlates. The centroid dimension is traditionally considered as reflecting timbral brightness. However, the question of whether multiple continuous acoustical and/or categorical cues influence brightness perception has not been addressed comprehensively. A triangulation approach was used to examine the dimensionality of timbral brightness, its robustness across different psychoacoustical contexts, and relation to perception of the sounds' source-cause. Listeners compared 14 acoustic instrument sounds in three distinct tasks that collected general dissimilarity, brightness dissimilarity, and direct multi-stimulus brightness ratings. Results confirmed that brightness is a robust unitary auditory dimension, with direct ratings recovering the centroid dimension of general dissimilarity. When a two-dimensional space of brightness dissimilarity was considered, its second dimension correlated with the attack-time dimension of general dissimilarity, which was interpreted as reflecting a potential infiltration of the latter into brightness dissimilarity. Dissimilarity data were further modeled using partial least-squares regression with audio descriptors as predictors. Adding predictors derived from instrument family and the type of resonator and excitation did not improve the model fit, indicating that brightness perception is underpinned primarily by acoustical rather than source-cause cues. © *2020 Acoustical Society of America*.
https://doi.org/10.1121/10.0002275
(Received 21 January 2020; revised 8 August 2020; accepted 30 September 2020; published online 21 October 2020)
[Editor: Jonas Braasch] Pages: 2256–2266

I. INTRODUCTION

The auditory attribute of brightness is among the most studied aspects of timbre perception, and arguably among the most important perceptual attributes actively shaped by music performers, composers, and audio engineers. It systematically emerges as a major dimension across different types of sounds and analytical approaches towards the study of timbre dissimilarity (McAdams, 2019) and timbre semantics (Saitis and Weinzierl, 2019). The word "bright" was shown to be in the top five most frequently mentioned attributes of instrumental timbre across 11 orchestration texts (Wallmark, 2019) and in the top three most commonly used descriptions of sound effects processing among audio production professionals (Pearce *et al.*, 2017). In singing voice pedagogy, the concept of *chiaroscuro*, or bright-dark tone, is central to the *bel canto* style, describing the ideal singing voice as having "a bright edge as well as a dark round quality in a complex texture of vocal resonances" (Stark, 2003, p. 33). Timbral brightness has also been shown to be an important factor in assessing concert hall acoustics (Lokki *et al.*, 2011; Weinzierl *et al.*, 2018). Despite the major role of brightness in music creation and perception, research has not yet delineated its detailed perceptual and cognitive structure. Here, a triangulation approach was used to comprehensively examine the dimensionality of brightness as an attribute of timbre, how it behaves across different psychoacoustical contexts, and whether it is influenced by the ability of the listener to identify the sounds' source-cause.

Musical timbre has most often been studied via "timbre spaces." These are geometrical configurations resulting from multidimensional scaling (MDS) of pairwise dissimilarity ratings among a set of sounds (for more detail and a recent review, see McAdams, 2019). Using recordings of musical instrument notes or synthetic sounds, previous MDS studies have repeatedly identified at least two robust perceptual dimensions of timbre (Caclin *et al.*, 2005; Grey, 1977; Krimphoff *et al.*, 1994; Lakatos, 2000; McAdams *et al.*, 1995). These dimensions correlate well with the attack time and with the spectral centroid (SC) of the sounds, respectively. The attack time is defined as the (logarithm) of the duration between the onset of a sound and its more stable part. The SC is defined as the amplitude-weighted mean frequency and can be interpreted as the center of gravity of the spectral envelope or the frequency that divides the spectrum into two regions with equal energy (Caetano *et al.*, 2019). The SC has also been shown to correlate with direct brightness ratings of musical instrument

[a)]Present address: Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom. Electronic mail: c.saitis@qmul.ac.uk, ORCID: 0000-0002-6860-9723.

https://doi.org/10.1121/10.0002275

tones (Almeida *et al.*, 2017; Schubert and Wolfe, 2006; Zacharakis *et al.*, 2014). In timbre spaces, the dimension most strongly correlated with the SC is then considered as reflecting timbral brightness. However, we are not aware of any study on whether the brightness dimension of timbre spaces can be recovered from direct brightness ratings.

It is further to be noted that spectral envelopes of sounds can vary in manifold ways, certainly more than can be comprehensively described by the one dimension of the SC. For instance, using synthetic tones with formant-like characteristics, Siedenburg (2018) demonstrated consistent shifts of perceived brightness between tones with highly similar SC values. In timbre semantics, sounds that are described as thick, dense, or rich are also described as less bright or brilliant, indicating an interplay between spectral energy distribution and spectral detail or richness (Saitis and Weinzierl, 2019). However incomplete the SC may be, it may still act as an effective summary descriptor for quantifying brightness perception (cf. Siedenburg *et al.*, 2016a). This perspective also motivates a question on the nature (or dimensionality) of brightness as an auditory attribute: could brightness be a lump sum of multiple (spectrally-based) attributes that are collectively associated with brightness but separate if considered in greater detail? In this study, we sought to address this question by considering brightness perception with the same methods as general timbre dissimilarity.

When it comes to listeners' strategies for sorting a set of sounds, Lemaitre *et al.* (2010) proposed to distinguish between *acoustical similarity* (similarity of acoustical properties), *causal similarity* (of the identified physical source-cause of the sound), and *semantic similarity* (of some knowledge or meaning associated to the sound or its source-cause). For instance, listeners may group a guitar and a violin pizzicato sounds together because they have similar temporal envelopes (acoustical similarity); because they both were made by plucking a vibrating string coupled with a wooden resonator (causal similarity); or, related to the temporal envelope and plucked string cues, because they both sound "abrupt" (semantic similarity). Causal similarity corresponds to what we here refer to as similarity in terms of source-cause cues. Timbre studies using dissimilarity ratings rely on the implicit assumption that as a task of qualitative comparison they are underpinned by acoustical rather than causal or semantic similarity. This justifies positioning sounds in a continuous space by assuming that dimensions such as brightness are continuously varying perceptual attributes. On the contrary, it has recently been suggested (Siedenburg *et al.*, 2016b) that dissimilarity ratings can be infiltrated by information from the sounds' source-cause that is partially independent of acoustical similarity. In judging the dissimilarity between, say, a marimba and a vibraphone, the fact that both are familiar percussion instruments and are excited in identical ways may shrink dissimilarity ratings.

Comparing two sounds on timbral brightness could be open to a similar bias. For instance, in a go/no-go categorization task of short (12.5–200 ms) sound excerpts comprising speech, musical instruments, and human environmental sounds (Ogg *et al.*, 2017), as the median SC value increased, listeners were more likely to categorize the stimuli as human environmental sounds and less likely to consider the sounds as coming from musical instruments. Furthermore, geometric spaces derived from dissimilarity ratings and ratings along verbal scales have been known to share many configurational and dimensional similarities (Faure *et al.*, 1996; Samoylenko *et al.*, 1996; Zacharakis *et al.*, 2015). Given these similarities between dissimilarity-based and verbally-based approaches to timbre, it does not seem far fetched to hypothesize that brightness ratings could show a similar influx of source-cause categories compared to general timbre dissimilarity.

In this study, we examined brightness perception for musical instrument sounds by posing three important, yet unexplored questions motivated above: the first question concerned the dimensionality of brightness as an attribute of timbre. Specifically, we wondered about the dimensionality that timbral brightness would exhibit as an auditory attribute in and of itself if considered through the empirical angle of pairwise dissimilarity ratings of a set of sounds. The second related question concerned the robustness (or stability) of brightness judgments across different tasks. Specifically, we wondered about the extent to which direct brightness ratings of a set of sounds would recover their ordering along the SC dimension obtained from general timbre dissimilarity ratings of the same sounds. The third question concerned the relation of brightness to source-cause categories. Specifically, we wondered whether brightness dissimilarity ratings of instrumental sounds would be affected by categorical stimulus features related to instrument family membership and the type of resonator and excitation.

These questions were approached by using three different experimental tasks that collected general timbre dissimilarity ratings, brightness dissimilarity ratings, and direct multi stimulus brightness ratings of the same set of musical instrument sounds. We carried out hierarchical clustering and MDS analyses of dissimilarity ratings and quantified the dimensional similarity between the general timbre space, timbral brightness space, and direct brightness ratings. We then conducted an exploratory regression analysis that enabled us to compare the contributions of source-cause categorical descriptors to general timbre and brightness dissimilarity ratings.

II. METHOD

A. Participants

Forty listeners with substantial experience in music and audio were recruited from the MSc program in Audio Communication and Technology at the Technical University of Berlin and the Tonmeister programme at the Berlin University of the Arts [average age = 29.5 years; standard deviation (SD) = 5.6 years; range = 23–49 years]. They were German native speakers or spoke German

fluently, and reported no hearing impairments. Participants received course credit whenever possible, and otherwise a monetary compensation of 10 EUR. All participants gave written informed consent in accordance with the Declaration of Helsinki (World Medical Association, 2013).

B. Stimuli and apparatus

Stimuli consisted of the same 14 recordings of single tones from Western orchestral instruments used by Siedenburg *et al.* (2016b): bass clarinet (BCL), bassoon (BSN), flute (FLT), harpsichord (HCD), horn (HRN), harp (HRP), marimba (MBA), piano (PNO), trumpet (TRP), bowed cello (VCE), cello pizzicato (VCP), vibraphone (VIB), bowed violin (VLI), and violin pizzicato (VLP), all played at mezzo-forte without vibrato. Piano and harpsichord samples were taken from Logic Professional 7; all other samples came from the Vienna Symphonic Library,[1] and only left channels were used. All sounds had a fundamental frequency of 311 Hz (E♭4) and a duration of 500 ms. Because the actual durations of the sound samples varied and were slightly longer than 500 ms, a raised cosine ramp from 480 to 500 ms was used as a fade-out to maintain the same duration for all stimuli.

Six expert listeners had previously (Siedenburg *et al.*, 2016b) equalized the perceived loudness of the 14 stimuli against a reference sound (MBA), using a protocol designed in PsiExp,[2] last accessed July 22, 2020) for the music-programming environment Pure Data.[3] Stimuli were presented through a Grace m904 amplifier, and listeners used a slider on the computer screen to adjust the loudness of the test sound until it matched that of the reference sound. Loudness was then normalized across all sounds on the basis of the median loudness adjustments.

Listeners were tested individually in a quiet room. Stimuli were presented on Sennheiser HD 800 S headphones using a Windows PC with digital-to-analog conversion on a Focusrite Scarlett 18i20 audio interface at an audio sampling rate of 44.1 kHz. Responses to the different tasks (see below) were collected by means of a graphical user interface programmed in the MATLAB software environment. The average presentation level was fixed at a comfortable level by the experimenter, which amounted to 86.1 dB sound pressure level (SPL) (SD = 2.1; range = 82.5–89) as measured with a Norsonic type 110 sound-level meter (A-weighting) with a Brüel and Kjær type 4152 artificial ear to which the headphones were coupled.

C. Design and procedure

Each participant attended a single experimental session, which included three tasks and lasted around one hour. All participants first listened to all sounds in pseudorandom order to familiarize themselves with the different sounds in the set. In each task, participants could listen to each stimulus or pair of stimuli as many times as desired but were encouraged to move at a reasonable pace. At the end of the third task, participants provided demographic and musical training information.

1. General and brightness dissimilarity ratings

The first part of the experiment comprised two dissimilarity rating tasks. In each trial, two stimuli were presented successively with an interstimulus interval of 300 ms and participants were asked to rate how dissimilar the two sounds were based on general dissimilarity (hereafter referred to as the GEdissim task) and based on brightness dissimilarity (hereafter referred to as the BRdissim task). In the GEdissim task, participants were asked to provide ratings simply in terms of how dissimilar they perceived the two sounds to be without specifying further what that entailed. Four example trials were given in the beginning of the task for training purposes. In the BRdissim task, listeners were instructed to judge the dissimilarity of the two sounds only with respect to their brightness. Given the goals of the study, no explanation was offered as to what brightness might refer to acoustically. Instead, participants were given two example trials pairing the bowed cello with low-pass and high-pass filtered versions of itself, in addition to the same four example trials as in the GEdissim task.

The order of presentation of the two dissimilarity tasks was counterbalanced across participants. Dissimilarity ratings were provided through a continuous scale with marks between "identical" and "very dissimilar" at the extremes. Each stimulus pair was presented once in one order (AB or BA for sounds A and B) and the order of presentation was counterbalanced across individuals. Pairs of identical stimuli were included, yielding 105 trials in total per block. We did not present the full 14 × 14 matrix of pairwise comparisons including both orders of pairs (AB and BA for sounds A and B) as dissimilarity ratings of the same set of instrumental sounds have been previously shown to be reliably symmetric (Siedenburg *et al.*, 2016b).

2. Direct multi stimulus brightness ratings

The second part of the experiment involved direct brightness ratings of the same 14 sounds in two steps (hereafter referred to as the BRdirect task). The design of this part took inspiration from the standardized multi stimulus test with hidden reference and anchor (MUSHRA) procedure developed for the perceptual evaluation of audio codecs, whereby listeners are allowed to switch between multiple stimuli presented in parallel as often as they want (ITU-R BS.1534-3; ITU, 2015). In MUSHRA, listeners do not only perform a direct rating of each stimulus, but also a ranking and inherently also pairwise comparisons.

Each step consisted of a graphical interface with nine sliders corresponding to nine sounds. Participants listened to each sound by pressing a button at the bottom of each slider. They rated each sound with the different sliders on a continuous scale with marks between "very bright" and "not bright at all" at the extremes. These nine stimuli comprised half of the tested sounds plus two "anchors," that is, hidden

https://doi.org/10.1121/10.0002275

repetitions of two of the tested stimuli (cf. Lemaitre *et al.*, 2015). The two anchors were expected to stabilize the brightness scaling of all 14 sounds across the two steps. This approach of splitting the task across two steps was conceived to be better manageable for participants compared to having to rate all 14 stimuli in parallel, which is usually the case in MUSHRA tests. The order of presentation of the stimuli within and across trials was counterbalanced across individuals. The interface was locked until a participant had listened to every sound at least once and positioned at least one slider to a value other than the minimal possible value.

D. Audio content descriptors of timbre

For acoustical modeling of the dissimilarity data, thirty-four audio descriptors of timbre (Table I) were extracted from the temporal and spectral envelopes of the acoustic signals using the Timbre Toolbox (Peeters *et al.*, 2011). Temporal descriptors model global features such as attack time (see Sec. I) and energy modulation (Elliott *et al.*, 2013), and time-varying energy. The latter is computed for each 25 ms time frame, as are spectral descriptors derived from an ERB-spaced gammatone filter bank decomposition of the signal (Equivalent Rectangular Bandwidth, Glasberg and Moore, 1990; Patterson *et al.*, 1992). These include, among others, the first four statistical moments of the spectrum, such as the SC, and estimates of local spectral change over time, such as the spectral variation or flux. Time-varying descriptors were summarized through the robust statistics of median and interquartile range as measures of central tendency and variability, respectively.

III. RESULTS

Prior to the main body of analysis, inter-listener agreement was assessed by calculating inter-rater correlations (IRC) for each of the GEdissim, BRdissim, and BRdirect tasks. Figure 1 shows the corresponding IRC distributions. The brightness dissimilarity ratings exhibited the lowest ICRs with a mean of around 0.63, while those of general dissimilarity ratings had a mean of around 0.72 and were clearly below the IRCs of direct brightness ratings with a mean of almost 0.8, indicating that the latter exhibited most agreement across participants.

Moreover, brightness ratings were extremely consistent across the two BRdirect steps, as indicated by a high correlation between the profile of group averages of ratings across the first and second stimuli subsets [$r(13) = 0.99, p < 0.0001$]. This confirmed the validity of collecting MUSHRA-like brightness ratings for one half of the 14 sounds at a time versus all in parallel.

To assess an effect of task ordering, we compared separately general and brightness dissimilarity matrices between listeners who first did the GEdissim task and then the BRdissim one (half of the participants) and those who did the two tasks in the reverse order. Within tasks, the corresponding dissimilarity matrices correlated almost perfectly (both $r = 0.99, p < 0.0001$), suggesting that the two tasks of GEdissim and BRdissim were perceptually separated by the listeners.

TABLE I. List of extracted audio content descriptors from the Timbre Toolbox (Peeters *et al.*, 2011). Temporal descriptors are computed from the signal energy (temporal) envelope and spectral descriptors from the ERB gammatone filterbank representation. For spectral descriptors and the root-mean-square (rms) envelope, medians (med), and interquartile range (IQR) are computed over time frames of 25 ms.

Spectral	Temporal
Centroid (med, IQR)	Attack time
Spread (med, IQR)	Decay time
Skewness (med, IQR)	Release
Kurtosis (med, IQR)	LAT
Slope (med, IQR)	Attack slope
Decrease (med, IQR)	Decrease slope
Rolloff (med, IQR)	Centroid
Variation (med, IQR)	Effective duration
Frame energy (med, IQR)	Frequency of energy modulation
Flatness (med, IQR)	Amplitude of energy modulation
Crest (med, IQR)	rms envelope (med, IQR)

A. Dissimilarity clusters

In order to visualize the basic grouping structure of the dissimilarity data, agglomerative hierarchical cluster analyses were computed on averaged dissimilarity data, using the complete-linkage method. The latter is based on a function that iteratively computes the distance of the two elements (one in each cluster) that are the farthest away from each other. Figure 2 depicts the resulting clusters for the GEdissim and BRdissim ratings. The threshold for overall grouping (indicated by color-coded clusters) was 70% of the maximum linkage (the default value of the used

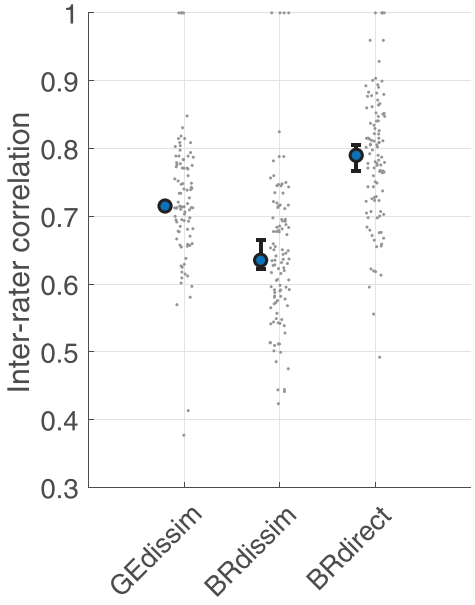

FIG. 1. (Color online) IRC for the three tasks. Errorbars correspond to 95% confidence intervals obtained via bootstrapping, grey dots to individual IRCs.

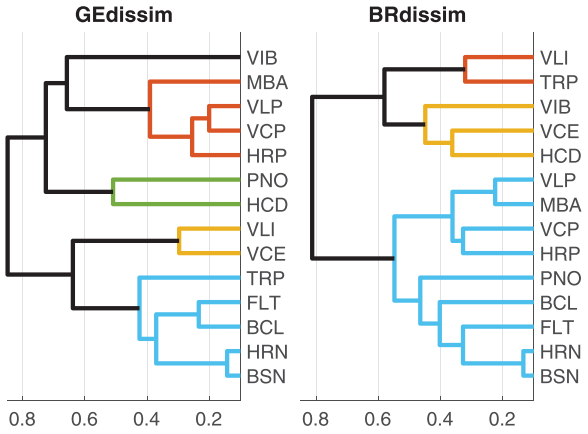

FIG. 2. (Color online) Hierarchical complete-linkage clustering of general (left) and brightness (right) dissimilarity ratings.

dendrogram function provided by MATLAB). Cophenetic correlation coefficients (the linear correlations between the cophenetic tree distances and the original dissimilarities) were 0.80 for general dissimilarity clusters and 0.43 for brightness dissimilarity clusters.

GEdissim data yielded five clusters, including the vibraphone singleton. As expected (Siedenburg *et al.*, 2016b), these corresponded to familiar musical instrument families, suggesting a partial influence of source-cause categories on general timbre dissimilarity. Wind instruments clustered together (light blue), as did bowed strings (orange). With respect to impulsively excited instruments, keyboard type strings (green) clustered separately from hand-plucked strings (red), and so did the wooden marimba (wooden bars) from the vibraphone (metal bars). However, the BRdissim tree is harder to interpret in the light of source-cause categories. In contrast to general dissimilarity ratings, each BRdissim cluster consisted of both continuously and impulsively excited instruments with little to no causal similarity but a grouping structure that clusters sounds according to brightness differences (VIB, HCD, VCE vs VLI, TRP, vs remaining instruments; see the BRdirect ratings in Fig. 4).

B. Scaling of dissimilarity and direct ratings

Next, the two sets of dissimilarity ratings were analyzed using nonmetric MDS (Kruskal, 1964b; Shepard, 1962), whereby it is assumed that only the ranks of a set of dissimilarities are known. Hence, nonmetric MDS produces distances that approximate these ranks, the latter being a nonlinear but monotonic transformation of the dissimilarities. The nonmetric approach has been proven robust in recovering the metric information of proximity data, even when random error is present (Young, 1970).

Figure 3 shows the evolution of the σ_1 (Stress-1; Kruskal, 1964a) and R^2 (the square of Pearson's r) goodness-of-fit measures for MDS solutions of between one and eight dimensions. Both measures exhibited clear knee points at two dimensions (2D) for the general dissimilarity ratings, but a smooth evolution for brightness dissimilarity. In fact, from a parsimonious perspective, the latter should thus be described using a one-dimensional (1D) solution (for the lack of a clear knee point). This result was in agreement with the coarser clustering structure observed in the BRdissim ratings, which had yielded three clusters of which a single cluster contained nine of the 14 sounds. Taken together, the data presented in Figs. 2 and 3 already suggested a clear qualitative difference in the underlying dimensionality of the general and brightness dissimilarity ratings.

Nevertheless, in order to scrutinize the intrinsic perceptual structure of brightness, we chose to inspect both 2D and 1D MDS solutions of the brightness dissimilarity ratings to facilitate comparisons with the 2D space representing general timbre dissimilarity and the 1D ordination from the direct brightness ratings, respectively (Fig. 4). For the 2D spaces, the order of dimensions reflects the order of columns in the respective MDS solution matrices. The first dimension of the GEdissim 2D space clearly separated impulsive from sustained sounds, which is in agreement with the literature (McAdams, 2019). The ordering of the 14 sounds along the first dimension of the BRdissim 2D space appeared to be spectral envelope based and moreover quite similar to that along the second dimension of the GEdissim 2D space. The second dimension in the BRdissim 2D space seemed to retain a temporal envelope based organization of the stimuli, but with much lower variance and the somewhat unexpected positioning of the bowed cello. Finally, BRdissim 1D and BRdirect yielded highly similar scalings of brightness across the tested sounds.

To examine the relation of brightness to general timbre dissimilarity, the relationships between the different dimensions in Fig. 4 were assessed by means of Pearson correlations (Table II). Standard errors (SE) for each coefficient (given in parentheses) were evaluated via 10 000 bootstrap replications (Efron and Tibshirani, 1994). The correlation between the first dimension of the brightness space (BRdissim 2D.1) and the second dimension of the general timbre space (GEdissim 2D.2) was high ($r = 0.83, p < 0.001$), as was that between the second brightness dimension (BRdissim 2D.2) and the first general timbre dimension (GEdissim 2D.1; $r = 0.87, p < 0.0001$). When brightness dissimilarities were scaled along a single dimension (BRdissim 1D) the stimuli configuration was equal to BRdissim 2D.1 ($r = 1.00, p < 0.0001$) but bore little relation to BRdissim 2D.2 ($r = 0.04, p = 0.88$). This reflected the lack of a clear knee point observed for BRdissim in Fig. 3. Furthermore, BRdissim 1D correlated well with GEdissim 2D.2 ($r = 0.81, p < 0.001$) but not with GEdissim 2D.1 ($r = 0.27, p = 0.36$). Direct ratings (BRdirect) correlated almost exactly with BRdissim 1D and BRdissim 2D.1 (both $r = 0.98, p < 0.0001$). Their correlation with GEdissim 2D.2 was comparable ($r = 0.77$, $p = 0.001$). Furthermore, the relationship of BRdirect to the two GEdissim dimensions was comparable to that between the latter and BRdissim 1D and BRdissim 2D.1.

These relationships were inspected further by looking at how the audio content descriptors of log attack time (LAT)

https://doi.org/10.1121/10.0002275

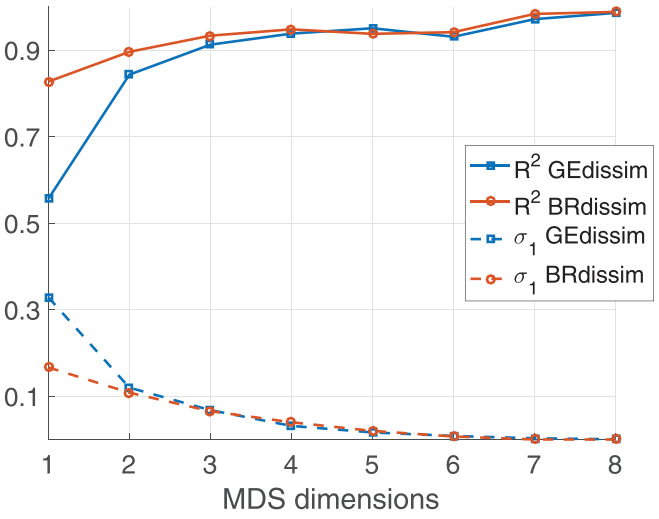

FIG. 3. (Color online) Goodness-of-fit measures for different MDS dimensionalities for general timbre and timbral brightness dissimilarity.

and SC correlated with the dimensions of general and brightness dissimilarity, and with the direct brightness ratings. The two descriptors were selected primarily for confirmatory purposes—they have been shown to account for a large portion of the variance in general dissimilarity tasks across a wide variety of sounds (Caclin *et al.*, 2005; Lakatos, 2000; McAdams *et al.*, 1995).

Pearson's coefficients and their SEs (obtained from 10 000 bootstrap replications) are reported in Table III. As expected, the GEdissim 2D.1 and 2D.2 dimensions correlated well with LAT ($r = -0.73, p < 0.01$) and SC ($r = 0.84, p < 0.001$), respectively. SC correlated even more strongly with the first dimension of the BRdissim 2D space ($r = 0.93, p < 0.0001$) and the 1D spaces of brightness dissimilarity ($r = 0.93, p < 0.0001$) and direct ratings ($r = 0.87, p < 0.0001$). However, the second dimension of BRdissim 2D did not correlate with SC ($r = 0.01$, $p = 0.975$). An examination of other spectral or spectrotemporal descriptors did not reveal any such correlates either (not reported here). Instead, BRdissim 2D.2 correlated well with LAT ($r = -0.64, p = 0.014$), which reflected its strong similarity with the first dimension of the general dissimilarity space. However, BRdissim 1D and BRdirect showed no correlation with LAT (both $r = 0.09, p = 0.75$).

C. Dissimilarity models

To examine whether source-cause categories exert an effect on timbral brightness perception, the general timbre and brightness dissimilarity data were analyzed using a modeling approach analogous to the one used by Siedenburg *et al.* (2016b). First, average ratings from each of the two dissimilarity tasks were predicted using a partial least-squares regression (PLSR) model that takes audio descriptors as regressors. The full set of spectral and temporal descriptors described in Sec. II D and Table I was used. PLSR is a generalization of multiple linear regression (MLR) that projects the predicted and observable variables onto respective sets of latent variables of maximum covariance (Wold, 1975; Wold *et al.*, 2001). Unlike MLR, PLSR can handle strongly collinear predictors, which is the case with the type of audio descriptors used here (Peeters *et al.*, 2011). For any single audio descriptor and stimulus pair, the absolute distance between the respective descriptor values was used as a predictor of dissimilarity. The dependent variable contained the 105 mean (general or brightness) dissimilarity ratings for the tested sounds.

It was then tested whether adding predictors derived from sound source-cause categories improved the model fit. Categorical predictors were based on the type of resonator (string, air column, bar), two types of resonator excitation (continuous, impulsive; blown, bowed, struck, plucked), and common instrument families in the western orchestra (woodwinds, brass, keyboards, strings, percussion). For all categorical descriptors, dissimilarity between instruments was treated as a binary code (Giordano *et al.*, 2013), encoding whether both sounds of a stimulus pair shared the same category (0) or not (1).

Here we used PLSR as implemented in the plsregress function provided by MATLAB, which uses the SIMPLS algorithm (de Jong, 1993). The significance of the regression coefficients was estimated by bootstrapping 95% confidence intervals; if intervals overlapped with zero, a variable's contribution was considered to be not significant (Mehmood *et al.*, 2012). To prevent overfitting of the response variable, six-fold cross-validation (Wold *et al.*, 2001) indicated a clear knee point for a model with three components, which was used in all subsequent analyses. Variables were z-normalized prior to entering the model.

Figure 5 displays the predicted and observed GEdissim and BRdissim data for three regression models (acoustical, categorical, combined) together with the corresponding proportions of explained variance (R^2). For general dissimilarity (upper row panels), the acoustical model yields a good fit with 85% of the overall variance in general dissimilarity data shared. There is one marked outlier on the right hand side of the regression line (coordinates $x = 0.86$, $y = 0.51$), which corresponds to an overestimation of dissimilarity by the acoustical model. This outlier corresponds to the instrument pair vibraphone-marimba, both of which are likely recognized as percussion instruments by the musician participants and thus judged as similar, even though there are drastic acoustical differences between the two tones (e.g., wooden versus metal bars). Hence, this outlier is indicative of the important role of source-cause categorical cues in general dissimilarity judgments. The model using only categorical variables well predicts the data, but not as accurately as the acoustical model, sharing 63% of the variance with the general dissimilarity data. Importantly, the combination of both models yields an improved fit ($R^2 = 0.92$) without possessing any strong outliers (Fisher's two-tailored z-test on the difference of correlations, $z = 2.36, p = 0.0183$).

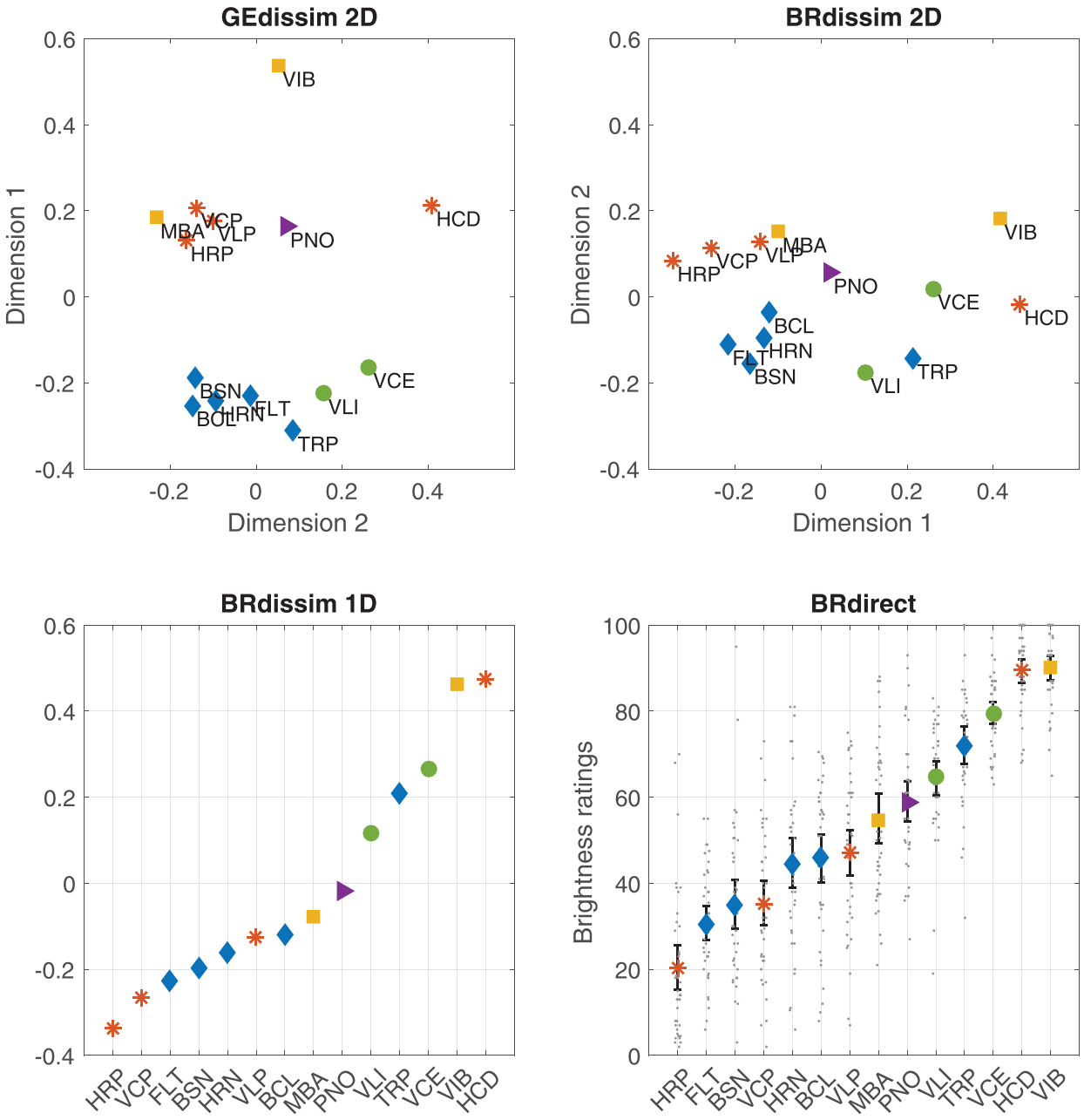

FIG. 4. (Color online) Top left: 2D MDS configuration for general timbre dissimilarity. Top right: 2D MDS configuration for timbral brightness dissimilarity. The order of dimensions reflects the order of columns in the respective MDS solution matrices. Bottom left: 1D MDS configuration for timbral brightness dissimilarity. Bottom right: Average direct brightness ratings; small dots correspond to individual ratings; error bars indicate 95% confidence intervals obtained via bootstrapping. Rhombus, blown air column; square, struck bar; circle, bowed string; star, plucked string; triangle, struck string.

For the brightness dissimilarity ratings (lower row panels in Fig. 5), the situation appears to be different. On the one hand, the acoustical model yields quantitatively the same fit as for general dissimilarity ($R^2 = 0.85$). However, the explanatory power of the categorical variables appears to be much weaker and they only share 40% of variance with the brightness dissimilarity data. When both acoustical and categorical descriptors were used, the combined model did not improve substantially over the acoustical model ($R^2 = 0.86$, Fisher's z-test: $z = 0.27, p = 0.78$). Note that the categorical model tends to yield predictions that group vertically because it only uses four predictors, each of which assigns binary dissimilarity values.

These observations were further confirmed by bootstrapped R^2 values (Efron and Tibshirani, 1994) as shown in Fig. 6. Specifically, every model was instantiated 10 000 times and per instance 14 stimuli were drawn at random with replacement from the set of the 14 original stimuli. The (general or brightness) dissimilarities of the resulting stimulus pairs were then predicted by the different model types (acoustical, categorical, combined), which provides a distribution of the resulting R^2 values. As a baseline, R^2 values

https://doi.org/10.1121/10.0002275

TABLE II. Pearson correlations r between the MDS dimensions of timbral brightness and general timbre dissimilarity, and between those and direct brightness ratings. See Fig. 4 for labels; X indicates the respective MDS plane dimension; () report SEs estimated by bootstrap with 10 000 runs.

Brightness MDS dimensions and direct ratings[a]	
BRdissim 1D–BRdissim 2D.1	1.00 (0.00) ***
BRdissim 1D–BRdissim 2D.2	0.04 (0.27)
BRdissim 1D–BRdirect	0.98 (0.01) ***
BRdissim 2D.1–BRdirect	0.98 (0.01) ***
BRdissim 2D.2–BRdirect	0.09 (0.27)
General MDS dimensions and brightness direct ratings[a]	
GEdissim 2D.1–BRdirect	0.26 (0.30)
GEdissim 2D.2–BRdirect	0.77 (0.11) *
General and brightness MDS dimensions[a]	
GEdissim 2D.1–BRdissim 1D	0.27 (0.32)
GEdissim 2D.1–BRdissim 2D.1	0.23 (0.32)
GEdissim 2D.1–BRdissim 2D.2	0.87 (0.06) ***
GEdissim 2D.2–BRdissim 1D	0.81 (0.10) **
GEdissim 2D.2–BRdissim 2D.1	0.83 (0.09) **
GEdissim 2D.2–BRdissim 2D.2	−0.24 (0.21)

[a]* $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$.

from a random model obtained by randomly shuffling the predictors of the combined model were included. Any of the three descriptor sets (acoustical, categorical, combined) improves over the random model. However, whereas the combined model for GEdissim generates R^2 values superior to the respective acoustical model, the distributions of R^2 values from these two models for BRdissim coincide.

IV. DISCUSSION

In a comprehensive examination of brightness perception of orchestral instrument sounds, by contrasting different methodological concepts we focused on the dimensionality of timbral brightness, its robustness across methods, and its relation to instrument categories. The present findings both have a confirmatory relation to the present state of knowledge on timbre perception and expand on it by providing answers to important yet previously unexplored questions concerning the perceptual and cognitive processes that determine timbral brightness perception.

TABLE III. Pearson correlations r between the two audio descriptors of SC and LAT, and individual dimensions of general dissimilarity, brightness dissimilarity, and direct brightness ratings. See Fig. 4 for labels; () report SEs estimated by bootstrap with 10 000 runs.

Dimension	SC[a]	LAT[a]
GEdissim 2D.1	0.30 (0.34)	−0.73 (0.10) *
GEdissim 2D.2	0.84 (0.10) **	0.31 (0.28)
BRdissim 2D.1	0.93 (0.05) ***	0.10 (0.28)
BRdissim 2D.2	0.01 (0.29)	−0.64 (0.15) *
BRdissim 1D	0.93 (0.05) ***	0.09 (0.28)
BRdirect	0.87 (0.07) ***	0.09 (0.27)

[a]* $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$.

The first question that steered the present research concerned the intrinsic dimensionality of timbral brightness as a perceptual attribute of musical instrument sounds when considered through the same empirical angle as general timbre perception, namely, dissimilarity ratings. Hierarchical clustering (Fig. 2) and MDS (Figs. 3 and 4) of general timbre dissimilarity ratings and brightness dissimilarity ratings suggested that the latter were less complex than the former. Brightness dissimilarity could be adequately described on the basis of a single dimension correlated with the SC of the tested stimuli (BRdissim 1D in Fig. 4), whereas at least 2D were needed for general timbre dissimilarity, one temporal (attack time) and a SC one, in agreement with the literature. The two SC dimensions further correlated strongly with each other and with the ordering of the tested sounds obtained from direct ratings of brightness (Table II), confirming the view that brightness, as modeled by the SC, is a relatively robust unitary perceptual dimension for acoustic instrument sounds.

When a 2D space of brightness dissimilarity was considered (BRdissim 2D in Fig. 4), its second dimension correlated with the attack-time dimension of the general timbre space (Tables II and III). Given that half of the participants performed the BRdissim task after having done the GEdissim task, it could be argued that they have shown a transfer effect from general to brightness dissimilarity, potentially resulting in attack time playing a small role in the latter. However, the practically perfect correlation between the brightness dissimilarity matrices from those listeners who did the BRdissim task first (half of the participants) and those who did it following the GEdissim task renders that hypothesis unlikely.

Instead, this finding could also suggest a leakage of general timbre dissimilarity into timbral brightness dissimilarity. Because participants may not be able to focus on specifically rating dissimilarity in terms of brightness, ratings may also reflect aspects of general timbre dissimilarity, and hence attack time. That is, brightness perception itself may not be substantially influenced by attack time as the correlation between the two MDS planes would appear to suggest, but the brightness dissimilarity ratings were potentially infiltrated by general dissimilarity. This view is corroborated by considering the observed stability of brightness judgments across dissimilarity ratings and MUSHRA-inspired direct multi stimulus ratings (BRdirect in Fig. 4), pertaining to the second question posed by the present study. Whereas both BRdissim and BRdirect tasks largely recovered the SC or "brightness" dimension of GEdissim, average IRC was highest for BRdirect and lowest for BRdissim (Fig. 1). This indicates that in the latter task, listeners lost a common frame of reference likely afforded by the combination of direct rating, ranking, and multiple comparison in the BRdirect task. This might further relate to the susceptibility of pairwise dissimilarity ratings to conflate other processes (Melara *et al.*, 1992; Siedenburg *et al.*, 2016b).

Another critical point to consider is that there exists an inherent correlation of temporal and spectral features in

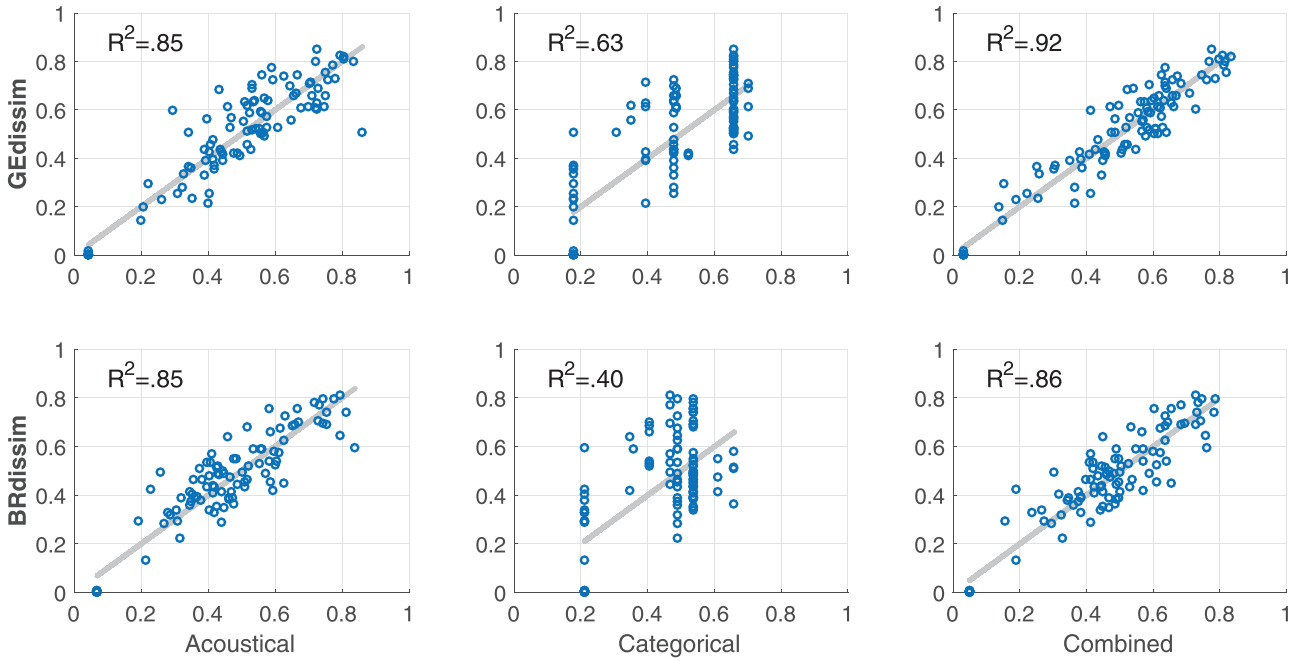

FIG. 5. (Color online) Scatterplot of normalized model prediction (x-axes) and average empirical data (y-axis) for general dissimilarity ratings (top) and brightness dissimilarity ratings (bottom). Columns correspond to models relying on acoustical descriptors (left), categorical descriptors of source-cause categories (middle), and the combined set of acoustical and categorical descriptors (right).

natural acoustic stimuli, such as musical instrument sounds (Elliott *et al.*, 2013; Patil *et al.*, 2012; Thoret *et al.*, 2017). Furthermore, perceptual dimensions of timbre have been described as interactive (Caclin *et al.*, 2007). Examining the "leakage" scenario against that of a potential temporal dimension for timbral brightness perception would thus require the use of synthetic sounds carefully controlled along disassociated temporal and spectral properties.

The third question of this study touched on another important issue, namely, the way in which cognitive processes related to the formation of source-cause categories of instrumental sounds intertwine in timbral brightness perception. This was addressed by means of dissimilarity models using PLSR (Figs. 5 and 6). Spectral and temporal scalar descriptors of the acoustic signal provided good predictions of both general timbre and brightness dissimilarity ratings. By using a *post hoc* inclusion of a set of categorical predictors that described an instrument's family membership and facts about source and excitation mechanisms, predictions of GEdissim improved by around seven percentage points compared to the solely acoustical model. On the contrary, correlations between observed and predicted BRdissim improved only slightly from the solely acoustical to the combined model.

These results replicate the findings from Siedenburg *et al.* (2016b), indicating that musicians integrate knowledge about source-cause categories in dissimilarity ratings of acoustic instrument tones. However, the present results demonstrate that this effect is specific to general dissimilarity: when listeners were instructed to rate dissimilarity based only on brightness, source-cause categories appeared to lose predictive power. This suggests that brightness as a perceptual attribute is underpinned primarily by acoustical rather than causal similarity (as per the terminology proposed by Lemaitre *et al.*, 2010). More generally, natural acoustic stimuli such as musical instrument sounds exhibit an inherent coupling of continuous acoustical dimensions and source-cause categories, which is what allows listeners

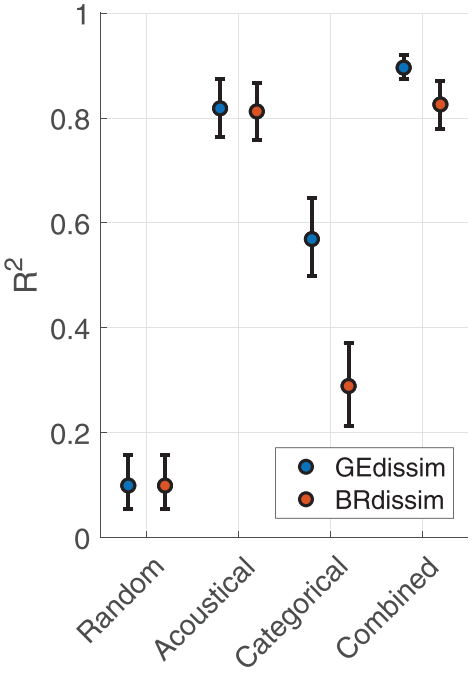

FIG. 6. (Color online) Bootstrapped R^2 values for PLSR models of general and brightness dissimilarity ratings using randomized descriptors, acoustic and categorical descriptors, and their combination.

familiar with such sounds to infer their source-cause in the first place. The present results thus indicate that categorical effects on general dissimilarity can be diminished when instructing listeners to base dissimilarity solely upon more constrained perceptual dimensions such as brightness.

V. CONCLUSION

In this paper, we studied brightness perception for musical instrument sounds by focusing on its dimensionality as an auditory attribute, its stability across different psychoacoustical contexts, and its relation to source-cause categories of acoustic instruments. Triangulating general timbre dissimilarity ratings with brightness dissimilarity ratings and direct multistimulus ratings of brightness corroborated that brightness is a salient component of (general) timbre perception. Results confirm the view that timbral brightness, as modeled by the SC, is a relatively robust unitary auditory dimension. However, an observed correlation between brightness dissimilarity ratings and the attack time dimension of the general dissimilarity space seems to suggest that brightness dissimilarity may have been infiltrated by general timbre dissimilarity, a finding that warrants further investigation. Finally, a PLSR model of timbre dissimilarity was used to compare the contributions of source-cause categories to general timbre and brightness dissimilarity ratings. When binary descriptors related to acoustic instrument family and excitation mechanisms were combined with audio descriptors, correlations with observed dissimilarities improved substantially for general timbre dissimilarity, but not for brightness dissimilarity. We interpret this as evidence that brightness perception is underpinned primarily by acoustical rather than source-cause cues.

ACKNOWLEDGMENTS

C.S. wishes to thank the Alexander von Humboldt Foundation for support through a Humboldt Research Fellowship (2016–2018). K.S. has received funding from the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014–2020) under the Marie Skłodowska-Curie Grant Agreement No. 747124. K.S. has also received funding from a Freigeist Fellowship of the Volkswagen Foundation.

[1]http://vsl.co.at (Last viewed January 3, 2020).
[2]http://psiexp.music.mcgill.ca/psiexp/ (Last viewed July 22, 2020).
[3]http://puredata.info (Last viewed July 22, 2020).

Almeida, A., Schubert, E., Smith, J., and Wolfe, J. (**2017**). "Brightness scaling of periodic tones," Atten. Percept. Psychophys. **79**(7), 1892–1896.
Caclin, A., Giard, M.-H., Smith, B. K., and McAdams, S. (**2007**). "Interactive processing of timbre dimensions: A Garner interference study," Brain Res. **1138**, 159–170.
Caclin, A., McAdams, S., Smith, B. K., and Winsberg, S. (**2005**). "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," J. Acoust. Soc. Am. **118**, 471–482.
Caetano, M., Saitis, C., and Siedenburg, K. (**2019**). "Audio content descriptors of timbre," in *Timbre: Acoustics, Perception, and Cognition*, edited by K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay (Springer, New York), pp. 297–333.
de Jong, S. (**1993**). "SIMPLS: An alternative approach to partial least squares regression," Chemometr. Intell. Lab. Syst. **18**(3), 251–263.
Efron, B., and Tibshirani, R. J. (**1994**). *An Introduction to the Bootstrap* (CRC Press, Boca Raton, FL).
Elliott, T. M., Hamilton, L. S., and Theunissen, F. E. (**2013**). "Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones," J. Acoust. Soc. Am. **133**(1), 389–404.
Faure, A., Mcadams, S., and Nosulenko, V. (**1996**). "Verbal correlates of perceptual dimensions of timbre," in *Proceedings of the 4th International Conference on Music Perception and Cognition*, August 11–15, Montreal, Canada, pp. 79–84.
Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., and Belin, P. (**2013**). "Abstract encoding of auditory objects in cortical activity patterns," Cereb. Cortex **23**, 2025–2037.
Glasberg, B. R., and Moore, B. C. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear Res. **47**(1–2), 103–138.
Grey, J. M. (**1977**). "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am. **61**, 1270–1277.
ITU (**2015**). ITU-R BS 1534-3, *Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems* (International Telecommunication Union, Geneva, Switzerland).
Krimphoff, J., McAdams, S., and Winsberg, S. (**1994**). "Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique" ("Characterization of the timbre of complex sounds. II. Acoustical analysis and psychophysical quantification"), J. Phys. IV **4**(C5), 625–628.
Kruskal, J. B. (**1964a**). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," Psychometrika **29**(1), 1–27.
Kruskal, J. B. (**1964b**). "Nonmetric multidimensional scaling: A numerical method," Psychometrika **29**(2), 115–129.
Lakatos, S. (**2000**). "A common perceptual space for harmonic and percussive timbres," Percept. Psychophys. **62**(7), 1426–1439.
Lemaitre, G., Houix, O., Misdariis, N., and Susini, P. (**2010**). "Listener expertise and sound identification influence the categorization of environmental sounds," J. Exp. Psychol. Appl. **16**(1), 16–32.
Lemaitre, G., Vartanian, C., Lambourg, C., and Boussard, P. (**2015**). "A psychoacoustical study of wind buffeting noise," Appl. Acoust. **95**, 1–12.
Lokki, T., Pätynen, J., Kuusinen, A., Vertanen, H., and Tervo, S. (**2011**). "Concert hall acoustics assessment with individually elicited attributes," J. Acoust. Soc. Am. **130**, 835–849.
McAdams, S. (**2019**). "The perceptual representation of timbre," in *Timbre: Acoustics, Perception, and Cognition*, edited by K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay (Springer, New York), pp. 23–57.
McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (**1995**). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," Psychol. Res. **58**(3), 177–192.
Mehmood, T., Liland, K. H., Snipen, L., and Sæbø, S. (**2012**). "A review of variable selection methods in partial least squares regression," Chem. Intell. Lab. Syst. **118**, 62–69.
Melara, R. D., Marks, L. E., and Lesko, K. E. (**1992**). "Optional processes in similarity judgments," Percept. Psychophys. **51**(2), 123–133.
Ogg, M., Slevc, L. R., and Idsardi, W. J. (**2017**). "The time course of sound category identification: Insights from acoustic features," J. Acoust. Soc. Am. **142**(6), 3459–3473.
Patil, K., Pressnitzer, D., Shamma, S., and Elhilali, M. (**2012**). "Music in our ears: The biological bases of musical timbre perception," PLoS Comp. Biol. **8**(11), e1002759.
Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (**1992**). "Complex sounds and auditory images," in *Auditory Physiology and Perception*, edited by Y. Cazals, L. Demany, and K. Horner (Pergamon Press, Oxford, UK), pp. 429–446.
Pearce, A., Brookes, T., and Mason, R. (**2017**). "Timbral attributes for sound effect library searching," in *Proceedings of the 2017 AES International Conference on Semantic Audio*, June 22–24, Erlangen, Germany.
Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (**2011**). "The Timbre Toolbox: Extracting audio descriptors from musical signals," J. Acoust. Soc. Am. **130**(5), 2902–2916.
Saitis, C., and Weinzierl, S. (**2019**). "The semantics of timbre," in *Timbre: Acoustics, Perception, and Cognition*, edited by K. Siedenburg, C. Saitis,

S. McAdams, A. N. Popper, and R. R. Fay (Springer, New York), pp. 119–149.
Samoylenko, E., McAdams, S., and Nosulenko, V. (**1996**). "Systematic analysis of verbalizations produced in comparing musical timbres," Int. J. Psychol. **31**(6), 255–278.
Schubert, E., and Wolfe, J. (**2006**). "Does timbral brightness scale with frequency and spectral centroid?," Acust. Acta united Ac. **92**(5), 820–825.
Shepard, R. N. (**1962**). "The analysis of proximities: Multidimensional scaling with an unknown distance function. II," Psychometrika **27**(3), 219–246.
Siedenburg, K. (**2018**). "Timbral Shepard-illusion reveals perceptual ambiguity and context sensitivity of brightness perception," J. Acoust. Soc. Am. **143**(2), EL93–EL98.
Siedenburg, K., Fujinaga, I., and McAdams, S. (**2016a**). "A comparison of approaches to timbre descriptors in music information retrieval and music psychology," J. New Music Res. **45**(1), 27–41.
Siedenburg, K., Jones-Mollerup, K., and McAdams, S. (**2016b**). "Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds," Front. Psychol. **6**, 1977.
Stark, J. (**2003**). *Bel Canto: A History of Vocal Pedagogy* (University of Toronto Press, Toronto, Canada).
Thoret, E., Depalle, P., and McAdams, S. (**2017**). "Perceptually salient regions of the modulation power spectrum for musical instrument identification," Front. Psychol. **8**, 587.
Wallmark, Z. (**2019**). "A corpus analysis of timbre semantics in orchestration treatises," Psychol. Music **47**, 585–605.
Weinzierl, S., Lepa, S., and Ackermann, D. (**2018**). "A measuring instrument for the auditory perception of rooms: The room acoustical quality inventory (RAQI)," J. Acoust. Soc. Am. **144**(3), 1245–1257.
Wold, H. (**1975**). "Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach," in *Perspectives in Probability and Statistics*, *Papers in Honour of M. S. Bartlett*, edited by J. Gani (Academic Press, London), pp. 117–142.
Wold, S., Sjöström, M., and Eriksson, L. (**2001**). "PLS-regression: A basic tool of chemometrics," Chemometr. Intell. Lab. Syst. **58**(2), 109–130.
World Medical Association (**2013**). "World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects," J. Am. Med. Assoc. **310**(20), 2191–2194.
Young, F. W. (**1970**). "Nonmetric multidimensional scaling: Recovery of metric information," Psychometrika **35**(4), 455–473.
Zacharakis, A., Pastiadis, K., and Reiss, J. D. (**2014**). "An interlanguage study of musical timbre semantic dimensions and their acoustic correlates," Music Percept. **31**, 339–358.
Zacharakis, A., Pastiadis, K., and Reiss, J. D. (**2015**). "An interlanguage unification of musical timbre: Bridging semantic, perceptual, and acoustic dimensions," Music Percept. **32**(4), 394–412.