Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre

Etienne Thoret, Baptiste Caramiaux, Philippe Depalle, Stephen Mcadams

▶ **To cite this version:**

Etienne Thoret, Baptiste Caramiaux, Philippe Depalle, Stephen Mcadams. Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre. Nature Human Behaviour, 2020, 5, pp.369-377. 10.1038/s41562-020-00987-5 . hal-03033757

HAL Id: hal-03033757
https://hal.science/hal-03033757v1

Submitted on 25 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre

Authors: Etienne Thoret[1,2,3*)], Baptiste Caramiaux[4], Philippe Depalle[1], Stephen McAdams[1]

*) Corresponding author: etienne.thoret@mail.mcgill.ca

[1] Schulich School of Music, McGill University, Montreal, Canada

[2] Aix Marseille Univ, CNRS, PRISM, LIS, Marseille, France

[3] Institute of Language Communication and the Brain (ILCB), Marseille, France

[4] Université Paris-Saclay, CNRS, Inria, LRI, Gif-sur-Yvette, France

Abstract

Humans excel at using sounds to make judgements about their immediate environment. In particular, timbre is an auditory attribute that conveys crucial information about the identity of a sound source, especially for music. While timbre has been primarily considered to occupy a multidimensional space, unravelling the acoustic correlates of timbre remains a challenge. Here we re-analyze 17 datasets from published studies between 1977 and 2016 and observe that original results are only partially replicable. We use a data-driven computational account to reveal the acoustic correlates of timbre. Human dissimilarity ratings are simulated with metrics learned on acoustic spectrotemporal modulation models inspired by cortical processing. We observe that timbre has both generic and experiment-specific acoustic correlates. These findings provide a broad overview of former studies on musical timbre and identify its relevant acoustic substrates according to biologically inspired models.

Introduction

The human auditory system processes acoustic information through several levels of increasing complexity[1,2]. At the first levels, transformations of the physical vibrations into neural activity conveyed by the peripheral auditory system are reasonably well known[3]. Yet, nonlinear transformations carried out at higher cortical levels remain poorly understood from a computational point of view. This is all the more striking when considering the perception of complex and rich sounds such as those from musical instruments, which involves a complex combination of both bottom-up and top-down processes. As such, achieving a unified comprehension of how human auditory perception responds to such complex sounds has the potential to critically advance our global understanding of the processing of acoustic information by the auditory system. We propose a meta-analysis of 17 published datasets and apply metric distance learning to acoustic models of spectrotemporal modulations similar to those represented in primary auditory cortex. Our main contributions are thus: 1) to address theoretically the historical question of defining the acoustical basis of musical instrument timbre, and 2) to develop a data-driven method to reveal the acoustic correlates of sound perception and interpret these correlates with spectrotemporal modulation (STM) models. More generally, this study implements a data-driven procedure to address the relationship between the physical properties of complex sounds and their perceptual ratings.

Musical instrument sounds have perceptually salient properties like loudness (perceived energy), pitch (perceived frequency), and timbre (related to their "sound quality"). Although the first two properties are well understood, the timbre of musical instrument sounds remains complex and ill defined[4]. Wherein lies the difference between the sounds produced by two musical instruments playing the same pitch at the same loudness? Participants are usually asked to rate the dissimilarity between pairs of sounds (Figure 1A) from which relevant perceptual dimensions of a timbre space are revealed by multidimensional scaling (MDS) (Figure 1B) [5,6,7,8,9,10,11,12,13,14,15,16,17]. MDS produces a low-dimensional parametric space (typically with two or three dimensions) in which sounds are assigned coordinates, and the distances between sounds reflect their perceptual dissimilarities. Relevant dimensions of these timbre spaces are then correlated with audio descriptors computed from the sound signal, enabling a psychoacoustic interpretation of musical timbre perception. However, this approach suffers from severe limitations. First, it has not been tested simultaneously across a wide variety of musical instrument timbre datasets, which renders

its generalizability uncertain. Second, among a plethora of audio descriptors, only two descriptors have been found to correlate well with the two first dimensions of timbre spaces: the attack time, corresponding to the initial onset portion of the temporal envelope, and the centroid of the spectral content[11,18]. Despite more than 40 years of research on timbre, the acoustic correlates for higher dimensions remain unclear[15]. Finally, and most problematically, by reducing timbre spaces to two main dimensions, the subtle differences between musical instrument timbres are lost. A rigorous argument comes from a series of experiments done on cochlear implanted patients. A study[19] indeed observed that dissimilarity rating experiments run with hearing-impaired participants led to the same main two dimensions in timbre spaces, although these listeners are clearly unable to perceive the subtleties of musical timbre.

Here, we present the meta-analysis on 17 datasets, comprising a wide range of stimuli and dissimilarity measures, stemming from eight historical studies on timbre between 1977 and 2016[7,8,10,11,12,13,14,16]. The studies employed very different stimuli ranging from recorded notes to synthesized and hybrid sounds (see Extended Data Figure 1 for details on the datasets). We trained an interpretable distance metric that could reveal latent acoustic dimensions involved in sound perception. In other words, rather than determining the salient dimensions of dissimilarity measures and correlating these with scalar acoustic features[20], we computationally train a distance kernel function k (similar to a weighted Euclidean distance) between acoustic models to maximally correlate with human dissimilarity measures (Figure 1C). The trained weights of this function highlight acoustic information fitting human perceptual distance.

We apply metric learning to acoustic models as achieved by primary auditory cortices, highlighting the spectrotemporal modulations of a sound event[21]. Spectrotemporal modulation (STM) models have been shown to be relevant to the study of musical sounds[14,22]. It is however unknown which parts of these acoustic models are relevant from a perceptual point of view.

Results

Multidimensional scaling analysis. To assess the replicability of previous studies, we first re-analyzed the 17 datasets with the classical MDS-based approach. Although parameters are usually adjusted differently in each study, here we used the same method for all the datasets (see Methods for details). We evaluated the Spearman correlations of the positions along the dimensions of the

MDS solutions with the two most basic acoustic descriptors (hereafter expressed as ρ^2): the logarithm of the attack time (LAT) and the spectral centroid (SC)[20]. Globally, the LAT correlates moderately with the first dimension of the timbre spaces [ρ^2: Mdn = .41, IQR = .29, Extended Data Figure 2] and moderately with the second dimension [ρ^2: Mdn = .20, IQR = .32, Extended Data Figure 2]. Mdn is the median and IQR is the Interquartile Range of the Spearman correlations for the 17 datasets. The degrees of freedom ranges between 9 and 18. The detailed full statistics (significance, statistical power and 95% Confidence Intervals) for each dataset are reported within the Supplementary Table 1. It must be noted that datasets including sounds whose temporal envelopes have been manipulated (Iverson & Krumhansl, 1993, Onset dataset; Iverson & Krumhansl, 1993, Remainder dataset)[7,10] failed to provide high correlations with MDS dimensions, whereas those including unmanipulated stimuli had higher correlations. The correlations between SC and the first dimension are generally poor [ρ^2: Mdn = .06, IQR=.07, Extended Data Figure 2 and Supplementary Table 1 for the full statistics], whereas it correlates well with the second dimension [ρ^2: Mdn = .61, IQR=.45, Extended Data Figure 2 and see Supplementary Table 1 for the full statistics]. In addition, when available, the original acoustical analyses can be replicated more or less well depending on the study (Figure 2, Extended Data Figure 3). In order to optimize the correlation between LAT and SC with one of the two first dimensions, we chose a method that automatically finds the best set of MDS parameters, i.e. the number of projected dimensions and the optimal rotation of the MDS dimensions (see Methods). Nevertheless, replicating the published correlation values remains very difficult. This is a major limitation of the MDS-based approach. Indeed, it critically depends on implicit decisions made by the authors, often not accurately and/or comprehensively reported in the published papers: the experimenter always makes choices to optimize the final outcomes, e.g., by choosing one particular type of MDS, adapting its parameters, or hand-tuning the audio descriptors used for the correlations. Taken together, these observations show the limitation of the dimensional approach and the need for an alternative to reveal the acoustical substrates of musical timbre.

Spectro-Temporal Modulations of musical sounds. Spectro-Temporal Modulation (STM) models are produced by mathematical tools that mimic the output of the processing of sounds by primary auditory cortical neurons. These models refer to Spectro-Temporal Receptive Fields (STRFs)[23,24], which are tuned to fire for specific STM patterns in sounds. Practically, the auditory spectrogram, representing the cochlear processing of a sound, is projected into a four-dimensional

representation space characterized by time, frequency (in Hz), temporal modulations (rate in Hz), and spectral modulations (scale in cycles/octave) (Figure 1). This STM representation (STMF: Spectro Temporal Modulation transfer Function) reflects the multiresolution analysis of the spectrotemporal information[21,25], which has already provided insights into the identification of musical instrument sounds[14,26,27,28,29] and correlates with human brain activity[30,31,32]. In this study, 128 frequency channels, 22 rates, and 11 scales were chosen to compute the representations (see Methods). As these representations are high-dimensional, they are averaged over time leading to tensors of dimension 128 × 22 × 11 (called the Full STMF below; further technical details on computations are provided in Methods) (Figure 1C, left box). In order to understand the role of the three acoustic dimensions embedded in the global 3D tensor, representations projected onto each pair of the three dimensions will also be considered, i.e., scale/rate (averaged over frequency), frequency/rate (averaged over scale) and frequency/scale (averaged over rate) (Figure 1C, middle box).

Optimized metrics simulating human dissimilarity ratings. The human ratings are perceptual distances between sounds. To understand the extent to which these ratings are derived from the acoustic difference between sounds, we fit a distance kernel (radial basis function) between STM representations that best approximate the perceptual distances (see Methods and Supplementary Figure 1). This method seeks the parts of the STMFs that are the most relevant to reproduce the ratings of the stimuli. On median, across the 17 datasets, 78% of the variance in dissimilarity ratings is explained by the optimized metrics [r^2: Mdn=.78, IQR=.58] (see Table 1 and Supplementary Table 2). We control for overfitting of each metric with the leave-one-sound-out cross-validation method (see Methods, Extended Data Figure 4 and Supplementary Table 3). Notably, the 2D projections scale-rate, freq-rate and freq-scale do not explain a large proportion of variance alone [scale-rate: r^2: Mdn=.20, IQR=.25; freq-rate: r^2: Mdn=.32, IQR=.49; freq-scale: r^2: Mdn=.55, IQR=.57] (see Supplementary Table 4, Supplementary Table 5, Supplementary Table 6). Lastly, as frequency and rate have more dimensions than scale, we tested for another potential overfitting in these two dimensions by running the previous analysis after having downsampled the rates and frequencies to the same number of dimensions as the scales (10 frequencies × 10 rates × 10 scales, see Methods for details). The explained variances obtained with this test are lower on average than with the higher-dimensional representation (see comparison in Table 1, Supplementary Table 7, Supplementary Table 8, Supplementary Table 9, Supplementary Table

10) but respect the same trends. As a first control, we ran the optimization process with the time-averaged auditory spectrum, which revealed lower proportion of variance explained [r^2: Mdn = .19, IQR = .48] (see Supplementary Table 11). As a second control, we computed the Euclidean distance between Full STMFs of pairs of stimuli to show the interest of learning the distance. This does not allow us to accurately simulate the human dissimilarities [r^2: Mdn = .11, IQR = .20] (see Table 1 and Supplementary Table 12, Supplementary Table 13). These results support the relevance of the distance metric learning approach to suggest a link between human behavioral data with acoustic representations of the stimuli.

Acoustic interpretations of the metrics. We first assess whether the kernels fitted to each individual dataset show similarities between datasets. We computed pairwise Spearman (ρ) correlations between the fitted weights w (see Methods). Spearman correlations were computed as the different fitted metrics may have different scales that may bias Pearson's correlations values. A summary of the full statistics is available in Supplementary Table 14. The fitted kernels for the Full STMF do not generalize well across the different datasets [ρ^2(30,974): Mdn=.25, IQR=.17]. Nevertheless, if we inspect each 2D projection of the metrics, as observed in Figure 4, for example, the fitted weights appear to have qualitatively similar traits. In particular, the weights fitted on the scale/rate projection are very similar across all of the datasets [ρ^2(240): Mdn=.68, IQR=.33]. In addition, for each metric, most of the energy is centered on low temporal modulations (< 15 Hz) and spectral modulations around 1 cycle/oct. Conversely, the weights fitted on the frequency/rate and frequency/scale projections are specific to each dataset [ρ^2(2,814): Mdn=.50, IQR=.23, and ρ^2(1406): Mdn=.28, IQR=.29, respectively].

As a control, we also equalized the dimensionalities of each representation to evaluate a potential influence of this factor on the correlation. Overall, it confirms our conclusions by showing very similar trends between the different projections [Downsampled Full STMF (10×10×10)— ρ^2(98): Mdn=.30, IQR=.18; scale/rate—ρ^2(98): Mdn=.58, IQR=.32; freq/rate—ρ^2(98): Mdn=.55, IQR=.26; freq/scale—ρ^2(98): Mdn=.45, IQR=.37]. We further refined this analysis by considering only the 10 datasets with stimuli that share the same fundamental frequency (Eb4 = 311 Hz). This confirms the previous observations: scale-rate metrics generalize between the datasets better than frequency-rate and frequency-scales [Full STMF: ρ^2(30,974): Mdn=.30, IQR=.18; scale/rate—

ρ^2(240): Mdn=.74, IQR=.25; freq/rate—ρ^2(2,814): Mdn=.55, IQR=.22; frequency/scale—ρ^2(1406): Mdn=.38, SD=.31].

Clustering of optimized metrics. In order to understand in depth how the stimuli of the experiments drive the properties of the metrics, clustering analyses were run based on the previous correlational analyses (Figure 3). For the Full STMF and each of its three projections, a cluster analysis was conducted based on the pairwise Spearman correlations. This analysis aims to understand whether: 1) the stimulus sets of similar stimuli group together; 2) edited, recorded and synthesized stimuli group together; 3) fundamental frequency affects grouping; and 4) which representation captures general and more stimulus-set-specific features. Based on the clusters determined with a hierarchical clustering of the Spearman correlation coefficients with a complete linkage method, the first two factors do not affect grouping in the case of the datasets under consideration. It's worth noting that 10 out of 17 datasets have the same fundamental frequency (311Hz) which partly limits the potential conclusions regarding the effect of fundamental frequency. Further research should therefore be conducted to investigate more systematically the relationship between fundamental frequency and timbre. However, scale-rate is the representation that captures most of the generality across the datasets as shown by the tight clustering of all but five datasets. This reinforces the idea that timbre has generic dimensions embedded in the scale-rate projection and that the subtler aspects of timbre that are specific to the stimulus set are embedded in the frequency-rate and frequency-scale projections.

Timbre perceptual metrics are experiment-specific. In order to understand the links between the stimuli and the metrics for each dataset, Pearson correlations between the fitted metrics and the sample-wise standard deviations of the stimuli were computed. Practically, for each dataset and each metric, we first compute the standard deviation for each bin of the representation across the stimuli, i.e., the standard deviation "pixel per pixel". Then, the Pearson correlation between the vectorized arrays of standard deviations and the vectorized version of the metrics was computed. This revealed that the full STMF is strongly correlated with the stimulus standard deviations [r^2(30,974): Mdn=.75, IQR=.12] (Figure 4, Supplementary Figures 2-18, Supplementary Table 15). If we inspect these correlations for each of the three 2D representations, all three strongly correlate with the stimulus sample-wise standard deviations [scale-rate—r^2(240): Mdn=.90, IQR=.11; frequency-rate—r^2(2,814): Mdn=.83, IQR=.04; frequency-scale—r^2(1406): Mdn=.79, IQR=.10] (Figure 4, Supplementary Table 16, Supplementary Table 17, Supplementary

Table 18). This result strongly suggests that perceptual metrics are experiment-dependent and can be derived from the stimulus variability. In order to understand more precisely the link between the optimized metrics and the stimuli of each experiment, we performed a multiple linear regression between the fitted metrics and the stimuli for the four representations (Full STMF, scale-rate, freq-rate, freq-scale) for each dataset. We observed that each metric can be decomposed as a linear combination of the different stimuli representations [Full STMF—R^2: Mdn=.86, IQR=.05; scale-rate—R^2: Mdn=.98, IQR=.04; freq-rate—R^2: Mdn=.96, IQR=.05;freq-scale— R^2: Mdn=.92, IQR=.05]. Secondly, in order to understand whether the four representations led to the same linear combination of stimuli, we computed the pairwise Spearman correlations between the regression weights obtained with the four different projections for each dataset. High correlations reflect that the global metric (Full STMF) is decomposed in the same way as the other considered metric. We observed that the weights of the linear combinations explaining the Full STMF metric correlate moderately with the weights of the regressions obtained with the freq-rate and freq-scale representations [Full STMF/freq-rate—ρ^2: Mdn=.59, IQR=.33; Full STMF/freq-scale—ρ^2: Mdn=.55, IQR=.38] (Supplementary Table 19). Conversely, these weights have a lower correlation with those obtained with the scale-rate projections [Full STMF/scale-rate—ρ^2: Mdn=.17, IQR=.31] (Supplementary Table 19). Lastly, the weights obtained with freq-rate and with freq-scale projections correlates moderately with each other [freq-rate/freq-scale—ρ^2: Mdn=.31, IQR=.25] and do not correlate with the weights obtained with the scale-rate projection [scale-rate/freq-rate—ρ^2: Mdn=.08, IQR=.18; scale-rate/freq-scale—ρ^2: Mdn=.05, IQR=.24] (Supplementary Table 19). This analysis reinforces the fact that the metrics fitted for each dataset is principally driven by the information embedded in the range of freq-rate and freq-scale projections in each dataset. Conversely, scale-rate, which is more generic across the different datasets, does not share common properties with freq-rate and freq-scale decompositions.

Discussion

In this paper, we addressed a historical question about musical instrument timbre perception: is the timbre of sounds best represented as a rigid acoustical space, or is it composed of both generic acoustical correlates embedded in the spectral and temporal envelopes and stimulus-set-specific ones embedded in the frequency/scale and frequency/rate patterns of STM representations. Through a computational meta-analysis of 17 experiments, this study showed that the classical

MDS-based approach of correlating positions of stimuli along a given perceptual dimension with univariate acoustic descriptors is only partially replicable for most of the studies, and that it is only adapted to investigate the generic acoustic correlates of timbre. With the distance metric learning method, we determine which information in the STM representations of sounds best simulates human dissimilarity ratings of musical instrument timbres. This study revealed that STM representations are crucial tools, and in particular we uncovered the role of each of their three projections for the perception of musical instrument timbre. This makes a significant step forward in understanding the acoustic correlates of musical instrument timbres by showing how listeners use the variability of the frequency/scale and frequency/rate representations of the stimuli to make their perceptual ratings. More globally, this study supports a generic data-driven procedure to address the relationship between the physical properties of sounds and their perceptual ratings.

Most of the studies in the last 40 years have considered timbre as a fixed perceptual space with dimensions correlating with acoustic descriptors[7,8,10,11,12,14], see [4] for a review. Although it has been efficient for the two first dimensions, this led to a never-ending debate concerning the number of perceptually relevant dimensions of the underlying musical instrument timbre spaces[15,33]. In addition, as shown in the MDS-based meta-analysis, the replicability of these studies is not obvious and is sensitive to the MDS model parameters and acoustic descriptors used by the authors of each study. Nevertheless, the present results confirm the relevance of the temporal and spectral envelopes: the fitted metrics indeed generalize in the scale/rate representation. But here, rather than focusing on a reduced dimensional space, we observed that human dissimilarity ratings are explained by the variability of the STM representations of the stimuli in each experiment, whereas the MDS-based approach excludes significant information by reducing a complex spectrotemporal morphology to a small number of dimensions and scalar descriptors. We argue here that in addition to these main acoustic dimensions, listeners adjust the range of their dissimilarity ratings based on the variability of more subtle sub-representations of the STM representation: the frequency/scale and frequency/rate projections, which can be directly observed by way of the optimization of Gaussian kernels. Practically, this adjustment of the dissimilarity range by the listeners may depend on the specific stimulus set, either during the training preceding each experiment or even over the course of the experiment itself.

The previous acoustic considerations support the use of STM representations and this notably relates to neurophysiological observations. In particular, the neuronal populations of

primary auditory cortex—the spectrotemporal receptive fields—are indeed plastic and driven by attention[34,35,36,37,38]. Brain imaging studies also support the hypothesis that musical timbre is specifically encoded in the human auditory cortex[30,31,39]. Human dissimilarity ratings could be adjusted based on this plasticity and through attentional processes that focus on different acoustic factors, which may depend on selective attention processes. These findings support the hypothesis that relevant STM cues can thus be opportunistically probed by attentional processes according to the listening situation and to the task, here judging the dissimilarity between two sounds. The STM representation thus provides a relevant unbiased acoustic input to more complex modelling of auditory processes as those performed by higher cortical levels. More complex tasks would indeed need more complex models given that the acoustic correlates feeding the Gaussian kernel metrics are not complex enough to reflect the globality of the cortical processes' nonlinear behaviours[2].

Understanding which acoustic information is relevant to perform an auditory task is a crucial issue in auditory cognitive neuroscience that goes well beyond musical instrument perception to speech[22,40,41], environmental sounds[42], and even animal bioacoustics[43]. The method implemented here allows one to determine: (1) whether an acoustic representation is relevant to simulate a given dissimilarity rating perceptual task, and (2) which information is used to reproduce the task. In the present case, STM representations were used. While the gold-standard approach has mainly used combinations of large numbers of audio descriptors to explain auditory tasks[44], the tools clearly lack interpretability regarding which biologically plausible representations are useful to model auditory cognition[45]. Here we implement a framework in which we interpret the weights of a Gaussian kernel on STM representations. The strength of this approach can be generalized with other metrics[46], to other acoustic models, e.g., statistical representations[47,48], and to other parametric models, e.g., speech prosody[41]. The method implemented here thus aims to provide a way to investigate the links between low-level acoustic representations and high-level auditory judgements, which was not easily achieved with the statistical methods used up to this point. It nevertheless captures a significant part of the variance in human dissimilarity ratings and links it to a potential acoustical substrate. This approach has the potential to be applied in many other domains of cognitive neuroscience, in particular to make non-intuitive hypotheses on a possible link between high-dimensional models and global dissimilarity data.

Methods

Ethics. Our study reports an analysis of quantitative experimental data collected from human participants who rated dissimilarity between pairs of sounds. No participants were actually run in this study as the data were provided by the original authors. For details on ethical regulations and editorial board references, see the original papers.

Acoustic representations of sounds. To enhance the speed of processing in the modelling effort, the sound files of the different studies, initially sampled at 44.1 kHz, were first down-sampled to 16 kHz. The auditory representations were computed with an adapted version of the NSL Tools[21]. Cochlear processing is modelled as a bank of 128 constant-Q asymmetric bandpass filters equally spaced on a logarithmic frequency scale spanning 5.3 octaves. Inner hair cell potentials and lateral inhibitory networks are modelled by a high-pass/low-pass filter and spectral sharpening, respectively. The loss in phase locking operating at the midbrain level is performed by using a short-term integration of 4 ms. These computations result in the auditory spectrogram, a two-dimensional time-frequency array. The auditory spectrum is computed by averaging the auditory spectrogram along the temporal dimension. The spectrotemporal modulation model (STM) is computed by applying a spectrotemporal modulation filterbank to the spectrogram. A detailed description is provided by Patil and colleagues[14]. Practically, a two-dimensional Fourier transform is first applied to the spectrogram, which results in a two-dimensional array, also called the Modulation Power Spectrum (MPS)[28,40] whose dimensions are spectral modulation (scale) and temporal modulation (rate). Because of the spectral symmetry of the spectrogram, only positive scale values are kept, whereas both positive and negative rates are kept. Finally, the Spectro-Temporal Modulation transfer Function (STMF) representation is a multiresolution analysis of the spectrogram which is derived by filtering the MPS according to different rates and scales and then coming back to the time-frequency domain. The final STMF representation can be seen as a series of spectrograms filtered according to different rates and scales. Here, we chose the following scale (s) and rate (r) center values as 2D Gaussian filters to generate the STM representation: s = [0.25, 0.35, 0.50, 0.71, 1.00, 1.41, 2.00, 2.83, 4.00, 5.66, 8.00 cyc/oct], r = ± [4.0, 5.7, 8.0, 11.3, 16.0, 22.6, 32.0, 45.3, 64.0, 90.5, 128.0 Hz]. The resulting representation thus corresponds to a 4D matrix with time, frequency, scale, and rate dimensions. This representation is then averaged across time to make the Full STMF representation (frequency, scale, rate). The Full STMF is then

averaged across each of its three dimensions to make the 2D scale/rate, frequency/rate, and frequency/scale representations. Decimated representations were also used with the following parameters: s = [0.25, 0.50, 0.71, 1.00, 1.41, 2.00, 2.83, 4.00, 5.66, 8.00 cyc/oct], r = ± [4.0, 11.3, 22.6, 45.3, 128.0 Hz] and frequency channels = [1, 15, 29, 43, 57, 72, 86, 100, 114, 128].

Stimulus datasets. We investigated 17 published timbre spaces that used different kinds of sounds: natural musical instrument sounds that have been analyzed and then resynthesized with simplifications or systematic modifications (Grey, 1977; Grey & Gordon, 1978; Barthet et al., 2010; Siedenburg et al., 2016, Exp. 2A Sets 2 and 3), imitations and hybrids of musical instruments synthesized with a frequency-modulation algorithm available on a commercial synthesizer (McAdams et al., 1995), and recorded and edited natural sounds (Iverson & Krumhansl, 1993; Lakatos, 2000; Patil et al., 2012; Siedenburg et al., 2016, Exp. 2A Set 1).

MDS-based analysis. For each dataset, eight standard non-metric MDS analyses were performed with Matlab (The MathWorks). The choice of dimensionality onto which the MDS is projected is generally based on the curve representing stress as a function of number of dimensions. Here we chose to use eight values, from 2 to 10 dimensions, which is clearly higher than the average number of dimensions generally sufficient to fit an MDS solution to a perceptual dataset. In addition, this allows us to be less user-specific. For each dataset and each MDS dimension, we then evaluated the Spearman correlation of the first two dimensions with the two most classical descriptors: the logarithm of the attack time (LAT) and the spectral centroid (SC). We finally assessed the correlation between the descriptors and the first two perceptual dimensions by choosing the solution from among the eight with the maximum correlation (ρ^2). As the non-metric MDS is based on ranks we here report the Spearman correlation which provides a non-parametric assessment of correlation between two variables. The LAT was computed as in Patil et al. (2012) by taking the logarithm of the time to increase the temporal envelope from -40 dB to -12 dB relative to the maximum waveform amplitude. The SC was computed as the centroid of the Fast-Fourier Transform averaged over all time frames.

Optimization of Gaussian kernels. For each dataset, we learn a distance metric mimicking human ratings from a STM representation. Distance metric learning is a well-known problem in machine learning[49,50], which aims to learn the coefficients of a distance in order to fit with a given distance. These coefficients can be interpreted as weights on the representations of the stimuli. Here the

learned distance, the kernel, is a radial basis function (see below) between STMFs of two sounds x and y:

$$k(x,y) = \exp\left(-\sum_{i=1}^{N} \frac{|x_i - y_i|^2}{{w_i}^2}\right) \text{ (Eq. 1)}$$

where x and y are the vectorized version of the STMFs defined in the previous section. Practically, each stimulus is initially a three-dimensional array with 22 rates, 11 scales, and 128 frequencies, which is reshaped into a one-dimensional array of size N = 22 x 11 x 128. w is the array of coefficients (weights). From a generic point of view, the distance metric learning here can be interpreted as a way to optimize weights on a Euclidean distance passed through an exponential function. The main interest here is that the relative importance of each feature can be directly observed as masks on the STMFs, which makes the method fully interpretable. More precisely, the learned coefficients w array has the same dimensionality as the STMF and can then be directly observed by reshaping it to the original 22 x 11 x 128 three-dimensional matrix.

The loss function used to fit the kernel (learning weights w) is the correlation between the simulated distances between sounds and the mean dissimilarity ratings. In other words, the objective is to learn the weights w by maximizing the Pearson correlation between the kernel distances k and the human ratings r, which can be formulated as:

$$J = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(k(x_i,x_j) - \underline{\mathbf{k}}\right)\left(d(x_i,x_j) - \underline{\mathbf{d}}\right)}{\sigma_k \sigma_d} \text{ (Eq. 2)}$$

where $\underline{\mathbf{k}}$ and $\underline{\mathbf{d}}$ and (σ_k and σ_d, respectively) are the means (standard deviations) of the representations and human dissimilarity ratings across sounds, respectively. Optimizing the weights w thus allows one to observe how the kernel emphasizes certain dimensions of the STM representation to mimic the human ratings (Figure 4 & Supplementary Figure 2-18).

The optimization of the Gaussian kernel weights, i.e., w_i in Eq. 1, is performed by maximizing the correlation between the human dissimilarities and the kernel distance. In practice we considered the log-kernel, which simplified the computation of the first- and second-order derivatives and stabilized the optimization. The optimization is performed using the limited-memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) algorithm. This is a particularly fast gradient-descent algorithm that is adapted for parameter estimation in machine learning with very large numbers of variables[51].

In this paper, we use the metric optimized on the whole set of sounds within each dataset. However, this method may be prone to overfitting, which we assessed as follows. We tested the metric learning approach using a leave-one-sound-out cross-validation protocol. More precisely, for a dataset with N sounds, N folds were considered. For each fold, the metric was fitted considering N-1 sounds from the dataset, and tested on the remaining sound. Early stopping was used when the testing correlation reach a local minimum and then went up during 200 iterations. Early stopping thus maximizes the cross-validation of the metric within each dataset. We found that the median of the N-1 training correlations is strongly correlated to the testing correlations across datasets ($r^2(15)=.93$: CI95%=[.82;.98]; $p < .001$, power=1). Then, the fitted weights on each fold for a given dataset were correlated to the fitted weights on the whole set of sounds within the same dataset (r^2: Mdn=.92, IQR=.109, Extended Data Figure 4, Supplementary Figure 19). We found that the metrics strongly correlate within the N folds (r^2: Mdn=.85, IQR=.210, Extended Data Figure 4, Supplementary Table 3) showing that the metric learned on all the sounds within a dataset doesn't overfit. We then used this metric for the subsequent analyses.

Statistics

All the correlations reported in the paper are supported by Full Statistics: degrees of freedom (df), 95% Confidence Interval (95%CI) or 95% Confidence Interval range (95%CI range), and statistical power (power). They are either reported in the main article or in the supplementary materials. All the statistical tests were two-tailed. The results can also be replicated with the scripts openly available on: https://github.com/EtienneTho/musical-timbre-studies

The corresponding author can be contacted for further details in order to replicate the results or to use the scripts.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon request and on: https://github.com/EtienneTho/musical-timbre-studies

Code Availability Statement

Custom codes that supports the findings of this study are available from the corresponding author upon request and on: https://github.com/EtienneTho/musical-timbre-studies

References

1. Huang, N., Slaney, M., & Elhilali, M. Connecting deep neural networks to physical, perceptual, and electrophysiological auditory signals. Frontiers in Neuroscience **12**, 532 (2018).
2. Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron **98**, 630-644 (2018).
3. Moore, B. C. An introduction to the psychology of hearing. 6th Edition. (Bingley, UK: Emerald, 2012).
4. Siedenburg, K., & McAdams, S. Four distinctions for the auditory "wastebasket" of timbre. Frontiers in Psychology **8**, 1747 (2017).
5. Plomp, R. Timbre as a multidimensional attribute of complex tones. In R. Plomp & G. F. Smoorenburg (Eds.), Frequency analysis and periodicity detection in hearing (pp. 397-414). Leiden: Sijthoff (1970).
6. Wessel, D. L. Timbre space as a musical control structure. Computer Music Journal **3**, 45-52. (1979).
7. Grey, J. M., & Gordon, J. W. Perceptual effects of spectral modifications on musical timbres. The Journal of the Acoustical Society of America **63**, 1493-1500 (1978).
8. Grey, J. M. Multidimensional perceptual scaling of musical timbres. The Journal of the Acoustical Society of America **61**, 1270-1277 (1977).
9. Krumhansl, C. L. Why is musical timbre so hard to understand? In S. Nielzen & O. Olsson (Eds.), Structure and perception of electroacoustic sound and music (pp. 43-53). Amsterdam, The Netherlands: Excerpta Medica (1989).
10. Iverson, P., & Krumhansl, C. L. Isolating the dynamic attributes of musical timbre. The Journal of the Acoustical Society of America **94**, 2595-2603 (1993).
11. McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. Psychological Research **58**, 177-192 (1995).
12. Lakatos, S. A common perceptual space for harmonic and percussive timbres. Perception & Psychophysics **62**, 1426-1439 (2000).

13. Barthet, M., Guillemain, P., Kronland-Martinet, R., & Ystad, S. From clarinet control to timbre perception. Acta Acustica united with Acustica **96**, 678-689 (2010).
14. Patil, K., Pressnitzer, D., Shamma, S., & Elhilali, M. Music in our ears: the biological bases of musical timbre perception. PLoS Computational Biology **8**, e1002759 (2012).
15. Elliott, T. M., Hamilton, L. S., & Theunissen, F. E. Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. The Journal of the Acoustical Society of America **133**, 389-404 (2013).
16. Siedenburg, K., Jones-Mollerup, K., & McAdams, S. Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds. Frontiers in Psychology **6**, 1977 (2016).
17. Ogg, M., & Slevc, L. R. Acoustic Correlates of Auditory Object and Event Perception: Speakers, Musical Timbres and Environmental Sounds. Frontiers in psychology **10**, 1594 (2019).
18. McAdams, S. The perceptual representation of timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), Timbre: Acoustics, perception, and cognition (pp. 23-57). Cham, Switzerland: Springer International Publishing (2019).
19. Macherey, O., & Delpierre, A. Perception of musical timbre by cochlear implant listeners: a multidimensional scaling study. Ear and Hearing **34**, 426-436 (2013).
20. Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. The timbre toolbox: Extracting audio descriptors from musical signals. The Journal of the Acoustical Society of America **130**, 2902-2916 (2011).
21. Chi, T., Ru, P., & Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America **118**, 887-906 (2005).
22. Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. Science **367**, 1043-1047 (2020).
23. Theunissen, F. E., Sen, K., & Doupe, A. J. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. Journal of Neuroscience **20**, 2315-2331 (2000).
24. Shamma, S. On the role of space and time in auditory processing. Trends in Cognitive Sciences **5**, 340-348 (2001).

25. Chi, T., Gao, Y., Guyton, M. C., Ru, P., & Shamma, S. Spectro-temporal modulation transfer functions and speech intelligibility. The Journal of the Acoustical Society of America, **106**, 2719-2732 (1999).
26. Suied, C., Dremeau, A., Pressnitzer, D., & Daudet, L. Auditory sketches: sparse representations of sounds based on perceptual models. In M. Aramaki et al. (Eds.) International Symposium on Computer Music Modeling and Retrieval 2012, LNCS 7900 (pp. 154-170). Berlin, Heidelberg: Springer. (2013).
27. Isnard, V., Taffou, M., Viaud-Delmon, I., & Suied, C. Auditory sketches: very sparse representations of sounds are still recognizable. PloS one **11**, e0150313 (2016).
28. Thoret, E., Depalle, P., & McAdams, S. Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments. The Journal of the Acoustical Society of America **140**, EL478-EL483 (2016).
29. Thoret, E., Depalle, P., & McAdams, S. Perceptually salient regions of the modulation power spectrum for musical instrument identification. Frontiers in Psychology **8**, 587 (2017).
30. Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. Behavioral and neural correlates of perceived and imagined musical timbre. Neuropsychologia **42**, 1281-1292 (2004).
31. Allen, E. J., Burton, P. C., Olman, C. A., & Oxenham, A. J. Representations of pitch and timbre variation in human auditory cortex. Journal of Neuroscience **37**, 1284-1293 (2017).
32. Ogg, M., Moraczewski, D., Kuchinsky, S. E., & Slevc, L. R. Separable neural representations of sound sources: Speaker identity and musical timbre. Neuroimage **191**, 116-126 (2019).
33. Terasawa, H., Slaney, M., & Berger, J. The thirteen colors of timbre. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005, New Paltz, NY, (pp. 323-326). (2005).
34. Fritz, J., Shamma, S., Elhilali, M., & Klein, D. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nature Neuroscience **6**, 1216-1223 (2003).

35. Kraus, N., Skoe, E., Parbery-Clark, A., & Ashley, R. Experience-induced malleability in neural encoding of pitch, timbre, and timing: Implications for language and music. Annals of the New York Academy of Sciences **1169**, 543-557 (2009).
36. David, S. V., Fritz, J. B., & Shamma, S. A. Task reward structure shapes rapid receptive field plasticity in auditory cortex. Proceedings of the National Academy of Sciences **109**, 2144-2149 (2012).
37. Mesgarani, N., & Chang, E. F. Selective cortical representation of attended speaker in multi-talker speech perception. Nature **485**, 233-236 (2012).
38. Kaya, E. M., & Elhilali, M. Modelling auditory attention. Philosophical Transactions of the Royal Society B: Biological Sciences **372**, 1-10 (2017).
39. Allen, E. J., Moerel, M., Lage-Castellanos, A., De Martino, F., Formisano, E., & Oxenham, A. J. Encoding of natural timbre dimensions in human auditory cortex. Neuroimage **166**, 60-70 (2018).
40. Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., & Poeppel, D. Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. Nature Human Behaviour **3**, 393-405 (2019).
41. Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J. J. Cracking the social code of speech prosody using reverse correlation. Proceedings of the National Academy of Sciences **115**, 3972-3977 (2018).
42. Nelken, I., & De Cheveigne, A. An ear for statistics. Nature Neuroscience **16**, 381 (2013).
43. Bregman, M. R., Patel, A. D., & Gentner, T. Q. Songbirds use spectral shape, not pitch, for sound pattern recognition. Proceedings of the National Academy of Sciences **113**, 1666-1671 (2016).
44. Lartillot, O., Toiviainen, P., & Eerola, T. A matlab toolbox for music information retrieval. In Preisach,C., Burkhardt, H., Schmidt-Thieme, L. and Decker, R. (Eds) Data analysis, machine learning and applications (pp. 261-268). Springer, Berlin, Heidelberg (2008).
45. Aucouturier, J. J., & Bigand, E. Seven problems that keep MIR from attracting the interest of cognition and neuroscience. Journal of Intelligent Information Systems **41**, 483-497 (2013).

46. Bellet, A., Habrard, A., & Sebban, M. A survey on metric learning for feature vectors and structured data. Preprint at arXiv https://arxiv.org/abs/1306.6709 (2013).
47. McDermott, J. H., & Simoncelli, E. P. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron **71**, 926-940 (2011).
48. Anden, J., Lostanlen, V., & Mallat, S. Joint time-frequency scattering. IEEE Transactions on Signal Processing **67**, 3704-3718 (2019).
49. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. Matching networks for one shot learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg and I. Guyon and R. Garnett (Eds.) Advances in neural information processing systems (pp. 3630-3638), Curran Associates, Inc. (2016).
50. Goldberger, J., Hinton, G. E., Roweis, S. T., & Salakhutdinov, R. R. Neighbourhood components analysis. In L. K. Saul and Y. Weiss and L. Bottou (Eds.) Advances in neural information processing systems (pp. 513-520), MIT Press (2005).
51. Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software (TOMS) **23**, 550-560 (1997).

Acknowledgments. This work was supported by the Canadian Natural Sciences and Engineering Research Council awarded to SMc (grants RGPIN-2015-05280 and RGPAS 478121-15) and to PD (RGPIN- 2018-05662), as well as a Canada Research Chair (grants 950-223484 and 950-231872) awarded to SMc. ET was funded through an ILCB/BLRI grant ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and the Excellence Initiative of Aix-Marseille University (A*MIDEX), BC was founded through EU Marie Skłodowska-Curie fellowship (Project MIM, H2020-MSCA-IF-2014, GA no. 659232). BC acknowledges STMS IRCAM-CNRS-Sorbonne Université in Paris where he was a Marie Sklodowska Curie research fellow at the beginning of the project. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the authors of the different studies re-analyzed for providing the stimuli and data from their experiments, Guillaume Mestdagh, Emmanuel Ponsot, and Benjamin Morillon for helpful discussions on earlier versions of the

manuscript, Mounya Elhilali and Daniel Pressnitzer for help in the initial implementation of the optimization framework.

Author Contributions

ET: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. BC: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - Review & Editing. PD: Conceptualization, Methodology, Writing - Review & Editing. SMc: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Competing Interests

The authors declare no competing interests.

Figure Legends

Figure 1. Two different approaches to investigate the auditory perception of musical instrument timbre. Each approach aims to fit a model (green box) to human participants' dissimilarity ratings between pairs of sounds (blue box) on the basis of the acoustic information (pink box). (A) Experiments: listeners make dissimilarity ratings on all pairs of sounds to produce a dissimilarity matrix. (B) Historical approach: dissimilarity ratings are analyzed with multidimensional scaling and the dimensions are correlated with scalar audio descriptors computed from basic acoustic representations. This has several limitations; in particular, the audio descriptors and the MDS parameters are hand-tuned by experimenters leading to difficulties in replicating the findings in the literature. (C) Distance metric learning on the STM representation: this data-driven approach simulates human dissimilarity ratings from the STM representations of sounds, which can then be interpreted. Computation of the STM representation. The waveform is first transformed into a time-frequency representation—an auditory spectrogram—and then into a STM representation embedding frequency, rate (temporal modulations), and scale (spectral modulations). A distance k(x,y) is optimized by learning the weighting coefficients w_i that best mimic the human dissimilarity ratings, providing an interpretable metric revealing the relevant information embedded in the STM representations. For each dataset, the generalizability of the fitted kernel is first cross-validated with the leave-one-sound-out method. The metric is then refitted on the whole dataset and the correlation with the cross-validated set is evaluated.

Figure 2. Replicability of the MDS-based approach. Spearman correlation (ρ^2) with the Logarithm of the Attack Time (LAT) (A) and the Spectral Centroid (SC) (B) in the meta-analysis vs. explained variance in the original study. The computed correlations in the meta-analysis are generally lower than the explained variance provided in the original studies. Eleven studies are considered for the LAT and nine for the spectral centroid corresponding to those available in the original papers (Extended Data Figure 3) – B2010: Barthet et al. (2010); P2012A3: Patil et al. (2012) A3 dataset; P2012GD4: Patil et al. (2012) GD4 dataset; P2012DX4: Patil et al. (2012) DX4 dataset; L2000P: Lakatos (2000) Percussive dataset; L2000H: Lakatos (2000) Harmonic dataset; L2000C: Lakatos (2000) Combined dataset; I1993W: Iverson & Krumhansl (1993) Whole dataset; I1993O: Iverson & Krumhansl (1993) Onset dataset; I1993R: Iverson & Krumhansl (1993) Remainder dataset; McA1995: McAdams et al. (1995).

Figure 3. Generalizability of the metrics learned for the different dataset. Fitted weights are correlated between all pairs of datasets. In the dissimilarity matrices, yellow indicates a perfect correlation (identity on the diagonal) and dark blue a zero correlation. For each representation: (A) Full STMF, (B) scale-rate, (C) freq-rate, and (D) freq-scale, dendrograms with complete linkage from the Spearman correlations were computed. Dendrograms are presented on the top of the correlation matrices and represent the similarity structure of the optimized metrics among the datasets. The fundamental frequency of the stimuli in each dataset are mentioned at the right of each label.

Figure 4. Correspondence between fitted metrics and standard deviations of the stimuli. Projections of optimized metrics (upper panels) and standard deviations between stimuli (lower panels) (from the A3 dataset in Patil et al., 2012) in the three projections of the STM representation, scale/rate (S/R - cyc/oct vs. Hz), frequency/rate (F/R - Hz vs. Hz) and frequency/scale (F/S - Hz vs. cyc/oct). The similarity of these representations is particularly high [$r^2(30,974) = .83$ in the case of the Patil et al., 2012, A3 dataset and $r^2(30,974) = .74$ on average across all the datasets] (Supplementary Figures 2-18 for the other datasets and Supplementary Table 2).

Tables

Table 1. Explained variance (r^2) of the human ratings by the optimized Gaussian kernels.

Study	Dataset name	Degrees of freedom	Auditory Spectrum	Euclidean distance on Full STMF[a]	Full STMF[a]	Scale/ Rate[a]	Freq/ Rate[a]	Freq/ Scale[a]
Grey, 1977	-	118	.47	.17 (.17)	.78 (.61)	.10 (.10)	.46 (.45)	.57 (.48)
Grey & Gordon, 1978	-	118	.11	.00 (.00)	.30 (.15)	.09 (.08)	.12 (.10)	.18 (.09)
Iverson & Krumhansl, 1993	Whole	118	.16	.26 (.26)	.83 (.48)	.35 (.37)	.40 (.24)	.61 (.25)
	Onset	118	.08	.01 (.01)	.21 (.16)	.06 (.08)	.07 (.05)	.13 (.08)
	Remainder	118	.03	.02 (.04)	.27 (.14)	.04 (.03)	.06 (.04)	.10 (.04)
McAdams et al., 1995	-	151	.30	.08 (.13)	.73 (.52)	.20 (.18)	.25 (.22)	.42 (.26)
Lakatos, 2000	Harmonic	134	.19	.06 (.02)	.83 (.48)	.35 (.37)	.40 (.24)	.61 (.25)
	Percussive	151	.18	.02 (.01)	.29 (.23)	.08 (.08)	.19 (.16)	.17 (.15)
	Combined	188	.14	.08 (.09)	.33 (.20)	.00 (.00)	.17 (.15)	.24 (.16)

Barthet et al., 2010	-	103	.74	.57 (.57)	.97 (.92)	.83 (.82)	.87 (.83)	.88 (.85)
Patil et al., 2012	A3	53	.62	.31 (.27)	.93 (.69)	.36 (.25)	.65 (.49)	.76 (.53)
	DX4	53	.69	.25 (.20)	.98 (.80)	.26 (.09)	.76 (.53)	.84 (.60)
	GD4	53	.45	.25 (.23)	.91 (.74)	.49 (.17)	.64 (.51)	.67 (.60)
Siedenburg et al., 2016	Exp 2A Set 1	89	.58	.18 (.21)	.90 (.80)	.20 (.22)	.55 (.49)	.73 (.65)
	Exp 2A Set 2	89	.70	.19 (.22)	.94 (.85)	.35 (.35)	.75 (.72)	.76 (.75)
	Exp 2A Set 3	89	.10	.04 (.05)	.48 (.26)	.19 (.16)	.14 (.10)	.16 (.13)
	Exp 2B (2A Set 3)	89	.07	.04 (.04)	.42 (.21)	.21 (.18)	.11 (.07)	.12 (.10)
Median			.19	.11 (.13)	.78 (.51)	.20 (.16)	.32 (.22)	.55 (.26)

[a]The values in parentheses correspond to the explained variance with the downsampled representations.