RESEARCH

Time–Frequency Scattering Accurately Models Auditory Similarities Between Instrumental Playing Techniques

Vincent Lostanlen[1], Christian El-Hajj[2], Mathias Rossignol[3], Grégoire Lafay[3], Joakim Andén[4,5] and Mathieu Lagrange[2*]

*Correspondence:
mathieu.lagrange@ls2n.fr
[2]LS2N, CNRS, Centrale Nantes,
Nantes University, 1, rue de la
Noe, 44000, Nantes, France
Full list of author information is
available at the end of the article

Abstract

Instrumental playing techniques such as vibratos, glissandos, and trills often denote musical expressivity, both in classical and folk contexts. However, most existing approaches to music similarity retrieval fail to describe timbre beyond the so-called "ordinary" technique, use instrument identity as a proxy for timbre quality, and do not allow for customization to the perceptual idiosyncracies of a new subject. In this article, we ask 31 human subjects to organize 78 isolated notes into a set of timbre clusters. Analyzing their responses suggests that timbre perception operates within a more flexible taxonomy than those provided by instruments or playing techniques alone. In addition, we propose a machine listening model to recover the cluster graph of auditory similarities across instruments, mutes, and techniques. Our model relies on joint time–frequency scattering features to extract spectrotemporal modulations as acoustic features. Furthermore, it minimizes triplet loss in the cluster graph by means of the large-margin nearest neighbor (LMNN) metric learning algorithm. Over a dataset of 9346 isolated notes, we report a state-of-the-art precision at rank five (P@5) of $99.0\% \pm 1$. An ablation study demonstrates that removing either the joint time–frequency scattering transform or the metric learning algorithm noticeably degrades performance.

Keywords: audio databases; continuous wavelet transform; demodulation; distance learning; human–computer interaction; music information retrieval.

Introduction

Music information retrieval (MIR) operates at two levels: symbolic and auditory [1]. By relying on a notation system, the symbolic level allows the comparison of musical notes in terms of quantitative attributes, such as duration, pitch, and intensity at the source. Timbre, in contrast, is a qualitative attribute of music, and is thus not reducible to a one-dimensional axis [2]. As a result, symbolic representations describe timbre indirectly, either via visuotactile metaphors (e.g., bright, rough, and so forth [3]) or via an instrumental playing technique (e.g., bowed or plucked) [17].

  Despite their widespread use, purely linguistic references to timbre fail to convey the intention of the composer. On the one hand, adjectives such as *bright* or *rough* are prone to misunderstanding, as they do not prescribe any musical gesture that is capable of achieving them [5]. On the other hand, the sole mention of a playing technique does not specify its effect in terms of auditory perception. For instance, although the term *breathy* alludes to a playing technique that is peculiar to wind

instruments, a cellist may accomplish a seemingly breathy timbre by bowing near the fingerboard, i.e., *sul tasto* in the classical terminology. Yet, such acoustical similarity at the scale of the entire instrumentarium is not directly reflected by the semantic similarity between playing technique denominations; and discovering it requires musical expertise [6].

Although a notation-based study of playing techniques in music has research potential in music information retrieval [7], the prospect of modeling timbre perception necessarily exceeds the symbolic domain. Instead, it involves a cognitive process which arises from the subjective experience of listening [8]. The simulation of this cognitive process amounts to the design of a multidimensional feature space wherein some distance function evaluates pairs of stimuli. Rather than merely discriminating instruments as mutually exclusive categories, this function must reflect judgments of acoustic dissimilarity, all other parameters—duration, pitch, and intensity—being equal [9].

Behind the overarching challenge of coming up with a robust predictive model for listening behaviors in humans, the main practical application of timbre similarity retrieval lies in the emerging topic of computer-assisted orchestration [10]. In such context, the composer queries the software with an arbitrary audio signal. The outcome is another audio signal which is selected from a database of instrumental samples and perceptually similar to the query. The advantage of this search is that, unlike the query, the retrieved sound is precisely encoded in terms of duration, pitch, intensity, instrument, and playing technique. Thus, following the aesthetic tradition of spectralism in contemporary music creation, the computer serves as a bridge from the auditory level to the symbolic level, i.e., from a potentially infinite realm of timbral sensations to a musical score of predefined range [11].

Goal

This article proposes a machine listening system which computes the dissimilarity in timbre between two audio samples. Crucially, this dissimilarity is not evaluated in terms of acoustic tags, but in terms of *ad hoc* clusters, as defined by a human consensus of auditory judgments. Our system consists of two stages: unsupervised feature extraction and supervised metric learning. The feature extraction stage is a nonlinear map which relies on the joint time–frequency scattering transform [12, 18], followed by per-feature Gaussianization [13]. It encodes patterns of spectrotemporal modulation in the acoustic query while offering numerical guarantees of stability to local deformations [14]. The metric learning stage is a linear map, optimized via large-margin nearest neighbors (LMNN) [15]. It reweights scattering coefficients so that pairwise distances between samples more accurately reflect human judgments on a training set. These human judgments may be sourced from a single subject or the intersubjective consensus of multiple subjects [16].

Figure 1 summarizes our experimental protocol: it illustrates how visual annotations (top) can inform feature extraction (center) to produce a nearest-neighbor search engine which is consistent with human judgments of timbre similarity (bottom).

Contributions

The main contribution of this article can be formulated as the intersection between three topics. To the best of our knowledge, prior literature has addressed these topics separately, but never in combination.

First, our model encompasses a broad range of extended playing techniques, well beyond the so-called "ordinary" mode of acoustic production. Specifically, we fit pairwise judgments for 78 different techniques arising from 16 instruments, some of which include removable timbre-altering devices such as mutes.

Secondly, we purposefully disregard the playing technique metadata underlying each audio sample during the training phase of our model. In other words, we rely on listeners, not performers, to define and evaluate the task at hand.

Thirdly, we supplement our quantitative benchmark with visualizations of time–frequency scattering coefficients in the rate–scale domain for various typical samples of instrumental playing techniques. These visualizations are in line with visualizations of the modulation power spectrum in auditory neurophysiology [19], while offering an accelerated algorithm for scalable feature extraction.

Outline

The related work section reviews recent research in the computational modeling of musical timbre. The data collection section describes our computer–human interface for annotating auditory similarities across playing techniques. The methods section presents the technical components of our machine listening system, namely joint time–frequency scattering and large-margin nearest neighbors. The results section describes our experimental benchmark. The discussion section conducts an ablation study of our best performing system. The conclusion summarizes the implications of our findings for future research in music information retrieval.[1]

Related work

Timbre involves multiple time scales in conjunction, from a few microseconds for an attack transient to several seconds for a sustained tone. Therefore, computational models of timbre perception must summarize acoustic information over a long analysis window, typically amounting to 10^4 digital audio samples or more [20]. Mapping this input to a feature space in which distances denote timbral dissimilarity requires a data-driven stage of dimensionality reduction. In this respect, the scientific literature exhibits a methodological divide as regards the collection of human-annotated data [21]: while the field of MIR mostly encodes timbre under the form of "audio tags", music psychology mostly measures timbre similarity directly from pairwise ML: similarity judgments.

Timbre modeling in music information retrieval (MIR)

On the one hand, most publications in music information retrieval cast timbre modeling as an audio classification problem [22, 23, 24, 25, 26, 27, 28, 29, 30, 31]. In this context, the instrumentation of each musical excerpt serves as an unstructured

---

[1]For the sake of research reproducibility, the source code for the experimental protocol of this paper is available online, alongside anonymized data from human subjects: https://github.com/mathieulagrange/lostanlen2020jasmp

set of "audio tags," encoded as binary outputs within some predefined label space. Because such tags often belong to the metadata of music releases, the process of curating a training set for musical instrument classification requires little or no human intervention. Although scraping user-generated content from online music platforms may not always reflect the true instrumentation with perfect accuracy, it offers a scalable and ecologically valid insight onto the acoustic underpinnings of musical timbre.

Furthermore, supplementing user-generated content with the outcome of a crowd-sourced annotation campaign allows an explicit verification of instrument tags. For instance, the Open-MIC dataset [32], maintained by the Community for Open and Sustainable Music Information Research (COSMIR) [33], comprises a vast corpus of 20k polyphonic music recordings spanning 20 instruments as a derivative of the Free Music Archive (FMA) dataset [34]. Another example is the Medley-solos-DB dataset [35], which comprises 21k monophonic excerpts from eight instruments as a derivative of the MedleyDB dataset of multitrack music [36].

Over the past decade, the availability of large digital audio collections, together with the democratization of high-performance computing on dedicated hardware, has spurred the development of deep learning architectures in music instrument recognition [37, 38, 39]. Notwithstanding the growing accuracy of these architectures in the large-scale data regime, it remains unclear how to extend them from musical instrument recognition to playing technique recognition, where labeled samples are considerably more scarce [40]. We refer to [41] for a recent review of the state of the art in this domain.

Timbre modeling in music cognition research

On the other hand, the field of music psychology investigates timbre with the aim of discovering its physiological and behavioral foundations [42]. In this setting, prior knowledge, however accurate, of instrumentation does not suffice to conduct a study. Rather, each excerpt must be played back to multiple human listeners. Yet, collecting subjective responses to acoustic stimuli is a tedious and unscalable procedure, which restricts the size of the musical corpus under study. These small corpus sizes hampers the applicability of optimization algorithms for representation learning, such as stochastic gradient descent in deep neural networks.

While training artificial neurons is prone to statistical overfitting, advanced methods in electrophysiology allow to observe the firing patterns of biological neurons in the presence of controlled stimuli. This observation, originally carried out on the ferret, has led to a comprehensive mapping of the primary auditory cortex in terms of its spectrotemporal receptive fields (STRFs) [43]. The STRF of a neuron is a function of time and frequency which represents the optimal predictor of its post-stimulus time histogram during exposure to a diverse range of auditory stimuli [44]. The simplest method to compute it in practice is by reverse correlation, i.e. by averaging all stimuli that trigger an action potential [45]. Historically, STRFs were defined by their Wigner–Ville distribution [46], thereby sparing the choice of a tradeoff in time–frequency localization, but eliciting cross-term interferences [47]. Since then, the STRF of a neuron was redefined as a spectrographic representation of its spike-triggered average [48].

Although this new definition is necessarily tied to a choice of spectrogram parameters, it yields more interpretable patterns than a Wigner–Ville distribution. In particular, a substantial portion of spectrographic STRFs exhibit a ripple-like response around a given region (t, λ) of the time–frequency domain [49]. This response can be approximately described by a pair of scalar values: a temporal modulation rate α in Hertz and a frequential modulation rate β in cycles per octave.

Interestingly, both α and β appear to be arranged in a geometric series and independent from the center time t and center frequency λ. This observation has led auditory neuroscientists to formulate an idealized computational model for STRF, known as the "full cortical model" [50], which densely covers the rate–scale domain (α, β) using geometric progressions. Because they do not require a data-driven training procedure, STRF yield a useful form of domain-specific knowledge for downstream machine listening applications, especially when the number of annotated samples is relatively small.

Spectrotemporal receptive fields (STRFs) as a feature extractor

Over recent years, several publications have employed the full cortical model as a feature extractor for a task of musical instrument classification, both in isolated recordings [19] and in solo phrases [51]. These biologically inspired features outperform the state of the art, especially in the small data regime where deep learning is inapplicable. Furthermore, the confusion matrix of the full cortical model in the label space of musical instruments is strongly correlated with the confusion matrix between a human listener and the ground truth. Another appeal of the full cortical model is that the three-way tensor of frequency λ, rate α, and scale β can be segmented into contiguous regions of maximal perceptual relevance for each instrument [52]. This is unlike fully end-to-end learning architectures, whose post hoc interpretability requires advanced techniques for feature inversion [53]. Lastly, beyond the realm of supervised classification, a previous publication [54] has shown that query-by-example search with STRFs allows to discriminate categories of environmental soundscapes, even after temporal integration and unsupervised dimensionality reduction.

The reasons above make STRFs an appealing feature extractor for a perceptual description of timbral similarity across instrumental playing techniques. Nonetheless, current implementations of STRF suffer from a lack of scalability, which explains why they have found few applications in MIR thus far. Indeed, the full cortical model is usually computed via two-dimensional Fourier transforms over adjacent time–frequency regions, followed by averaging around specific rates and scales. This approach requires a uniform discretization of the scalogram, and thus an oversampling of the lower-frequency subbands to the Nyquist frequency of the original signal. In contrast, joint time–frequency scattering offers a faster extraction of spectrotemporal modulations while preserving properties of differentiability [55] and invertibility [56]. Such acceleration is made possible by discretizing the wavelet transforms involved in time–frequency scattering according to a multirate scheme, both along the time and the log-frequency variables [18]. In this multirate scheme, every subband is discretized at its critical sample rate, i.e., in proportion to its center frequency. As a by-product, the multirate approach draws an explicit

connection between scattering networks and deep convolutional networks, because both involves numerous convolutions with small kernels, pointwise rectifying non-linearities, and pooling operations [57].

Moreover, a quantitative benchmark over Medley-solos-DB has demonstrated that joint time–frequency scattering, unlike purely temporal scattering, outperforms deep convolutional networks in supervised musical instrument classification, even in a relatively large data regime with 500 to 5k samples per class [18]. However, it remains to be seen whether joint time–frequency scattering is capable of fine-grained auditory categorization, involving variability in instrument, mute, and playing technique. In addition, previous publications on joint time–frequency scattering lack a human-centric evaluation, independently from any classification task.

Claim of originality

The contributions of this paper strive to fill the gap in scholarship between MIR and music cognition approaches to timbre, in the context of computer-assisted spectralist orchestration with extended playing techniques. From the standpoint of MIR, the model presented here offers an efficient and generic multidimensional representation for timbre similarity, alongside theoretical guarantees of robustness to elastic deformations in the time–frequency domain. Conversely, from the standpoint of music cognition, our model offers a scalable and biologically plausible surrogate for stimulus-based collection of acoustic dissimilarity judgments, which is also readily tailored to subjective preferences.

Data collection

The European symphonic orchestra encompasses four families of instruments: strings, woodwinds, brass, and percussion. In this article, we focus on the first three, and leave the question of learning auditory similarities between percussion instruments to future research. We refer to [58] and [59] for reviews of the recent literature on the timbre modeling of percussive instruments, from the standpoints of MIR and music cognition, respectively.

Dataset

We consider a list of 16 instruments: violin (Vn), viola (Va), cello (Vc), contrabass (Cb), concert harp (Hp), Spanish guitar (Gtr), accordion (Acc), flute (Fl), soprano clarinet (BbCl), alto saxophone (ASax), oboe (Ob), bassoon (Bn), trumpet in C (TpC), French horn (Hn), tenor trombone (TTbn), and bass tuba (BBTb). Among this list, the first six are strings, the next six are woodwind, and the last four are brass. Some of these instruments may be temporarily equipped with timbre-altering mutes, such as a rubber sordina on the bridge of a violin or an aluminium "wah-wah", also known as *harmon*, inside in the bell of a trumpet. Once augmented with mutes, the list of 16 instruments grows to 33. Furthermore, every instrument, whether equipped with a mute or not, affords a panel of playing techniques ranging in size between 11 (for the accordion) and 41 (for the bass tuba). In the rest of this paper, we abbreviate instrument–mute–technique by means of the acronym "IMT". One example of IMT is `TpC+S-ord`, i.e., trumpet in C with a straight mute played in the ordinary technique. Another example of IMT is `Vn-pont`, i.e., violin without any mute played in the *sul ponticello* technique (bowing near the bridge).

Performers can play each IMT at various pitches according to the tessitura of their instrument. This tessitura may depend on the choice of playing technique but is independent of the choice of mute. Among the 16 instruments in this study, the two instruments with widest and narrowest tessituras, in their respective ordinary techniques, are the accordion (81 semitones) and the trumpet in C (32 semitones) respectively. Lastly, each IMT may be played at up to five intensity dynamics, ranging from quietest to loudest as: pianissimo (*pp*), piano (*p*), mezzo forte (*mf*), forte (*f*), and fortissimo (*ff*). The resort to a non-ordinary playing technique may restrict both the tessitura and the dynamics range of the instrument–mute pair under consideration. For example, the pitch of pedal tones in brass instruments is tied to the fundamental mode of the bore, i.e., usually B♭ or F. Likewise, the intensity of key clicks in the oboe is necessarily *pp*, while the intensity of snap pizzicato *à la Bartók* in plucked strings is necessarily *ff*.

In summary, audio signals from isolated musical notes may vary across three categorical variables (instrument, mute, and technique) and two quantitative variables (intensity and pitch). The Studio On Line dataset (SOL), recorded at Ircam in 1998, offers a joint sampling of these variables. The version of SOL that we use throughout this paper, named "0.9 HQ", amounts to a total of 25444 audio signals. Beyond playing techniques, we should note that SOL erases other factors of acoustic variability, such as identity of performer, identity of instrument manufacturer, audio acquisition equipment, and room response characteristics, which are all restricted to singletons. Addressing these factors of variability is beyond the scope of this paper, which focuses on the influence of playing technique. Despite this restriction, the SOL dataset remains impractically large for collecting human similarity judgments. Our protocol addresses this problem by means of three complementary approaches: disentanglement of factors, expert pre-screening, and the use of an efficient annotation interface.

Disentanglement of factors

First, we purposefully disentangle categorical variables (IMTs) from continuous variables (pitch and intensity) in the SOL dataset. Indeed, under first approximation, the perception of timbre is invariant to pitch and intensity. Therefore, we select auditory stimuli according to a reference pitch and a reference intensity; in our case, middle C (C_4) and *mf*. After this selection, every IMT triplet contains a single acoustic exemplar, regarded as canonical in the following. The number of canonical stimuli for the entire SOL dataset is equal to 235. We should note, however, that the proposed pitch and intensity cannot be strictly enforced across all IMTs. Indeed, as explained above, a fraction of IMTs can only be achieved at restricted values of pitch and intensity parameters, e.g., pedal tones or key clicks. Therefore, at a small cost of consistency, we only enforce the pitch–intensity reference (i.e., C_4 and *mf*) when practically feasible, and fall back to other pitches and intensities if necessary.

Expert pre-screening

Secondly, we reduce the number of IMTs in our study by focusing on those which are deemed to be most relevant. Here, we define the relevance of an IMT as the possibility of imitating it by means of another IMT from a different instrument.

One example of such imitation is the acoustic similarity between slap tonguing in reed instruments and a snap pizzicato in string instruments. To collect perceptual ratings of relevance, we recruited two professors in music composition at the Paris Conservatory (CNSMDP[2]). Each of them inspected the entire corpus of 235 IMTs and annotated them in terms of relevance according to a Likert scale with seven ticks. In this Likert scale, the value 1 (least relevant) denotes that the IMT under consideration has a timbre that is idiosyncratic, and that therefore, it is unlikely that humans will pair it with other IMTs. Conversely, the value 7 (most relevant) denotes that the IMT under consideration bears a strong similarity with some other IMT from the corpus.

Once both experts completed their annotations, we retained all IMTs whose average score was judged equal to 3 or higher, thus resulting in a shortlist of $N = 78$ IMTs (see Tables 1 and 2 in the Appendix). It is worth noting that, according to both experts, the timbre of the accordion was judged too idiosyncratic to be relevant for this experiment, regardless of playing technique. Indeed, the accordion is the only instrument in the aforementioned list of instrument to have free reeds, keyboard-based actuation, or handheld airflow. Consequently, regardless of mute and technique, the set of instruments $\mathcal{I}$ in our study contains 15 elements.

Efficient annotation interface

Thirdly, we design a graphical user interface for partitioning a corpus of short audio samples. The need for such an interface arises from the unscalability of Likert scales in the context of pairwise similarity judgments. Assuming that similarity is a symmetric quantity, collecting a dense matrix of continuously valued ratings of similarity among a dataset of N items would require $\frac{1}{2}(N^2 - N)$ Likert scales. In the case of $N = 78$ IMTs, the task would amount to about 3k horizontal sliders, i.e., several hours of cumbersome work for the human annotator.

Engaging as many participants as possible in our study called for a more streamlined form of human–computer interaction, even if it sacrificed the availability of quantitative ratings. To this end, we implemented a web application, named Cyberlioz, in which the user can spontaneously listen and arrange sounds into clusters of timbre similarity.[3] The name Cyberlioz is a portmanteau between the prefix cyber- and the French composer Hector Berlioz. The choice is by no means coincidental: Berlioz is famous for having, in his *Treatise on Orchestration* (1844), shed a particular focus on the role of timbre as a parameter for musical expression.

Cyberlioz consists of a square panel on which is displayed a collection of circular grey dots, each of them corresponding to one of the IMTs, and initially distributed uniformly at random. Hovering the screen pointer onto each dot results in a playback of a representative audio sample of this IMT, i.e., C_4 and *mf* in most cases. Furthermore, each dot can be freely placed on the screen by clicking, dragging, and dropping. Lastly, the user can assign a color to each dot among a palette of 20 hues.

---

[2]CNSMDP: Conservatoire National Supérieur de Musique et de Danse de Paris. Official website: https://www.conservatoiredeparis.fr

[3]The Web application for efficient audio annotation, as well as the raw anonymized responses of all 31 participants to our study, is available at: https://soundthings.org/research/cyberlioz/

The goal of the Cyberlioz interface is to form clusters of timbre similarity between IMTs, expressed by sameness of color.

In comparison with a traditional web-based form, Cyberlioz offers a more intuitive and playful user experience, while limiting the acquisition of similarity judgments to a moderate duration of 30 to 60 minutes for each participant. In May and June 2016, we recruited volunteers to use Cyberlioz on their own computers, via a web browser, and equipped with a pair of earphones. To this end, we publicized this study on the internal mailing list of students at CNSMDP, as well as two international mailing lists for research in music audio processing: AUDITORY and ISMIR Community.[4] Within two months, $K = 31$ subjects accessed Cyberlioz and completed the task.

Hypergraph partitioning

Once the data collection campaign was complete, we analyzed the color assignments of each subject k and converted them into a cluster graph $\mathcal{G}_k$, where the integer k is an anonymized subject index, ranging between 1 and K. For a given k, the graph $\mathcal{G}_k$ contains N vertices, each representing a different IMT in the corpus. In $\mathcal{G}_k$, an edge connects any two vertices m and n if the corresponding dots in Cyberlioz have the same color. Otherwise, there is no edge connecting m and n. Thus, $\mathcal{G}_k$ contains as many connected components as the number of similarity clusters for the subject k, i.e., the number of distinct colors on the Cyberlioz interface in the response of k.

We aggregate the similarity judgments from all K subjects by embedding them into a hypergraph $\mathcal{H}$, that is, a graph whose edges may connect three or more vertices at once. Specifically, $\mathcal{H}$ contains N vertices, each representing an IMT; and each "hyperedge" in $\mathcal{H}$ corresponds to some connected component in one of the graphs $\mathcal{G}_1, \ldots, \mathcal{G}_K$. Then, we convert the hypergraph $\mathcal{H}$ back into a conventional graph $\mathcal{G}_0$ by means of a combinatorial optimization algorithm known as hypergraph partitioning [60].

For a particular subject k, let us denote by C_k the number of clusters in the graph $\mathcal{G}_k$. Across all $K = 31$ subjects, C_k varies between 3 and 19, with a median value of 10. To construct $\mathcal{G}_0$, we select a number of clusters that is equal to the maximal value of the C_k's; that is, $C_0 = 19$. Then, we run hypergraph partitioning on $\mathcal{H}$ to assign each vertex i to one of the C_0 clusters in $\mathcal{G}_0$. Intuitively, hypergraph partitioning optimizes a tradeoff between two objectives: first, balancing the size of all clusters in terms of their respective numbers of vertices; and secondly, keeping most hyperedges enclosed within as few distinct clusters as possible [61, 62].

While the graphs $\mathcal{G}_1, \ldots, \mathcal{G}_K$ encode the subjective similarity judgments of participants 1 to K, the graph $\mathcal{G}_0$ represents a form of consensual judgment that is shared across all participants while discarding intersubjective variability. Although the rest of our paper focuses on the consensus $\mathcal{G}_0$, it is worth pointing out that the same technical framework could apply to a single subject k, or to a subgroup of the $K = 31$ subjects. This remark emphasizes the potential of our similarity learning method as a customizable tool for visualizing and extrapolating the timbre similarity space of a new subject.

---

[4]For more information about these mailing lists, please visit: http://www.auditory.org/ and http://ismir.net/

Inter-instrument similarity

Before addressing the question of modeling timbre in the audio domain, we conduct an exploratory study on the structure of similarity judgments themselves. Specifically, given a corpus of IMT samples, we ask whether human listeners tend to cluster these samples by instrument. To answer this question, we derive from the consensus clustering graph $\mathcal{G}_0$ an instrument-wise similarity matrix $\mathbf{A}_{\mathcal{I}}$ whose rows and columns are defined on the finite set $\mathcal{I}$ of instruments. For every instrument–instrument pair (i, j), we set the value of $\mathbf{A}_{\mathcal{I}}$ at row i and column j equal to the number of edges in $\mathcal{G}_0$ connecting IMTs from instruments i and j, renormalized by the number of IMT samples from either instrument i or instrument j.

Let us denote the existence of an edge in $\mathcal{G}_0$ connecting two IMT samples m and n by the binary relation $m \overset{\mathcal{G}_0}{\sim} n$. The mathematical definition of the matrix $\mathbf{A}_{\mathcal{I}}$ is:

$$\mathbf{A}_{\mathcal{I}}(i, j) = \frac{\operatorname{card}\left\{(m, n) \mid \text{Instrument}(m) = i; \text{Instrument}(n) = j; m \overset{\mathcal{G}_0}{\sim} n\right\}}{\operatorname{card}\left\{n \mid \text{Instrument}(n) \in \{i, j\}\right\}}. \quad (1)$$

Then, we run Ward's method [?] to cluster instruments in $\mathcal{I}$ according to the similarity matrix $\mathbf{A}_{\mathcal{I}}$. This method yields a permutation $\sigma_{\mathcal{I}} : \mathcal{I} \to \mathcal{I}$ of the rows and columns in $\mathbf{A}_{\mathcal{I}}$ such that the distance $|\sigma_{\mathcal{I}}(i) - \sigma_{\mathcal{I}}(j)|$ after permutation is small if and only if the similarity $\mathbf{A}_{\mathcal{I}}(i, j)$ is large. Figure 2 displays the rearranged similarity matrix $\widetilde{\mathbf{A}}_{\mathcal{I}} : (i, j) \in \mathcal{I}^2 \mapsto \mathbf{A}_{\mathcal{I}}(\sigma_{\mathcal{I}}^{-1}(i), \sigma_{\mathcal{I}}^{-1}(j))$ as a result of such agglomerative hierarchical clustering procedure.

Interestingly, the matrix $\widetilde{\mathbf{A}}_{\mathcal{I}}$ exhibits a block diagonal structure, which roughly reflects the classical taxonomy of musical instruments: four woodwinds (BbCl, Fl, ASax, Ob) are clustered together, followed by a cluster containing one woodwind (Bn) and four brass (BBTb, Hn, TTbn, and TpC), followed by a cluster of all strings (Cb, Va, Vn, Vc, Gtr, Hp). Furthermore, the largest cross-instrument similarity arises for the only pair of purely plucked instruments, namely, harp and guitar.

Inter-technique similarity

From the consensus clustering graph $\mathcal{G}_0$, we also derive a technique-wise similarity matrix $\mathbf{A}_{\mathcal{T}}$ whose rows and columns are defined on the finite set $\mathcal{T}$ of playing techniques. As for $\mathbf{A}_{\mathcal{I}}$, we define the similarity between any two playing techniques u and v in $\mathcal{T}$ as the following ratio of set cardinalities:

$$\mathbf{A}_{\mathcal{T}}(u, v) = \frac{\operatorname{card}\left\{(m, n) \mid \text{Technique}(m) = u; \text{Technique}(n) = v; m \overset{\mathcal{G}_0}{\sim} n\right\}}{\operatorname{card}\left\{n \mid \text{Technique}(n) \in \{u, v\}\right\}}. \quad (2)$$

Again, we run Ward's method to cluster playing techniques in $\mathcal{T}$ according to their similarity $\mathbf{A}_{\mathcal{T}}$. Figure 3 displays the rearranged similarity matrix $\widetilde{\mathbf{A}}_{\mathcal{T}} : (u, v) \in \mathcal{T}^2 \mapsto \mathbf{A}_{\mathcal{T}}(\sigma_{\mathcal{T}}^{-1}(u), \sigma_{\mathcal{T}}^{-1}(v))$.

The block diagonal structure of the matrix $\widetilde{\mathbf{A}}_{\mathcal{T}}$ is less salient than in $\widetilde{\mathbf{A}}_{\mathcal{I}}$. Nevertheless, these clusters correspond to some basic attributes of qualitative timbre: by and large, the top-left region of the matrix contains sustained sounds whereas the bottom-right region contains percussive sounds. Interestingly, the ordinary technique (`ord`) appears at the center of the matrix, i.e., at the intersection between

sustained sounds and percussive sounds. This is because the notion of "ordinariness" does not prescribe the same gesture for all instruments: e.g., an ordinary guitar sound is expected to be percussive whereas an ordinary flute sound is expected to be sustained.

Methods

The previous section described our protocol for collecting timbral similarity judgments between instrumental playing techniques. In this section, we aim to recover these similarity judgments from digital audio recordings according to a paradigm of supervised metric learning. To this end, we present a machine listening system composing joint time–frequency scattering and large-margin nearest neighbors (LMNN).

Joint time–frequency scattering transform

Let $\boldsymbol{\psi} \in \mathbf{L}^2(\mathbb{R}, \mathbb{C})$ a complex-valued filter with zero average, dimensionless center frequency equal to one, and an equivalent rectangular bandwidth (ERB) equal to $1/Q$. We define a constant-Q wavelet filterbank as the family $\boldsymbol{\psi}_\lambda : t \mapsto \lambda \boldsymbol{\psi}(\lambda t)$. Each wavelet $\boldsymbol{\psi}_\lambda$ has a center frequency of λ, an ERB of λ/Q, and an effective receptive field of $(2\pi Q/\lambda)$ in the time domain. In practice, we define $\boldsymbol{\psi}$ as a Morlet wavelet:

$$\boldsymbol{\psi} : t \longmapsto \exp\left(-\frac{t^2}{2\sigma_\psi^2}\right)\left(\exp\left(2\pi \mathrm{i} t\right) - \kappa_\psi\right), \tag{3}$$

where the Gaussian width σ_ψ grows in proportion with the quality factor Q and the corrective term κ_ψ ensures that $\boldsymbol{\psi}$ has a zero average. Moreover, we discretize the frequency variable λ according to a geometric progression of common ratio $2^{\frac{1}{Q}}$. Thus, the base-two logarithm of center frequency, denoted by $\log_2 \lambda$, follows an arithmetic progression. We set the constant quality factor of the wavelet filterbank $(\boldsymbol{\psi}_\lambda)_\lambda$ to $Q = 12$, thus matching twelve-tone equal temperament in music.

Convolving the wavelets in this filterbank with an input waveform $\boldsymbol{x} \in \mathbf{L}^2(\mathbb{R})$, followed by an application of the pointwise complex modulus yields the wavelet scalogram

$$\mathbf{U_1}\boldsymbol{x}(t, \lambda) = \left|\boldsymbol{x} * \boldsymbol{\psi}_\lambda\right|(t) = \left|\int_{\mathbb{R}} \boldsymbol{x}(t - t')\, \boldsymbol{\psi}_\lambda(t')\, \mathrm{d}t'\right|, \tag{4}$$

which is discretized similarly to the constant-Q transform of [63]. Then, we define a two-dimensional Morlet wavelet $\boldsymbol{\Psi}$ of the form

$$\Psi : (t, u) \longmapsto \exp\left(-\frac{t^2 + u^2}{2\sigma_\Psi^2}\right)\left(\exp\left(2\pi \mathrm{i}(t + u)\right) - \kappa_\Psi\right), \tag{5}$$

taking two real variables t and u as input. In the rest of this paper, we shall refer to $\boldsymbol{\Psi}$ as a time–frequency wavelet. The former is the time variable while the latter is the base-two logarithm of frequency: $u = \log_2 \lambda$. Note that u roughly corresponds to the human perception of relative pitch [64].

We choose the Gaussian width $\sigma_{\boldsymbol{\Psi}}$ in Equation 5 such that the quality factor of the Morlet wavelet $\boldsymbol{\Psi}$ is equal to one, both over the time dimension and over the log-frequency dimension. Furthermore, the corrective term $\kappa_{\boldsymbol{\Psi}}$ ensures that $\boldsymbol{\Psi}$ has a zero average over $\mathbb{R}^2$, similarly to Equation 3. From $\boldsymbol{\Psi}$, we define a two-dimensional wavelet filterbank of the form:

$$\boldsymbol{\Psi}_{\alpha,\beta} : (t, u) \longmapsto \alpha\,\beta\,\boldsymbol{\Psi}(\alpha t, \beta u). \tag{6}$$

In the equation above, α is a temporal modulation rate and β is a temporal modulation scale, following the terminology of spectrotemporal receptive fields (STRF, see related work section). While α is measured in Hertz and is strictly positive, β is measured in cycles per octaves and may take positive as well as negative values. Both α and β are discretized by geometric progressions of common ratio equal to two. Furthermore, the edge case $\beta = 0$ corresponds to $\boldsymbol{\Psi}_{\alpha,\beta}$ being a Gaussian low-pass filter over the log-frequency dimension, while remaining a Morlet band-pass filter of center frequency α over the time dimension. We denote this low-pass filter by $\boldsymbol{\phi}_F$ and set its width to $F = 2$ octaves.

We now convolve the scalogram $\mathbf{U_1}$ with time–frequency wavelets $\boldsymbol{\Psi}_{\alpha,\beta}$ and apply the complex modulus, yielding the four-way tensor

$$
\begin{aligned}
\mathbf{U_2}\boldsymbol{x}(t, \lambda, \alpha, \beta) &= \big|\mathbf{U_1}\boldsymbol{x} \circledast \boldsymbol{\Psi}_{\alpha,\beta}\big|(t, \lambda) \\
&= \left|\iint_{\mathbb{R}^2} \mathbf{U_1}\boldsymbol{x}(t - t', \log \lambda - u')\,\boldsymbol{\Psi}_{\alpha,\beta}(t', u')\,\mathrm{d}t'\,\mathrm{d}u'\right|,
\end{aligned} \tag{7}
$$

where the circled asterisk operator $\circledast$ denotes a joint convolution over time and log-frequency. In the equation above, the sample rate of time t is proportional to α. Conversely, the sample rate of log-frequency $u = \log_2 \lambda$ is proportional to $|\beta|$ if $\beta \neq 0$ and proportional to F^{-1} otherwise.

Let $\boldsymbol{\phi}_T$ be a Gaussian low-pass filter. We define the joint time–frequency scattering coefficients of the signal $\boldsymbol{x}$ as the four-way tensor

$$
\begin{aligned}
\mathbf{S_2}(t, \lambda, \alpha, \beta) &= \mathbf{U_2}\boldsymbol{x} \circledast \big(\boldsymbol{\phi}_T \otimes \boldsymbol{\phi}_F\big)(t, \lambda) \\
&= \iint_{\mathbb{R}^2} \mathbf{U_1}\boldsymbol{x}(t - t', \log \lambda - u', \alpha, \beta)\,\boldsymbol{\phi}_T(t')\,\boldsymbol{\phi}_F(u')\,\mathrm{d}t'\,\mathrm{d}u',
\end{aligned} \tag{8}
$$

where the symbol $\otimes$ denotes the outer product over time and log-frequency. In the equation above, the sample rate of time t is proportional to T^{-1} and the sample rate of log-frequency $u = \log_2 \lambda$ is proportional to F^{-1}. Furthermore, the rate α spans along a geometric progression ranging from T^{-1} to λ/Q. In the following, we set the time constant to $T = 1000\,\mathrm{ms}$ unless specified otherwise.

The tensor $\mathbf{S_2}$ bears a strong resemblance with the idealized response of an STRF at the rate α and the scale β. Nevertheless, in comparison with the "full cortical model" [19], joint time–frequency scattering enjoys a thirty-fold reduction in dimensionality while covering a time span that is four times larger (1000 ms) and an acoustic bandwidth that is also four times larger (0–16 kHz). This is due to the multirate discretization scheme applied throughout the application of wavelet convolutions and pointwise modulus nonlinearities.

In addition to second-order scattering coefficients (Equation 8), we compute joint time–frequency scattering at the first order by convolving the scalogram $\mathbf{U_1}\boldsymbol{x}$ (Equation 4) with the low-pass filter $\boldsymbol{\phi}_T$ over the time dimension with wavelets $\boldsymbol{\psi}_\beta$ ($\beta \geq 0$) over the log-frequency dimension, and by applying the complex modulus:

$$
\begin{aligned}
\mathbf{S_1}\boldsymbol{x}(t, \lambda, \alpha = 0, \beta) &= \big|\mathbf{U_1}\boldsymbol{x} \circledast \big(\boldsymbol{\phi}_T \otimes \boldsymbol{\psi}_\beta\big)\big|(t, \lambda) \\
&= \left|\iint_{\mathbb{R}^2} \mathbf{U_1}\boldsymbol{x}(t - t', \log \lambda - u')\,\boldsymbol{\phi}_T(t')\,\boldsymbol{\psi}_\beta(u')\,\mathrm{d}t'\,\mathrm{d}u'\right|, \quad (9)
\end{aligned}
$$

where the time constant T is the same as in Equation 8, i.e., $T = 1000\,\mathrm{ms}$ by default.

Over the time variable, we set the modulation rate α of $\mathbf{S_1}$ to zero in the equation above. Conversely, over the log-frequency variable, the edge case $\beta = 0$ corresponds to replacing the wavelet $\boldsymbol{\psi}_\beta$ by the low-pass filter $\boldsymbol{\phi}_F$. We refer to [18] for more details on the implementation of joint time–frequency scattering.

We adopt the multi-index notation $p = (\lambda, \alpha, \beta)$ as a shorthand for the tuple of frequency, rate, and scale. The tuple p is known as a scattering path (see [65]), and may apply to index both first-order ($\mathbf{S_1}$) and second-order ($\mathbf{S_2}$) coefficients. Given an input waveform $\boldsymbol{x}$, we denote by $\mathbf{S}\boldsymbol{x}$ the feature vector resulting from the concatenation of $\mathbf{S_1}\boldsymbol{x}$ and $\mathbf{S_2}\boldsymbol{x}$:

$$
\mathbf{S}\boldsymbol{x}\big(t, p = (\lambda, \alpha, \beta)\big) = \left\{ \begin{array}{ll} \mathbf{S_1}\boldsymbol{x}(t, \lambda, \alpha, \beta) & \text{if } \alpha = 0, \\ \mathbf{S_2}\boldsymbol{x}(t, \lambda, \alpha, \beta) & \text{otherwise.} \end{array} \right. \tag{10}
$$

Visualization

At the second order, joint time–frequency scattering coefficients depend upon four variables: time t, log-frequency λ, temporal modulation rate α in Hertz, and frequential modulation rate β in cycles per octave. From a data visualization standpoint, rendering the four-dimensional tensor $\mathbf{S_2}\boldsymbol{x}(t, \lambda, \alpha, \beta)$ is impossible. To address this limitation, a recent publication has projected this tensor into a two-dimensional "slice", thus yielding an image raster [18]. In accordance with their protocol, we compute the following matrix:

$$
\mathbf{V}\boldsymbol{x}(\alpha, \beta) = \iint_{\mathbb{R}^2} \mathbf{U_2}\boldsymbol{x}(t - t', \log \lambda - u')\,\mathrm{d}t'\,\mathrm{d}u', \tag{11}
$$

Observe that the equation above is a limit case of $\mathbf{S_2}$ (Equation 8) in which the constants T and F tend towards infinity. As a result, $\mathbf{V}\boldsymbol{x}$ depends solely upon scale α and rate β. In the scientific literature on spectrotemporal receptive fields (STRF), the matrix $\mathbf{V}\boldsymbol{x}$ is called "cortical output collapsed on the rate-scale axes" [67].

Previous publications on STRFs have demonstrated the interest of visualizing the slice $\mathbf{V}\boldsymbol{x}$ in the case of speech [68], lung sounds [69], and music [19]. However, the visualization of musical sounds has been restricted to some of most common playing techniques, i.e., piano played staccato and violin played pizzicato. Furthermore, prior publications on time–frequency scattering have displayed slices of $\mathbf{S_2}\boldsymbol{x}$ in the case of synthetic signals; but there is a gap in literature as regards the interpretability of the scale–rate domain in the case of real-world signals.

To remedy this gap, we select twelve isolated notes from the SOL dataset from two instruments: violin (Figure 4) and flute (Figure 5). By and large, we find that joint time–frequency scattering produces comparable patterns in the scale–rate domain as the "cortical output" of the STRF. For example, Figure 4 (a) shows that a violin note played in the *ordinario* technique has a local energy maximum at the rate $\alpha = 6\,\mathrm{Hz}$. A visual inspection of $\mathbf{U_1}$ demonstrates that this rate coincides with the rate of vibrato of the left hand in the violin note. As seen in Figure 4 (b), this local energy maximum is absent when the playing technique is denoted as *nonvibrato*. Furthermore, Figure 4 (c) shows that the local energy maximum is displaced to a higher rate ($\alpha = 12\,\mathrm{Hz}$) when the vibrato is replaced by a tremolo.

The visual interpretation of playing techniques in terms of their joint time–frequency scattering coefficients is not restricted to periodic modulation, such as vibrato or tremolo: rather, it also encompasses the analysis of attack transients. Figures 4 (d), (e), and (f) show the matrix $\mathbf{V}\boldsymbol{x}$ for three instances of impulsive violin sounds: sforzando, pizzicato, and staccato respectively. These three techniques create ridges in the scale–rate domain (α, β), where the cutoff rate α is lowest with sforzando and highest with staccato. These variations in cutoff rate coincide with perceptual variations in "hardness", i.e. impulsivity, of the violin sound. Moreover, in the case of staccato, we observe a slight asymmetry in the frequential scale parameter β. This asymmetry could be due to the fact that higher-order harmonics decay faster than the fundamental, thus yielding a triangular shape in the time–frequency domain.

Figure 5 shows six playing techniques of the flute. Similarly to the violin (Figure 4), we observe that periodic modulations, such as a trill (Figure 5 b) or a beating tone (c), cause local energy maxima whose rate α is physically interpretable. Likewise, impulsive flute sounds such as sforzando (Figure 5 d), key click (e), and vibrato (f) create ridges in the scale–rate domain of varying cutoff rates α. We distribute the implementation of these figures as part of the MATLAB library *scattering.m*, which is released under the MIT license[5].

Median-based logarithmic compression and affine standardization

Now, we apply a pointwise nonlinear transformation on averaged joint time-frequency scattering coefficients. The role of this transformation, which is adapted to the dataset in an unsupervised way, is to Gaussianize the histogram of amplitudes of each scattering path p. We consider a collection $\mathcal{X}$ of N waveforms $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. For every path p in the joint time–frequency scattering transform operator $\mathbf{S}$, we average the response of each scattering coefficient $\mathbf{S}\boldsymbol{x}_n$ over time and take its median value across all samples n from 1 to N:

$$\boldsymbol{\mu}(p) = \underset{1 \leq n \leq N}{\mathrm{median}} \int_{\mathbb{R}} \mathbf{S}\boldsymbol{x}_n(t, p)\,\mathrm{d}t. \tag{12}$$

If the collection is split between a training set and a test set (see discussion section), we compute $\boldsymbol{\mu}$ on the training set only. Then, to match a decibel-like perception

---

[5]Link to download the *scattering.m* library: https://github.com/lostanlen/scattering.m

of loudness, we apply the following adaptive transformation, which composes a median-based renormalization and a logarithmic compression:

$$\widetilde{\mathbf{S}}\boldsymbol{x}_n(p) = \log\left(1 + \frac{\int_{\mathbb{R}} \mathbf{S}\boldsymbol{x}_n(t,p)\,\mathrm{d}t}{\varepsilon\boldsymbol{\mu}(p)}\right) \tag{13}$$

where ε is a predefined constant. The offset of one before the application of the pointwise logarithm ensures that the transformation is nonexpansive in the sense of Lipschitzian maps. On a dataset of environmental audio recordings, a previous publication has shown empirically that Equation 13 brings the histogram of $\widetilde{\mathbf{S}}\boldsymbol{x}_n(p)$ closer to a Gaussian distribution [13]. Since then, this finding has also been confirmed in the case of musical sounds [17].

Lastly, we standardize every feature $\widetilde{\mathbf{S}}\boldsymbol{x}_n$ to null mean and unit variance, across the dataset $\mathcal{X} = \{\boldsymbol{x}_1 \dots \boldsymbol{x}_N\}$, independently for each scattering path p. Again, if $\mathcal{X}$ is split between training and test sets, we measure means and variances over the training set only and propagate them as constants to the test set. With a slight abuse of notation, we still denote by $\widetilde{\mathbf{S}}\boldsymbol{x}_n(p)$ the standardized log-scattering features at path p for sample n, even though its value differs from Equation 13 by an affine transformation.

Metric learning with large-margin nearest neighbors (LMNN)

Let $\boldsymbol{x}$ some arbitrary audio sample in the dataset $\mathcal{X}$. Let $\mathcal{G}$ a cluster graph with $N = \operatorname{card} \mathcal{X}$ vertices and C clusters in total. We denote by $\mathcal{G}(\boldsymbol{x})$ the cluster to which the sample $\boldsymbol{x}$ belongs. Given another sample $\boldsymbol{y}$ in $\mathcal{X}$, $\boldsymbol{y}$ is similar to $\boldsymbol{x}$ if and only if belongs to the cluster $\mathcal{G}(\boldsymbol{x})$. Because $\mathcal{G}$ is a disjoint union of complete graphs, this relation is symmetric: $\boldsymbol{x} \in \mathcal{G}(\boldsymbol{y})$ is equivalent to $\boldsymbol{y} \in \mathcal{G}(\boldsymbol{x})$.

In our protocol, $\boldsymbol{x}$ contains the sound of an isolated musical note and the cluster graph $\mathcal{G}$ encodes auditory similarities within the dataset $\mathcal{X}$. Moreover, we take $\mathcal{G}$ to be the equal to the "consensus" cluster graph $\mathcal{G}_0$, i.e., arising from the partition of a hypergraph $\mathcal{H}$ which contains the judgments of all K subjects from our perceptual study (see data collection section).

We denote by $\widetilde{\mathbf{S}}\boldsymbol{x}$ the feature vector of joint time–frequency scattering resulting from $\boldsymbol{x}$. This vector includes both first-order and second-order scattering coefficients, after median-based logarithmic compression and affine standardization (see subsections above). Furthermore, we denote by $\mathcal{Y}_R(\boldsymbol{x})$ the list of R nearest neighbors to $\widetilde{\mathbf{S}}\boldsymbol{x}$ in the feature space of joint time–frequency scattering coefficients according to the Euclidean metric. Unlike cluster similarity, this relationship is not symmetric: $\boldsymbol{y} \in \mathcal{Y}_R(\boldsymbol{x})$ does not necessarily imply $\boldsymbol{x} \in \mathcal{Y}_R(\boldsymbol{y})$. Note that the dependency of $\mathcal{Y}_R(\boldsymbol{x})$ upon the operator $\mathbf{S}$ is left implicit. In all of the following, we set the constant R to 5; this is in accordance with our chosen evaluation metric, average precision at rank 5 (AP@5, see results section).

Let P be the number of scattering paths in the operator $\mathbf{S}$. The large-margin nearest neighbor (LMNN) algorithm learns a matrix $\mathbf{L}$ with P rows and P columns by minimizing an error function of the form:

$$\mathcal{E}(\mathbf{L}) = \frac{1}{2}\mathcal{E}_{\mathrm{pull}}(\mathbf{L}) + \frac{1}{2}\mathcal{E}_{\mathrm{push}}(\mathbf{L}) \tag{14}$$

where, intuitively, $\mathcal{E}_{\text{pull}}$ tends to shrink local Euclidean neighborhoods in feature space while $\mathcal{E}_{\text{push}}$ tends to penalize small distances between samples that belong to different clusters in $\mathcal{G}$.

The definition of $\mathcal{E}_{\text{pull}}$ is:

$$\mathcal{E}_{\text{pull}}(\mathbf{L}) = \sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{\boldsymbol{y}\in\mathcal{Y}_R(\boldsymbol{x})} \left\|\mathbf{L}\widetilde{\mathbf{S}}\boldsymbol{x} - \mathbf{L}\widetilde{\mathbf{S}}\boldsymbol{y}\right\|^2, \tag{15}$$

Note that the error term $\mathcal{E}_{\text{pull}}$ is unsupervised, in the sense that it does not depend on the cluster assignments of $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{G}$.

While the the term $\mathcal{E}_{\text{pull}}$ operates on pairs of samples, the term $\mathcal{E}_{\text{push}}$, operates on triplets $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathcal{X}^3$. The first sample, $\boldsymbol{x}$, is known as an "anchor." The second sample, $\boldsymbol{y}$, is known as a "positive", and is assumed to belong to the Euclidean neighborhood of the anchor: $\boldsymbol{y} \in \mathcal{Y}_R(\boldsymbol{x})$. The third sample, $\boldsymbol{z}$, is known as a "negative", and is assumed to belong to a different similarity cluster as the anchor: $\boldsymbol{z} \notin \mathcal{G}(\boldsymbol{x})$. The term $\mathcal{E}_{\text{push}}$ penalizes $\mathbf{L}$ unless the positive-to-anchor distance is smaller than the negative-to-anchor distance by a margin of at least 1.

The definition of $\mathcal{E}_{\text{push}}$ is:

$$\mathcal{E}_{\text{push}}(\mathbf{L}) = \sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{\boldsymbol{y}\in\mathcal{Y}_R(\boldsymbol{x})} \sum_{\boldsymbol{z}\notin\mathcal{G}(\boldsymbol{x})} \rho\Big(1 + \left\|\mathbf{L}\widetilde{\mathbf{S}}\boldsymbol{x} - \mathbf{L}\widetilde{\mathbf{S}}\boldsymbol{y}\right\|^2 - \left\|\mathbf{L}\widetilde{\mathbf{S}}\boldsymbol{x} - \mathbf{L}\widetilde{\mathbf{S}}\boldsymbol{z}\right\|^2\Big), \tag{16}$$

where the function $\rho : u \mapsto \max(0, u)$ denotes the activation function of the rectified linear unit (ReLU), also known as hinge loss. The cost function described in the equation above is known in deep learning as "triplet loss" and has recently been applied to train large-vocabulary audio classifiers in an unsupervised way [70]. We refer to [71] for a review of the state of the art in metric learning.

Results

The previous section described our methods for extracting spectrotemporal modulations in audio signals, as well as learning a non-Euclidean similarity metric between them. We now turn to apply these methods to the problem of allocating isolated musical notes to clusters of some timbre similarity graph $\mathcal{G}$. In practice, for training purposes, the cluster graph $\mathcal{G}$ represents the consensus of the $K = 31$ clustering provided by the users interacting with the Cyberlioz web application, which was described in the data collection section ($\mathcal{G} = \mathcal{G}_0$). However, for evaluation purposes, this cluster graph corresponds to the subjective preferences of a single user $k \geq 1$, in which case we take $\mathcal{G} = \mathcal{G}_k$.

Semi-supervised label propagation

In order to suit the practical needs of contemporary music composers, computer-assisted orchestration must draw from a diverse realm of instruments and techniques. Therefore, whereas our data collection procedure for timbre similarity judgments focused on a single pitch (middle C) and a single intensity level (mf), we formulate our machine listening experiment on an expanded dataset of audio samples, containing variations in pitch and dynamics.

Given an audio stimulus $\boldsymbol{x}_n$ from our perceptual study, we seek its position in the cluster graph $\mathcal{G}_0$. Then, we identify its instrument–mute–technique (IMT) triplet, scrape for audio samples in SOL matching this triplet, and assign them all to the same cluster as the original audio stimulus. We repeat the same procedure for all $N = 78$ nodes in $\mathcal{G}_0$, resulting in $N' = 9346$ samples in total. This is a form of semi-supervised label propagation: from a limited amount of human annotation in order to curate a relatively large subset of SOL, amounting to about one third of the entire dataset (9346 out of 25444 samples).

Evaluation metric

Let us denote by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N'}$ the N' audio samples associated with our annotated dataset after semi-supervised label propagation. Given a sample n and a human subject k, we denote by $\mathcal{G}_k$ the cluster graph associated to the subject k, and by $\mathcal{G}_k(n)$ the cluster to which the sample $\boldsymbol{x}_n$ belongs. Our machine listening system takes the waveform $\boldsymbol{x}_n$ as input and returns a ranked list of nearest neighbors: $\boldsymbol{\Phi}_1(\boldsymbol{x}_n)$, $\boldsymbol{\Phi}_2(\boldsymbol{x}_n)$, $\boldsymbol{\Phi}_3(\boldsymbol{x}_n)$, and so forth.

In the context of browsing an audio collection by timbre similarity, $\boldsymbol{x}_n$ is a user-defined query while the function $\boldsymbol{\Phi}$ plays the role of a search engine. We consider the first retrieved sample, $\boldsymbol{\Phi}_1(\boldsymbol{x}_n)$, to be relevant to user k if and only if it belongs to the same cluster as $\boldsymbol{x}_n$ in the cluster graph $\mathcal{G}_k$; hence the Boolean condition $\boldsymbol{\Phi}_1(\boldsymbol{x}_n) \in \mathcal{G}_k(n)$. Likewise, the second retrieved sample is relevant if and only if $\boldsymbol{\Phi}_2(\boldsymbol{x}_n) \in \mathcal{G}_k(n)$. To evaluate $\boldsymbol{\Phi}$ on the query $\boldsymbol{x}_n$, we measure the relevance of all nearest neighbors $\boldsymbol{\Phi}_r(\boldsymbol{x}_n)$ up to some fixed rank R and average the result:

$$p_{\boldsymbol{\Phi}}(n, k, R) = \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}\left(\boldsymbol{\Phi}_r(\boldsymbol{x}_n) \in \mathcal{G}_k(n)\right). \tag{17}$$

In the equation above, the indicator function $\mathbb{1}$ converts Booleans to integers, i.e., $\mathbb{1}(b)$ returns one if b is true and return zero if b is false. Thus, the function $p_{\boldsymbol{\Phi}}$ takes fractional values between 0 and 1, which are typically expressed in percentage points.

The precision at rank R of the system $\boldsymbol{\Phi}$ is defined as the average value taken by the function $p_{\boldsymbol{\Phi}}$ over the entire corpus of N' audio samples:

$$\mathrm{P}_{\boldsymbol{\Phi}}(k, R) = \frac{1}{N'} \sum_{n=1}^{N'} p_{\boldsymbol{\Phi}}(n, k, R) \tag{18}$$

Lastly, the "average precision at rank R" (henceforth, AP@R) is the average value of $\mathrm{P}_{\boldsymbol{\Phi}}$, for constant R, across all $K = 31$ subjects from our perceptual study:

$$\mathrm{AP}_{\boldsymbol{\Phi}}(R) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{P}_{\boldsymbol{\Phi}}(k, R) \tag{19}$$

It appears from the above that an effective system $\boldsymbol{\Phi}$ should retrieve sounds whose IMT triplets are similar according to all of the K cluster graphs $\mathcal{G}_1 \ldots \mathcal{G}_K$.

In the rest of this paper, we set R to 5. This is in accordance with the protocol of [17], in which the authors trained a metric learning algorithm on the SOL dataset to

search for similar instruments and playing techniques, yet without the intervention of a human subject.

Best performing system

Our best performing system comprises five computational blocks:

1. Joint time–frequency scattering up to a maximal time scale of $T = 1000\,\text{ms}$,
2. Temporal averaging at the scale of the whole musical note,
3. Median-based logarithmic compression,
4. Affine standardization so that each feature has zero mean and unit variance, and
5. Nearest-neighbor search according to priorly learned non-Euclidean metric.

Note that the non-Euclidean metric is learned via LMNN (see methods section) on the "consensus" cluster graph $\mathcal{G}_0$. Therefore, the system $\boldsymbol{\Phi}$ performs timbre similarity retrieval in a user-agnostic way, and can serve as a convenient default algorithm for newcoming users. That being said, it is conceivable to replicate the five-stage protocol above on the cluster graph $\mathcal{G}_k$ of a specific user k, instead of the cluster graph $\mathcal{G}_0$. This operation would lead to a new configuration of the search engine $\boldsymbol{\Phi}$ that is better tailored to the perceptual idiosyncrasy of user k in terms of timbre similarity.

Composers may personalize their search engine $\boldsymbol{\Phi}$ by manually defining a cluster graph via the Cyberlioz interface (see data collection section). This preliminary annotation lasts 30 to 60 minutes, which is relatively short in comparison with the duration of the $N' = 9346$ audio samples in $\mathcal{G}_k$ after label propagation: i.e., roughly three hours of audio.

Within the default setting ($\mathcal{G} = \mathcal{G}_0$), our system $\boldsymbol{\Phi}$ achieves an average precision at rank five (AP@5) of **99.0%**, with a standard deviation across $K = 31$ subjects of the order of 1%. This favorable result suggests that joint time–frequency scattering provides a useful feature map for learning similarities between instrumental playing techniques. In doing so, it is in line with a recent publication [72], in which the authors successfully trained a supervised classifier on joint time–frequency scattering features in order to detect and classify playing techniques from the Chinese bamboo flute *(dizi)*. However, the originality of our work is that $\boldsymbol{\Phi}$ relies purely on auditory information (i.e., timbre similarity judgments), and does not require any supervision from the symbolic domain. In particular, it does not assume the metadata (instrument, mute, technique, pitch, dynamics, and so forth) of any musical sample $\boldsymbol{x}_n$ to be observable, in part or in full, at training time.

Discussion

The previous section described our dataset of audio samples with diverse instruments, mutes, and playing techniques, as well as the method we adopt to evaluate a given system for timbre similarity retrieval. Over a cohort of $K = 31$ subjects, we reported an average precision at rank five (AP@5) of $99.0\% \pm 1$. We now turn to alter certain key choices in the design of the above-described computational blocks, and discuss their respective impacts on downstream performance.

Figure 6 summarizes our results. Interestingly, the system $\boldsymbol{\Phi}$ is not only best on average, but also best for every subject in the cohort. Specifically, replacing

$\mathbf{\Phi}$ by a simpler model $\mathbf{\Phi}'$ (see subsections below for examples of such models) results in $\mathrm{P}_{\mathbf{\Phi}'}(k, 5) < \mathrm{P}_{\mathbf{\Phi}}(k, 5)$ for every k. Borrowing from the terminology of welfare economics, $\mathbf{\Phi}$ can be said to be uniquely Pareto-efficient [**?**]. This observation suggests that the increase in performance afforded by the state-of-the-art model with respect to the baseline does not come at the detriment of user fairness.

Is the dataset large enough for evaluation? Is the system overfitting?

First of all, it is worth noting that the presented AP@5 figure of $99.0\% \pm 1$ does not abide to a conventional training set vs. test set paradigm, as is most often done in machine learning research. Rather, the LMNN algorithm is trained on all available samples ($N' = 9346$ isolated notes) with the "consensus" cluster graph as training objective ($\mathcal{G} = \mathcal{G}_0$). Then, it is evaluated on the same samples with individual cluster graphs as ground truth ($\mathcal{G} = \mathcal{G}_k$ for $k \geq 1$). The reason behind this choice is that, in the context of computer-assisted orchestration, the collection of audio samples has a fixed size, as it is shipped alongside the software itself. This is the case, for example, of Orchidea,[6] which comes paired with its own version of SOL, named OrchideaSOL [73]. Henceforth, our main goal was to evaluate the generalization ability of our metric learning algorithm beyond the restricted set of samples for which human annotations are directly available (see data collection section); that is, beyond one pitch class (middle C) and one intensity level (mf).

Despite this caveat, we may adopt a "query-by-example" (QbE) paradigm by partitioning the database and audio samples in half ($\frac{1}{2}N' = 4673$), training the LMNN algorithm on the first half, and querying it with samples from the other half. In this evaluation framework, our system reaches an AP@5 of **96.2%** ± 2. Interestingly, querying the system with samples from the training set leads to an AP@5 of **96.5%** ± 2, i.e., roughly on par with the test set. Thus, it appears that the gap in performance between the evaluation presented in the results section ($99.0\% \pm 1$) and query-by-example evaluation ($96.2\% \pm 2$) is primarily attributable to a reduction of the size of training set by half, whereas statistical overfitting of the training set with respect to the test set likely plays a minor role.

Is metric learning necessary?

Replacing the LMNN metric learning algorithm by linear discriminant analysis (LDA) leads to a P@5 of **76.6%** ± 11. Moreover, we evaluate the nearest neighbor algorithm $\mathbf{\Phi}$ in the absence of any metric learning at all. This corresponds to using a Euclidean distance to compare scattering coefficients, i.e., to set $\mathbf{L}$ to the identity matrix. Note that the runtime complexity of Euclidean nearest-neighbor search is identical to LMNN search. We report an average precision at rank five (AP@5) of **92.9%** ± 3, which is noticeably worse than the best performing system. This gap in performance demonstrates the importance of complementing unsupervised feature extraction by supervised metric learning in the design of computational models for timbre similarity between instrumental playing techniques.

---

[6]Link to Orchidea software and OrchideaSOL dataset: `www.orch-idea.org`

How important is the amount of temporal context T?

Our best performing system operates with joint time–frequency scattering coefficients as spectrotemporal modulation features. These features are extracted within a temporal context of duration equal to $T = 1000\,\text{ms}$. This value is considerably larger than the frame size of purely spectral features, such as spectral centroid, spectral flux, or mel-frequency cepstral coefficients (MFCC). Indeed, the frame size of spectral features for machine listening is typically set to $T = 23\,\text{ms}$, i.e., $2^{10} = 1024$ samples at a sampling rate of $44.1\,\text{kHz}$.

As a point of comparison, we set the maximum time scale of joint time–frequency scattering coefficients to $T = 25\,\text{ms}$, hence a 40-fold reduction in context size. Over our cohort of $K = 31$ subjects, we report a precision at rank five (P@5) of **90.9%** $\pm\,4$, which is noticeably worse than the best performing system. This gap in performance extends the findings of a previous publication [17], which reported that metric learning with temporal scattering coefficients tends to improve with growing values of T, until reaching a plateau of performance around $T \sim 500\,\text{ms}$.

Is joint time–frequency scattering necessary?

Let us recall the full definition of second-order joint time–frequency scattering (see methods section):

$$\mathbf{S_2}\boldsymbol{x}(t, \lambda, \alpha, \beta) = \left(\left| \left| \boldsymbol{x} * \boldsymbol{\psi}_\lambda \right| \circledast \boldsymbol{\Psi}_{\alpha,\beta} \right| \circledast \left(\boldsymbol{\phi}_T \otimes \boldsymbol{\phi}_F \right) \right)(t, \lambda), \qquad (20)$$

where the $\boldsymbol{\psi}_\lambda$ denotes a Morlet wavelet of center frequency λ and resp. $\boldsymbol{\phi}_T$ denotes a Gaussian low-pass filter of width T. Besides the *joint* time–frequency scattering transform, the generative grammar of scattering transforms [74] also encompasses the *separable* time–frequency scattering transform:

$$\mathbf{S_2^{\text{sep}}}\boldsymbol{x}(t, \lambda, \alpha, \beta) = \left(\left| \left| \left| \boldsymbol{x} * \boldsymbol{\psi}_\lambda \right| * \boldsymbol{\psi}_\alpha \right| \circledast \left(\boldsymbol{\phi}_T \otimes \boldsymbol{\psi}_\beta \right) \right| \circledast \left(\boldsymbol{\phi}_T \otimes \boldsymbol{\phi}_F \right) \right)(t, \lambda), \ (21)$$

where the wavelet $\boldsymbol{\psi}_\lambda$ has a quality factor of $Q = 12$ whereas the wavelets $\boldsymbol{\psi}_\alpha$ and $\boldsymbol{\psi}_\beta$ have a quality factor of one. Previous publications have successfully applied separable time–frequency scattering in order to classify environmental sounds [75] as well as playing techniques from the Chinese bamboo flute [76].

In comparison with its joint counterpart, separable time–frequency scattering contains about half as many coefficients. This is because the temporal wavelet transform with $\boldsymbol{\psi}_\alpha$ and the frequential wavelet transform with $\boldsymbol{\psi}_\beta$ are separated by an operation of complex modulus. Hence, $\boldsymbol{\psi}_\beta$ operates on a real-valued input. Because separable time–frequency scattering cannot distinguish ascending chirps from descending chirps, flipping the sign of the scale variable β is no longer necessary. Moreover, separable time–frequency scattering has a lower algorithmic complexity than joint time–frequency scattering. Indeed, in Equation 21, the frequential wavelet transform with $\boldsymbol{\psi}_\beta$ operates on a tensor whose time axis is subsampled at a fixed rate T^{-1}, thus allowing vectorized computations. Conversely, in Equation 20, the frequential wavelet transform must operate on a multiresolution input, whose sample rate varies depending on α: it ranges between T^{-1} and the sample rate of $\boldsymbol{x}$ itself.

Yet, despite its greater simplicity, separable time–frequency scattering suffers from known weaknesses in its ability to represent spectrotemporal modulations. In particular, a previous publication has shown that frequency-dependent time shifts affect joint time–frequency scattering coefficients while leaving separable time–frequency scattering coefficients almost unchanged [18]. The same observation was made by [77] in the case of joint and separable Gabor filterbank features (GBFB), which bear some resemblance with joint and separable time–frequency scattering coefficients respectively.

Over our cohort of $K = 31$ subjects, separable time–frequency scattering achieves a precision at rank five (P@5) of **91.9%** $\pm$ 4. This figure is noticeably worse than joint time–frequency scattering (**99.0%** $\pm$ 1), all other things being equal. This gap in performance, together with the theory of spectrotemporal receptive fields (STRF) in auditory neuroscience (see related work section), demonstrates the importance of joint spectrotemporal modulations in the modeling of timbre similarity across instrumental playing techniques.

What is the baseline performance?

Lastly, we train a baseline system in which joint time–frequency scattering coefficients are replaced by mel-frequency cepstral coefficients (MFCC). Specifically, we extract a 40-band mel-frequency spectrogram by means of the RASTAMAT library; apply the pointwise logarithm; and compute a discrete cosine transform (DCT) over the mel-frequency axis.[7] This operation results in a 40-dimensional feature vector over frames of duration $T = 25$ ms. Over our cohort of $K = 31$ subjects, we report a precision at rank five (P@5) of **81.8%** $\pm$ 7.

Arguably, this figure is not directly comparable with our best performing system (P@5 of 99.0% $\pm$ 1), due to the mismatch in dimensionality between MFCC and joint time–frequency scattering coefficients. In order to clarify the role of feature dimensionality in our computational pipeline, we apply a feature engineering technique involving multiplicative combinations of MFCC. We construct the following Gram matrix:

$$\mathbf{G}\boldsymbol{x}(\alpha, \beta) = \int_0^{+\infty} \mathrm{MFCC}(\boldsymbol{x})(t, \alpha) \mathrm{MFCC}(\boldsymbol{x})(t, \beta) \, \mathrm{d}t, \tag{22}$$

where α and β represent different dimensions ("quefrencies") of the MFCC feature vector. The symmetric matrix $\mathbf{G}\boldsymbol{x}$ contains 40 rows and 40 columns, hence 800 unique coefficients. Concatenating these coefficients to the 40 averaged MFCC features results in a feature vector of 840 coefficients. This dimension is of the same order of magnitude as the dimension of our joint time–frequency scattering representation ($d = 1180$, see methods section).

Training the LMNN algorithm on this 840-dimensional representation is analogous to a "kernel trick" in support vector machines [78]. In our case, the implicit similarity kernel is a homogeneous quadratic kernel. Despite this increase in representational power, we obtain a P@5 of 81.5% $\pm$ 7, i.e. essentially the same as MFCC under a linear kernel. Therefore, it appears that the gap in performance between MFCC

---

[7]Link to RASTAMAT library: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat

$(81.8\%\pm7)$ and joint time–frequency scattering $(99.0\%\pm1)$ is primarily attributable to the multiscale extraction of joint spectrotemporal modulations, whereas high-dimensional embedding likely plays a minor role.

Conclusion

We see from the ablation study conducted above that each step of the proposed model is necessary for its satisfactory performance. Among other things, it indicates the joint time–frequency scattering transform itself should not be used directly for similarity comparisons, but is best augmented by a learned stage. This explains the relative success of such models [13, 17] over others where distances on raw scattering coefficients were used for judging audio similarity.

On the other hand, we note that the complexity of the learned model does not have to be high. Indeed, for this task, a linear model on the scattering coefficients is sufficient. There is no need for deep networks with large numbers of parameters to accurately represent the similarity information. In other words, the scattering transform parametrizes the signal structure in a way that many relevant quantities (such as timbre similarity) can be extracted through a simple linear mapping. This is in line with other classification results, where linear support vector machines applied to scattering coefficients have achieved significant success [12, 18].

We also see the necessity of a fully *joint* time–frequency model for accessing timbre, as opposed to a purely spectral model or one that treats the two axes in a separable manner. This fact has also been observed in other contexts, such as the work of Patil et al. [19]. A related observation is the need for capturing large-scale structure. Indeed, reducing the window size to 25 ms means that we lose a great deal of time–frequency structure, bringing results closer to that of the separable model.

The success of the above system in identifying timbral similarities has immediate applications in browsing music databases. These are typically organized based on instrumental and playing techniques taxonomies, with additional keywords offering a more flexible organization. Accessing the sounds in these databases therefore requires some knowledge of the taxonomy and keywords used. Furthermore, the user needs to have some idea of the particular playing technique they are searching for.

As an alternative, content-based searches allow the user to identify sounds based on similarity to some given query sound. This query-by-example approach provides an opportunity to search for sounds without having a specific instrument of playing technique in mind, yielding a wider range of available sounds. A composer with access to such a system would therefore be able to draw on a more diverse palette of musical timbre.

The computational model proposed in this work is well suited to such a query-by-example task. We have shown that it is able to adequately approximate the timbral judgments of a wide range of subjects included in our study. Not only that, but the system can be easily retrained to approximate an individual user's timbre perception by having that user perform the clustering task on the reduced set of 78 IPTs and running the LMNN training step on those clustering assignments. This model can then be applied to new data, or, alternatively, be retrained with these new examples if the existing model proves unsatisfactory.

We shall note, however, that the current model has several drawbacks. First, it is only applied to instrumental sounds. While this has the advantage of simplifying the interpretation of the results, the range of timbre under consideration is necessarily limited (although less restricted than only considering *ordinario* PTs). This also makes applications such as *query-by-humming* difficult, since we cannot guarantee that the timbral similarity measure is accurate for such sounds.

That being said, the above model is general enough to encompass a wide variety of recordings, not just instrumental sounds. Indeed, we have strong assumptions that the tools used (scattering transforms and LMNN weighting matrices) do not depend on the type of sound being processed. ML: Future work will demonstrate that they work for more general classes of sounds. To extend the model, it is only necessary to retrain the LMNN weighting matrix by supplying it with new cluster assignments. These can again be obtained by performing a new clustering experiment with one or more human subjects.

Another aspect is the granularity of the similarity judgments. In the above method, we have used *hard* clustering assignments to build our model. A more nuanced similarity judgment would ask users to rate the similarity of a pair of IPTs on a more graduated scale, which would yield a finer, or *soft*, assignment. This however, comes with additional difficulties in providing a consistent scale across subjects, but could be feasible if the goal is to only adapt the timbral space to a single individual. An approach not based on clustering would also have to replace the LMNN algorithm with one that accounts for such soft assigments.

Author details
[1]Music and Audio Research Lab, New York University, 370 Jay Street, New York, NY, USA. [2]LS2N, CNRS, Centrale Nantes, Nantes University, 1, rue de la Noe, 44000, Nantes, France. [3]Lonofi, 57 rue Letort, 75018, Paris, France. [4]Department of Mathematics, KTH Royal Institute of Technology Lindstedtsvägen 25 SE-100 44, Stockholm Sweden. [5]Center for Computational Mathematics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA.

References
1. Downie, J.S.: Music information retrieval. Annual review of information science and technology **37**(1), 295–340 (2003)
2. Siedenburg, K., Saitis, C., McAdams, S.: In: Siedenburg, K., Saitis, C., McAdams, S., Popper, A.N., Fay, R.R. (eds.) The Present, Past, and Future of Timbre Research, pp. 1–19. Springer, Cham (2019)
3. Faure, A., McAdams, S., Nosulenko, V.: Verbal correlates of perceptual dimensions of timbre. In: Proceedings of the International Conference on Music Perception and Cognition (ICMPC) (1996)
4. Kostka, S.: Materials and Techniques of Post-tonal Music. Taylor & Francis, ??? (2016)
5. Antoine, A., Miranda, E.R.: Musical acoustics, timbre, and computer-aided orchestration challenges. In: Proc. ISMA (2018)
6. Kolozali, S., Barthet, M., Fazekas, G., Sandler, M.B.: Knowledge representation issues in musical instrument ontology design. In: Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference (2011)
7. Calvo-Zaragoza, J., Hajič Jr., J., Pacha, A.: Understanding optical music recognition. ACM Computing Surveys (2020)
8. Erickson, R.: Sound Structure in Music. University of California Press, ??? (1975)
9. Thoret, E., Caramiaux, B., Depalle, P., McAdams, S.: Human dissimilarity ratings of musical instrument timbre: A computational meta-analysis. Journal of the Acoustical Society of America **143**(3), 1745–1746 (2018)
10. Maresz, Y.: On computer-assisted orchestration. Contemporary Music Review **32**(1), 99–109 (2013)

11. Caetano, M., Zacharakis, A., Barbancho, I., Tardón, L.J.: Leveraging diversity in computer-aided musical orchestration with an artificial immune system for multi-modal optimization. Swarm and Evolutionary Computation **50**, 100484 (2019)
12. Andén, J., Lostanlen, V., Mallat, S.: Joint time-frequency scattering for audio classification. In: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6 (2015). IEEE
13. Lostanlen, V., Lafay, G., Andén, J., Lagrange, M.: Relevance-based quantization of scattering features for unsupervised mining of environmental audio. EURASIP Journal on Audio, Speech, and Music Processing **2018**(1), 15 (2018)
14. Andén, J., Mallat, S.: Scattering representation of modulated sounds. In: Proceedings of the International Conference on Digital Audio Effects (DAFx) (2012)
15. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research **10**(Feb), 207–244 (2009)
16. 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**(3), 177–192 (1995)
17. Lostanlen, V., Andén, J., Lagrange, M.: Extended playing techniques: the next milestone in musical instrument recognition. In: Proceedings of the International Conference on Digital Libraries for Musicology (DLfM), pp. 1–10 (2018). ACM
18. Andén, J., Lostanlen, V., Mallat, S.: Joint time–frequency scattering. IEEE Transactions on Signal Processing **67**(14), 3704–3718 (2019)
19. Patil, K., Pressnitzer, D., Shamma, S., Elhilali, M.: Music in our ears: The biological bases of musical timbre perception. PLoS Computational Biology **8**(11), 1002759 (2012)
20. Joder, C., Essid, S., Richard, G.: Temporal integration for audio classification with application to musical instrument classification. IEEE Transactions on Audio, Speech, and Language Processing **17**(1), 174–186 (2009)
21. Siedenburg, K., Fujinaga, I., McAdams, S.: A comparison of approaches to timbre descriptors in music information retrieval and music psychology. Journal of New Music Research **45**(1), 27–41 (2016)
22. Martin, K.D., Kim, Y.E.: Musical instrument identification: A pattern recognition approach. In: Proceedings of the Acoustical Society of America (1998)
23. Brown, J.C.: Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. The Journal of the Acoustical Society of America **105**(3), 1933–1941 (1999)
24. Eronen, A., Klapuri, A.: Musical instrument recognition using cepstral coefficients and temporal features. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2000)
25. Herrera Boyer, P., Peeters, G., Dubnov, S.: Automatic classification of musical instrument sounds. Journal of New Music Research **32**(1), 3–21 (2003)
26. Wieczorkowska, A.A., Żytkow, J.M.: Analysis of feature dependencies in sound description. Journal of Intelligent Information Systems **20**(3), 285–302 (2003)
27. Livshin, A., Rodet, X.: Musical instrument identification in continuous recordings. In: Proceedings of the International Conference on Digital Audio Effects (DAFx) (2004)
28. Krishna, A.G., Sreenivas, T.V.: Music instrument recognition: from isolated notes to solo phrases. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2004)
29. Kaminskyj, I., Czaszejko, T.: Automatic recognition of isolated monophonic musical instrument sounds using kNNC. Journal of Intelligent Information Systems **24**(2-3), 199–221 (2005)
30. Benetos, E., Kotti, M., Kotropoulos, C.: Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2006)
31. Bhalke, D.G., Rao, C.B.R., Bormane, D.S.: Automatic musical instrument classification using fractional Fourier transform based-MFCC features and counter propagation neural network. Journal of Intelligent Information Systems **46**(3), 425–446 (2016)
32. Humphrey, E., Durand, S., McFee, B.: Openmic-2018: an open dataset for multiple instrument recognition. In: Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference (2018)
33. McFee, B., Humphrey, E.J., Urbano, J.: A plan for sustainable MIR evaluation. In: Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference (2016)
34. Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X.: FMA: A dataset for music analysis. In: Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference (2017)
35. Lostanlen, V., Cella, C.E.: Deep convolutional networks on the pitch spiral for musical instrument recognition. In: Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference (2016)
36. Bittner, R.M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., Bello, J.P.: MedleyDB: A multitrack dataset for annotation-intensive MIR research. In: Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference (2014)
37. McFee, B., Humphrey, E., Bello, J.: A software framework for musical data augmentation. In: Proceedings of the International Society on Music Information Retrieval (ISMIR) (2015)
38. Pons, J., Slizovskaia, O., Gong, R., Gómez, E., Serra, X.: Timbre analysis of music audio signals with convolutional neural networks. In: 25th European Signal Processing Conference (EUSIPCO), pp. 2744–2748 (2017)
39. Gururani, S., Summers, C., Lerch, A.: Instrument activity detection in polyphonic music using deep neural networks. In: Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference (2018)
40. Loureiro, M.A., de Paula, H.B., Yehia, H.C.: Timbre classification of a single musical instrument. In: Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference (2004)

41. Han, Y., Kim, J., Lee, K.: Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE Transactions on Audio, Speech, and Language Processing **25**(1), 208–221 (2017)

42. McAdams, S., Giordano, B.L.: The perception of musical timbre, pp. 72–80 (2009)

43. Depireux, D., Simon, J., Klein, D., Shamma, S.: Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. Journal of Neurophysiology **85**(3), 1220–1234 (2001)

44. Aertsen, A., Johannesma, P.: The spectro-temporal receptive field. Biological Cybernetics **42**(2), 133–143 (1981)

45. De Boer, E., Kuyper, P.: Triggered correlation. IEEE Transactions on Biomedical Engineering (3), 169–179 (1968)

46. Flandrin, P.: Time-frequency/time-scale Analysis. Academic press, ??? (1998)

47. Eggermont, J.: Wiener and volterra analyses applied to the auditory system. Hearing research **66**(2), 177–201 (1993)

48. Klein, D., Depireux, D., Simon, J., Shamma, S.: Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. Journal of Computational Neuroscience **9**(1), 85–111 (2000)

49. Theunissen, F.E., Sen, K., Doupe, A.J.: Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. Journal of Neuroscience **20**(6), 2315–2331 (2000)

50. Chi, T., Ru, P., Shamma, S.A.: Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America **118**(2), 887–906 (2005)

51. Patil, K., Elhilali, M.: Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases. EURASIP J. Audio Speech Music Process. **2015**(1), 27 (2015)

52. 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**(6), 478–483 (2016)

53. Mishra, S., Sturm, B.L., Dixon, S.: Understanding a deep machine listening model through feature inversion. In: Proceedings of the International Society on Music Information Retrieval (ISMIR) Conference, pp. 755–762 (2018)

54. Hemery, E., Aucouturier, J.: One hundred ways to process time, frequency, rate and scale in the central auditory system: a pattern-recognition meta-analysis. Frontiers in computational neuroscience **9** (2015)

55. Andreux, M., Angles, T., Exarchakis, G., Leonarduzzi, R., Rochette, G., Thiry, L., Zarka, J., Mallat, S., Belilovsky, E., Bruna, J., et al.: Kymatio: Scattering transforms in python. Journal of Machine Learning Research **21**(60) (2020)

56. Lostanlen, V., Hecker, F.: The shape of remixxxes to come: Audio texture synthesis with time–frequency scattering. In: Proceedings of the Digital Audio Effects Conference (DAFX) (2019)

57. Mallat, S.: Understanding deep convolutional networks. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **374**(2065), 20150203 (2016)

58. Wu, C.-W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., Muller, M., Lerch, A.: A review of automatic drum transcription. IEEE Transactions on Audio, Speech, and Language Processing **26**(9), 1457–1483 (2018)

59. Pearce, A., Brookes, T., Mason, R.: Modelling timbral hardness. Applied Sciences **9**(3), 466 (2019)

60. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. The Bell system technical journal **49**(2), 291–307 (1970)

61. Han, E.-H., Karypis, G., Kumar, V.: Scalable Parallel Data Mining for Association Rules vol. 26. ACM, ??? (1997)

62. Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research **3**(Dec), 583–617 (2002)

63. Schörkhuber, C., Klapuri, A.: Constant-q transform toolbox for music processing. In: Proceedings of the Sound and Music Computing (SMC) Conference (2010)

64. Lostanlen, V., Sridhar, S., Farnsworth, A., Bello, J.P.: Learning the helix topology of musical pitch. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2020)

65. Mallat, S.: Group invariant scattering. Communications in Pure and Applied Mathematics **65**(10), 1331–1398 (2012)

66. Oyallon, E., Sifre, L., Mallat, S.: Generic deep networks with wavelet scattering. In: Proceedings of the ICLR Workshop (2014)

67. Elhilali, M., Chi, T., Shamma, S.A.: A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. Speech communication **41**(2-3), 331–348 (2003)

68. Bellur, A., Elhilali, M.: Detection of speech tokens in noise using adaptive spectrotemporal receptive fields. In: Proceedings of the Annual Conference on Information Sciences and Systems (CISS), pp. 1–6 (2015). IEEE

69. Emmanouilidou, D., Patil, K., West, J., Elhilali, M.: A multiresolution analysis for detection of abnormal lung sounds. In: Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), pp. 3139–3142 (2012). IEEE

70. Jansen, A., Plakal, M., Pandya, R., Ellis, D.P., Hershey, S., Liu, J., Moore, R.C., Saurous, R.A.: Unsupervised learning of semantic audio representations. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 126–130 (2018). IEEE

71. Bellet, A., Habrard, A., Sebban, M.: Metric Learning. Morgan & Claypool Publishers, ??? (2015)

72. Wang, C., Lostanlen, V., Benetos, E., Chew, E.: Playing technique recognition by joint time–frequency scattering. (2020). Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

73. Cella, C.-E., Ghisi, D., Lostanlen, V., Lévy, F., Fineberg, J., Maresz, Y.: Orchideasol: A dataset of extended instrumental techniques for computer-aided orchestration. In: Proceedings of the International Computer Music Conference (ICMC) (2020)

74. Lostanlen, V.: On time-frequency scattering and computer music. In: Schafhausen, N., Müller, V.J. (eds.) Florian Hecker: Halluzination, Perspektive, Synthese. Sternberg Press, Berlin (2019)

75. Baugé, C., Lagrange, M., Andén, J., Mallat, S.: Representing environmental sounds using the separable

scattering transform. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8667–8671 (2013). IEEE

76. Wang, C., Benetos, E., Lostanlen, V., Chew, E.: Adaptive time–frequency scattering for periodic modulation recognition in music signals. (2019). Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference

77. Schädler, M.R., Kollmeier, B.: Separable spectro-temporal gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition. The Journal of the Acoustical Society of America **137**(4), 2047–2059 (2015)

78. Chang, Y.-W., Hsieh, C.-J., Chang, K.-W., Ringgaard, M., Lin, C.-J.: Training and testing low-degree polynomial data mappings via linear svm. Journal of Machine Learning Research **11**(Apr), 1471–1490 (2010)
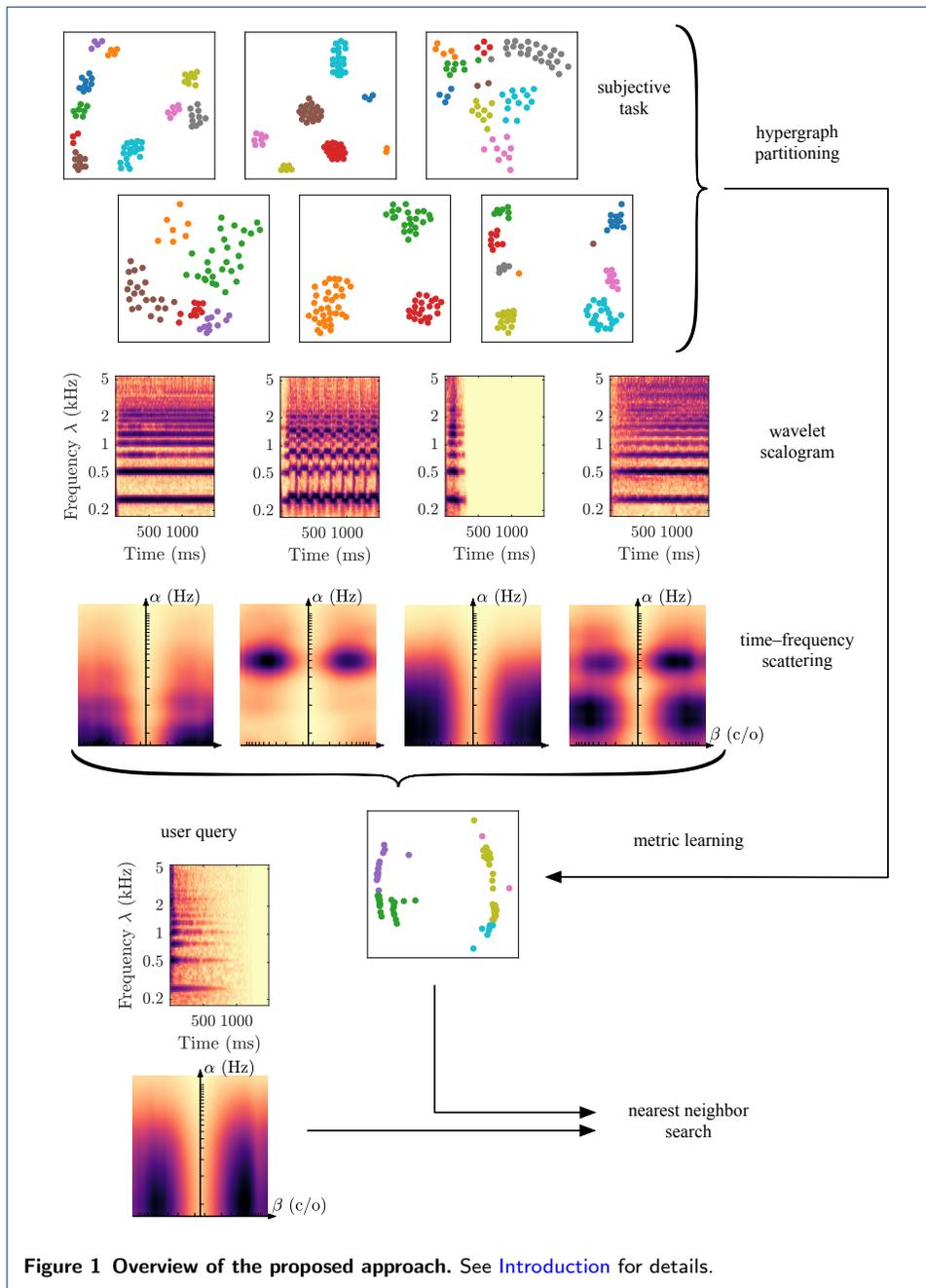
Figures

Figure 1 Overview of the proposed approach. See Introduction for details.

Table 1 Full list of audio stimuli (1/2). In each instrument, a blank space in the rightmost column denotes the ordinary playing technique (ordinario).

name	instrument	mute	playing technique
ASax-key-cl-C4-p	Alto saxophone		key click
ASax-ord-C4-mf	Alto saxophone		
ASax-ord-hi-reg-C6-mf	Alto saxophone		
ASax-slap-C4-mf	Alto saxophone		slap tonguing
ASax-slap-unp-C4-p	Alto saxophone		unpitched slap tonguing
BbCl-key-cl-C4-pp	Soprano clarinet		key click
BbCl-ord-hi-reg-A6-ff	Soprano clarinet		
BBTb-explo-slap-C#1-mf	Bass tuba		explosive slap
BBTb-pdl-tone-F1-mf	Bass tuba		pedal tone
BBTb-slap-F1-mf	Bass tuba		slap tonguing
BBTb-slap-unp-mf-1	Bass tuba		unpitched slap tonguing
Bn-key-cl-C3-mf	Bassoon		key click
Bn-ord-C4-mf	Bassoon		
Cb-pizz-bartok-C4-ff-1c	Contrabass		pizzicato à la Bartók (snap)
Cb-pizz-lv-C4-mf-1c	Contrabass		pizzicato & *laissez vibrer*
Cb-pizz-sec-C4-mf-1c	Contrabass		pizzicato secco
Cb-pont-C4-mf-1c	Contrabass		sul ponticello
Fl-key-cl-C4-f	Flute		key click
Fl-ord-C4-mf	Flute		
Fl-tng-ram-C4-mf	Flute		tongue ram
Gtr-ord-C4-mf-2c	Spanish guitar		
Gtr-ord-hi-reg-E5-mf-3c	Spanish guitar		
Gtr-pizz-bartok-C4-ff-2c	Spanish guitar		pizzicato Bartók
Gtr-pizz-C4-mf-2c	Spanish guitar		pizzicato
Hn-ord-C4-mf	French horn		
Hn-slap-C4-mf	French horn		slap tonguing
Hp-harm-fngr-C4-f	Harp		harmonic fingering
Hp-ord-C4-m4	Harp		
Hp-pizz-bartok-C4-mf	Harp		pizzicato Bartók
Hp-xyl-C4-p	Harp		xylophonic
Ob-blow-no-reed-C4	Oboe		blow without reed
Ob-key-cl-C4-pp	Oboe		key click
Ob-ord-C4-mf	Oboe		
TpC-ord-C4-mf	Trumpet in C		
TpC-pdl-tone-F3-mf	Trumpet in C		pedal tone
TpC-slap-C4-p	Trumpet in C		slap tonguing
TpC+C-ord-C4-mf	Trumpet in C	cup	
TpC+H-ord-C4-mf	Trumpet in C	harmon	
TpC+S-ord-C4-mf	Trumpet in C	straight	
TpC+W-ord-closed-C4-mf	Trumpet in C	wah (closed)	
TpC+W-ord-open-C4-mf	Trumpet in C	wah (open)	
TTbn-ord-C4-mf	Tenor trombone		
TTbn+C-ord-C4-mf	Tenor trombone	cup	
TTbn+H-ord-C4-mf	Tenor trombone	harmon	
TTbn+S-ord-C4-mf	Tenor trombone	straight	
TTbn+W-ord-closed-C4-mf	Tenor trombone	wah (closed)	
TTbn+W-ord-open-C4-mf	Tenor trombone	wah (open)	
Va-art-harm-C5-mf-4c	Viola		artificial harmonic
Va-legno-batt-C4-mf-3c	Viola		col legno battuto
Va-ord-C4-mf-3c	Viola		
Va-pizz-bartok-C4-ff-3c	Viola		pizzicato Bartók
Va-pizz-lv-C4-mf-3c	Viola		pizzicato & *laissez vibrer*
Va-pizz-sec-C4-mf-3c	Viola		pizzicato secco
Va-pont-C4-mf-3c	Viola		sul ponticello
Va+S-ord-C3-mf-3c	Viola	sordina	
Va+SP-ord-D4-mf-2c	Viola	piombo	
Vc-art-harm-C4-mf	Cello		artificial harmonic
Vc-legno-batt-C4-mf-1c	Cello		col legno battuto
Vc-legno-tratto-C4-mf-1c	Cello		col legno tratto
Vc-nonvib-C4-mf-1c	Cello		nonvibrato
Vc-ord-C4-mf-1c	Cello		
Vc-pizz-bartok-C4-ff-1c	Cello		pizzicato Bartók
Vc-pizz-lv-C4-mf-1c	Cello		pizzicato & *laissez vibrer*
Vc-pizz-sec-C4-mf-1c	Cello		pizzicato secco
Vc-pont-C4-mf-2c	Cello		sul ponticello
Vc-tasto-C4-mf-1c	Cello		sul tasto
Vc+S-ord-C4-mf-1c	Cello	sordina	
Vc+SP-ord-C4-mf-1c	Cello	piombo	

Table 2 Full list of audio stimuli (2/2). In each instrument, a blank space in the rightmost column denotes the ordinary playing technique (ordinario).

name	instrument	mute	playing technique
Vn-art-harm-G5-mf-4c	Violin		artificial harmonic
Vn-legno-batt-C4-mf-4c	Violin		col legno battuto
Vn-nonvib-C4-mf-4c	Violin		nonvibrato
Vn-ord-C4-mf-4c	Violin		
Vn-pizz-bartok-C4-ff-4c	Violin		pizzicato Bartók
Vn-pizz-lv-C4-mf-4c	Violin	pizzicato & *laissez vibrer*	
Vn-pizz-sec-C4-mf-4c	Violin	pizzicato secco	
Vn-pont-C4-mf-4c	Violin		sul ponticello
Vn+S-ord-C4-mf-4c	Violin	sordina	
Vn+SP-ord-C4-mf-4c	Violin	piombo	

Figure 2 Matrix $\widetilde{\mathbf{A}}_{\mathcal{I}}$ of perceived timbre similarities between instruments from our dataset of $N = 78$. samples. Darker shades indicate higher frequencies of co-occurrence in the cluster graph $\mathcal{G}_0$ (see Equation 1), obtained from a consensus of $K = 31$ subjects. The rows and columns in $\widetilde{\mathbf{A}}_{\mathcal{I}}$ were re-arranged according to Ward's minimum variance method. Observe that the block diagonal structure in $\widetilde{\mathbf{A}}_{\mathcal{I}}$ reflects an organological taxonomy of musical instruments: woodwinds, brass, and strings. See data collection section for details, including the meaning of each instrument abbreviation.

Figure 3 Matrix $\widetilde{\mathbf{A}}_{\mathcal{T}}$ of perceived timbre similarities between playing techniques from our dataset of $N = 78$ samples. Darker shades indicate higher frequencies of co-occurrence in the cluster graph $\mathcal{G}_0$ (see Equation 2), obtained from a consensus of $K = 31$ subjects. The rows and columns in $\widetilde{\mathbf{A}}_{\mathcal{T}}$ were re-arranged according to Ward's minimum variance method. See data collection section for details,. See Tables 1 and 2 for an English description of each playing technique abbreviation.

Figure 4 Six playing techniques of the violin. Subfigures: (a) ordinario (with vibrato), (b) nonvibrato, (c) tremolo, (d) sforzando, (e) pizzicato (*laissez vibrer*, i.e., let ring), and (f) staccato. In each subfigure, the top image shows the wavelet scalogram as a function of time t (in seconds) and frequency λ (in Hertz). Conversely, the bottom image shows the average time–frequency scattering coefficients, as a function of temporal modulation rate α (in Hertz) and frequential modulation scale β (in cycles per octave). Darker shades denote greater values of acoustic energy. See methods section for details.

Figure 5 Six playing techniques of the flute. Subfigures: (a) ordinario, (b) trill $(C_4 - D_4)$, (c) interference (play $C^\sharp_4$ while singing C_4), (d) sforzando, (e) key click, (f) staccato. In each subfigure, the top image shows the wavelet scalogram as a function of time t (in seconds) and frequency λ (in Hertz). Conversely, the bottom image shows the average time–frequency scattering coefficients, as a function of temporal modulation rate α (in Hertz) and frequential modulation scale β (in cycles per octave). Darker shades denote greater values of acoustic energy. See methods section for details.

Figure 6 Impact of different processing architecture or protocol designs. For each condition, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The performance achieved for each reference clustering is depicted by a lozenge whose color is chosen arbitrarily but consistently across conditions. See discussion section for details.