Extended playing techniques: The next milestone in musical instrument recognition

Vincent Lostanlen
New York University
35 W 4th St
New York, NY, USA 10014
vincent.lostanlen@nyu.edu

Joakim Andén
Flatiron Institute
162 5th Ave
New York, NY, USA 10010
janden@flatironinstitute.edu

C. El Hajj, M. Lagrange
École Centrale de Nantes, CNRS
1, rue de la Noë
44321 Nantes, France 43017-6221
mathieu.lagrange@cnrs.fr

ABSTRACT

The expressive variability in producing a musical note conveys information essential to the modeling of orchestration and style. As such, it plays a crucial role in computer-assisted browsing of massive digital music corpora. Yet, although the automatic recognition of a musical instrument from the recording of a single "ordinary" note is considered a solved problem, the ability of a computer to precisely identify instrumental playing techniques (IPTs) remains largely underdeveloped. We conduct a benchmark of machine listening systems for query-by-example browsing among 143 extended IPTs for 16 instruments, amounting to 469 triplets of instrument, mute, and technique. We identify and discuss three necessary conditions for significantly outperforming the traditional mel-frequency cepstral coefficient (MFCC) baseline: the addition of second-order scattering coefficients to account for amplitude modulation; the incorporation of long-range temporal dependencies; and post-processing using large-margin nearest neighbors (LMNN), a supervised metric learning method that reduces intra-class variability in feature space. Evaluating this system on the Studio On Line (SOL) dataset, we obtain a precision at rank t of 99.7% for instrument recognition (baseline at 89.0%) and of 61.0% for IPT recognition (baseline at 44.5%). We interpret this gain through a qualitative assessment of practical usability and visualization using nonlinear dimensionality reduction.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

ACM proceedings, LaTeX, text tagging

The source code to reproduce the experiments of this paper is made available at: https://www.github.com/mathieulagrange/dlfm2018

DLfM, Sep. 2018, Paris, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference format:
Vincent Lostanlen, Joakim Andén, and C. El Hajj, M. Lagrange. 2018. Extended playing techniques: The next milestone in musical instrument recognition. In *Proceedings of DLfM, Paris, France, Sep. 2018,* 11 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

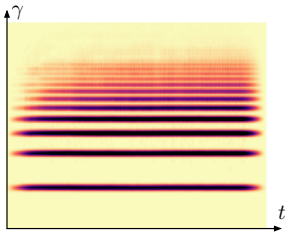
1 INTRODUCTION

Mathieu: fix ccs and keywords

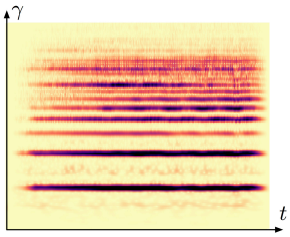
The gradual diversification of the timbral palette in Western classical music since the dawn of the 20th century is reflected in five concurrent trends: the addition of new instruments to the symphonic instrumentarium, either by technological inventions (e.g. theremin) or importation from non-Western musical cultures (e.g. marimba) [54, epilogue]; the creation of novel instrumental associations, as epitomized by *Klangfarbenmelodie* [55, chapter 22]; the temporary alteration of resonant properties through mutes and other "preparations" [19]; a more systematic usage of extended instrumental techniques, such as artificial harmonics, *col legno batutto*, or flutter tonguing [33, chapter 11]; and the resort to electronics and digital audio effects [65]. The first of these trends has somewhat stalled. To this day, most Western composers rely on an acoustic instrumentarium that is only marginally different from the one that was available in the Late Romantic period. Nevertheless, the remaining trends in timbral diversification have been adopted on a massive scale in post-war contemporary music. In particular, an increased concern for the concept of musical gesture [25] has liberated many unconventional instrumental techniques from their figurativistic connotations, thus making the so-called "ordinary" playing style merely one of many compositional – and improvisational – options.

Far from being exclusive to contemporary music, extended playing techniques are also commonly found in oral tradition; in some cases, they even stand out as a distinctive component of musical style. Four well-known examples are the snap pizzicato ("slap") of the upright bass in rockabilly, the growl of the tenor saxophone in rock'n'roll, the shuffle stroke of the violin ("fiddle") in Irish folklore, and the glissando of the clarinet in Klezmer music. Consequently, the organology (the instrumental *what?*) of a recording, as opposed to its chironomics (the gestural *how?*), is a poor organizing principle for browsing and recommendation in large music databases.

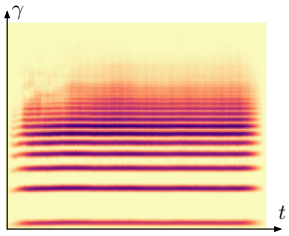
Yet, past research in music information retrieval (MIR), and especially in machine listening, rarely acknowledges the benefits of integrating the influence of performer gesture into a coherent taxonomy of musical instrument sounds. Instead, gesture is often framed as a spurious form of intra-class variability between instruments without delving into its interdependencies with pitch and intensity. In other works, it is conversely used as a probe for the



(a) Trumpet note (*ordinario*).

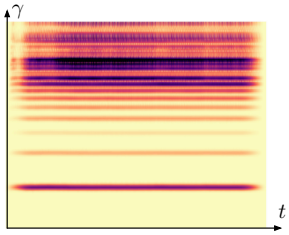


(b) Instrument (violin).

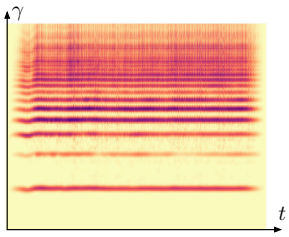


(c) Pitch (G3).

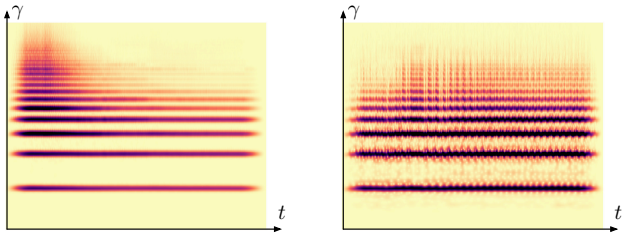
(d) Intensity (*pianissimo*).



(e) Mute (*harmon*).



(f) Tone quality (brassy).



acoustical study of a given instrument without emphasis on the broader picture of orchestral diversity.

One major cause of this gap in research is the difficulty of collecting and annotating data for contemporary instrumental techniques. Fortunately, this obstacle has recently been overcome, owing to the creation of databases of instrumental samples for music orchestration in spectral music [44]. In this work, we capitalize on the availability of this data to formulate a new line of research in MIR, namely the joint retrieval of organological ("*what* instrument is being played in this recording?") and chironomical information ("*how* is the musician producing sound?"), while remaining invariant to other factors of variability deliberately regarded as contextual. These include at what pitch and intensity the music was recorded, but also where, when, why, by whom, and for whom it was created.

Figure 1a shows the constant-Q wavelet scalogram (i.e. the complex modulus of the constant-Q wavelet transform) of a trumpet musical note, as played with an ordinary technique. Unlike most existing publications on instrument classification (e.g. 1a vs. 1b), which exclusively focus on intra-class variability due to pitch (Figure 1c) and intensity (Figure 1d), and mute (1e), this work aims to also account for the presence of instrumental playing techniques (IPTs), such as changes in tone quality (Figure 1f), attack (Figure 1g), tonguing (Figure 1h), and articulation (Figure 1e). These factors are considered either as intra-class variability, for the instrument recognition task, or as inter-class variability, for the IPT recognition task. The analysis of IPTs whose definition involves more than a single musical event, such as phrasing (Figure 1j), is beyond the scope of this paper. Joakim: What about 1h and 1i?

Section 2 reviews the existing literature on the topic. Section 3 defines taxonomies of instruments and gestures from which the IPT classification task is derived. Section 4 describes how two topics in machine listening, namely characterization of amplitude modulation and incorporation of supervised metric learning, are relevant to address this task. Section 5 reports the results from an IPT classification benchmark on the Studio On Line (SOL) dataset.

2 RELATED WORK

This section reviews recent MIR literature on the audio analysis of IPTs with a focus on the datasets available for the various classification tasks considered.

2.1 Isolated note instrument classification

The earliest works on musical instrument recognition restricted their scope to individual notes played with an ordinary technique, eliminating most factors of intra-class variability due to the performer [7, 12, 21, 28, 31, 45, 61]. These results were obtained on datasets such as MUMS [51], MIS,[1] RWC [26], and samples from the Philharmonia Orchestra.[2] This line of work culminated with the development of a support vector machine classifier trained on spectrotemporal receptive fields (STRF), which are idealized computational models of neurophysiological responses in the central auditory system [15]. Not only did this classifier attain a near-perfect mean accuracy of 98.7% on the RWC dataset, but the confusion matrix of its predictions was close to that human listeners [53].

[1] http://theremin.music.uiowa.edu/MIS.html
[2] http://www.philharmonia.co.uk/explore/sound_samples

Therefore, supervised classification of instruments from recordings of ordinary notes could arguably be considered a solved problem; we refer to [9] for a recent review of the state of the art.

2.2 Solo instrument classification

A straightforward extension of the problem above is the classification of solo phrases, encompassing some variability in melody [34], for which the accuracy of STRF models is around 80% [52]. Since the Western tradition of solo music is essentially limited to a narrow range of instruments (e.g. piano, classical guitar, violin) and genres (sonatas, contemporary, free jazz, folk), datasets of solo phrases, such as solosDb [30], are exposed to strong biases. This issue is partially mitigated by the recent surge of multitrack datasets, such as MedleyDB [10], which has spurred a renewed interest in single-label instrument classification [63]. In addition, the cross-collection evaluation methodology [36] reduces the risk of overfitting caused by the relative homogeneity of artists and recording conditions in these small datasets [11]. To date, the best classifiers of solo recordings is are the joint time-frequency scattering transform [1] and the spiral convolutional network [39] trained on the Medley-solos-DB dataset [38], i.e., a cross-collection dataset which aggregates MedleyDB and solosDb following the procedure of [20]. We refer to [27] for a recent review of the state of the art.

2.3 Multilabel classification in polyphonic mixtures

Because most publicly released musical recordings are polyphonic, the generic formulation of instrument recognition as a multilabel classification task is the most relevant for many end-user applications [13, 46]. However, it suffers from two methodological caveats. First, polyphonic instrumentation is not independent from other attributes, such as geographical origin, genre, or key. Second, the inter-rater agreement decreases with the number of overlapping sources [23, chapter 6]. These problems are all the more troublesome since there is currently no annotated dataset of polyphonic recordings diverse enough to be devoid of artist bias. The Open-MIC initiative, from the newly created Community for Open and Sustainable Music and Information Research (COSMIR), is working to mitigate these issues in the near future [47]. We refer to [29] for a recent review of the state of the art.

2.4 Solo playing technique classification

Finally, there is a growing interest for studying the role of the performer in musical acoustics, from the perspective of both sound production and perception. Apart from its interest in audio signal processing, this topic is connected to other disciplines, such as biomechanics and gestural interfaces [49]. The majority of the literature focuses on the range of IPTs afforded by a single instrument. Recent examples include clarinet [41], percussion [57], piano [8], guitar [14, 22, 56], violin [64], cello [17, chapter 6], and erhu [62]. Some publications frame timbral similarity in a polyphonic setting, yet do so according to a purely perceptual definition of timbre – with continuous attributes such as brightness, warmth, dullness, roughness, and so forth – without connecting these attributes to the discrete latent space of IPTs (i.e., through a finite set of instructions, readily interpretable by the performer) [4]. We refer to [35] for a recent review of the state of the art.

In the following, we define the task of retrieving musical timbre parameters across a range of instruments found in the symphonic orchestra. These parameters are explicitly defined in terms of sound production rather than by means of perceptual definitions.

3 TASKS

In this section, we define a taxonomy of musical instruments and another for musical gestures, which are then used for defining the instrument and IPT query-by-example tasks. We also describe the dataset of instrument samples used in our benchmark.

3.1 Taxonomies

The Hornbostel-Sachs taxonomy (H-S) organizes musical instruments only according to their physical characteristics and purposefully ignores sociohistorical background [50]. Since it offers an unequivocal way of describing any acoustic instrument without any prior knowledge of its applicable IPTs, it serves as a *lingua franca* in ethnomusicology and museology, especially for ancient or rare instruments which may lack available informants. The classification of the violin in H-S (321.322-71), as depicted in Figure 2, additionally encompasses the viola and the cello. The reason is that these three instruments possess a common morphology. Indeed, both violin and viola are usually played under the jaw and the cello is held between the knees, these differences in performer posture are ignored by the H-S classification. Accounting for these differences begs to refine H-S by means a vernacular taxonomy. Most instrument taxonomies in music signal processing, including MedleyDB [10] and AudioSet [24], adopt the vernacular level rather than conflating all instruments belonging to the same H-S class. A further refinement includes potential alterations to the manufactured instrument – permanent or temporary, at the time scale one or several notes – that affect its resonant properties, e.g., mutes and other preparations [19]. The only node in the MedleyDB taxonomy which reaches this level of granularity is *tack piano* [10] . In this work, we will not consider variability due to the presence of mutes as discriminative, both for musical instruments and IPTs.

Unlike musical instruments, which are amenable to a hierarchical taxonomy of resonating objects, IPTs result from a complex synchronization between multiple gestures, potentially involving both hands, arms, diaphragm, vocal tract, and sometimes the whole body. As a result, they cannot be trivially incorporated into H-S, or indeed any tree-like structure [32]. Instead, an IPT is described by a finite collection of categories, each belonging to a different "namespace." Figure 3 illustrates such namespaces for the case of the violin. It therefore appears that, rather than aiming for a mere increase in granularity with respect to H-S, a coherent research program around extended playing techniques should formulate them as belonging to a meronomy, i.e., a modular entanglement of part-whole relationships, in the fashion of the Visipedia initiative in computer vision [6]. In recent years, some works have attempted to lay the foundations of such a modular approach, with the aim of making H-S relevant to contemporary music creation [42, 60]. However, such considerations are still in large part speculative

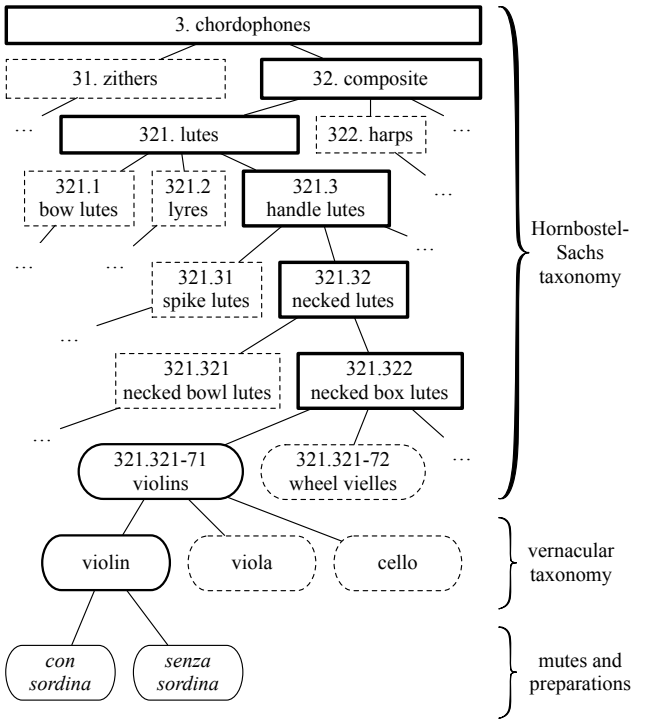


Figure 2: Taxonomy of musical instruments.

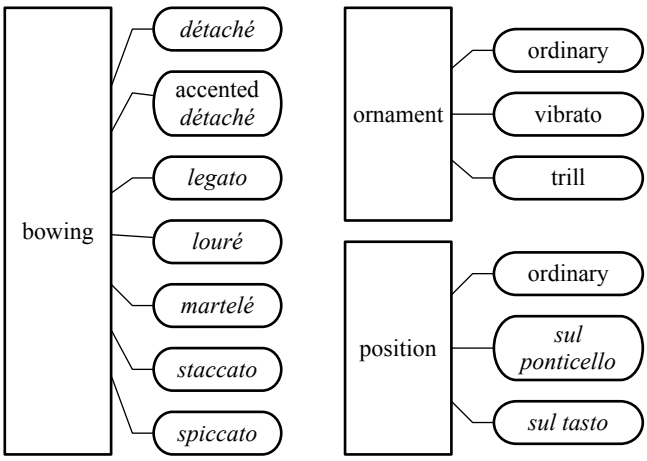


Figure 3: Namespaces of violin playing techniques.

and offer no definitive procedure for evaluating, let alone training, information retrieval systems.

3.2 Application setting and evaluation

In what follows, we adopt a middle ground position between the two aforementioned approaches: neither a supervised classifier (as in a hierarchical taxonomy), nor a caption generator (as in a meronomy), our system is a query-by-example search engine in a large database of isolated notes. Joakim: This doesn't make any sense to me. We're not "adopting a middle ground". We're effectively using a taxonomy for both instruments and IPTs, but we're not considering the classification problem and instead looking at the retrieval problem. The ground truth could be something else but in our case, we are using the taxonomy. Given a query recording $\boldsymbol{x}(t)$, such a system retrieves a small number k of recordings judged similar to the query. In our system, we implement this using a k-nearest neighbors (k-NN) algorithm. The nearest neighbor search is not performed in the raw waveform domain of $\boldsymbol{x}(t)$, but in a feature space of translation-invariant, spectrotemporal descriptors. In what follows, we use averaged mel-frequency cepstral coefficients (MFCCs) as a baseline, which we extend using second-order scattering coefficients [3, 43]. Joakim: Averaged MFCCs? The baseline k-NN algorithm is applied using the standard Euclidean distance in feature space. To improve performance, we also apply it using a weighted Euclidean distance with a learned weight matrix.

In the context of music creation, the query $\boldsymbol{x}(t)$ may be an instrumental or vocal sketch, a sound event recorded from the environment, a computer-generated waveform, or any mixture of the above [44]. Upon inspecting the recordings returned by the search engine, the composer may decide to retain one of the retrieved notes. Its attributes (pitch, intensity, and playing technique) are then readily available for inclusion in the musical score.

Faithfully evaluating such a system is a difficult procedure, and ultimately depends on its practical usability as judged by the composer. Nevertheless, a useful quantitative metric for this task is the precision at k (P@k) of the test set with respect to the training set, either under an instrument taxonomy and an IPT taxonomy. This metric is defined as the proportion of "correct" recordings returned for a given query, averaged over all queries in the test set. For our purposes, a returned recording is correct if it is of the same class as the query for a specific taxonomy. In all subsequent experiments, we report P@k for the number of retrieved items $k = 5$.

3.3 Studio On Line dataset (SOL)

The Studio On Line dataset (SOL) was recorded at IRCAM in 2002 and is freely downloadable as part of the Orchids software for computer-assisted orchestration.[3] It comprises 16 musical instruments playing 25444 isolated notes in total. The distribution of these notes, shown in Figure 4, spans the full combinatorial diversity of intensities, pitches, preparations (i.e., mutes), and all applicable playing techniques. The distribution of playing techniques – the most common of which are shown in Figure 5 – is heavy-tailed (average 178, standard deviation 429). Joakim: What does this mean? What distribution? This is because some playing techniques are shared between many instruments (e.g., *tremolo*) whereas other are instrument-specific (e.g., *xylophonic*, which is specific to the harp). The SOL dataset has 143 IPTs in total, and 469 applicable instrument-mute-technique triplets. Joakim: Why is this last number important? Also, since we don't seem to consider "mutes" in any taxonomy, why mention them here?

4 METHODS

In this section, we describe the scattering transform used to capture amplitude modulation structure and supervised metric learning

[3] http://forumnet.ircam.fr/product/orchids-en/

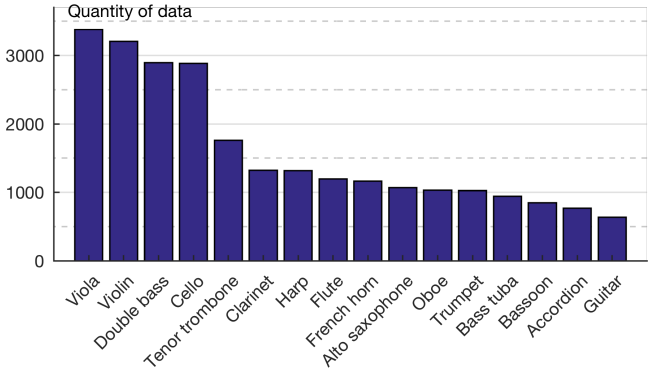


Figure 4: Instruments in the SOL dataset.

which constructs a similarity measure suited for our query-by-example task.

4.1 Scattering transform

The scattering transform is a cascade of constant-Q wavelet transforms alternated with modulus operators [3, 43]. Given a signal $\boldsymbol{x}(t)$, its first layer outputs the first-order scattering coefficients $\mathrm{S}_1\boldsymbol{x}(\lambda_1, t)$, which captures the intensity of $\boldsymbol{x}(t)$ at frequency λ_1. Its frequency resolution is logarithmic in λ_1 and is sampled using $Q_1 = 12$ bins per octave. The second layer of the cascade yields the second-order scattering coefficients $\mathrm{S}_2\boldsymbol{x}(\lambda_1, \lambda_2, t)$, which extract amplitude modulation at frequency λ_2 in the subband of $\boldsymbol{x}(t)$ at frequency λ_1. Both first- and second-order coefficients are averaged in time over intervals of size T. The modulation frequencies λ_2 are logarithmically spaced with $Q_2 = 1$ bin per octave. In the following, we denote by $\mathrm{S}\boldsymbol{x}(\lambda, t)$ the concatenation of all scattering coefficients, where λ corresponds to either a single λ_1 for first-order coefficients or a pair (λ_1, λ_2) for second-order coefficients.

The first-order scattering coefficients are equivalent to the mel-frequency spectrogram which forms a basis for MFCCs [3]. Second-order coefficients, on the other hand, characterize common non-stationary structures in sound production, such as tremolo, vibrato, and dissonance [2, section 4]. As a result, these coefficients are better suited to model extended IPTs. We refer to [3] an introduction on scattering transforms for audio signals and to [37, sections 3.2 and 4.5] for a discussion on its application to musical instrument classification in solo recordings and its connections to STRFs.

To match a decibel-like perception of loudness, we apply the adaptive, quasi-logarithmic compression

$$\widetilde{\mathrm{S}}\boldsymbol{x}_i(\lambda, t) = \log\left(1 + \frac{\mathrm{S}\boldsymbol{x}_i(\lambda, t)}{\varepsilon \times \boldsymbol{\mu}(\lambda)}\right) \quad (1)$$

where $\varepsilon = 10^{-3}$ and $\boldsymbol{\mu}(\lambda)$ is the median of $\mathrm{S}\boldsymbol{x}_i(\lambda, t)$ across t and i.

4.2 Metric learning

Linear metric learning algorithms construct a matrix $\mathbf{L}$ such that the weighted distance

$$\mathrm{D}_{\mathbf{L}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\mathbf{L}(\widetilde{\mathrm{S}}\boldsymbol{x}_i - \widetilde{\mathrm{S}}\boldsymbol{x}_j)\|_2 \quad (2)$$

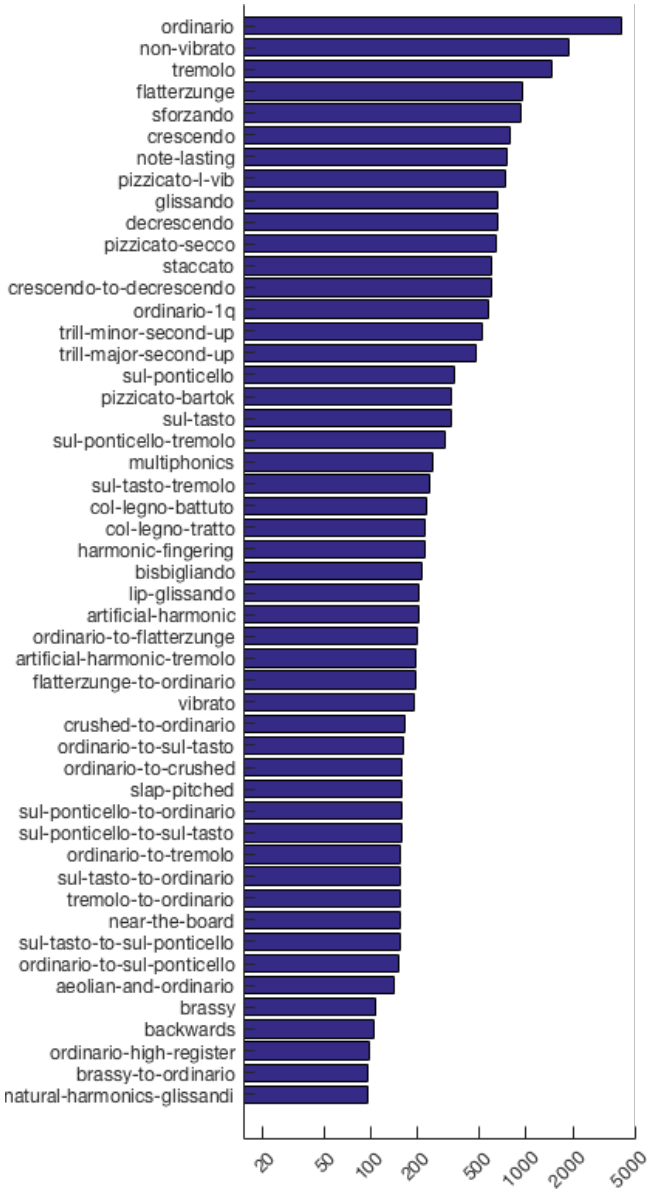


Figure 5: The 50 most common IPTs in the SOL dataset.

between all pairs of samples $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ optimizes some objective function. We refer to [5] for a review of the state of the art. In the following, we shall consider the large-margin nearest neigbors (LMNN) algorithm. It attempts to construct L such that for every signal $\boldsymbol{x}_i(t)$ the distance $\mathrm{D}_{\mathbf{L}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to $\boldsymbol{x}_j(t)$, one of its k nearest neighbors, is small if $\boldsymbol{x}_i(t)$ and $\boldsymbol{x}_j(t)$ belong to the same class and large otherwise. The matrix $\mathbf{L}$ is obtained by applying the special-purpose solver of [59, appendix A]. In subsequent experiments, disabling LMNN is equivalent to setting $\mathbf{L}$ to the identity matrix, which yields the standard Euclidean distance on the scattering coefficients $\widetilde{\mathrm{S}}\boldsymbol{x}(\lambda, t)$.

Compared to a class-wise generative model, such a Gaussian mixture model, a global linear model ensures some robustness to

minor alterations of the taxonomy, which is important in the context of IPT. For example, one performer's *slide* is another's *glissando*. Joakim: I don't understand this argument. Why does a global linear model ensure robustness? Is it because of fewer parameters so less risk of overfitting? A major drawback of LMNN is its dependency on the standard Euclidean distance for determining nearest neighbors [48]. However, this is alleviated for scattering coefficients, since the scattering transform $\mathbf{S}\boldsymbol{x}(t, \lambda)$ is Lipschitz continuous to elastic deformation in the signal $\boldsymbol{x}(t)$ [43, Theorem 2.16]. In other words, the Euclidean distance between the scattering transform of $\boldsymbol{x}(t)$ and a deformed version of the same signal is bounded by the extent of that deformation.

5 EXPERIMENTAL EVALUATION

In this section, we study a query-by-example browsing system for the SOL dataset based on nearest neigbors. We discuss how the performance of the system is affected by the choice of feature (MFCCs or scattering transforms) and distance (Euclidean or LMNN), both quantitatively and qualitatively. Finally, we visualize the two feature spaces using non-linear dimensionality reduction.

5.1 Instrument recognition

In the task of instrument recognition, we provide a query $\boldsymbol{x}(t)$ and the system retrieves k recordings $\boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_k(t)$. We consider a retrieved recording to be relevant to the query if it corresponds to the same instrument, regardless of pitch, intensity, mute, and IPT. We therefore apply the LMNN with instruments as class labels. This lets us compute the precision at rank 5 (P@5) for a system by counting the number of relevant recordings for each query.

We compare scattering features to a baseline of MFCCs, defined as the 13 lowest coefficients of the discrete cosine transform (DCT) applied to the logarithm of the 40-band mel-frequency spectrum. For the scattering transform, we vary the maximum time scale T of amplitude modulation from 25 ms to 1 s. In the case of the MFCCs, $T = 25$ ms corresponds to the inverse of the lowest audible frequency ($T^{-1} = 40$ Hz). Therefore, increasing the frame duration T has no effect on the value of the MFCCs, because the mel-spectrogram is equivalent to a local averaging of the scalogram at the time scale T, leaving unchanged the global averaging of $\mathbf{S}\boldsymbol{x}(\lambda, t)$ at the time scale of whole musical notes [2, section II.B]. Joakim: I don't understand this part. How does changing T not affect the MFCCs? If it's larger, we average more, no? There are some details missing here, it seems. Mathieu: ce n'est pas la meme chose de calculer des mfccs a 25 ms et de moyenner sur 30 sec, que de faire une dct sur 30 sec, mais en pratique je n'ai jamais observÃľ de difference, voir figure 6. peut etre nuancer le propos

The left column of Figure 6 summarizes our results. MFCCs reach a relatively high P@5 of 89%. Keeping all 40 DCT coefficients rather than the lowest 13 brings P@5 down to 84%, because the DCT coefficients are most affected by spurious factors of intra-class variability, such as pitch and spectral flatness [37, subsection 2.3.3].

At the smallest time scale $T = 25$ ms, the scattering transform reaches a P@5 of 89%, thus matching the performance of the MFCCs. This is expected since there is little amplitude modulation below this scale, corresponding to λ_2 over 40 Hz, so the scattering transform is dominated by the first order, which is equivalent to MFCCs [3]. Moreover, disabling median renormalization degrades P@5 down to 84%, while disabling logarithmic compression altogether degrades it to 76%. This is consistent with [40], which applies scattering transform to a query-by-example retrieval task for acoustic scenes.

On one hand, replacing the canonical Euclidean distance by a distance learned by LMNN marginally improves P@5 for the MFCC baseline, from 89.3% to 90.0%. Applying LMNN to scattering features, on the other hand, significantly improves their performance with respect to the Euclidean distance, from 89.1% to 98.0%.

The gain in precision afforded by scattering coefficients over MFCCs could simply be caused by a higher number of dimensions which could give more flexibility to the LMNN algorithm. To refute this hypothesis, we supplement the 13 coefficients resulting from a global averaging at the time scale of full musical notes by higher-order summary statistics, namely polynomial combination of those features of degrees 2 and 3. Instrument retrieval in the resulting feature space, whose dimension (494) is comparable to the number of scattering coefficients, has a P@5 of 91%, i.e., slightly above the baseline. Therefore, it is more likely the multiresolution structure of scattering coefficients, rather than its dimensionality, that causes a strong boost in performance. Joakim: I don't understand the objection or the argument here. How could the better performance be caused by a "higher number of dimensions"? If I add a bunch of coefficients to MFCCs that are just Gaussian white noise, does that improve performance? What if I copy it several times over? How does our particular dimensionality increasing strategy refute that claim? We haven't looked at all possible ways of increasing the dimensions here. Mathieu: l'idee ici est de montrer que ce n'est pas le lmnn avec un nombre de degrÃľ de libertÃľ suffisant qui overfit. Si le nombre de features est egal au nombre de points le probleme est trivial. J'ai reformuler un peu le par. Peut etre a liser un peu

Finally, increasing T from 25 ms up to 1 s – i.e., including all amplitude modulations between 1 Hz and 40 Hz – brings LMNN to a near-perfect P@5 of 99.7%. Not only does this result confirm that straighforward techniques in audio signal processing (here, wavelet scattering and metric learning) are sufficient to retrieve the instrument from a single ordinary note, it also demonstrates that the results remain satisfactory despite large intra-class variability in terms of pitch, intensity, usage of mutes, and extended IPTs. In other words, the monophonic recognition of Western instruments is, all things considered, indeed a solved problem.

5.2 Playing technique recognition

The situation is different when considering IPT, rather than instrument, as the reference for evaluating the query-by-example system. In this setting, a retrieved item is considered relevant if and only if it shares the same IPT as the query, regardless of instrument, mute, pitch, or dynamics. Therefore, we apply the LMNN with IPTs instead of instruments as class labels, yielding a different distance function optimized to distinguish playing techniques. Joakim: Need to say before that we train it with instruments as labels. Mathieu: done The right column if Figure 6 summarizes our results. The MFCC baseline has a low P@5 of 44.5%, indicating that its coarse description of the short-term spectral envelope is not sufficient to model acoustic similarity in IPT. Perhaps more surprisingly, we find that optimal performance is only achieved by

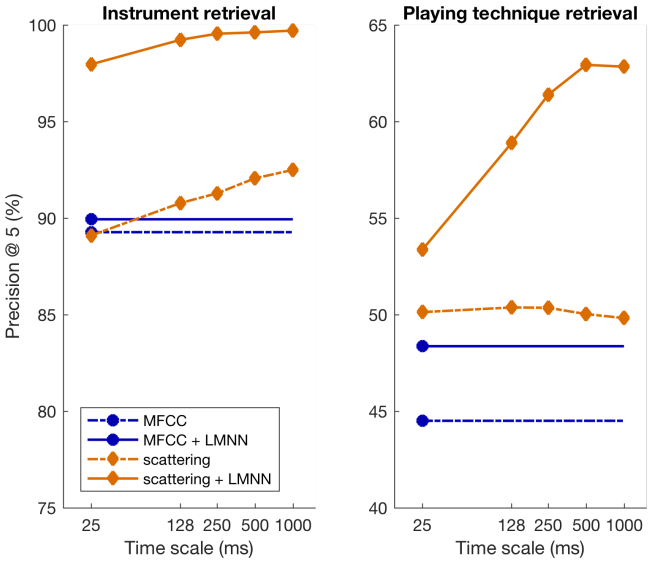


Figure 6: Summary of results on the SOL dataset.

combining all proposed improvenments: log-scattering coefficients with median renormalization, $T = 500$ ms, and LMNN. This yields a P@5 of 63.0%. Indeed, an ablation study of that system reveals that, all other things being equal, reducing T to 25 ms brings the P@5 to 53.3%, disabling LMNN reduces it to 50.0%, and replacing scattering coefficients by MFCCs yields 48.4%. This result contrasts with the instrument recognition setting: whereas the improvements brought by the three aforementioned modifications are approximately additive in P@5 for musical instruments, they interact in a super-additive manner for IPTs. In particular, it appears that increasing T above 25 ms is only beneficial to IPT similarity retrieval if combined with LMNN.

5.3 Qualitative error analysis

For demonstration purposes, we select an audio recording $\boldsymbol{x}(t)$ to query two versions of the proposed query-by-example system. The first version uses MFCCs with $T = 25$ ms and LMNN; it has a P@5 of 90.0% for instrument retrieval and 48.4% for IPT retrieval. The second version uses scattering coefficients with $T = 1$ s, logarithmic transformation with median renormalization (see Equation 1), and LMNN; it has a P@5 of 99.7% for instrument retrieval and 63.0% for IPT retrieval. Both versions adopt IPT labels as reference for training LMNN. Joakim: If so, we shouldn't cite the instrument retrieval numbers above, since that concerns a different system. Mathieu: je suis d'accord. The main difference between the two versions is the choice of spectrotemporal features.

Figure 7 shows the constant-Q scalograms of the five retrieved items for both versions of the system as queried by the same audio signal $\boldsymbol{x}(t)$: a violin note from the SOL dataset, played with ordinary playing technique on the G string with pitch G4 and *mf* dynamics. Both versions correctly retrieve five violin notes which vary from the query in pitch, dynamics, string, and use of mute. Therefore, both systems have an instrument retrieval P@5 of 100% for this query. However, although the scattering-based version is also 100%

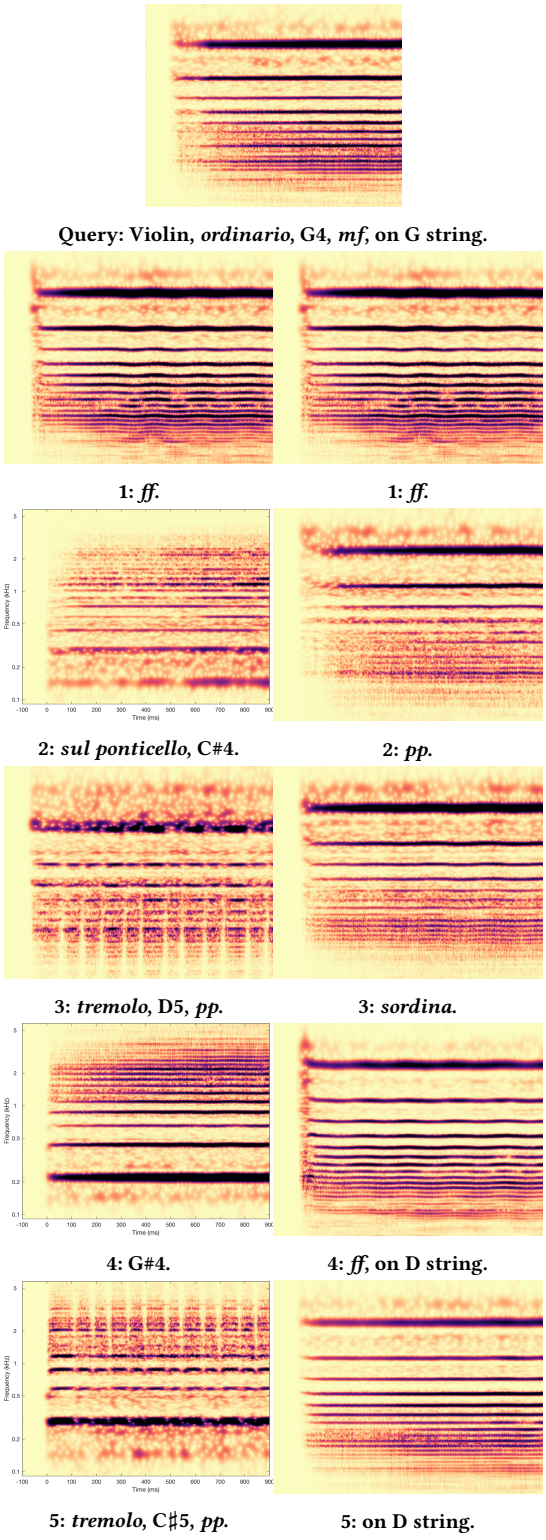


Figure 7: Five nearest neighbors of the same query (a violin note with ordinary playing technique, at pitch G4, *mf* dynamics, played on the G string), as retrieved by two different versions of our system: with MFCC features (left) and with scattering transform features (right). The captions denote the musical attribute(s) that differ from those of the query: mute, playing technique, pitch, and dynamics.

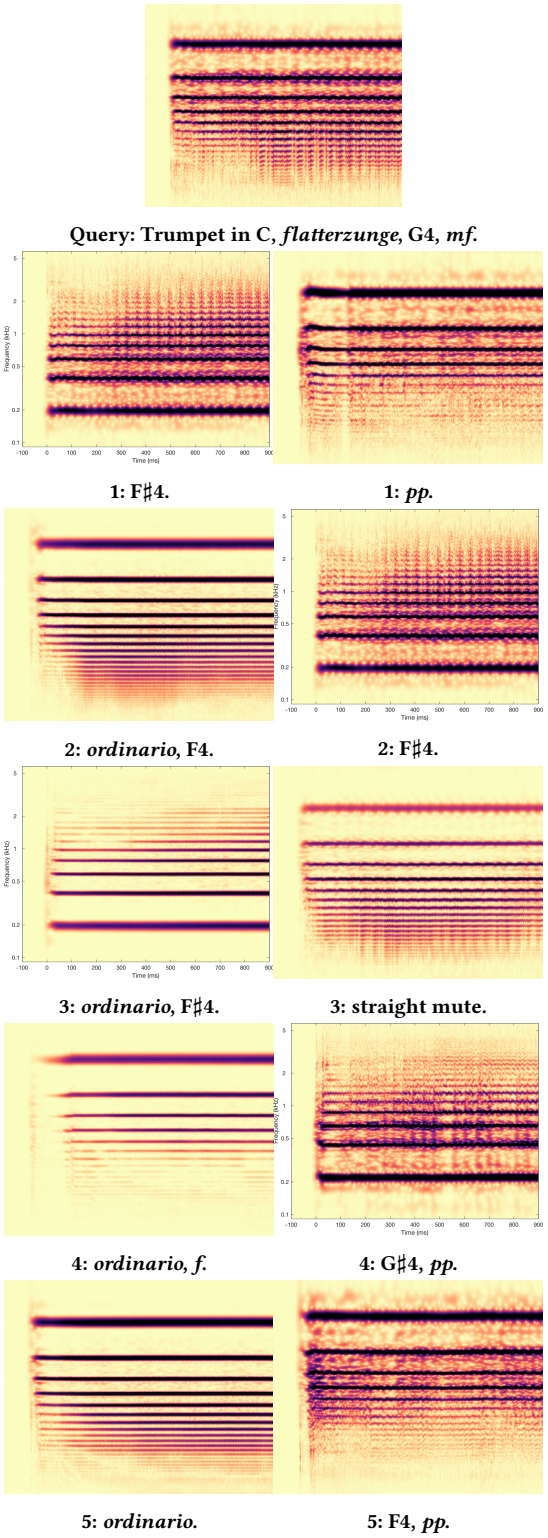


Figure 8: Five nearest neighbors of the same query (a trumpet note with *flatterzunge* technique, at pitch G4, *mf* dynamics), as retrieved by two different versions of our system: with MFCC features (left) and with scattering transform features (right). Joakim: Combine both figures into one big one?

correct in terms of IPT retrieval (i.e., it retrieves five *ordinario* notes), the MFCC-based version is only 40% correct. Indeed, three recordings exhibit on of the *tremolo* or *sul ponticello* playing techniques. We hypothesize that the confusion between *ordinario* and *tremolo* is caused by the presence of vibrato in the ordinary query since MFCCs cannot distinguish amplitude modulations (tremolo) from frequency modulations (vibrato) for the same modulation frequency [2]. These differences, however, are perceptually small and in some musical contexts vibrato and tremolo are used interchangeably.

The situation is different when querying both systems with recording $\boldsymbol{x}(t)$ exhibiting an extended rather than ordinary IPT. Figure 8 is analogous to Figure 7 but with a different audio query. The query is a trumpet note from the SOL dataset, played with the *flatterzunge* (flutter-tonguing) technique, pitch G4, and *mf* dynamics. Again, the scattering-based version retrieves five recordings with the same instrument (trumpet) and IPT (*flatterzunge*) as the query. In contrast, four out of the five items retrieved by the MFCC system have an *ordinario* IPT instead of *flatterzunge*. This shortcoming has direct implications on the usability of the MFCC query-by-example system for contemporary music creation. More generally, this system is less reliable when queried with extended IPTs.

Unlike instrument similarity, IPT similarity seems to depend on long-range temporal dependencies in the audio signal. In addition, it is not enough to capture the raw amplitude modulation provided by the second-order scattering coefficients. Instead, an adaptive layer on top of this is needed to extract the discriminative elements from those coefficients. Here, that layer consists of the LMNN metric learning algorithm, but other methods may work equally well.

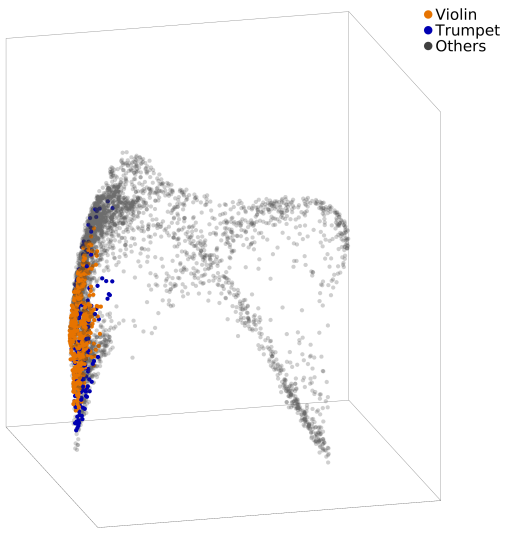
5.4 Feature space visualization

To visualize the feature space generated by MFCCs and scattering transforms, we embed them using diffusion maps. These embeddings preserve local distances while reducing dimensionality by forming a graph from those distances and calculating the eigenvectors of its graph Laplacian [18]. Diffusion maps have previously been used to successfully visualize scattering coefficients [16, 58].

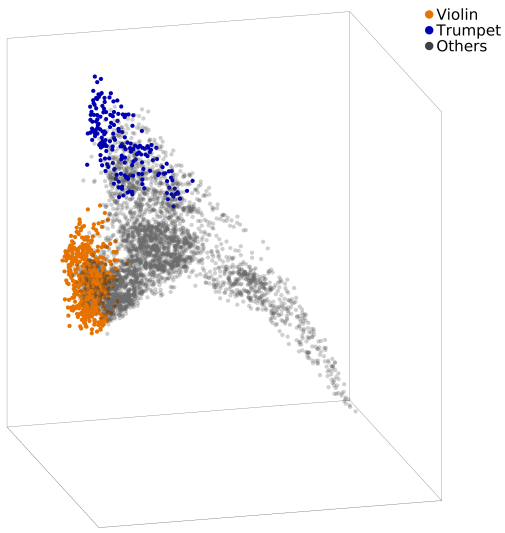
Figure 9 shows embeddings of MFCCs and scattering coefficients, both post-processed using LMNN, for different subsets of recordings. In Figure 9a, we see how the MFCCs fail to separate violin and trumpet notes for the *ordinario* playing technique. Scattering coefficients, on the other hand, successfully separate the instruments as seen in Figure 9b. Similarly, Figures 9c and 9d show how, restricted to bowed instruments (violin, viola, violoncello, and contrabass), MFCCs do not separate the *ordinario* from *tremolo* playing techniques, while scattering coefficients discriminates well. These visualizations provide motivation for our choice of scattering coefficients to represent single notes.

6 CONCLUSION

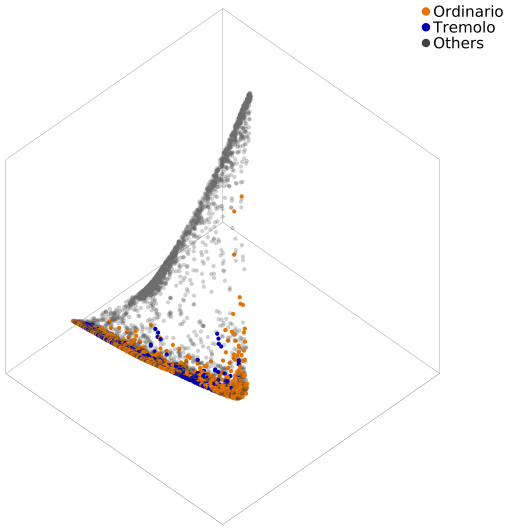
Whereas the MIR literature abounds on the topic of musical instrument recognition for so-called “ordinary” isolated notes and solo performances, little is known about the problem of retrieving the instrumental playing technique from an audio query within a fine-grained taxonomy. Yet the knowledge of IPT is a precious source of musical information, not only to characterize the physical interaction between player and instrument, but also in the realm



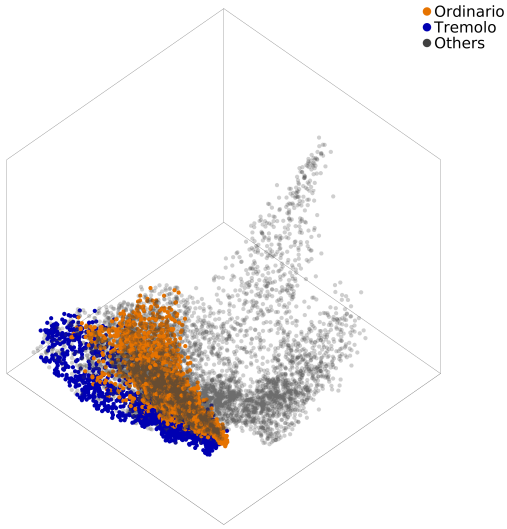
(a) Instrument embedding with MFCC.



(b) Instrument embedding with scattering transform.



(c) Playing technique embedding with MFCC.



(d) Playing technique embedding with scattering transform.

Figure 9: Diffusion maps produce low-dimensional embeddings of MFCC features (left) vs. scattering transform features (right). In the two top plots, each dot represents a different musical note, after restricting the SOL dataset to the *ordinario* playing technique of each of the 31 different instrument-mute couples. Blue (resp. orange) dots denote violin (resp. trumpet in C) notes, including notes played with a mute: *sordina* and *sordina piombo* (resp. *cup*, *harmon*, *straight*, and *wah*). In the two bottom plots, each dot corresponds to a different musical note, after restricting the SOL dataset to 4 bowed instruments (violin, viola, violoncello, and contrabass), and keeping all 38 applicable techniques. Blue (resp. orange) dots denote tremolo (resp. ordinary) notes. In both experiments, the time scales of both MFCC and scattering transform are set equal to $T = 1\,\mathrm{s}$, and features are post-processed by means of the large-margin nearest neighbor (LMNN) metric learning algorithm, using playing technique labels as reference for reducing intra-class neighboring distances.

of contemporary music creation. It also bears an interest for organizing digital libraries as a mid-level descriptor of musical style. To the best of our knowledge, this paper is the first to benchmark query-by-example MIR systems according to a large-vocabulary, multi-instrument IPT reference (143 classes) instead of an instrument reference. We find that this new task is considerably more challenging than musical instrument recognition as it amounts to characterizing spectrotemporal patterns at various scales and comparing them in a non-Euclidean way. Although the combination of methods presented here – wavelet scattering and large-margin nearest neighbors – outperforms the MFCC baseline, its accuracy on the SOL dataset certainly leaves room for future improvements. For example, we could replace the standard time scattering transform with the more discriminative joint time-frequency scattering transform [1].

The evaluation methodology presented here uses ground truth IPT labels to quantify the relevance of returned items. This approach is useful in that the labels are unambigious, but it might be too coarse to reflect practical use. Indeed, as it is often the case in MIR, some pairs of labels are subjectively more similar than others. For example, *slide* is evidently closer to *glissando* than to *pizzicato-bartok*. The collection of subjective ratings for IPT similarity, and its comparison with automated ratings, is left as future work. Another promising avenue of research is to formulate a structured prediction task for isolated musical notes, simultaneously estimating the pitch, dynamics, instrument, and IPT to construct a unified machine listening system, akin to a caption generator in computer vision.

ACKNOWLEDGMENTS

The authors wish to thank Philippe Brandeis, Étienne Graindorge, Stéphane Mallat, Adrien Mamou-Mani, and Yan Maresz for fruitful discussions on contemporary music creation as part of the TICEL research project (“Traité instrumental collaboratif en ligne”), Andrew Farnsworth and Grant Van Horn for fruitful discussions on Visipedia, and Katherine Crocker for helpful suggestions on the title of this article. This work is supported by the ERC InvariantClass grant 320959, the NSF award 1633259 (BIRDVOX), the Leon Levy Foundation, and a Google faculty award.

REFERENCES

[1] Joakim Andén, Vincent Lostanlen, and Stéphane Mallat. 2018. Classification with Joint Time-Frequency Scattering. (Jul 2018). arXiv:1807.08869

[2] Joakim Andén and Stéphane Mallat. 2012. Scattering representation of modulated sounds. In *Proc. DAFx*.

[3] Joakim Andén and Stéphane Mallat. 2014. Deep scattering spectrum. *IEEE Trans. Sig. Proc.* 62, 16 (2014), 4114–4128.

[4] Aurélien Antoine and Eduardo R. Miranda. 2018. Musical Acoustics, Timbre, and Computer-Aided Orchestration Challenges. In *Proc. ISMA*.

[5] Aurélien Bellet, Amaury Habrard, and Marc Sebban. 2013. A survey on metric learning for feature vectors and structured data. (2013). arXiv:1306.6709

[6] Serge Belongie and Pietro Perona. 2016. Visipedia circa 2015. *Pattern Recognition Letters* 72 (2016), 15–24.

[7] Emmanouil Benetos, Margarita Kotti, and Constantine Kotropoulos. 2006. Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection. In *Proc. IEEE ICASSP*.

[8] Michel Bernays and Caroline Traube. 2013. Expressive production of piano timbre: touch and playing techniques for timbre control in piano performance. In *Proc. SMC*.

[9] D.G. Bhalke, C.B. Rama Rao, and Dattatraya S. Bormane. 2016. Automatic musical instrument classification using fractional Fourier transform based-MFCC features and counter propagation neural network. *J. Intell. Inf. Syst.* 46, 3 (2016), 425–446.

[10] Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. 2014. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *Proc. ISMIR*.

[11] Dimitry Bogdanov, Alastair Porter, Perfecto Herrera Boyer, and Xavier Serra. 2016. Cross-collection evaluation for music classification tasks. In *Proc. ISMIR*.

[12] Judith C. Brown. 1999. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *J. Acoust. Soc. Am.* 105, 3 (1999), 1933–1941.

[13] Juan José Burred, Axel Robel, and Thomas Sikora. 2009. Polyphonic musical instrument recognition based on a dynamic model of the spectral envelope. In *Proc. IEEE ICASSP*. 173–176.

[14] Yuan-Ping Chen, Li Su, and Yi-Hsuan Yang. 2015. Electric Guitar Playing Technique Detection in Real-World Recording Based on F0 Sequence Pattern Recognition. In *Proc. ISMIR*.

[15] Taishih Chi, Powen Ru, and Shihab A. Shamma. 2005. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118, 2 (2005), 887–906.

[16] Václav Chudácek, Ronen Talmon, Joakim Andén, Stéphane Mallat, Ronald R. Coifman, et al. 2014. Low dimensional manifold embedding for scattering coefficients of intrapartum fetal heart rate variability. In *Proc. IEEE EMBC*. 6373–6376.

[17] Magdalena Chudy. 2016. *Discriminating music performers by timbre: On the relation between instrumental gesture, tone quality and perception in classical cello performance.* Ph.D. Dissertation. Queen Mary University of London.

[18] Ronald R. Coifman and Stéphane Lafon. 2006. Diffusion maps. *Appl. and Comput. Harmon. Anal.* 21, 1 (2006), 5–30.

[19] Tzenka Dianova. 2007. *John Cage's Prepared Piano: The Nuts and Bolts.* Ph.D. Dissertation. U. Auckland.

[20] Patrick J. Donnelly and John W. Sheppard. 2015. Cross-Dataset Validation of Feature Sets in Musical Instrument Classification. In *Proc. IEEE ICDMW*. 94–101.

[21] Antti Eronen and Anssi Klapuri. 2000. Musical instrument recognition using cepstral coefficients and temporal features. In *Proc. IEEE ICASSP*.

[22] Raphael Foulon, Pierre Roy, and François Pachet. 2013. Automatic classification of guitar playing modes. In *Proc. CMMR*. Springer.

[23] Ferdinand Fuhrmann. 2012. *Automatic musical instrument recognition from polyphonic music audio signals.* Ph.D. Dissertation. Universitat Pompeu Fabra.

[24] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, et al. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP*.

[25] Rolf Inge Godøy and Marc Leman. 2009. *Musical Gestures: Sound, Movement, and Meaning.* Taylor & Francis.

[26] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. 2003. RWC music database: music genre database and musical instrument sound database. (2003).

[27] Yoonchang Han, Jaehun Kim, and Kyogu Lee. 2017. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE Trans. Audio Speech Lang. Process.* 25, 1 (2017), 208–221.

[28] Perfecto Herrera Boyer, Geoffroy Peeters, and Shlomo Dubnov. 2003. Automatic classification of musical instrument sounds. *J. New Music Res.* 32, 1 (2003), 3–21.

[29] Eric Humphrey, Simon Durand, and Brian McFee. 2018. OpenMIC-2018: an open dataset for multiple instrument recognition. In *Proc. ISMIR*.

[30] Cyril Joder, Slim Essid, and Gaël Richard. 2009. Temporal integration for audio classification with application to musical instrument classification. *IEEE Trans. Audio Speech Lang. Process.* 17, 1 (2009), 174–186.

[31] Ian Kaminskyj and Tadeusz Czaszejko. 2005. Automatic recognition of isolated monophonic musical instrument sounds using kNNC. *J. Intell. Inf. Syst.* 24, 2-3 (2005), 199–221.

[32] Sefki Kolozali, Mathieu Barthet, György Fazekas, and Mark B. Sandler. 2011. Knowledge Representation Issues in Musical Instrument Ontology Design.. In *Proc. ISMIR*.

[33] Stefan Kostka. 2016. *Materials and Techniques of Post Tonal Music.* Taylor & Francis.

[34] A.G. Krishna and Thippur V. Sreenivas. 2004. Music instrument recognition: from isolated notes to solo phrases. In *Proc. IEEE ICASSP*.

[35] Marc Leman, Luc Nijs, and Nicola Di Stefano. 2017. *On the Role of the Hand in the Expression of Music.* Springer International Publishing, Cham, 175–192.

[36] Arie Livshin and Xavier Rodet. 2003. The importance of cross database evaluation in sound classification. In *Proc. ISMIR*.

[37] Vincent Lostanlen. 2017. *Convolutional operators in the time-frequency domain.* Ph.D. Dissertation. École normale supérieure.

[38] Vincent Lostanlen, Rachel M. Bittner, and Slim Essid. 2018. Medley-solos-DB: a cross-collection dataset of solo musical phrases. (2018).

[39] Vincent Lostanlen and Carmine Emanuele Cella. 2016. Deep convolutional networks on the pitch spiral for musical instrument recognition. In *Proc. ISMIR*.

[40] Vincent Lostanlen, Grégoire Lafay, Joakim Andén, and Mathieu Lagrange. 2018. Relevance-based Quantization of Scattering Features for Unsupervised Mining of Environmental Audio. *In review, EURASIP J. Audio Speech Music Process.* (2018).

[41] Mauricio A. Loureiro, Hugo Bastos de Paula, and Hani C. Yehia. 2004. Timbre Classification Of A Single Musical Instrument. In *Proc. ISMIR*.

[42] Thor Magnusson. 2017. Musical Organics: A Heterarchical Approach to Digital Organology. *J. New Music Res.* 46, 3 (2017), 286–303.

[43] Stéphane Mallat. 2012. Group invariant scattering. *Comm. Pure Appl. Math.* 65, 10 (2012), 1331–1398.

[44] Yan Maresz. 2013. On computer-assisted orchestration. *Contemp. Music Rev.* 32, 1 (2013), 99–109.

[45] Keith D. Martin and Youngmoo E. Kim. 1998. Musical instrument identification: A pattern recognition approach. In *Proc. ASA*.

[46] Luis Gustavo Martins, Juan José Burred, George Tzanetakis, and Mathieu Lagrange. 2007. Polyphonic instrument recognition using spectral clustering.. In *Proc. ISMIR*.

[47] Brian McFee, Eric J. Humphrey, and Julián Urbano. 2016. A plan for sustainable MIR evaluation. In *Proc. ISMIR*.

[48] Brian McFee and Gert R. Lanckriet. 2010. Metric learning to rank. In *Proc. ICML*.

[49] Cheryl D. Metcalf, Thomas A. Irvine, Jennifer L. Sims, Yu L. Wang, Alvin W.Y. Su, and David O. Norris. 2014. Complex hand dexterity: a review of biomechanical methods for measuring musical performance. *Front. Psychol.* 5 (2014), 414.

[50] Jeremy Montagu. 2009. It's time to look at Hornbostel-Sachs again. *Muzyka (Music)* 1, 54 (2009), 7–28.

[51] Frank J. Opolko and Joel Wapnick. 1989. McGill University Master Samples (MUMS). (1989).

[52] Kailash Patil and Mounya Elhilali. 2015. Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases. *EURASIP J. Audio Speech Music Process.* 2015, 1 (2015), 27.

[53] Kailash Patil, Daniel Pressnitzer, Shihab Shamma, and Mounya Elhilali. 2012. Music in our ears: the biological bases of musical timbre perception. *PLOS Comput. Biol.* 8, 11 (2012), e1002759.

[54] Curt Sachs. 2012. *The History of Musical Instruments.* Dover Publications.

[55] Arnold Schoenberg. 2010. *Theory of Harmony.* University of California.

[56] Li Su, Li-Fan Yu, and Yi-Hsuan Yang. 2014. Sparse Cepstral, Phase Codes for Guitar Playing Technique Classification.. In *Proc. ISMIR*.

[57] Adam R. Tindale, Ajay Kapur, George Tzanetakis, and Ichiro Fujinaga. 2004. Retrieval of percussion gestures using timbre classification techniques.. In *Proc. ISMIR*.

[58] Paul Villoutreix, Joakim Andén, Bomyi Lim, Hang Lu, Ioannis G. Kevrekidis, Amit Singer, and Stas Y. Shvartsman. 2017. Synthesizing developmental trajectories. *PLOS Comput. Biol.* 13, 9 (09 2017), 1–15.

[59] Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* 10, Feb (2009), 207–244.

[60] Stéphanie Weisser and Maarten Quanten. 2011. Rethinking musical instrument classification: towards a modular approach to the Hornbostel-Sachs system. *Yearb. Tradit. Music* 43 (2011), 122–146.

[61] Alicja A. Wieczorkowska and Jan M. Żytkow. 2003. Analysis of feature dependencies in sound description. *J. Intell. Inf. Syst.* 20, 3 (2003), 285–302.

[62] Luwei Yang, Elaine Chew, and Sayid-Khalid Rajab. 2014. Cross-cultural Comparisons of Expressivity in Recorded Erhu and Violin Music: Performer Vibrato Styles. In *Proc. Int. Workshop on Folk Music Analysis (FMA)*.

[63] Hanna Yip and Rachel M. Bittner. 2017. An accurate open-source solo musical instrument classifier. In *Proc. ISMIR, Late-Breaking / Demo session (LBD)*.

[64] Diana Young. 2008. Classification of Common Violin Bowing Techniques Using Gesture Data from a Playable Measurement System.. In *Proc. NIME*. Citeseer.

[65] Udo Zölzer. 2011. *DAFX: Digital Audio Effects.* Wiley.