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

Mathieu 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 which a musical note can be produced conveys some essential information 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 now considered a solved problem, the ability of a computer to precisely identify instrumental playing techniques (IPT) remains largely underdeveloped. We conduct a benchmark of machine listening systems for query-by-example browsing among 143 instrumental playing techniques, including the most contemporary, for 16 instruments in the symphonic orchestra, thus amounting to 469 triplets of instrument, mute, and technique. We identify and discuss three necessary conditions for significantly outperforming the classical mel-frequency cepstral coefficients (MFCC) baseline: the inclusion of second-order scattering coefficients to account for the presence of amplitude modulations; the inclusion of long-range temporal dependencies; and the resort to large-margin nearest neighbors (LMNN), a supervised metric learning method that reduces intra-class variability in feature space. We report a P@5 of 99.7% for instrument recognition (baseline at 89.0%) and of 61.0% for playing technique recognition (baseline at 44.5%). We interpret this quantitative gain by means of a qualitative assessment of practical usability as well as data visualizations resulting from nonlinear dimensionality reduction.

CCS CONCEPTS

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

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference format:
Vincent Lostanlen, Joakim Andén, and Mathieu Lagrange. 2018. Extended playing techniques: the next milestone in musical instrument recognition. In *Proceedings of DLfM, Paris, France, Sep. 2018,* 10 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

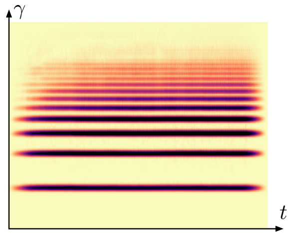
1 INTRODUCTION

The gradual diversification of the timbral palette in Western classical music at the turn 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) [49, epilogue]; the creation of novel instrumental associations, as epitomized by *Klangfarbenmelodie* [50, chapter 22]; the temporary alteration of resonant properties through mutes and other "preparations" [16]; a more systematic usage of extended instrumental techniques, such as artificial harmonics, *col legno batutto*, or flutter tonguing [29, chapter 11]; and the resort to electronics and digital audio effects [59]. 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 latter approaches to timbral diversification were massively adopted into post-war contemporary music. In particular, an increased concern for the concept of musical gesture [22] 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.

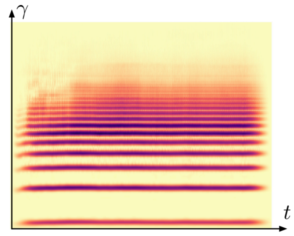
Far from being exclusive to erudite music, extended playing techniques are also commonly found in oral tradition; in some cases, they even stand out as a distinctive component of musical style. Four well-known examples are: the snap pizzicato ("slap") of the upright bass in rockabilly, the growl of the 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 mere knowledge of organology (the instrumental *what?* of music), as opposed to chironomics (its gestural *how?*), is a rather weak source of information for browsing and recommendation in large music databases.

Yet, past research in music information retrieval (MIR), and especially machine listening, rarely acknowledges the benefits of integrating the influence of performer gestures into a coherent taxonomy of musical instrument sounds. Instead, gestures are either framed as a spurious form of intra-class variability between instruments, without delving into its interdependencies with pitch and intensity; or, symmetrically, as a probe for the acoustical study of a given instrument, without enough emphasis onto 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, such obstacle has recently been overcome, owing to the creation of databases of instrumental samples in a perspective

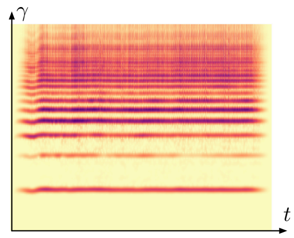


(a) Trumpet note (*ordinario*).

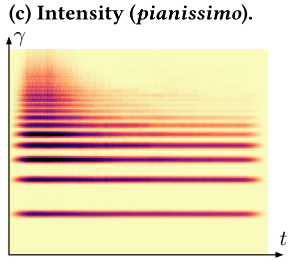


(b) Pitch (G3).

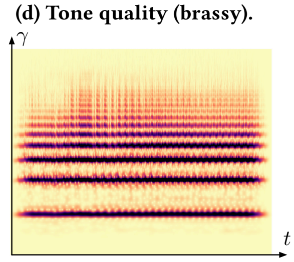
(c) Intensity (*pianissimo*).



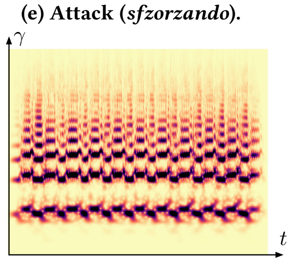
(d) Tone quality (brassy).



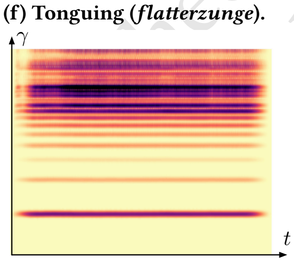
(e) Attack (*sfzorzando*).



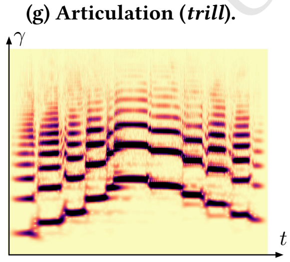
(f) Tonguing (*flatterzunge*).



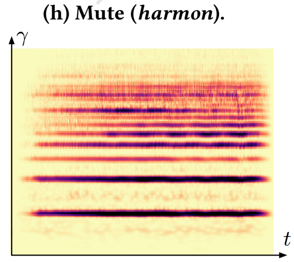
(g) Articulation (*trill*).



(h) Mute (*harmon*).



(i) Phrasing (*détaché*).



(j) Instrument (violin).

Figure 1: Ten factors of variations of a musical note: pitch (1b), intensity (1c), tone quality (1d), attack (1e), tonguing (1f), articulation (1g), mute (1h), phrasing (1i), and instrument (1j).

of spectralist music orchestration [39]. In this article, we capitalize on the availability of data to formulate a new line of research in MIR, namely the joint retrieval of organological information ("*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, which are deliberately regarded as contextual: at what pitches and intensities, but also where, when, why, by whom, and for whom was the music recorded.

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, which exclusively focus on pitch (Figure 1b) and intensity (Figure 1c) as the main factors of intra-class variability, this paper aims at accounting for the presence of instrumental playing techniques (IPT), such as changes in tone quality (Figure 1d), attack (Figure 1e), tonguing (Figure 1f), and articulation (Figure 1h), either as intra-class variability (instrument recognition task) or as inter-class variability (IPT recognition task). The analysis of IPTs whose definition necessarily involves more than a single musical event, such as phrasing (Figure 1i), is beyond the scope of this paper.

Section 2 reviews the existing literature on the topic. Section 3 derives the task of IPT classification from the definition of both a taxonomy of instruments and a taxonomy of gestures. Section 4 describes how two topics in machine listening, namely scattering transforms and 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 some of the recent MIR literature on the audio analysis of instrumental playing techniques, with a focus on the available datasets for each formulation of the problem at hand.

2.1 Classification of ordinary isolated notes

The earliest works on musical instrument recognition restricted their scope to individual notes played with an ordinary technique – with datasets such as MUMS [46], MIS, RWC [23], and Philharmonia – thus eliminating most factors of intra-class variability due to the performer [6, 11, 18, 25, 27, 40, 55]. These works have culminated with the development of a support vector machine (SVM) classifier trained on spectrotemporal receptive fields (STRF), which are idealized computational models of neurophysiological responses in the central auditory system [14]. Not only did it attain a near-perfect mean accuracy of 98.7% on the RWC dataset, but the confusion matrix of its automated predictions was closely similar to the confusion matrix of human listeners [48]. Therefore, the supervised classification of musical instruments from recordings of ordinary notes could arguably be considered a solved problem; we refer to [8] for a recent review of the state of the art.

2.2 Classification of solo recordings

One straightforward extension of the problem above is the classification of solo phrases, encompassing some variability in melody [30], for which the accuracy of STRF models is around 80% [47].

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 [26], are exposed to strong biases. This issue is partially mitigated by the recent surge of multitrack datasets, such as MedleyDB [9], which has spurred a renewed interest in single-label instrument classification [57]. In addition, the cross-collection evaluation methodology [32] allows to prevent the risk of overfitting caused by the relative homogeneity of these small datasets in terms of artists and recording conditions [10]. To this date, the best classifier of solo recordings is a spiral convolutional network [35] trained on the Medley-solos-DB dataset [34], i.e. a cross-collection dataset which aggregates MedleyDB and solosDb following the procedure of [17]. We refer to [24] 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 appropriate for large-scale deployment [12, 41]. However, it suffers from two methodological caveats: first, polyphonic instrumentation is not independent from other attributes of information, such as geographical origin, genre, or key; and secondly, the inter-rater agreement decreases with the number of overlapping sources [20, chapter 6]. Such issues are all the more troublesome that there is, to this date, no annotated dataset of polyphonic mixtures that is 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), might contribute to mitigating them in the near future [42].

2.4 Single-instrument playing technique classification

Lastly, there is a growing interest for studying the role of the performer in musical acoustics, from both perspectives of sound production and sound perception. Besides its interest in audio signal processing, this topic is connected to other disciplines, such as biomechanics and gestural interfaces [44]. The majority of the available literature focuses on the range of IPTs afforded by a single instrument: recent examples include clarinet [36], percussion [52], piano [7], guitar [13, 19, 51], violin [58], cello [15, chapter 6], and erhu [56]. 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 [3].

In this paper, we formulate the retrieval of expressive parameters of musical timbre at the scale of the symphonic orchestra at large, while expliciting these parameters in terms of sound production (i.e. through a finite set of instructions, readily interpretable by the performer) rather than by means of perceptual epithets only. We refer to [31] for a recent review of the state of the art.

3 TASKS

In this section, we distinguish taxonomies of musical instruments from taxonomies of musical gestures.

3.1 Taxonomies

The Hornbostel-Sachs taxonomy (H-S) strives to organize the diversity of musical instruments according to their manufactural characteristics only, and is purposefully unaffected by sociohistorical background [45]. Because it offers an unequivocal way of describing any acoustic instrument without any prior knowledge on 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 location of the violin in H-S (321.321-71), as depicted in Figure 2, also encompasses the viola and the cello in addition to the violin. This is because these three instruments, viewed as inert objects, share a common morphology, despite differences in posture for the performer: both violin and viola are usually played under the jaw whereas the cello is held between the knees. Accounting for these differences begs to refine H-S by means a vernacular taxonomy. Most instrument taxonomies in music signal processing, including MedleyDB and AudioSet [21], reach the vernacular level rather than conflating all instruments belonging to the same H-S node. In some cases, an even finer level of granularity is attained by the listing of potential alterations to the instrument – be them permanent or temporary, at the time scale of more than a single note – that affect its resonant properties after the end of the conventional manufacturing process, e.g. mutes and other preparations [16]. The only example of node in the MedleyDB taxonomy reaching this level is *tack piano* [9] .

Unlike musical instruments, which are approximately amenable to a hierarchical taxonomy of resonating objects, IPTs result from a complex synchronization between multiple gestures, which may involve both hands and arms, as well as diaphragm, vocal tract, and sometimes the whole body. As a result, there is no immediate way to interface them with H-S, or indeed any tree-like structure [28]. Instead, every playing technique is described by a finite collection of categories, each belonging to a different "namespace"; Figure 3 illustrates such namespaces in 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 [5]. In recent years, some publications have attempted to lay the foundations of such a modular approach, with the aim of making H-S relevant to contemporary music creation [37, 54]; yet, such considerations are still in large part speculative, 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. This system is meant to provide a small number k of nearest neighbors in the dataset of musical instrument

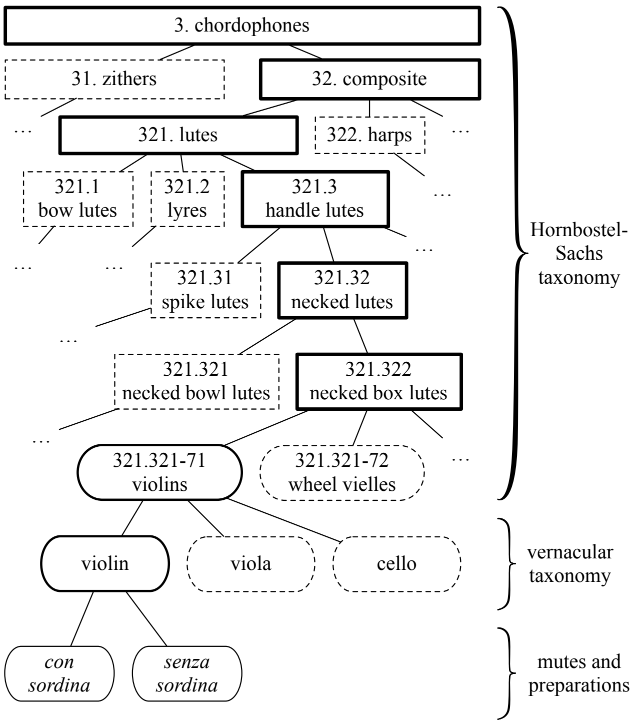


Figure 2: Taxonomy of musical instruments.

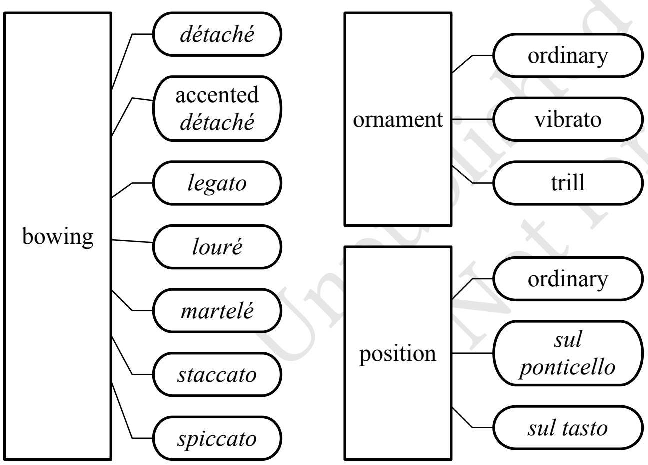


Figure 3: Namespaces of violin playing techniques.

samples to any user-defined audio query $\boldsymbol{x}(t)$. The search for nearest neighbors 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 (MFCC) as a baseline and scattering coefficients as an improvement upon this baseline. Furthermore, although our baseline adopts an Euclidean distance function to underlie the k-nearest neighbor algorithm in feature space, we will show that learning a non-Euclidean Mahalanobis metric as a replacement for

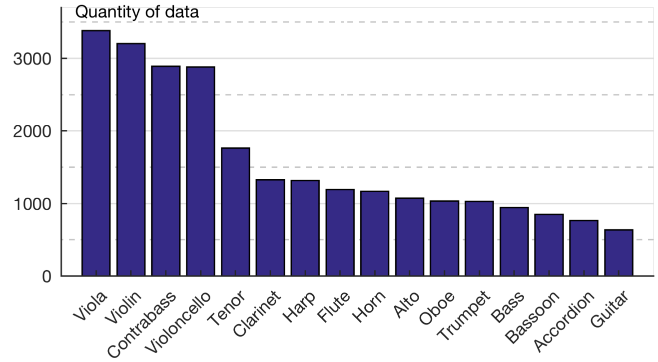


Figure 4: Instruments in the SOL dataset.

the canonical Euclidean metric also allows to improve upon the baseline.

In the context of contemporary music creation, $\boldsymbol{x}(t)$ may be an instrumental or vocal sketch; a sound event recorded from the environment; a computer-generated waveform; or any mixture of the above [39]. Upon inspecting the k nearest neighbors returned by the search engine, the composer may decide to retain one of the retrieved notes, in which case its attributes (pitch and intensity, but also the exact playing technique) are readily available and can be included into the musical score to approximate the query.

Faithfully evaluating such a system is a difficult procedure, and ultimately would rest on its practical usability, as judged by the composers themselves. 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, both under a instrument taxonomy and an IPT taxonomy. In all subsequent experiments, we report P@k after setting the number of retrieved items to $k = 5$.

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.[1] 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 applicable intensities, pitches, preparations (i.e. mutes), as well as all applicable playing techniques. The distribution of playing techniques – whose most common are shown in Figure 5 – is heavy-tailed (average 178, standard deviation 429): 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.

4 METHODS

In this section, we describe the scattering transform and supervised metric learning used to implement all query-by-example systems in our benchmark.

[1] Link to SOL dataset: http://forumnet.ircam.fr/product/orchids-en/

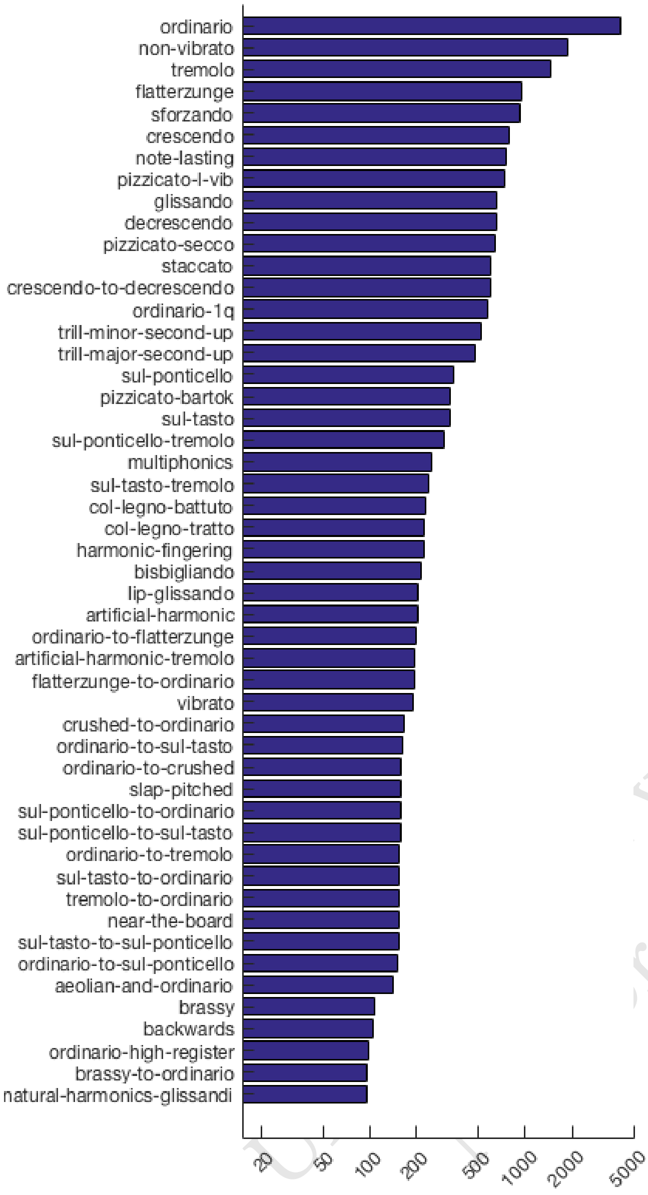


Figure 5: Playing techniques in the SOL dataset.

4.1 Scattering transform

The scattering transform is a cascade of two wavelet modulus operators, each followed by temporal averaging: the first layer extracts the average spectral envelope $\mathrm{S}_1\boldsymbol{x}(\lambda_1)$ of $\boldsymbol{x}(t)$ at frequencies λ_1, whereas the second layer $\mathrm{S}_2\boldsymbol{x}(\lambda_1, \lambda_2)$ extracts amplitude modulations of this spectral envelope at rates λ_2. The set of frequencies discretizes the auditory range according to the mel scale, with $Q_1 = 12$ bins per octave at topmost frequencies; whereas rates λ_2 follow a geometric sequence between λ_1 and some minimal rate T^{-1}, with $Q_2 = 1$ bin per octave. We refer to [2] for a general introduction to scattering transforms in audio classification, and to [33, sections 3.2 and 4.5] for a discussion on its application to musical instrument classification in solo recordings, as well as its close connections with STRF. The scattering transform is theoretically suited to model extended playing techniques, since various values of the rate λ_2 characterize some of the most common nonstationarities in sound production, including tremolo, vibrato, and dissonance [1, section 4]. In the following, we denote by $\mathbf{S}\boldsymbol{x}(\lambda)$ the concatenation of all scattering coefficients, whether the generic scattering path λ corresponds to a singleton (λ_1) or a pair (λ_1, λ_2).

In order to match a decibel-like perception of loudness, we apply the path-adaptive, quasi-logarithmic compression

$$\widetilde{\mathbf{S}}\boldsymbol{x}_i(\lambda) = \log\left(1 + \frac{\mathbf{S}\boldsymbol{x}_i(\lambda)}{\varepsilon \times \boldsymbol{\mu}(\lambda)}\right) \tag{1}$$

where $\varepsilon = 10^{-3}$ and $\boldsymbol{\mu}(\lambda)$ is the median value of the scattering coefficient $\mathbf{S}\boldsymbol{x}_i(\lambda)$ for path λ across samples i.

4.2 Metric learning

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

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

between all pairs of samples $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ optimizes some objective function. We refer to [4] for a review of the state of the art. In particular, the large-margin nearest neighbors (LMNN) algorithm aims at bringing all k nearest neighbors $\boldsymbol{x}_j$ of every $\boldsymbol{x}_i$ closer than the canonical Euclidean distance $\mathrm{D}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\widetilde{\mathbf{S}}\boldsymbol{x}_i - \widetilde{\mathbf{S}}\boldsymbol{x}_j\|_2$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same class, and further apart otherwise. The matrix $\mathbf{L}$ is obtained by applying the special-purpose solver of [53, appendix A]. In subsequent experiments, disabling LMNN is equivalent to setting $\mathbf{L}$ to the identity matrix, and retrieving a list of k nearest neighbors from $\boldsymbol{x}_i(t)$ according to the canonical Euclidean distance in feature space rather than the Mahalanobis distance.

As compared to a class-wise generative model (such as Gaussian mixtures), a global linear model ensures some robustness to minor alterations of the taxonomy, which is important in the context of IPT: e.g. what some instrumentists may call *slide*, others will call *glissando*. Furthermore, although one of its major drawback relies on its strong dependency on the Euclidean neighbors to determine intra-class variability [43], this drawback is alleviated in the case of a feature space based on scattering transform coefficients, whose Euclidean metric provably approximates the extent of elastic deformation needed to shear $\boldsymbol{x}_i(t)$ into $\boldsymbol{x}_j(t)$ in the time-frequency domain [38, Theorem 2.16].

5 EXPERIMENTAL RESULTS

In this section, we combine the aforementioned methods to the construction of a query-by-example browsing system in the Studio On Line (SOL) dataset; discuss the factors enabling to improve upon the state of the art; and provide both a qualitative and a quantitative comparison between MFCC-based and scattering-based feature spaces, in the context of timbral similarity between instrumental playing techniques.

5.1 Evaluation of instrument recognition

In the task of instrument recognition, each of the k elements $\boldsymbol{x}_j$ returned by the system is considered relevant to the query $\boldsymbol{x}_i$ if and only if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ correspond to the same instrument, regardless of pitch, intensity, mute, and playing technique.

We compare scattering features to a baseline of mel-frequency cepstral coefficients (MFCC), corresponding to the 13 lowest quefrencies after applying a discrete cosine transform (DCT) on the logarithm of the 40-band mel-frequency spectrum. In addition, we vary the maximum time scale T of amplitude modulation between 25 ms and 1 s. In the case of MFCC, $T = 25$ 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 MFCC, because the mel-spectrogram is equivalent to a local averaging of the wavelet scalogram at the time scale T, leaving unchanged the global averaging of $\mathbf{S}\boldsymbol{x}(\lambda)$ at the time scale of whole musical notes [1, section II.B].

Figure 6 (left) summarizes our results. We find that MFCC reach a relatively high P@5 of 89%. Keeping all 40 quefrencies rather than the lowest 13 brings the P@5 down to 84%, because the highest quefrencies are the most affected by some spurious factors of intra-class variability, namely pitch and spectral flatness [33, subsection 2.3.3].

At the smallest time scale $T = 25$ ms, the scattering transform reaches a P@5 of 89%, thus matching exactly the performance of MFCC. This is because the relatively few second-order scattering coefficients whose rate λ_2 exceeds 40 Hz have a negligible effect on Euclidean distances, as they carry very little energy [2]. Moreover, disabling median renormalization – i.e. setting $\boldsymbol{\mu}(\lambda) = 1$ for all scattering paths λ – degrades P@5 down to 84%, while disabling logarithmic compression altogether – i.e. the limit case $\varepsilon \rightarrow \infty$ – degrades it to 76%. These results are consistent with another publication [?], which applies scattering transform to a query-by-example retrieval task among environmental acoustic scenes.

On one hand, replacing the canonical Euclidean distance by a Mahalanobis distance learned by the LMNN algorithm marginally improve P@5 in the case of the MFCC baseline, from 89.3% to 90.0%. On the other hand, applying LMNN on scattering features strongly enhances their performance with respect to the Euclidean distance, from 89.1% to 98.0%.

The gain in precision afforded by scattering coefficients over MFCC could simply be caused by a higher number of dimensions. 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 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.

Lastly, increasing T from 25 ms up to 1 s – that is, including all amplitude modulations between 1 Hz and 40 Hz – brings LMNN to a near-perfect P@5 of 99.71%. Not only does this result confirm that well-established methods 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 excellent despite large intra-class variability within instruments: in pitch and intensity, but also in the usage of mutes

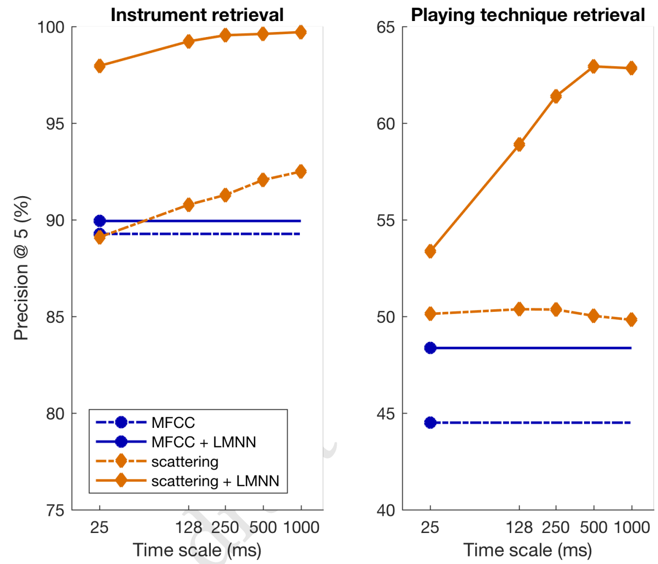


Figure 6: Summary of results on the SOL dataset.

and extended IPTs. In other words, the monophonic recognition of Western instruments is, all things considered, a solved problem indeed.

5.2 Evaluation of playing technique recognition

The situation is different when considering IPT, rather than instrument, as the reference for evaluating the reference of the query-by-example search engine. In this second evaluation 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, whereas we trained LMNN with instrument labels as classes in the previous subsection, this second experiment re-trains LMNN with IPTs as classes. In other words, the Mahalanobis metric is no longer optimized to distinguish instruments but to distinguish playing techniques.

Figure 6 (right) summarizes our results. The MFCC baseline has a relatively low P@5 of 44.5%, which indicates that a coarse description of the short-term spectral envelope is rarely ever sufficient to model acoustic similarity in IPT. Perhaps more surprisingly, we find that only the system combining all presented variations, i.e. log-scattering coefficients with median renormalization, $T = 500$ ms, and LMNN, strongly outperforms the MFCC baseline, with a state-of-the-art 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, 50.0%; and replacing scattering coefficients by MFCC, to 48.4%. This result contrasts with the previous evaluation setting: whereas the improvements brought by the three aforementioned variations are approximately additive and independent in terms of P@5 for musical instruments, they cause a super-additive interaction in terms of P@5 for IPTs. In particular, it appears that increasing T above 25 ms is only beneficial to IPT similarity retrieval if it is combined with metric learning.

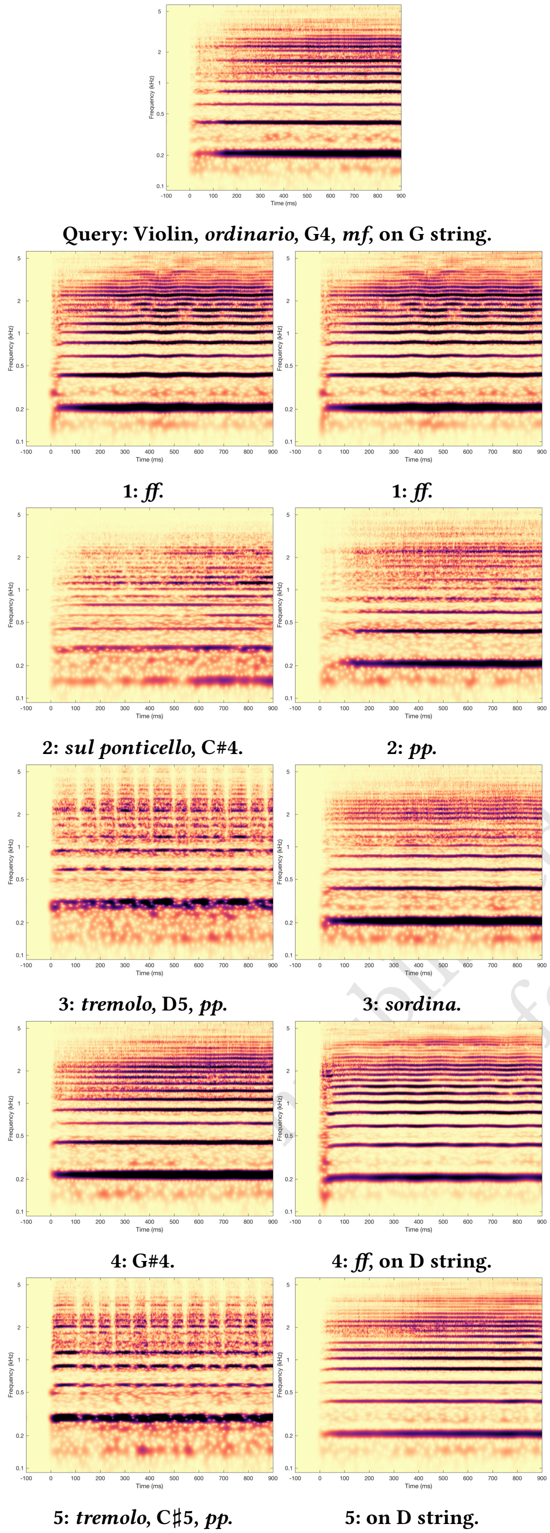


Figure 7: Five nearest neighbors of the same query (a violin note with ordinary playing technique, at pitch G4, *mf* dynamics, played on the G string), as retrieved by two different versions of our system: with MFCC features (left) and with scattering transform features (right).

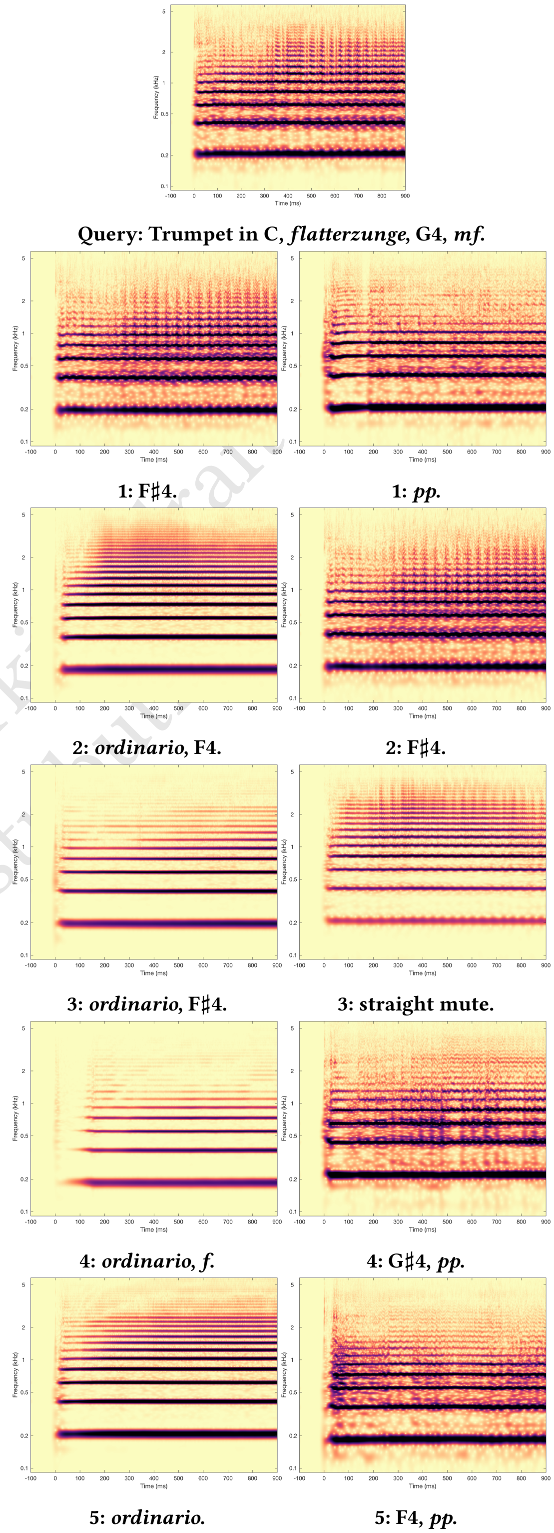


Figure 8: Five nearest neighbors of the same query (a trumpet note with *flatterzunge* technique, at pitch G4, *mf* dynamics), as retrieved by two different versions of our system: with MFCC features (left) and with scattering transform features (right).

5.3 Qualitative error analysis

For demonstration purposes, we arbitrarily select some audio recordings $\boldsymbol{x}(t)$, and run two versions of the proposed query-by-example search using $\boldsymbol{x}(t)$ as query. The first version uses MFCC features with $T = 25\,\mathrm{ms}$ and metric learning with LMNN; it has a P@5 of 90.0% for instrument retrieval and 48.4% for IPT retrieval. The second version uses scattering features with $T = 1\,\mathrm{s}$, logarithmic transformation with median renormalization (see Equation **??**), and metric learning with LMNN; it has a P@5 of 99.7% for instrument retrieval and 63.0% for IPT retrieval. The main difference between these two versions lies in the choice of spectrotemporal features.

: a violin note from the SOL dataset, played with ordinary playing technique, pitch G4, *mf* dynamics, on the G string.

Figure 8 shows the constant-Q scalograms of the five retrieved items

What stems from these observations is that, unlike instrument similarity, IPT similarity results from long-range temporal dependencies in the audio signal. In addition, the dissimilarity between two different playing techniques is not a matter of elastic deformation in the time-frequency domain – as approximated by Euclidean distance in the feature space of scattering coefficients – but also involves an adaptive process which combines the saliences of various acoustic frequencies and modulation rates in several nonuniform ways, thus producing a metric that favors certain factors of acoustic variability while mitigating others.

5.4 Feature visualization with diffusion maps

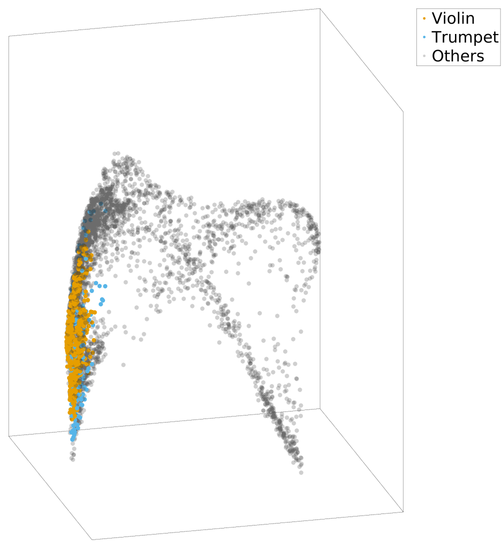
6 CONCLUSION

Whereas the MIR literature abounds on the topic of musical instrument recognition in so-called "ordinary" isolated notes and solo performances, little is known about the problem of retrieving the instrumental playing technique (IPT) of an audio query within a fine-grained taxonomy. Yet, the knowledge of IPT is a precious source of music information, not only to characterize the physical interaction between player and instrument, but also in the realm of contemporary music creation. In all likelihood, 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 in benchmarking query-by-example MIR systems according to a large-vocabulary 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 rates and comparing them in a highly non-Euclidean way. Although the combination of methods presented here – wavelet scattering and large-margin nearest neighbors – outperforms the MFCC baseline (even at comparable dimensionalities and number of learnable parameters), its accuracy on the SOL dataset certainly leaves some room for future improvements.

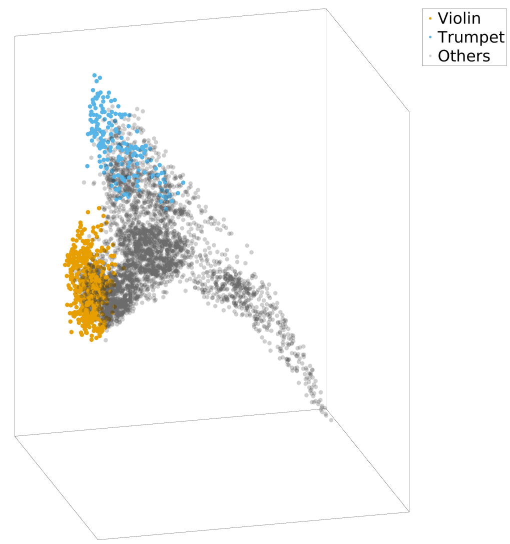
The evaluation methodology presented here uses ground truth IPT labels to quantify the relevance of returned items. Despite the advantage of unequivocality, it might be too harsh to reflect practical use. Indeed, as it is often the case in MIR, some pairs of labels are subjectively more similar than others: e.g. *slide* is evidently closer to *glissando* than to *pizzicato-bartok*. The collection of subjective ratings of IPT similarity, and its comparison with automated ratings, is left as future work.

ACKNOWLEDGMENTS

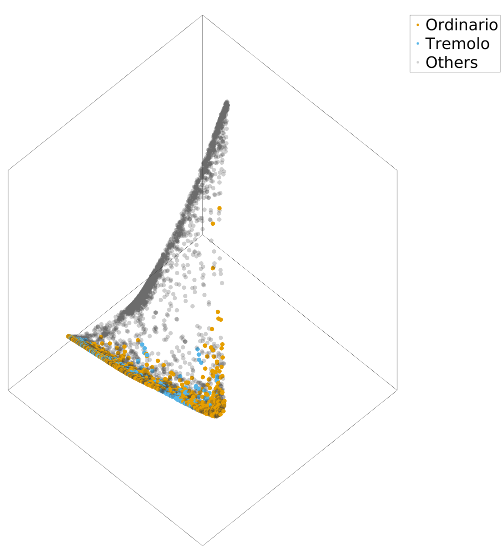
The authors wish to thank Philippe Brandeis, Étienne Graindorge, Stéphane Mallat, Adrien Mamou-Mani, and Yan Maresz, for fruitful discussions on contemporary music creation, as part of the TICEL research project ("Traité instrumental collaboratif en ligne"); Andrew Farnsworth and Grant Van Horn, for 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.



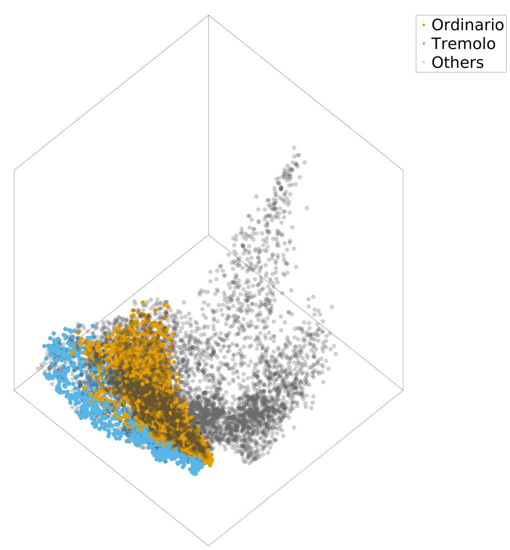
(a) Instrument embedding with MFCC.



(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 $T = 1$ 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.

REFERENCES

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

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

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

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

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

[6] 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*.

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

[8] 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. *Journal Int. Inform. Syst.* 46, 3 (2016), 425–446.

[9] 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*.

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

[11] 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.

[12] 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*. IEEE, 173–176.

[13] 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*.

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

[15] 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.

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

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

[18] Antti Eronen and Anssi Klapuri. 2000. Musical instrument recognition using cepstral coefficients and temporal features. In *Proc. IEEE ICASSP*, Vol. 2. IEEE, II753–II756.

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

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

[21] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP*.

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

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

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

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

[26] 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.

[27] 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.

[28] Sefki Kolozali, Mathieu Barthet, György Fazekas, and Mark B. Sandler. 2011. Knowledge Representation Issues in Musical Instrument Ontology Design.. In *Proc. ISMIR*. 465–470.

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

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

[31] 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. https://doi.org/10.1007/978-3-319-66881-9_11

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

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

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

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

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

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

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

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

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

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

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

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

[44] 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.

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

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

[47] 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.

[48] 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.

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

[50] Arnold Schoenberg. 2010. *Theory of Harmony* (100th anniversary edition ed.). University of California.

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

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

[53] 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.

[54] 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.

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

[56] 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)*.

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

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

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