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. 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. In this paper, 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%).

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,* 3 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

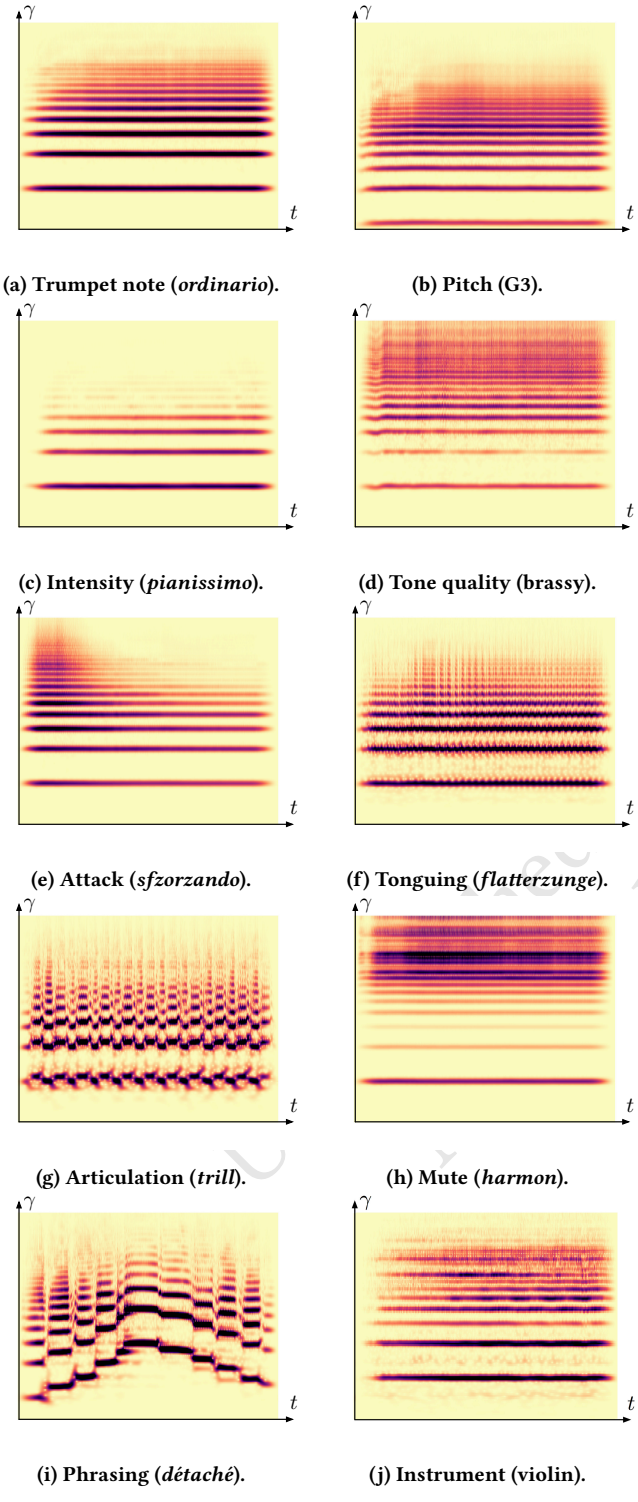
The gradual diversification of the timbral palette in Western classical music at the turn of the 20th century is reflected in five concurrent trends: the addition of new instruments to the symphonic instrumentarium, either by technological inventions (e.g. theremin) or importation from non-Western musical cultures (e.g. marimba) [5, epilogue]; the creation of novel instrumental associations, as epitomized by *Klangfarbenmelodie* [6, chapter 22]; the temporary alteration of resonant properties through mutes and other "preparations" [1]; a more systematic usage of extended instrumental techniques, such as artificial harmonics, *col legno batutto*, or flutter tonguing [3, chapter 11]; and the resort to electronics and digital audio effects [7]. 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 [2] 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

DLfM, Sep. 2018, Paris, France

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



(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.

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 of spectralist music orchestration [4]. 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 [**?**], MIS, RWC [**?**], and Philharmonia – thus eliminating most factors of intra-class variability due to the performer [**?** **?** **?** **?** **?** **?** **?**]. 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 [**?**]. 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 [**?**]. Therefore, the supervised classification of musical instruments from recordings of ordinary notes could arguably be considered a solved problem; we refer to [**?**] 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 [**?**], for which the accuracy of STRF models is around 80% [**?**]. 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 [**?**], are exposed to strong biases. This issue is partially mitigated by the recent surge of multitrack datasets, such as MedleyDB [**?**], which has spurred a renewed interest in single-label instrument classification [**?**]. In addition, the cross-collection evaluation methodology [**?**] allows to prevent the risk of overfitting caused by the relative homogeneity of these small datasets in terms of artists and recording conditions [**?**]. To this date, the best classifier of solo recordings is a spiral convolutional network [**?**] trained on the Medley-solos-DB dataset [**?**], i.e. a cross-collection dataset which aggregates MedleyDB and solosDb following the procedure of [**?**]. We refer to [**?**] 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 [**? ?**]. 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 [**?** , 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 [**?**].

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 [**?**]. The majority of the available literature focuses on the range of IPTs afforded by a single instrument: recent examples include clarinet [**?**], percussion [**?**], piano [**?**], guitar [**? ? ?**], violin [**?**], cello [**?** , chapter 6], and erhu [**?**]. 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 [**?**].

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 [**?**] 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 [**?**]. 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 **??**, 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 [**?**], 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 [1]. The only example of node in the MedleyDB taxonomy reaching this level is *tack piano* [**?**] .

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 [**?**]. Instead, every playing technique is described by a finite collection of categories, each belonging to a different "namespace"; Figure **??** 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 [**?**]. 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 [**? ?**]; yet, such considerations are

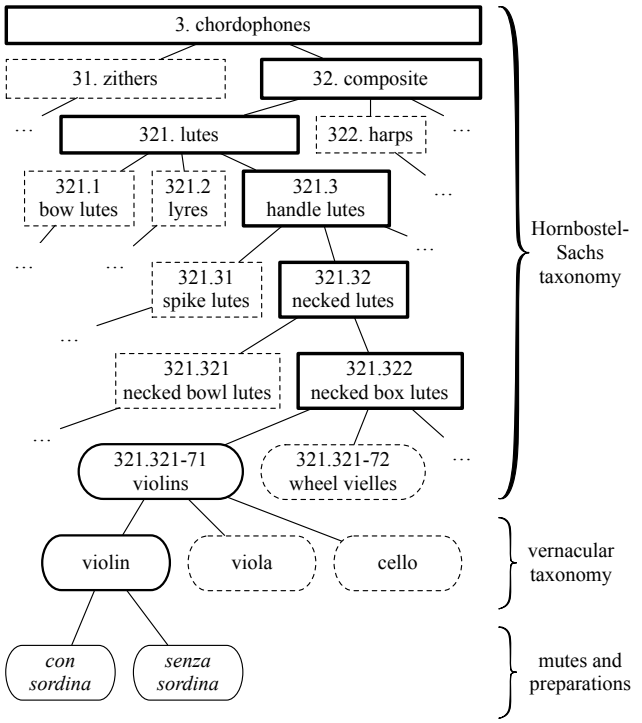


Figure 2: Taxonomy of musical instruments.

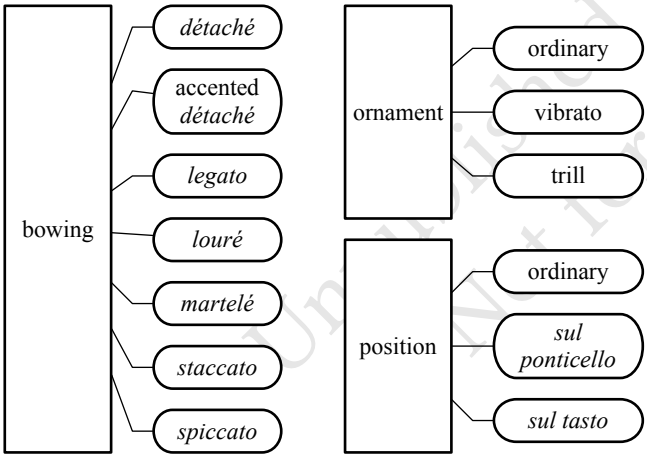


Figure 3: Namespaces of violin playing techniques.

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 samples to

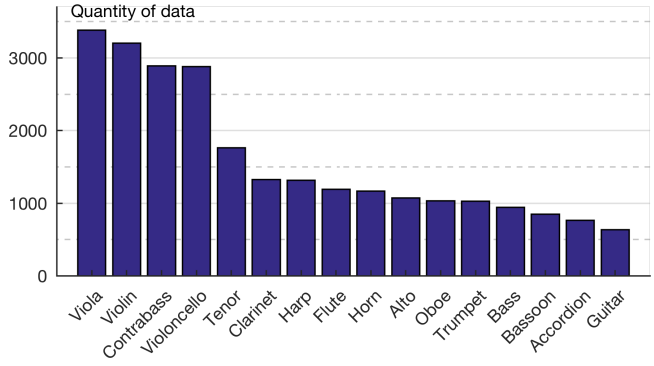


Figure 4: Instruments in the SOL dataset.

any user-defined audio query $\boldsymbol{x}(t)$. In the context of contemporary music creation, this $\boldsymbol{x}(t)$ may be an instrumental or vocal sketch; a sound event recorded from the environment; a computer-generated waveform; or any mixture of the above [4]. 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 **??**, 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 **??** – 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.

ACKNOWLEDGMENTS

The authors wish to thank Philippe Brandeis for fruitful discussions on contemporary music creation and Katherine Crocker for helpful suggestions on the title of this article.

REFERENCES

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

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

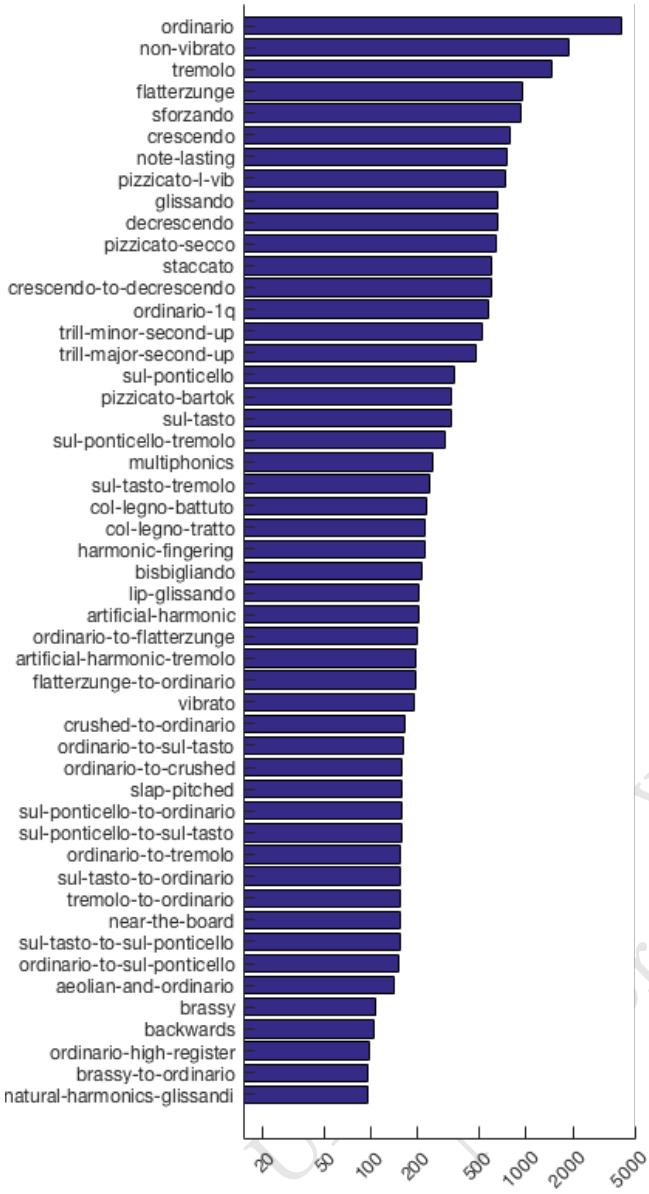


Figure 5: Playing techniques in the SOL dataset.

[2] Rolf Inge Godøy and Marc Leman. 2009. *Musical Gestures: Sound, Movement, and Meaning*. Taylor & Francis.
[3] Stefan Kostka. 2016. *Materials and Techniques of Post Tonal Music*. Taylor & Francis.
[4] Yan Maresz. 2013. On computer-assisted orchestration. *Contemp. Mus. Rev.* 32, 1 (2013), 99–109.
[5] Curt Sachs. 2012. *The History of Musical Instruments*. Dover Publications.
[6] Arnold Schoenberg. 2010. *Theory of Harmony* (100th anniversary edition ed.). University of California.
[7] Udo Zölzer. 2011. *DAFX: Digital Audio Effects*. Wiley.