EXTENDED PLAYING TECHNIQUES: THE NEXT MILESTONE IN MUSICAL INSTRUMENT RECOGNITION

First Author
Affiliation1
author1@ismir.edu

Second Author
Retain these fake authors in submission to preserve the formatting

Third Author
Affiliation3
author3@ismir.edu

ABSTRACT

Although the automatic recognition of a musical instrument from the recording of a single "ordinary" note is close to becoming solved problem, the ability of a computer to precisely identify instrumental playing techniques (IPT) within an extended taxonomy remains far below human accuracy. This article provides the first benchmark of machine listening systems for query-by-example browsing of instrumental playing techniques, including "extended" techniques such as reed slaps and , in the symphonic orchestra. We identify three necessary conditions for significantly outperforming the classical mel-frequency cepstral coefficients (MFCC) baseline: (1) the inclusion of second-order scattering coefficients to account for the presence of amplitude modulations ; (2) the inclusion of large temporal scales ; and (3) the resort to supervised metric learning.

1. INTRODUCTION

The progressive 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 Western symphonic instrumentarium, either by technological inventions (e.g. theremin) or importation from non-Western musical cultures (e.g. marimba) [**?**]; the creation of novel instrumental associations, as epitomized by *Klangfarbenmelodie* [**?**]; the temporary alteration of resonant properties through mutes and other "preparations" [**?**]; a more systematic usage of extended instrumental techniques, such as artificial harmonics, *col legno batutto*, or flutter tonguing [**?**]; and the resort to electronic and digital audio effects [**?**]. 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 [**?**] has liberated many unconventional instrumental techniques from their figurativistic connotations, thus making the so-called "ordinary" playing style merely one of many compositional 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 recommending music within large audio databases.

Yet, past research in music information retrieval (MIR), and especially machine listening, too 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 of spectralist music orchestration [**?**]. 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: where, when, why, by whom, and for whom was the music (in this recording) played.

Section 2 reviews the existing literature on the topic of retrieving information from instrumental playing techniques (IPT). 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 some of the recent MIR literature on the audio analysis of instrumental playing techniques.

For lack of more appropriate datasets, the earliest works on musical instrument recognition focused on isolated notes (Brown, 1999; Eronen and Klapuri, 2000; Herrera-Boyer, Peeters, and Dubnov, 2003; Kaminskyj and Czaszejko, 2005; Martin and Kim, 1998; Wieczorkowska and Zytkow, 2003). In this context, generalizing from the training set to the test set entails the construction of a representation of musical notes which is invariant to small variations in pitch and dynamics while remaining discriminative to qualitative timbre. More recently, Patil et al. (2012) have managed to classify isolated notes belonging to 11 instruments from the RWC dataset (Goto et al., 2003) with a mean accuracy of 98.7%. They used a collection of spectrotemporal receptive field (STRF) features, which are akin to time-frequency scattering features, and a support vector machine (SVM) classifier with a Gaussian kernel. Not only did they attain a near-perfect recognition rate, but they also found that the confusion matrix of the classifier was closely similar to the confusion matrix of human listeners.

Drawing on these findings, it seems that the supervised classification of musical instruments from recordings of isolated notes could now be considered a solved problem (Patil and Elhilali, 2015).

Timbre classification of a single musical instrument (clarinet): [**?**]. Retrieval of percussion gestures using timbre classification techniques: [**?**]. Polyphonic instrument recognition using spectral clustering: [**?**]. Knowledge representation issues in musical instrument ontology design: [**?**]. Guitar playing technique classification: [**?**]. MedleyDB: [**?**]. Audio Set: [**?**]. Visipedia: [**?**] Scattering transforms in musical instrument recognition: [**?**, **?**].

3. TASK

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

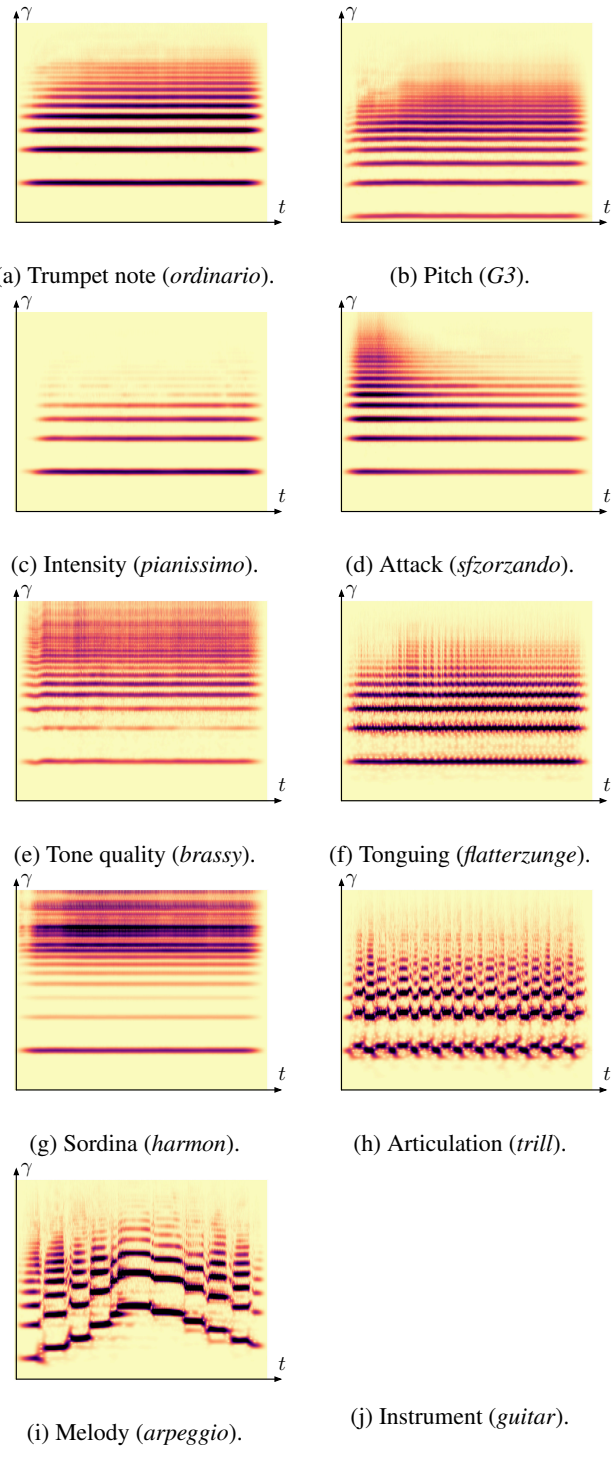
ML : also describe the evaluation methodology. Ranking for compositional use, query by example, etc

4. METHODS

In this section, we point out the theoretical limitations of mel-frequency cepstral coefficients (MFCC) in the representation of musical sounds comprising with extended instrumental playing techniques (IPT), and describe how both the scattering transform and supervised metric learning may overcome such limitations.

4.1 Limitations of mel-frequency cepstral coefficients

ML : not sure we need to discuss this. We should include the mfccs as baseline but more focus on what is needed: a representation of the data that is able to reliably consider larger frame sizes than usually considered



(a) Trumpet note (*ordinario*). (b) Pitch (*G3*).

(c) Intensity (*pianissimo*). (d) Attack (*sfzorzando*).

(e) Tone quality (*brassy*). (f) Tonguing (*flatterzunge*).

(g) Sordina (*harmon*). (h) Articulation (*trill*).

(i) Melody (*arpeggio*). (j) Instrument (*guitar*).

Figure 1: Ten variations of a musical note.

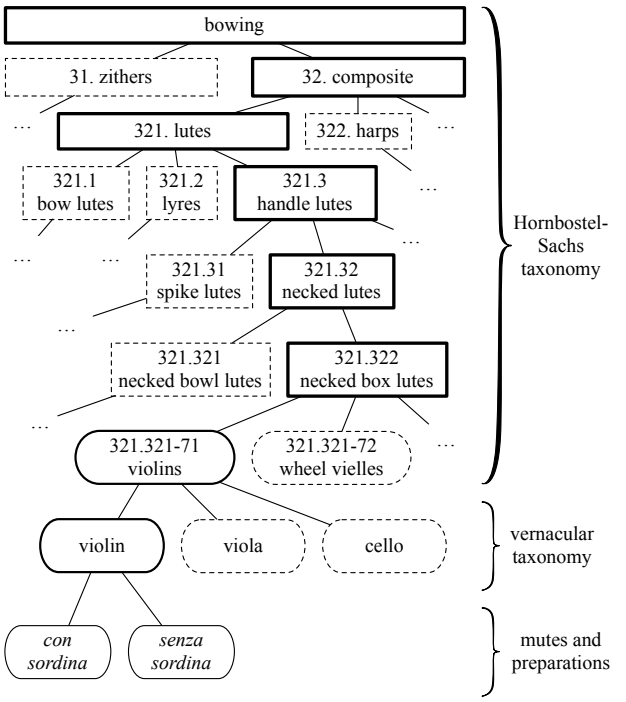


Figure 2: Instrument taxonomy.

4.2 Scattering transform

[**?**]

5. METRIC LEARNING

As shown in the experiments described in Section 6, it is helpful to consider a supervised projection of the scattering coefficients in order to select among the large dimensionality of the resulting feature space, which axes are relevant for the task at hand.

May approaches can be considered to achieve such a task []. One standard approach is the use of the linear discriminant analysis (LDA) that linearly projects ($x \rightarrow Lx$) the input data into a features space of C-1 dimensions that maximizes the amount of between-class variance relative to the amount of within-class variance. The linear transformation L is chosen to maximize the ratio of between-class to within-class variance, subject to the constraint that L defines a projection matrix.

A more flexible approach often taken in metric learning [**?**] is to optimize a malahanobis matrix that linearly projects the input data into another feature space of the same dimensionality. In this case:

$$M = L^T L \tag{1}$$

and the resulting distance cane be expressed as follows:

$$D_m(x, y) = (x_i - x_j)^T M (x_i - x_j) \tag{2}$$

x_i x_j being features vectors. In this study, the large-margin nearest neighbors (LMNN) approach [**?**, **?**] that optimizes the above described performance metric is considered. During the learning process, the following constraints are enforced: : the k nearest neighbors of any training instance should belong to the same class of the training instance (the "pull" constraint) while keeping away instances of other classes (the "push" constraint).

6. EXPERIMENTAL RESULTS

In this section, we apply the aforementioned methods to instrument playing techniques classification in the Studio On Line (SOL) dataset of musical instrument samples for spectralist orchestration.

6.1 Studio On Line (SOL) dataset

The Studio On Line (SOL) dataset have been recorded at Ircam and is part of the orchestration tools developed in the institute. It is composed of 25444 recordings of single tones played by 16 musical instruments: Accordion, Alto, Bass, Bassoon, Clarinet, Contrabass, Flute, Guitar, Harp, Horn, Oboe, Tenor, Trumpet, Viola, Violin, and Violoncello. The number of recordings per class of instrument is on average 1590 $\pm$936. Each instrument is played at different nuance and pitch if relevant using 143 different playing technique. The average number of number playing techniques is 178 $\pm$429. The large variance is due to the fact that some playing techniques may be considered for many instruments such as *ordinario*, whereas other are specific to some instruments such as *flatterzunge*. Figure **??** shows the playing techniques with the higher, lower and median number of recordings.

6.2 Musical instruments typology as reference

In this section, we report experimental results while considering the musical instruments as the reference. Thus the task aims at grouping together in the feature space, recordings that are played by the same musical instrument regardless of the nuance, pitch and playing technique.

Baseline choice: cut: keep lowest 13 over 40 std: remove mean divide by variance

features	cut	standardize	p (%)
mfcc	0	0	85
mfcc	0	1	84
mfcc	1	0	88
mfcc	1	1	**89**
mel		0	53
mel		1	50

Scattering setting:

median	compress	standardize	p (%)
0	0	0	64
0	0	1	76
0	1	0	84
0	1	1	83
1	0	0	77
1	0	1	76
1	1	0	**89**
1	1	1	89

6.3 Instrumental playing technique typology as reference

	features	cut	standardize	p (%)
	mfcc	0	0	35
	mfcc	0	1	32
Baseline choice:	mfcc	1	0	**46**
	mfcc	1	1	45
	mel		0	19
	mel		1	19

6.4 Discussion

7. CONCLUSION

Every quest for information is also a quest for invariance. [...]

Our main finding is that mel-frequency cepstral coefficients (MFCC), although highly informative for musical instrument recognition in so-called "ordinary" notes, fail to discriminate extended instrumental techniques for at least three reasons. First, they summarize spectral envelope without retaining its amplitude modulations. Secondly, the frame rate at which they are computed ($T = 25\,\mathrm{ms}$ or so) is short enough to assume local stationarity, but not long enough to encompass all informative local correlations. Thirdly, the discrete cosine transform (DCT) over mel frequencies does not, contrary to a widespread belief, make them invariant to the frequency transposition of complex sounds. To address each of these shortcomings, we propose a simple pipeline of three elements: a time scattering transform; unsupervised metric learning over frequencies and modulation rates with large-margin nearest neighbors (LMNN); and a support vector machine (SVM) classifier.

8. ACKNOWLEDGMENTS