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

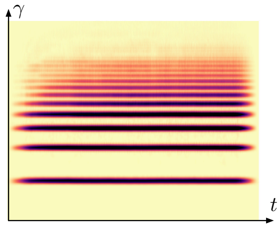
The expressive variability in which a musical note can be produced conveys some essential information to the modeling of orchestration and style. Yet, although the automatic recognition of a musical instrument from the recording of a single "ordinary" note is now considered a solved problem, the ability of a computer to precisely identify instrumental playing techniques remains largely underdeveloped.

In this paper, 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 supervised metric learning in order to boost the representativity of the scattering coefficients relevant for the task at hand.

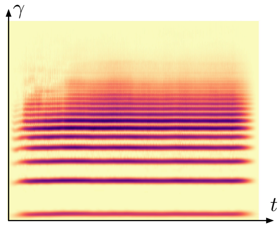
We report a P@5 of 99.7% for instrument recognition (baseline at 92.5%) and of 61.0% for playing technique recognition (baseline at 50.0%).

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 symphonic instrumentarium, either by technological inventions (e.g. theremin) or importation from non-Western musical cultures (e.g. marimba) [49]; the creation of novel instrumental associations, as epitomized by *Klangfarbenmelodie* [50]; 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]; and the resort to electronic and

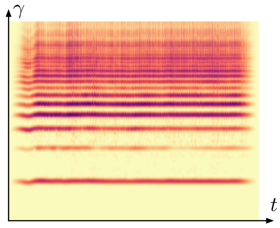


(a) Trumpet note (*ordinario*).

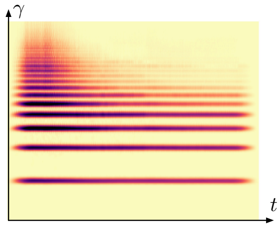


(b) Pitch (G3).

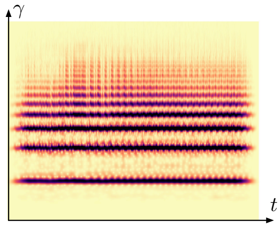
(c) Intensity (*pianissimo*).



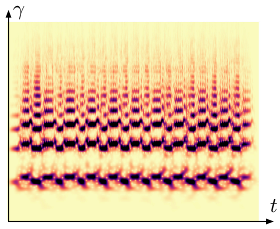
(d) Tone quality (brassy).



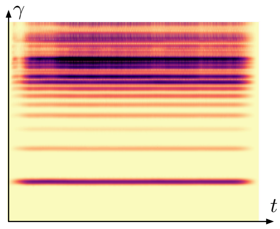
(e) Attack (*sfzorzando*).



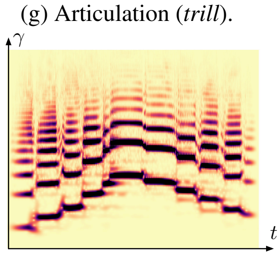
(f) Tonguing (*flatterzunge*).



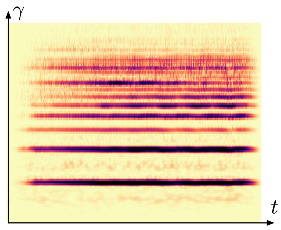
(g) Articulation (*trill*).



(h) Mute (*harmon*).



(i) Phrasing (*détaché*).



(j) Instrument (violin).

Figure 1: Ten factors of variations of a musical note.

digital audio effects [60]. 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 ML : and interpretative ? 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, 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 [40]. 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 ML : what about pitch and nuance ?.

Figure 1a shows the constant-Q wavelet transform (CQT) 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 playing techniques 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 some of the recent MIR literature on the audio analysis of instrumental playing techniques, with a focus on the available datasets afferent to each formulation of the problem.

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 [6,11,18,25,27,41,56]. 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, it seems that the supervised classification of musical instruments from recordings of ordinary notes could now 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 particularly difficult to design ML : why ?. This issue is partially mitigated by the recent surge of multitrack datasets, such as MedleyDB [9], which has spurred a renewed interest in monophonic instrumental recordings ML : datasets ? [58]. In addition, the cross-collection evaluation methodology [33] 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 [36] trained on the Medley-solos-DB dataset [35], 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 [12,42]. 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 [43].

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 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 playing techniques afforded by a single instrument: recent examples include clarinet [38], percussion [52], piano [7], guitar [13, 19, 51], violin [59], cello [15, chapter 6], and erhu [57]. Some publications frame timbral similarity in a polyphonic setting do so according to a purely perceptual definition of timbre – with continuous attributes such as brightness, warmth, dullness, roughness, and so forth [3] – yet without connecting these attributes to the discrete latent space of playing techniques.

In his 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 discrete set of instructions, readily interpretable by the performer) rather than by means of perceptual epithets only. We refer to [32] 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

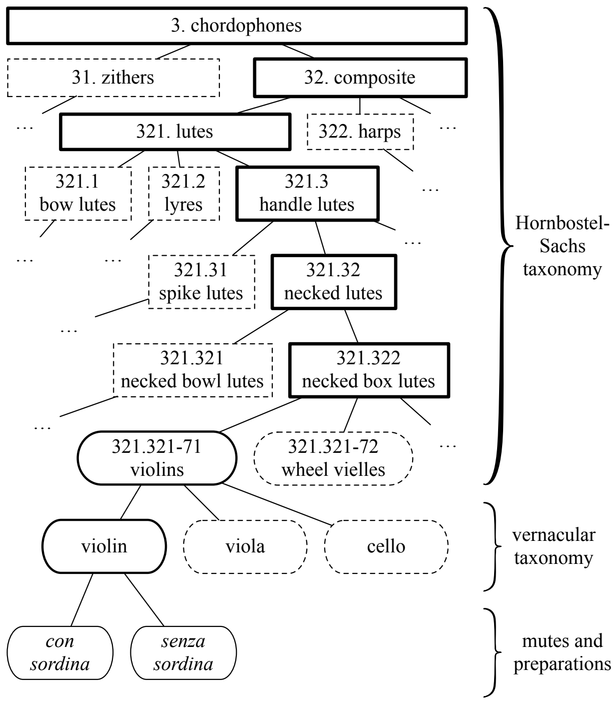


Figure 2: Taxonomy of musical instruments.

offers an unequivocal way of describing any acoustical instrument without any prior knowledge on its afferent playing techniques, 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 involves refining H-S with 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, playing techniques 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 any tree-like structure of that sort [28]. Instead, every playing technique is described by a finite collection of categories, each belonging to a different "namespace"; Figure 3 illustrates such namespaces

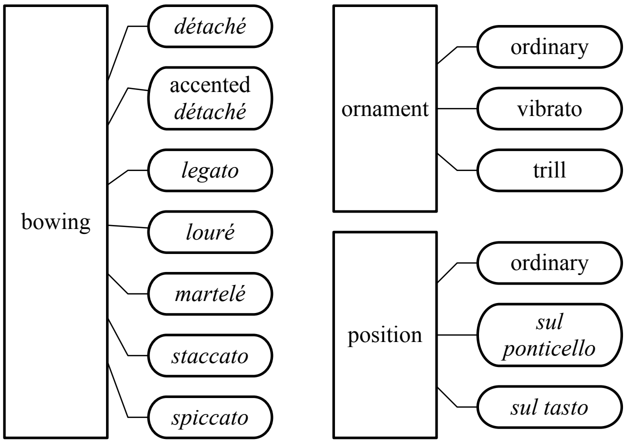


Figure 3: Namespaces of violin playing techniques.

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 [39,55] ; 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 samples to 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 [40]. 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.

Faithfully evaluating such a system is a difficult procedure, and ultimately would rest on its usability, as judged by the composers themselves.

ML : We formulate here the instrument / playing techniques recognition problem as a query by example one as we are *in fine* interested in an application scenario where the user specify a seed sound, be it played by a musical instrument or not, and the system returns several items from the database that have some level of similarity with the seed.

Designing such a system would requires the definition of a reference similarity space where the similarity among musical tones is defined perceptually. The aquisition of such reference is out of the scope of this paper and is left for future work. We thus aim here at validating the processing pipeline considered to implement the query by example system. We would like that such architecture is able to effectively rank items of the database according to 1) the partition of the musical instruments and 2) the partition of the playing techniques.

To evaluate the performance of the alternative implementations of the query by example system, the precision at rank k is considered. It is computed as the number of relevant items among the k items retrieved by the system divided by k. Revelance here is evaluated as the seed and the retrieved item having the same label in the reference partition.

When considering the whole corpus, each item of the database is considered as a seed, and the precision of the system under evaluation is computed for this seed. The precision for this system is the precision averaged over all the seeds. When considering a train /test split corpus, the seeds are taken from the test set and items of the train corpus similar to this seed are retrieved. The precision for this seed is computed among those retrieved items and the precision for the system under evaluation is the precision averaged over all the items of the test set.

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 is composed of 25444 recordings of isolated notes as played by 16 musical instruments. The distribution of these notes, shown in Figure 4, spans the 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 3 – 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 playing techniques in total.

4. METHODS

In this section, we describe the scattering transform and supervised metric learning used to implement the query by example processing pipeline.

4.1 Scattering transform

We refer to [2] for a general introduction to scattering transforms in audio classification, and to [34] for a discussion on its application to musical instrument classification in solo recordings. [1]

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

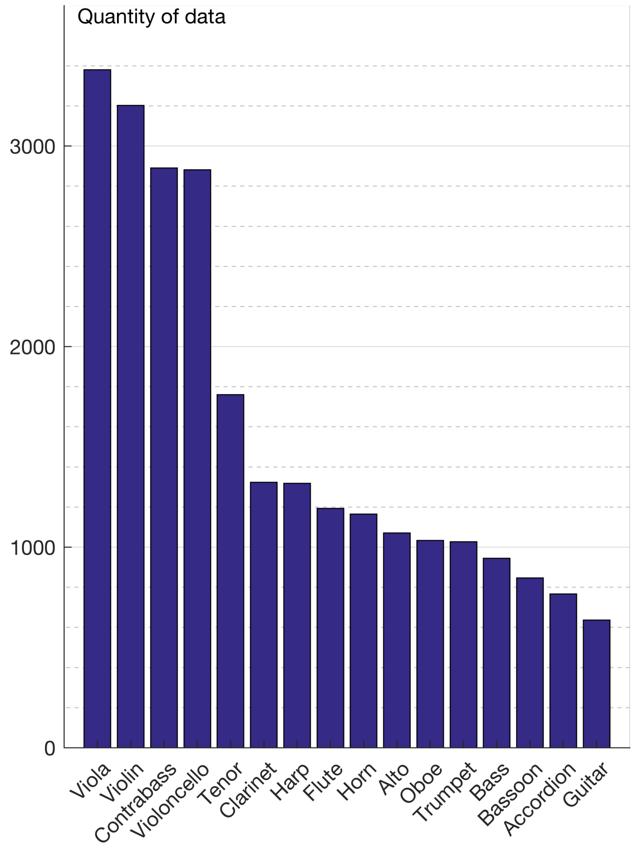


Figure 4: Instruments in the SOL dataset.

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

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

4.2 Metric learning

As shown in the experiments described in Section 5, 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 ($x \to Lx$) 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 [4] is to optimize a Mahalanobis matrix that linearly projects the input data into another feature space of the same dimensionality. In this case:

$$M = L^T L \qquad (2)$$

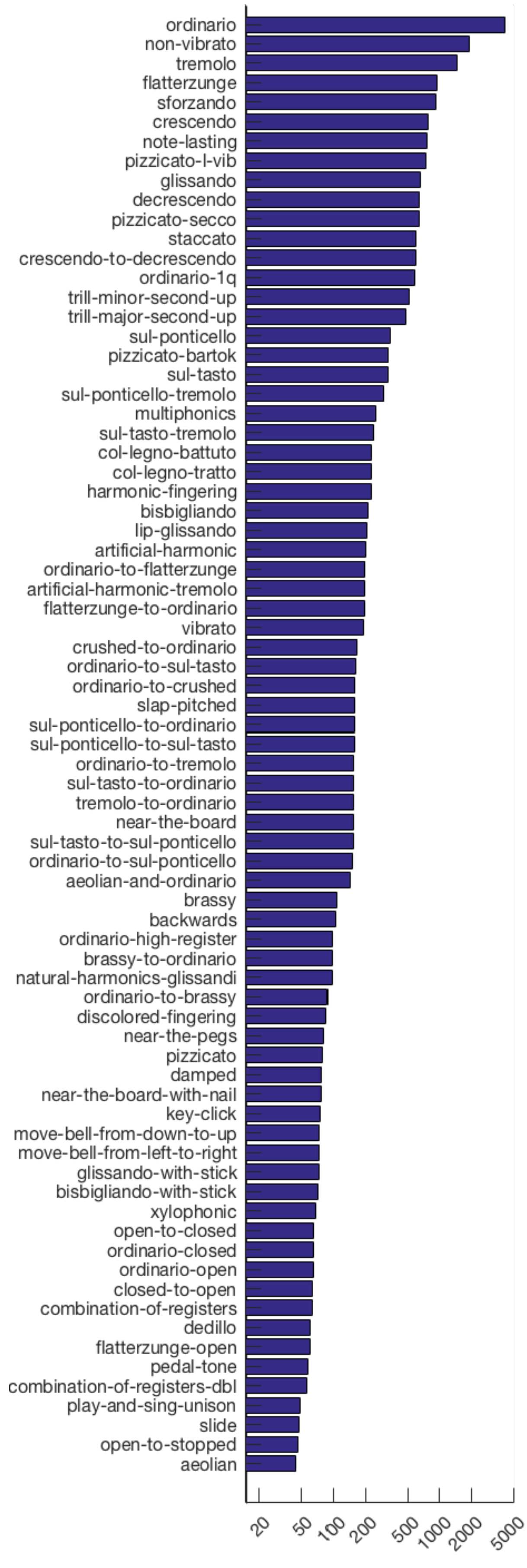


Figure 5: Playing techniques in the SOL dataset.

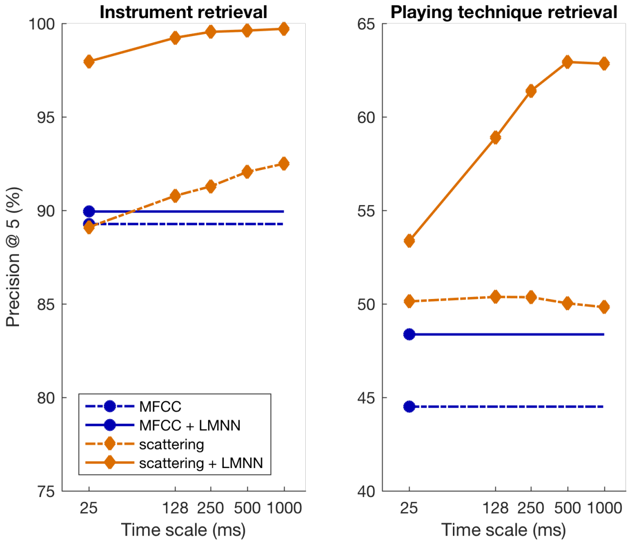


Figure 6: Summary of results on the SOL dataset.

and the resulting distance cane be expressed as follows:

$$D_m(x, y) = (x_i - x_j)^T M (x_i - x_j) \tag{3}$$

x_i x_j being features vectors. In this study, the large-margin nearest neighbors (LMNN) approach [53, 54] 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).

5. EXPERIMENTAL RESULTS

In this section, we apply the aforementioned methods to instrument classification and instrumental playing techniques classification in the Studio On Line (SOL) dataset.

5.1 Evaluation of instrument recognition

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.

MFCC: 89%. Keeping all 40 MFCC rather than the lowest 13 degrades accuracy down to 84%.

Scattering: 89%. Disabling median renormalization – i.e. setting 84%. Disabling logarithmic compression altogether: 76%. These results are consistent with [37].

5.2 Evaluation of playing technique recognition

MFCC: 45%.

6. 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; an unsupervised renormalization procedure; and a supervised metric learning over frequencies and modulation rates with large-margin nearest neighbors (LMNN).

7. REFERENCES

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

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

[3] Aurélien Antoine and Eduardo R. Miranda. Musical acoustics, timbre, and computer-aided orchestration challenges. In *Proc. ISMA*, 2018.

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

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

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

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

[8] DG Bhalke, CB Rama Rao, and Dattatraya S Bormane. Automatic musical instrument classification using fractional Fourier transform based-MFCC features and counter propagation neural network. *Journal Int. Inform. Syst.*, 46(3):425–446, 2016.

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

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

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

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

[13] Yuan-Ping Chen, Li Su, Yi-Hsuan Yang, et al. Electric guitar playing technique detection in real-world recording based on f0 sequence pattern recognition. In *Proc. ISMIR*, 2015.

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

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

[16] Tzenka Dianova. *John Cage's Prepared Piano: The Nuts and Bolts*. PhD thesis, U. Auckland, 2007.

[17] Patrick J Donnelly and John W Sheppard. Cross-dataset validation of feature sets in musical instrument classification. In *Proc. IEEE ICDMW*, pages 94–101. IEEE, 2015.

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

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

[20] Ferdinand Fuhrmann. *Automatic musical instrument recognition from polyphonic music audio signals*. PhD thesis, Universitat Pompeu Fabra, 2012.

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

[22] R.I. Godøy and M. Leman. *Musical Gestures: Sound, Movement, and Meaning*. Taylor & Francis, 2009.

[23] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. 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. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *Proc. Trans. Audio Speech Lang. Process.*, 25(1):208–221, 2017.

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

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

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

[28] Sefki Kolozali, Mathieu Barthet, György Fazekas, and Mark B Sandler. Knowledge representation issues in musical instrument ontology design. In *Proc. ISMIR*, pages 465–470, 2011.

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

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

[31] Mathieu Lagrange, Grégoire Lafay, Boris Défréville, and Jean-Julien Aucouturier. The bag-of-frames approach: a not-so-sufficient model for urban soundscapes. *J. Acoust. Soc. Am.*, 138(5):EL487–EL492, 2015.

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

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

[34] Vincent Lostanlen. *Convolutional operators in the time-frequency domain*. PhD thesis, 'Ecole normale supérieure, 2017.

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

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

[37] Vincent Lostanlen, Grégoire Lafay, Joakim Andén, and Mathieu Lagrange. Relevance-based quantization of scattering features for unsupervised mining of environmental audio. *Submitted to EURASIP J. Audio Speech Music Process.*, 2018.

[38] Mauricio A Loureiro, Hugo Bastos de Paula, and Hani C Yehia. Timbre classification of a single musical instrument. In *Proc. ISMIR*, 2004.

[39] Thor Magnusson. Musical organics: A heterarchical approach to digital organology. *J. New Music Research*, 46(3):286–303, 2017.

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

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

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

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

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

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

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

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

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

[49] Curt Sachs. *The History of Musical Instruments*. Dover Books on Music. Dover Publications, 2012.

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

[51] Li Su, Li-Fan Yu, and Yi-Hsuan Yang. Sparse cepstral, phase codes for guitar playing technique classification. In *Proc. ISMIR*, 2014.

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

[53] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Proc. NIPS*, pages 1473–1480, 2006.

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

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

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

[57] Luwei Yang, Elaine Chew, and Sayid-Khalid Rajab. Cross-cultural comparisons of expressivity in recorded erhu and violin music: Performer vibrato styles. 2014.

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

[59] Diana Young. Classification of common violin bowing techniques using gesture data from a playable measurement system. In *Proc. NIME*. Citeseer, 2008.

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