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

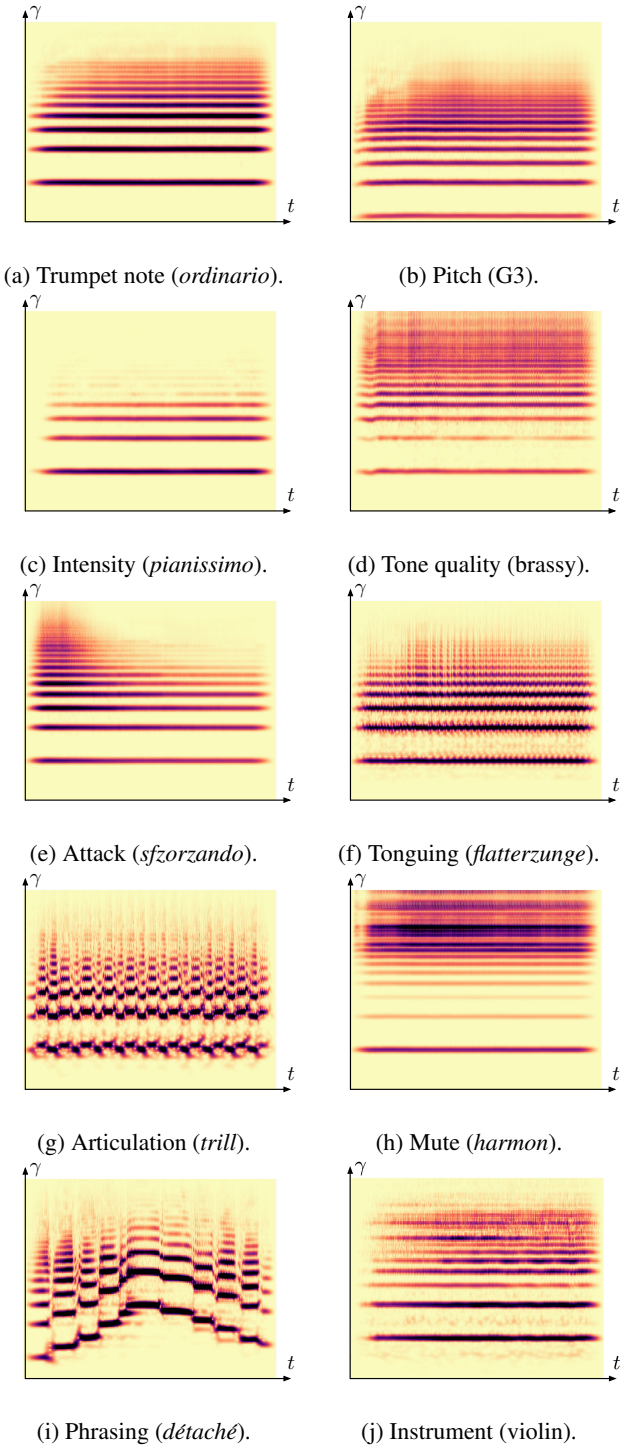
Third Author
Affiliation3
author3@ismir.edu

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 (IPT) remains largely underdeveloped, and far below human accuracy. This article provides the first benchmark of machine listening systems for query-by-example browsing among 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. On the Studio On Line (SOL) dataset, we report a P@5 of 99.7% for instrument recognition (baseline at 92.5%) and of 61.0% for instrumental 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) [45]; the creation of novel instrumental associations, as epitomized by *Klangfarbenmelodie* [46]; the temporary alteration of resonant properties through mutes and other "preparations" [15]; 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 digital audio effects [56]. The first of these trends has



(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.

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 [21] 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 [37]. 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.

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 [42], MIS, RWC [22], and Philharmonia – thus eliminating most factors of intra-class variability due to the performer [5,10,17,25,27,38,52]. 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 [13]. 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 [44]. 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 [7] 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% [43]. 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. This issue is partially mitigated by the recent surge of multitrack datasets, such as MedleyDB [8], which has spurred a renewed interest in monophonic instrumental recordings [54]. 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 [9]. 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 ag-

gregates MedleyDB and solosDB following the procedure of [16]. We refer to [23] 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 [11, 39]. 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 [19, 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 [40].

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. Outside of audio signal processing, this topic is connected to other disciplines, such as biomechanics and gestural interfaces [41]. Nevertheless, the vast majority of the available literature focuses on the range of playing techniques afforded by a single instrument: recent examples include clarinet [36], percussion [49], piano [6], guitar [12, 18, 48], violin [55], cello [14, chapter 6], and erhu [53]. On the other hand, the few publications which 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 [**?**] – yet without connecting these attributes to the discrete latent space of playing techniques. To the best of our knowledge, this paper is the first to 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 [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

Knowledge representation issues in musical instrument ontology design: [28]. Audio Set: [20]. Visipedia: [4]

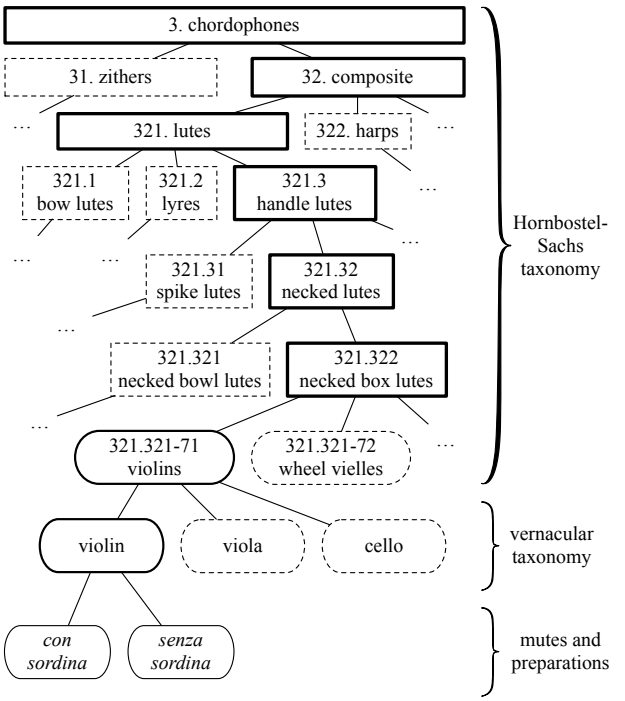


Figure 2: Taxonomy of musical instruments.

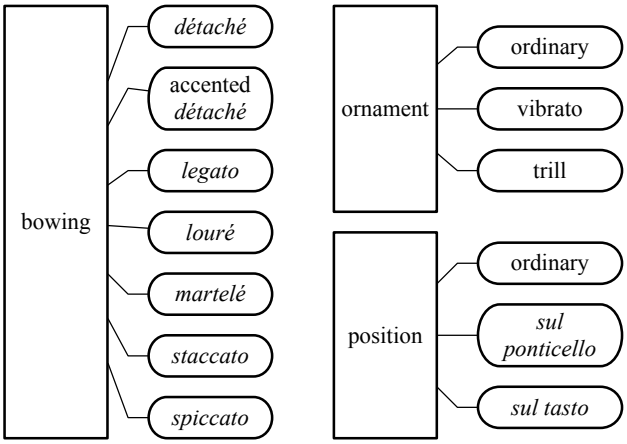


Figure 3: Taxonomies of playing techniques.

3.2 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.

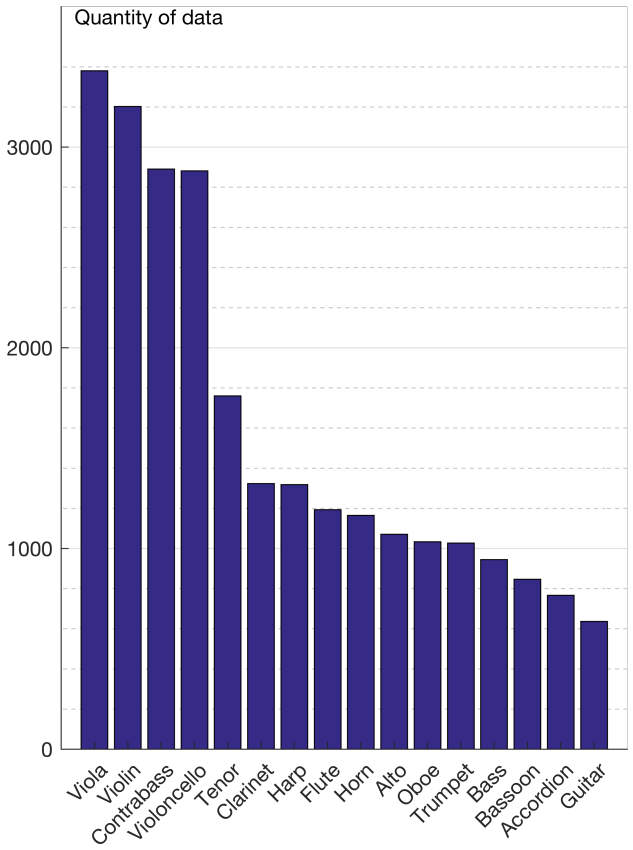


Figure 4: Instruments in the SOL dataset.

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 of these limitations.

4.1 Scattering transform

We refer to [2] for a general introduction to scattering transforms in audio classification, and to [33] for a discussion on its application to musical instrument classification in solo recordings. [1]

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 ($x \rightarrow Lx$) the input data into a features space of C-1 dimensions that maximizes the amount of between-class variance relative

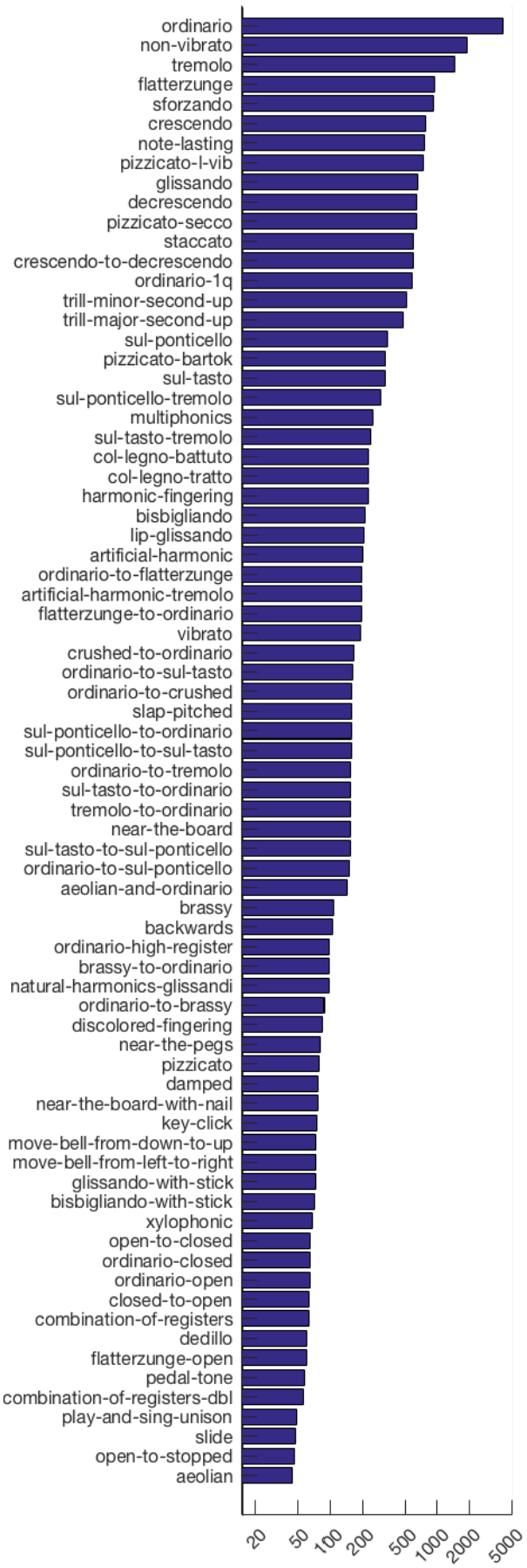


Figure 5: Playing techniques in the SOL dataset.

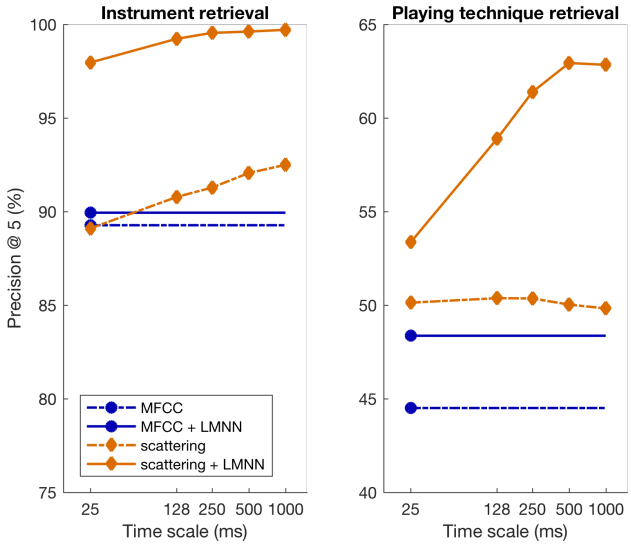


Figure 6: Summary of results on the SOL dataset.

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 [3] is to optimize a Mahalanobis matrix that linearly projects the input data into another feature space of the same dimensionality. In this case:

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

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 [50, 51] 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: 84%. Disabling logarithmic compression: 76%. These results are consistent with [33].

5.2 Evaluation of playing technique recognition

MFCC: 45%.

5.3 Pitch-adaptive metric learning

A well-known rule of thumb in psychoacoustics states that the "bandwidth for timbre invariance", i.e. the musical interval beyond which two notes are judged dissimilar, is of the order of one octave for untrained listeners [24] and up to 2.5 octaves for trained listeners [47].

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

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

[5] 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.

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

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

[16] 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.

[17] 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.

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

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

[20] 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.

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

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

[23] 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.

[24] Stephen Handel and Molly L Erickson. A rule of thumb: The bandwidth for timbre invariance is one octave. *Music Perception: An Interdisciplinary Journal*, 19(1):121–126, 2001.

[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] 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.

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

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

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

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

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

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

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

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

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

[41] 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.

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

[43] 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.

[44] 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.

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

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

[47] Kenneth M Steele and Amber K Williams. Is the bandwidth for timbre invariance only one octave? *Music Perception*, 23(3):215–220, 2006.

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

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

[50] 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.

[51] 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.

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

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

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

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

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