

PERCUSSION-RELATED SEMANTIC DESCRIPTORS OF MUSIC AUDIO FILES

PERFECTO HERRERA¹, VEGARD SANDVOLD², AND FABIEN GOUYON¹

¹ *Universitat Pompeu Fabra, Barcelona, Spain*

pherrera@iua.upf.es, fgouyon@iua.upf.es

² *University of Oslo, Oslo, Norway*

vegardsa@ifi.uio.no

Automatic extraction of semantic music content descriptors has traditionally focused on melodic, rhythmic and harmonic aspects. In the present paper, we will present several music content descriptors that are related to percussion instrumentation. The “percussion index” estimates the amount of percussion that can be found in a music audio file and yields a (numerical or categorical) value that represents the amount of percussion detected in the file. A further refinement is the “percussion profile”, which roughly indicates the existing balance between drums and cymbals. We finally present the percussivity descriptor, which represents the overall impulsiveness or abruptness of the percussive events. Data from initial evaluations, both objective and subjective will also be presented and discussed.

INTRODUCTION

Automatic extraction of music content metadata has traditionally focused on melodic, rhythmic and harmonic aspects. On the contrary, timbre or instrumentation descriptors have been traditionally missing from research agendas. Among the most content-informative instrumentation-related features, we find those that can be extracted by focusing on percussive events. The mainstream approach to music content processing from audio files is the one we term the *transcriptionist* approach. According to this approach, describing music content equates to extracting a score-like representation of the original audio. Source separation is also a natural strategy under this approach. But using a score as a ground truth for matching the output of a content processing system makes sense only when the intended user of the system is a musically-educated one. This is not the case with most of the users of existing music downloading systems, which amount probably more than ninety percent of the users of music retrieval systems. In contrast with this transcriptionist view, we advocate here for a *descriptionist* approach, which has also been advocated elsewhere by Martin et al. [1], or Carreras and Leman [2]. The descriptionist approach is an ecological and user-centred way of addressing the description of music contents from audio files. It is an ecological approach because the research context is that of a system in use, the structure and functionalities of which will pose specific problems and will shape the knowledge structures of the users. It is a user-centered approach because the attempted solutions spring from

the user needs and requirements, and not by a pre-existing musical theoretical construct.

According to this approach, we advance several percussion-related descriptors that we have named *percussion index*, *percussion profile*, *kick-snare crossings*, and *percussivity*. They do not correspond to solid musical theoretical entities, but we suggest that, on the other hand, they correspond to entities that are (or can be) represented in the minds of the users of music information retrieval systems. Because of that, they can be exploited, taken one by one or combining them synergistically, to define and refine query and retrieval operations of music files.

Although percussion has been traditionally the poor relative in music or signal processing research, in the last two years we have witnessed a growing wealth of papers focusing on it, mainly with a focus on transcription. Goto and Murakoa [3] studied drum sound classification in the context of source separation and beat tracking [4]. They implemented an “energy profile”-based snare-kick discriminator, though no effectiveness evaluation was provided. More recently, Zils et al. [5] reported very good performance rate at identifying kicks and snares in songs by means of a technique of analysis and incremental refinement of synthesis that was originally developed by Gouyon [6]. Jørgensen [7] attempted to use cross-correlation between sound templates extracted from isolated sound recordings and realistic drum-kit recordings. Using this technique only kicks and snares seem to be detected with some reliability. A very different motivation has been that of Kragtwijk et al. [8] who have presented a 3D virtual drummer that re-creates with synthetic images the playing movements of a real drummer after

analyzing the audio input coming from a real performance. Unfortunately, the audio analysis part of the system was finally underdeveloped.

Riskedal [9] developed a system that combined a drumloop-adapted onset detection procedure (taken from Klapuri [10]) with an adaptation of Independent Component Analysis [11], a source separation technique. Though the results seemed promising, only a few examples were presented by the author.

Orife [12] developed, again without providing systematic evaluation, a tool for extracting rhythmic information from loops and songs using Independent Subspace Analysis (ISA), a technique for source separation discussed by Casey and Westner [13]. More recently, Fitzgerald et al. [14] have also taken advantage of ISA and of some additional sub-band pre-processing, and reported a success rate of 89.5% when transcribing a database of 15 drum loops containing snare, kick and hi-hats. When knowledge about the sources was incorporated, then a Prior Subspace Analysis technique [15] allowed them to achieve 92.5% of correct identifications. When they extended the technique to music (i.e. mixtures of pitched instruments and drums) [16] they got 89.3% of correct identifications for a database of 25 drum loops. A possible drawback, though, is that the system needs human intervention, and is fit to a specific kind of sounds. Uhle et al. [17], also using ISA and descriptors like percussiveness, noise-likeness, spectral dissonance, spectral flatness, and the third order cumulant for classifying the segments after ISA decomposition, reported 95% percent of correctness for a database of only 9 music titles. Paulus and Klapuri [18] have used acoustic models and N-gram based models that allowed getting acceptable performance rates, though the approach seemed to be not general enough. More recently [19], they have improved the system by including rhythm information as an additional cue for label assignment. The system has been tested with 359 MIDI songs (that were rendered into audio), yielding up to 86% of correct decisions. Virtanen [20], using sparse coding, i.e. another source separation technique, achieved 66% of effectiveness when separating the kick and the snare from the rest of musical mixes from a database of 100 MIDI polyphonic songs (that were rendered into audio). A different approach, based on Parametric Vector Quantization, has been presented by Wang et al. [21]. Their system features computed on 3 sub-bands for clustering those events that could reliably be considered as “percussive”. Small scale evaluations yielded an average of 84.5% of correctness. Steelant et al. [22] have also addressed the transcription problem and their first results classifying kick and snare sound slices extracted from music CDs and MIDI songs achieved recall rates ranging from 85% to 96%, depending on the type of Support Vector Machine they used for inducing the classification models.

One major criticism in most of the existing research is that of using small databases (though in the current status our research also suffers from the same drawback). Another one is that of disregarding performance evaluations that are based on independent samples (i.e. not used during the analysis or training phase). As we discussed elsewhere [23], the estimations can therefore be as much as 10% overoptimistic when training and testing is done using the same sample of observations.

Our approach, in contrast to the previous ones, has purposely skipped sound separation techniques in order to keep computation requirements as low as possible, and has not focused on absolute transcription of percussive events, but on the creation of semantic descriptors related to them.

We suggest the interested readers a forthcoming publication [24] that will present and discuss in depth the details that this poster presentation is leaving aside.

1 PERCUSSION-RELATED DESCRIPTORS

1.1 Pre-processing

Most of the descriptors presented here are computed after an onset detection and percussion segment extraction process. We use an algorithm adapted from Klapuri [10]. After detecting onsets, audio slices of 100 milliseconds starting from the onset position are segmented. A high-pass filter is then applied to the extracted slices though the original version is also kept for being used by some of the classifiers used for the assignment of instrument labels.

1.2 Percussion Index

The *Percussion Index* (PI from now on) estimates the amount of percussive events that can be found in a music audio file, allowing a user to query for music titles containing from “no percussion at all” to “a lot of percussion”.

In order to classify the audio slices as containing a percussion sound or not, we take advantage of a long list of low-level timbre-related descriptors including MFCCs, energy in Bark bands, spectral centroid, skewness, flatness, their variances, etc. We use them as input for with inductive modelling techniques that take the decision on the presence/absence of percussion in the extracted segments. We have found several meta-learning approaches such as bagging [25] or boosting [26] yielding the best results (compared to other more traditional options such as neural networks, support vector machines, simple decision trees, or lazy learning). Once a dichotomic decision has been taken for each extracted segment, the PI is computed as the ratio between the events containing percussion against those that do not contain any.

A ground-truth evaluation database has been set up in order to estimate the reliability of the index for the task of song description. The database contained twenty-nine 20" segments extracted from songs coming from all kind of genres and epochs (from nineteen-forties to the current year). This amounts up to almost 1500 events to be classified. The objective evaluation of the H was done by comparing the PI obtained using the twenty-nine songs that were manually annotated with the PI that was obtained when the songs were automatically segmented and classified. The Pearson correlation coefficient was of 0.67 ($p < 0.001$), which is an evidence of the validity of the automatically computed PI (i.e. it is computing what is intended to be computed, under some error tolerance), but also is a hint that there is still some room for improvement.

The nature of the PI makes it quite robust to local misses of onsets. Even if the onset detector misses some of the percussive or non-percussive events, the computed value still keeps overall meaningfulness for the whole song. In addition, a certain amount of misclassifications can also be accepted. The robustness of the PI is even increased once we convert the real value into a discrete one: although the original computation of the PI yields a real-value which is bounded between 0 and 1, a quantization is needed in order to make it usable in the real-world. A JND or just noticeable difference can be asked for and, even though the evaluation is still being finished, it seems that it will yield between four and five discrete labels for a final user to exploit consistently the underlying concept (*no percussion, low amount of percussion, fair amount of percussion and lots of percussion*).

Bartok: Music for Strings, Percussion and Celesta IV Allegro molto	0.22
Mancini: Pink Panther Theme	0.46
Pet Shop Boys: It's a Sin	0.56
The Beatles: Help!	0.66
Portishead: Strangers	0.91

Table 1. An example of sorting popular titles according to the raw PI value (it was computed using only an excerpt of 20 seconds).

1.3 Percussion Profile

The *Percussion Profile* (PP from now on) is a further refinement of the PI. It has been devised in order to yield an index that gives relative information about specific classes of percussion instruments. The PP makes possible to describe songs as "having lot of cymbals" or "being short on snares". The practical requirement for achieving a reliable PP computation is the existence of reliable models for the specific classes we are interested in (for example, membranes versus plates and finer distinctions). Hence, the problem

addressed by this descriptor is much more complex than that of the PI, as here we need to discriminate between, for example, a snare and a hi-hat, in the context of a mixture containing also harmonic sounds. Even worse, sometimes we will find simultaneous occurrences of several classes of percussion instruments (i.e. kick + hihat, snare + hihat).

Three different strategies can be envisioned to achieve the computation of the PP: raw class modelling, standard pre-processing of the slices (i.e. using filtering or enhancements, or trying reverse-engineering of noise-reduction techniques), and source separation prior to attempting class decisions. We have opted for the raw classification, though a high-pass filtering is done by the classifier that is specialized in the cymbals. The effective computation of the PP is achieved by a hierarchical classifier that first separates percussion slices from non-percussion slices, then, percussion slices are divided into membranes and plates, then single hits or combinations of sounds are detected, and finally the kick and snare are further separated. All the computed decisions are then integrated and the final label assignment {kick, snare, cymbal, kick+cymbal, snare+cymbal, no-percussion} is computed. Each classifier uses a different set of features and the error rate of them, using a holdout database taken from a different song collection than the one used for learning the class models is kept under 15%. An absolute objective measurement (i.e. comparing the automatically assigned labels with manually annotated labels) is under way.

1.4 Kick-Snare Crossings

Kick and snare hits do not usually happen at the same time as they contribute to keep a contrastive running beat pattern. Their roles are different, and the ways their sounds are organized along time contribute to define specific rhythm and timbre patterns. For example, in the 80's pop music, the snare was to be found in every downbeat of a 4/4 bar (i.e. 2nd and 4th), whereas in reggae music the snare can be usually found mostly in the 3rd beat. Hence, the changes from kick to snare are expected to be different, on average, for each of these two genres. The average number of changes from kick to snare and the other way round is computed as the Kick-Snare Crossings (KSC) descriptor. As the current formulation of this descriptor does not yield a true semantic descriptor we are searching for ways to map it into a proper concept to be exploited for end-users of a MIR system.

1.5 Percussivity

Contrasting with the previous ones, the Percussivity descriptor includes a heavy perceptual quality in its definition. According to the previously presented descriptors, two songs can be quite similar, but one of

them having the percussion parts much more prominent than the other (for example, by mixing percussion sounds to be “slapping in the face” when compared with the rest “backing” instrumentation). Another example can be found when comparing a given rhythm pattern played with brushes or played with sticks. In the first case the percussive sensation is lower than in the second case. Percussivity can also be found whenever no percussion instruments are played at all. That is the case of a string section excerpt played *pizzicato*, which will sound more percussive than the same excerpt played *bowed*. We suggest quantifying this kind of sensation by means of defining a “percussivity” descriptor that is computed by adapting an approach taken from research on the perceived impulsiveness of environmental sounds [27]. Given the nature of the percussivity descriptor, it can only be evaluated by means of a listening experiment using human subjects.

2 DISCUSSION

The descriptors we have presented here are intended to complement other more usual descriptors of rhythm, tonality, melody, structure, or even those belonging to a more arcane *complexity* facet. All of them will soon be integrated in the prototypes developed under the SIMAC project. We believe that a truly usable MIR system for music audio titles needs to exploit descriptors addressed to most of the musical facets that can be computed, and this exploitation has to be done according to the knowledge level of the intended users. It is usually the case that their knowledge level does not have much to do with the constructs managed by music theories. Our descriptors seem to be valid and reliable, and keep a kind of coarseness that makes them learnable and usable in practical situations.

We still have important open issues such as the conversion of the KSC into a truly semantic descriptor, or the improvements of the PP. There is also the need to perform a truly large-scale evaluation using an independent (i.e. containing instances that were not used during the learning phase of the instrument models) test database.

A final additional refinement that is worth to be considered is that of distinguishing between *percussivity* and *percussiveness*. The later can be defined as a quality or possibility of a sound for being percussion-like (see Uhle et al. [17] for a formal definition that is not far from our concept). For example, there are synthetic sounds that, even though they are not being created by percussion instruments, can be considered as playing the role of them and/or mocking them (i.e. a blast of filtered noise, or some noises made by striking the keys of wind instruments). This kind of percussiveness would probably deserve a new descriptor too.

3 CONCLUSIONS

Extracting semantic descriptors from music audio files does not necessarily call for a transcription. Scores are only one of the multifarious semantic devices we can take advantage of when trying to describe music contents and most of the semantic descriptors that can be managed after transcribing a musical piece can only be exploited by people having been educated under occidental-music traditions. In order to find truly usable semantic descriptors for popular music retrieval systems (i.e. Kazaa, iTunes, Napster) we have to quest for extracting knowledge from users of them. The proposed descriptors can be considered the first captures we have got in this quest. The work we are reporting here is still unfinished and some improvements on the performance rates should be expected soon, in addition to other ones we have just started to identify.

ACKNOWLEDGEMENTS

This research has been partially funded by the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents). More information can be found at the project website <http://www.semanticaudio.org>.

REFERENCES

- [1] K. D. Martin, E.D.Scheirer, and B. L. Vercoe, "Music content analysis through models of audition," in *Proc. of the ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications*, Bristol UK (1998).
- [2] F. Carreras, and M. Leman, "Automatic harmonic description of musical signals using schema-based chord decomposition", *Journal of New Music Research*, 28, 4, pp. 310-333 (2000).
- [3] M. Goto and Y. Muraoka, "A sound source separation system for percussion instruments," *Transactions of the Institute of Electronics, Information and Communication Engineers D-II*, Vol. J77, No. 5, pp. 901-911 (1994).
- [4] M. Goto and Y. Muraoka, "A real-time beat tracking system for audio signals," *International Computer Music Conference*, pp. 171-174 (1995).
- [5] A. Zils, F. Pachet, O. Delerue, and F.Gouyon, "Automatic Extraction of Drum Tracks from Polyphonic Music Signals," *2nd International Conference on Web Delivering of Music (WedelMusic)* (2002).
- [6] F. Gouyon, F. Pachet, and O. Delerue, "On the use of zero-crossing rate for an application of

- classification of percussive sounds," *Proc. of the Conference on Digital Audio Effects* (2000).
- [7] M. Jørgensen, "Drumfinder: DSP project on recognition of drum sounds in drum tracks," <http://www.daimi.au.dk/~elmer/dsp>
- [8] M. Kragtwijk, *Percussive Music and the Automatic Generation of 3D Animations*, M.Sc. Thesis, University of Twente, The Netherlands, (2001).
- [9] E. Riskedal, *Drum Analysis*, M.Sc. Thesis, Dept. of Informatics, Univ. Bergen, Norway (2002).
- [10] A. Klapuri, "Onset detection by applying psychoacoustic knowledge," *International Conference on Acoustics, Speech, and Signal Processing* (1999).
- [11] A. J. Bell and T. J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, Vol. 7, No. 6, pp. 1129-1159 (1995).
- [12] I. Orife, *Riddim: A rhythm analysis and decomposition tool based on independent subspace analysis*, Master of Arts Thesis, Dartmouth College, Hanover, NH (2001).
- [13] M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," *Proc. of the International Computer Music Conference*, Berlin, Germany (2000).
- [14] D. FitzGerald, E. Coyle, and B. Lawlor, "Sub-Band Independent Subspace Analysis for Drum Transcription," *5th International Conference on Digital Audio Effects (DAFX-02)*, pp. 65-69 (2002).
- [15] D. FitzGerald, E. Coyle, and B. Lawlor, "Prior subspace analysis for drum transcription", *114th AES convention*, Amsterdam, The Netherlands (2003).
- [16] D. FitzGerald, B. Lawlor, and E. Coyle, "Drum transcription in the presence of pitched instruments using Prior Subspace Analysis", *Irish Signals and Systems Conference*, Limerick, Ireland (2003).
- [17] C. Uhle, C. Dittmar, T. Sporer, "Extraction of drum tracks from polyphonic music using Independent Subspace Analysis", *Proceedings of the International Conference on Independent Components Analysis*, Nara, Japan (2003).
- [18] J. Paulus, J. and A. Klapuri, "Conventional and Periodic N-grams in the Transcription of Drum Sequences", in *Proc. of IEEE International Conference on Multimedia and Expo*, Baltimore, USA (2003).
- [19] J. Paulus, and A. Klapuri, "Model-based event labelling in the transcription of percussive audio signals", in *Proceedings of the 6th Conference on Digital Audio Effects*, London, UK (2003).
- [20] T. Virtanen "Sound Source Separation Using Sparse Coding with Temporal Continuity Objective," *Proceedings of the International Computer Music Conference*, Singapore, Malaysia (2003).
- [21] Y. Wang, J. Tang, A. Ahmaniemi, M. Vaalgamaa, "Parametric vector quantization for coding percussive sounds in music". *Proceedings of the International Conference on Speech and Signal Processing*, Hong Kong (2003).
- [22] D. van Steelant, K. Tanghe, S. Degroeve, B. De Baets, M. Leman, and J.-P. Martens, "Classification of percussive sounds using Support Vector Machines", *Proceedings of the annual machine learning conference of Belgium and The Netherlands*, Brussels, Belgium (2004).
- [23] P. Herrera, A. Dehamel, and F. Gouyon, "Automatic labeling of percussion sounds", *114th AES convention*, Amsterdam, The Netherlands (2003).
- [24] P. Herrera, V. Sandvold, F. Gouyon, E. Pampalk, "Semantic interaction with music audio contents using percussion-related descriptors", *Journal of Intelligent Information Systems* (accepted).
- [25] L. Breiman, "Bagging predictors", *Machine Learning*, 24, pp. 123-140 (1996).
- [26] R. E. Schapire, "A brief introduction to boosting", *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* (1999).
- [27] T. H. Pedersen, "Objective method for measuring the prominence of impulsive sounds and for adjustment of LAeq", *Proceedings of the International Congress and Exhibition on Noise Control*, Den Haag, The Netherlands (2001).