FROM LOW-LEVEL TO SONG-LEVEL PERCUSSION DESCRIPTORS OF POLYPHONIC MUSIC

Martín Haro, Perfecto Herrera
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{martin.haro,perfecto.herrera}@upf.edu

ABSTRACT

We address here the automatic description of percussive events in real-world polyphonic music. By taking a pattern recognition approach we evaluate more than 2,450 object-level features. Three binary instrument-wise support vector machines (SVM) are built from a training set of more that 100 songs and 10 genres. Then, we use these binary models to build a drum transcription system achieving comparable results with state of the art algorithms. Finally, we present 17 song-level percussion descriptors computed from the imperfect output of the transcription algorithm. We evaluate the usefulness of the proposed descriptors in music information retrieval (MIR) tasks like genre classification, danceability estimation and Western vs. non-Western music discrimination. We conclude that the presented song-level percussion descriptors provide complementary information to "classic" descriptors, that can help in the previously mentioned MIR tasks.

1. INTRODUCTION

During the last decade the interest in the transcription of percussive instruments has grown and most of the work has focused on the problem of drum [1] transcription [1]. The aim of such systems is to obtain, from an audio signal, a representation of the type of percussion instrument played (instrument recognition), and when it has been played (temporal location).

The transcription of isolated or polyphonic drum sounds (i.e. without concurrent pitched sounds) can be considered a practically solved problem (e.g. see [2]). However, the automatic transcription of percussive events in polyphonic music is an open problem where there is still a lot of room for improvement.

Instead of focusing on a full transcription system, we consider that, when MIR of polyphonic music is addressed, an automatic music "description" approach should be taken. The main idea behind such an approach is to obtain "predicates" or labels that apply to a given music excerpt and usually this information goes beyond traditional music scores.

In this paper we present and evaluate several song-level percussion descriptors extracted from the output of an imperfect transcription system. The aim of these descriptors is to semantically describe general characteristics of within-song drum events such as drum-instrument degree of presence, drum-instrument relationships (i.e. inter-instrument ratios) and most-frequent inter-instrument intervals. Finally, we explore the usefulness of the proposed descriptors for some MIR tasks such as genre classification, danceability estimation and Western vs. non-Western music classification.

The paper is organized as follows: An overview on percussion transcription of polyphonic music is presented in section 2. In section 3 "full" and "relaxed" transcription systems are described. Next, song-level percussion descriptors are proposed and evaluated within several MIR tasks (section 4). Finally, section 5 presents some conclusions.

2. RELATED WORK

Most of the works on transcription of percussive events in polyphonic music have focused on transcription of drum kit sounds, specially on bass drum (BD), snare drum (SD) and hi-hat (HH) sounds. In Table 1 a summary of the most relevant works on drum transcription in polyphonic music is presented. It is worth to notice here that the presented results can not be directly compared since they were not evaluated on the same dataset. Sandvold et al. [3] used a combination of general and localized sound models. The correct classified instances obtained from the general model (C4.5 with AdaBoost) were manually parsed to a localized training set. Yoshii et al. [4] achieved the best results in the MIREX 2005 [2] audio drum detection contest by using matching template spectrograms. The main idea here was to obtain a template spectrogram representation of a particular percussion instrument from a large training database of sounds. Thus, when analyzing a song, a template-adaptation algorithm was applied on every onset. A distance measure for template matching was used to try to minimize the spectral overlapping of other sounds. Dittmar's [5] system was also evaluated within

[1] The word "drum" refers to a standard Rock/Pop drum kit found in Western music.

[2] See MIREX web site: http://www.music-ir.org/evaluation/mirex-results/audio-drum/index.html for more information.

Authors	# songs	Approach	Main Algorithms	Overall	BD	SD	HH
Sandvold et al. [3]	25	Patt-Rec	local. model	0.924	0.951	0.931	–
Yoshii et al. [4]	50	Sp. Temp	Temp. match.	0.670	0.728	0.702	0.574
Dittmar [5]	50	S. Sep+Sp. Temp	NN ICA	0.588	0.606	0.581	0.585
Tanghe et al. [6]	50	Patt-Rec	SVM	0.615	0.688	0.555	0.601
Gillet and Richard [7]	20	Patt-Rec	SVM+local. model	0.840	0.824	0.842	–
Paulus and Klapuri [8]	45	Patt-Rec	HMM	0.697	0.795	0.655	0.660
Gillet and Richard [9]	28	S. Sep+Patt-Rec	SVM	0.678	0.695	0.583	0.755

Table 1. Summary of drum transcription systems in polyphonic music. Almost all works classify bass drum (BD), snare drum (SD) and hi-hat (HH) sounds, except for Gillet and Richard [7] (where only BD and SD were detected) and Sandvold et al. [3] (where BD, SD and Cymbal were computed)."Approach" considers Pattern Recognition (Patt-Rec), Source Separation (S. Sep), and Spectral Template (Sp. Temp). F-measures results (except for Sandvold et al. where accuracy was measured). Since different datasets were used, results can not be directly compared.

the MIREX contest. It combined source separation (Non-Negative ICA) with template matching algorithms. Tanghe et al. [6], another MIREX participant, presented an onset-based classification system using N-binary SVM as recognition algorithm (being N the number of instruments to detect). Gillet and Richard [7] also used N-binary SVM as classification algorithm. This system performed a band-wise harmonic/noise decomposition as pre-processing step to enhance the presence of unpitched instruments. A localized adapted model like the one presented in [3] was also evaluated. Paulus and Klapuri [8] presented an evaluation using Hidden Markov Models with a combination of spectral features and temporal descriptors calculated from long narrow-band frames. In a recent work by Gillet and Richard [9] a combination of source-separation and pattern-recognition algorithms was proposed. Two set of features were computed, one from the original audio signal and the other from a "drum enhanced" track obtained by source separation. These feature vectors were then classified by three binary SVM.

As can be seen from the literature review, there is still a lot of room for improvement in the problem of drum transcription in polyphonic music. But, instead of pursuing a perfect transcription system, we decided to explore the potential of song-level percussion descriptors to describe real-world music. In order to achieve that we implemented a simple drum transcription system. This system could be used later as a baseline for more complex implementations (e.g. based on source separation or harmonic/noise decomposition). Thus, we chose to follow a standard pattern-recognition approach, trained on a large set of sounds and audio features. The potential of this kind of basic system to describe percussive events in polyphonic music was not previously assessed.

3. TRANSCRIPTION EXPERIMENTS

3.1 Datasets

We used three song collections with proper annotations of percussive events. Two of them are publicly available and were used in previous studies on drum transcription.

ENST-Drums database [10]: This is the largest publicly available drum database. Since we wanted to detect drum events in "real" music, we decided to mix the provided "wet" drums and their accompaniment tracks without further changes on amplitude (-6dB drum level). From the obtained collection of 64 songs we randomly selected 30 seconds excerpts of each song and their labels.

MAMI database [11]: This database is a collection of 52 music fragments (30 seconds length) extracted from commercial CDs. We managed to gather 48 songs and aligned them with the provided annotations. This database was one of the three databases used in the MIREX 2005 audio drum detection contest.

In-house database [12]: This is a database of 30 annotated music excerpts (20 seconds length), extracted from commercial CDs. Since HH events were not specifically annotated, we only used the BD and SD labels.

Due to the number of instances and the musical importance of each instrument within the drum kit we decided to work with the following instruments classes: BD, SD and HH (including open and closed HH sounds). Finally, we obtained a large set of polyphonic music excerpts adding up a total of 142 songs labeled with three, possibly concurrent, tags. The final number of instances per instrument was: 6,407 for BD, 5,655 for SD and 8,400 for HH.

3.2 Descriptors

First, we computed a set of frame-level descriptors (frame-size of 46 ms, hop-size of 12 ms) namely: temporal descriptors (zero-crossing rate and lpc coefficients), spectral descriptors (e.g. centroid, complexity, crest, decrease, dissonance, energy, flatness, flux, kurtosis, pitch, rms, rolloff, skewness, spread, strong-peak), perceptual descriptors (MFCCs, Bark-bands and Bark-bands kurtosis, skewness and spread) and tonal descriptors (Harmonic Pitch Class Profile). See [13] and [14, p. 20] for an overview on these descriptors.

After this first step we computed a set of object-level descriptors [3] from the time series of each frame-level descriptor (about 12 frames per object). The computed object-level descriptors were:

[3] The term "object" is considered here as: every sound event starting from an onset and finishing 150 ms after (or in the next onset if this new onset falls within the 150 ms interval).

a) Amplitude-related object descriptors: mean, variance, minimum, maximum, skewness and kurtosis.

b) Time-related object descriptors: temporal skewness, temporal kurtosis, temporal centroid, max. and min. normalized position (normalized temporal position of the maximum, or minimum, value of the time series), slope (arctangent of the slope of the linear regression of the data), attack and decay (slope descriptor from initial, or end, point to the maximum point) and amplitude-normalized attack and decay.

At the end of this process we obtained about 2,400 descriptors for every sound object. See [14, p. 46] for a detailed explanation on the computed descriptors.

3.3 Model training

Since we were working with three drum categories that can occur at the same time, we decided to train N-binary (SVM) classifiers instead of one model with 2^N possible classes. In this context we have each trained model in charge of detecting the presence or absence of one particular instrument (e.g. SD or not-SD).

In order to have a more representative database for training purposes we mixed the ENST and the MAMI databases. Taking into account that the final system has to label predetected onsets we decided to train our models with labeled onsets. Thus, we performed an onset detection (by using an implementation of the onset algorithm proposed by Brossier in [15]) and we assigned the corresponding labels to every detected onset. Finally we split the database leaving 90% for training and reserving 10% to be used as independent test set. We called these databases 90%MIX and 10%MIX. The In-House database was also reserved as second independent testing set.

To build the SVM models we first used the correlation based feature selection (CFS) [16] algorithm in 10-fold cross-validation (CV) to identify the most informative object-level descriptors from the 90%MIX database. We chose only those descriptors selected in all CVs (i.e. 10 times) obtaining 56 relevant descriptors for BD (e.g. low Bark-bands, MFCCs, spectral-energy low and spectral flux), 77 for SD (e.g. mid Bark-bands, temporal lpc, MFCCs and spectral flatness) and 38 for HH (e.g. high Bark-bands, temporal lpc, MFCCs, spectral spread and spectral flatness). Then, we trained the SVM models with the selected descriptors of the 90%MIX database and evaluated their performance using 10-fold CV. We also applied these models to the testing sets (i.e. 10%MIX and In-House). The classification results for every labeled instance and every model (after a grid search of SVM parameters) can be seen in Table 2. We obtained averaged F-measure results of 0.806 and 0.782 for the training and testing sets respectively.

3.4 Full transcription

Since up to this step we had worked only with labeled onsets, the next step was to evaluate the learned models against all the ground truth labels in the datasets. In order to do that we implemented a complete drum transcription system. The three previously described databases were analyzed (ENST, MAMI and In-house) adding up a total of 142 songs (20 to 30 seconds length).

Instrument Model	90%MIX	In-House	10%MIX
bass drum	0.834	0.812	0.835
snare	0.778	0.687	0.773
hi-hat	0.806	—	0.802

Table 2. F-measure classification results after grid search of SVM parameters. Models were trained with 90%MIX database. Results were evaluated using 10-fold CV on each dataset.

The experiment set-up for evaluating the transcription capabilities of our system was as follows: a) Perform an onset detection on the audio excerpts (we used the same onset detector as in the model training step). b) Compute the descriptors used by each model on every onset plus 150 ms (or until the next onset). c) Apply the models to every set of descriptors to obtain the predicted labels. d) Evaluate the predicted results against the ground truth annotations (as in the MIREX 2005 contest, a range of ± 30 ms from the true times was allowed). After evaluating all 142 song excerpts, we obtained an overall result of 0.659 (F-measure) and per instrument F-measure results of 0.699 for BD, 0.652 for SD and 0.626 for HH. If we compare our system with the fully automatic systems described in section 2 (i.e. [4, 6, 8, 9]) we can see that our system obtained near state-of-the-art drum transcription performance with a quite simple pattern recognition algorithm. Nevertheless, these performances are still far from reliable transcriptions.

3.5 Relaxed Transcription

Taking into account that state-of-the-art algorithms are still far from yielding perfect transcriptions and that our final goal was to derive song-level percussion descriptors, we decided to evaluate the capacity of our transcription system to estimate the total number of drum events in a song (e.g. how many BD, SD or HH strikes a particular song has). These descriptors could be used to characterize a song as having, for example, a lot of SD, no HH, etc., hence they contribute to bridge the semantic gap [17]. In this experiment we considered as a "correct" decision the total number of instrument instances (e.g. HH events) in the whole audio file discarding time-information [4]. Using the same datasets as in the full transcription experiments we obtained, as expected, better classification performance (F-measure) for all classes (BD = 0.822, HH = 0.794 and SD = 0.698). The overall performance of this "relaxed" transcription system was 0.771 (F-measure). These results encouraged us to investigate if useful song-wise percussion descriptors could be computed.

[4] We define correct transcription (CTR) as the $\arg\min(TR, GT)$, being TR = transcription and GT = ground truth labels per instrument. Then, we compute $P = CTR/TR$ and $R = CTR/GT$ and finally $F = 2PR/(P + R)$.

4. SONG-LEVEL PERCUSSION DESCRIPTORS

4.1 Computed Descriptors

In [12] and [18] two percussion-related descriptors were presented and evaluated with promising results namely: *Percussion Index* (a ratio between the total number of detected percussion events and the number of detected onsets) and *Percussion Profile* (the relative amount of BD, SD, cymbals, and non-percussion events normalized by the total number of onsets). Following this idea of percussion related descriptors we decided to compute and evaluate the following song-level percussion descriptors (some of these descriptors appeared as suggested future work in [12] but, up to our knowledge, they have not been implemented nor evaluated yet).

Computed song-level percussion descriptors:

- **Percussion Profile:** The ratio between the number of detected percussion events and the number of detected onsets [18]. Computed for BD, SD, HH and drum (D) [5] (e.g. BD/total, SD/total).
- **Inter-Instrument Ratio:** The ratio among all percussive instrument events namely: BD/SD, BD/HH and SD/HH.
- **Instrument Per Minute:** The number of detected events per minute for BD, SD, HH and D.
- **Inter-Instrument Interval (iii):** The first and second peak values of the histogram of the differences between successive events. Thus, we computed: first and second-iii-peak for BD, SD and HH.

At the end of this process we obtained 17 song-level percussion descriptors for each song.

4.2 Evaluation

To investigate the correlation between the proposed percussion descriptors and the ground truth values we computed the percussion descriptors both from the ground truth labels (labeled onsets) and from the output of our transcription system. Then, we built a fractional ranking [6] for each descriptor (for every song) and computed the Pearson's correlation coefficient between both rankings. The correlation results showed large correlation values (i.e. > 0.5) for 12 out of 17 proposed descriptors (only: BD/SD, second-iii-peak for the three instruments and first-iii-peak for HH presented correlation values below 0.5). These highly correlated values between our descriptors and descriptors computed from the ground truth labels were specially strong (i.e. > 0.7) for D/total, D/min, HH/total and HH/min. These results suggest that the proposed descriptors have potential to describe the percussive content of a song.

[5] In this context "drum" means the number of detected onsets labeled by the system as BD, SD or HH.

[6] If the ordered vector to rank is A,B,C,D and B is equal to C (i.e. tie) the fractional ranking assigns the same mean position value in both cases, i.e. 1,2.5,2.5,4.

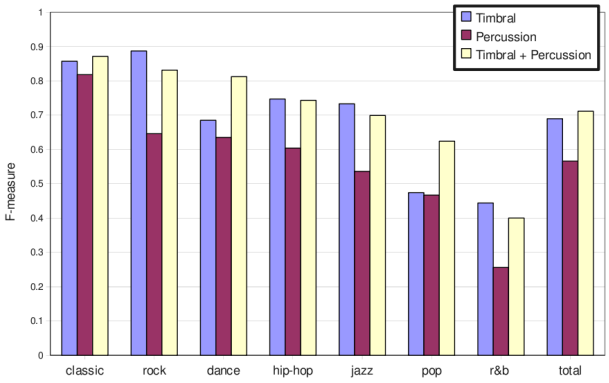

Figure 1. Genre classification results per genre and descriptor set. F-measure after 10-fold CV.

Next, we evaluated the usefulness of the percussion descriptors as features in several MIR tasks such as genre and sub-genre classification, danceability [7] and Western vs. non-Western music estimation. We decided to set-up a general methodology for evaluating the song-level percussion descriptors on every selected MIR task. Firstly, we computed, for each song in the dataset, the mean value of a set of "standard" descriptors to be used as baseline for the evaluation. Secondly, we computed the proposed song-level percussion descriptors on the same database. Thirdly, we selected a classification algorithm and we determined the "best" classification values for the "standard", "percussion" and "standard + percussion" descriptor sets. Finally, we evaluated the classification results by comparing F-measures and performing a Binomial test [20, p. 37] with 5% significance level (i.e. $\alpha = 0.05$). This Binomial test determines if the difference between correctly classified songs for each descriptor set is statistically significant or not. It is worth to notice that none of the songs used for training the SVM models were used in these evaluation experiments.

4.2.1 Genre

For genre classification we used an in-house database of 30 seconds excerpts extracted from 350 songs equally distributed among 7 genres: classic, dance, hip-hop, jazz, pop, r&b and rock. The computed "standard" descriptors were: Bark-bands, Bark-bands kurtosis, Bark-bands skewness, Bark-bands spread, spectral centroid, spectral crest, spectral decrease, spectral dissonance, spectral energy, spectral energy-band high, spectral energy-band low, spectral energy-band middle high, spectral energy-band middle low, spectral flatness, spectral flux, spectral hfc, spectral kurtosis, MFCCs, spectral skewness, spectral spread and temporal zero-crossing rate. We called "timbral" descriptors this set of 60 features. We used multi-class SVM as classification algorithm.

The genre classification results can be seen in Figure 1. From these results it is interesting to notice that by using the percussion descriptors only, good discrimination rates

[7] The easiness with which one can dance to a musical track [19].

Genre	T	P	T+P
ambient	0.531	0.433	0.588
drum'n bass	0.475	0.619	0.576
house	0.200	0.500	0.427
techno	0.369	0.269	0.380
trance	0.438	0.566	0.427
Average	**0.403**	**0.478**	**0.480**

Table 3. Electronic sub-genre classification. T = timbral, P = percussion and T+P = timbral+percussion. C4.5 classification algorithm. Results in F-measure after 10-fold CV.

can be achieved for classic ($F = 0.818$), rock, dance and hip-hop ($F \approx 0.600$) genres. The overall classification for the percussion-only data set was about 12 percentage points (pp) below "timbral" descriptors. When combining "timbral" and "percussion" descriptors a small improvement in the overall result was observed (+2.1 pp). It is worth to notice that big improvements were produced in dance (+12.7 pp) and pop (+15 pp) results, whereas results for rock and r&b decreased 5.6 and 4.4 pp respectively.

The Binomial test showed no statistically significant difference between "timbral + percussion" and "timbral" descriptors ($p = 0.1932$), but both sets evidenced significant differences with the "percussion" descriptor set ($p < 0.0001$ in both cases).

4.2.2 Electronic

For electronic sub-genre classification we performed our experiments on an in-house database of 270 songs (30 seconds length each) equally distributed among the following genres: ambient, drum'n bass, house, techno and trance. The computed descriptors were the same as in the genre experiment. Given that sub-genre and genre classification could be considered as very related tasks, we decided to try a different algorithm to gain some insight on the descriptors. Therefore, we used in this case the C4.5 decision tree algorithm for classification, since its output can be easily summarized into interpretable trees of descriptors.

Results for electronic sub-genre classification are depicted in Table 3. In this experiment we observed that classification results obtained by the "percussion" set outperformed "timbral" descriptors by 7.5 pp. The combination of "timbral" and "percussion" descriptors showed no significant difference with results from percussion-only descriptors in the overall classification result (although this combination seems to output more balanced classification rates among categories). In both "percussive" and "timbral + percussive" models the D/total and first and second iii-peak BD were the most informative descriptors.

The significance test corroborates the conclusions extracted from the F-measure results where "percussion" descriptors performed significatively better than "timbral" descriptors ($p = 0.0028$), "timbral + percussion" performed better than "timbral" ($p = 0.0058$) and no statistical difference between "percussion" and "percussion + timbral" descriptor sets was appreciated ($p = 0.4273$).

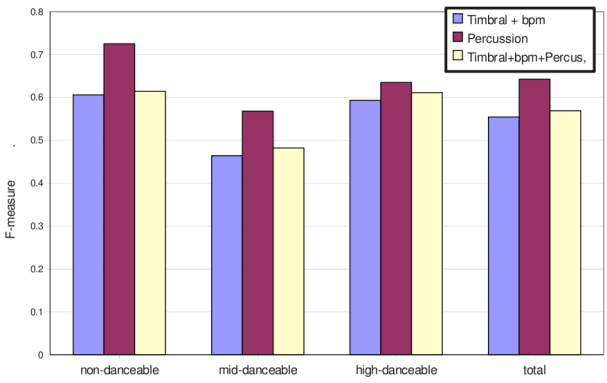

Figure 2. Danceability classification results after 10-fold CV.

4.2.3 Danceability

For danceability tests we used an in-house database of 374 song excerpts of 30 seconds equally distributed into three classes (i.e. non-dance., mid-dance. and high-dance.). We computed the same descriptors as in the genre experiment plus an estimation on the beats per minute (bpm) of the song [8], we called this descriptor set as "timbral + bpm". As in the genre experiments we decided to use the SVM algorithm (multi-class).

Results for Danceability tests are shown in Figure 2. From these results we can conclude that "percussion" descriptors performed better than both "timbral + bpm" and "timbral + bpm + percussion". Percussion-only descriptors outperformed by 8.9 pp and 7.4 pp "timbral + bpm" and "timbral + bpm + percussion" respectively, obtaining better results in all three categories. It is interesting to notice that percussion descriptors also outperformed obtained results by [19] which achieved an accuracy of 61.78% in classifying 225 songs into the same three categories by using a different and more complex approach.

The Binomial test on danceability results showed that "percussion" descriptors provided significatively better performance than the other two sets ($p = 0.0025$ for "timbral + bpm" and $p = 0.0074$ for "timbral + bpm + percussion"). The test also showed no statistical difference between "timbral + bpm" and "timbral + bpm + percussion" descriptor sets ($p = 0.3793$).

4.2.4 Western vs. non-Western music classification

For Western vs. non-Western experiments we used an in-house database of 139 Western songs from 16 genres including classical, jazz, rock, pop, religious and hip-hop, and 139 non-Western songs including songs from Africa, Arab States, Asia and the Pacific. The computed descriptors and classification algorithm were the same as in genre experiments.

The results for these experiments are shown in Table 4. Here we observed an almost linear increment in the classification rates starting by "timbral" descriptors with

[8] Since bpm is an important descriptor for danceability estimation we included it into the "standard" set. Otherwise, it would be too easy for our descriptors to outperform.

Class	T	P	T+P
Western	0,817	0,803	0,856
non-Western	0,747	0,828	0,833
Average	**0,782**	**0,816**	**0,844**

Table 4. Western vs. non-Western music classification. T = timbral, P = percussion and T+P = timbral+percussion. SVM classification algorithm, F-measure results after 10-fold CV.

$F = 0.782$ followed by "percussion" descriptors with $F = 0.816$ (+3.4 pp) and "timbral + percussion" with $F = 0.844$ (+2.8 pp from "percussion"). It seems clear that adding percussion descriptors helped in the process of Western vs. non-Western song discrimination. Is is also interesting to notice that classification results for non-Western music were much better when percussion descriptors were used (more than 8 pp above "timbral").

The significance test showed no statistically significant difference between "percussion + timbral" and "percussion" descriptors ($p = 0.1212$) and between "percussion" and "timbral" descriptors ($p = 0.1349$). The test also depicted statistical difference between "percussion + timbral" and "timbral" descriptors ($p = 0.0096$). See [21] for an in-depth study on Western vs. non-Western music classification.

5. CONCLUSIONS

Within the present work we have conducted several experiments in order to detect and describe percussive events in polyphonic music. Firstly, we built, by combining three databases, a large set of percussion-labeled songs. Secondly, we evaluated the capacity of an automatic drum transcription system, based on object-level features and three binary SVM models, to transcribe percussion events in polyphonic music. From the transcription results we extrapolated that our relatively simple algorithm can be placed among the top ranked ones, even though all these systems leave a lot of room for improvement. After performing "relaxed" transcription experiments we observed that our system can detect the total number of drum events in a song with an overall F-measure of 0.771. Finally, we presented 17 song-level percussion descriptors and we evaluated their usefulness among several MIR tasks. These preliminary results suggest that song-level percussion (i.e. "semantic") descriptors, even though they are based on imperfect transcriptions, can help in MIR tasks such as genre and sub-genre classification, danceability and Western vs. non-Western music estimation. It also seems clear that song-level percussion descriptors offer useful information that complements the one provided by classic "spectral" and "timbral" descriptors. This new information could also be exploited in music similarity tasks.

6. REFERENCES

[1] D. Fitzgerald and J. Paulus. *Unpitched Percussion Transcription*, chapter 5, pages 131–162. Signal Processing Methods for Music Transcription. Springer, 2006.

[2] P. Herrera, A. Dehamel, and F. Gouyon. Automatic labeling of unpitched percussion sounds. In *Proc. of AES 114th*, Amsterdam, The Netherlands, 2003.

[3] V. Sandvold, F. Gouyon, and P. Herrera. Percussion classification in polyphonic audio recordings using localized sound models. In *Proc. of ISMIR*, pages 537–540, Barcelona, Spain, 2004.

[4] K. Yoshii, M. Goto, and H. G. Okuno. Adamast: A drum sound recognizer based on adaptation and matching of spectrogram. In *MIREX*, London, UK, 2005.

[5] C. Dittmar. Drum detection from polyphonic audio via detailed analysis of the time frequency domain. In *MIREX*, London, UK, 2005.

[6] K. Tanghe, S. Degroeve, and B. De Baets. An algorithm for detecting and labeling drum events in polyphonic music. In *MIREX*, London, UK, 2005.

[7] O. Gillet and G. Richard. Drum track transcription of polyphonic music using noise subspace projection. In *Proc. of ISMIR*, pages 92–99, London, UK, 2005.

[8] J. Paulus and A. Klapuri. Combining temporal and spectral features in hmm-based drum transcription. In *Proc. of ISMIR*, pages 225–228, Vienna, Austria, 2007.

[9] O. Gillet and G. Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):529–540, 2008.

[10] O. Gillet and G. Richard. ENST-Drums: an extensive audio-visual database for drum. In *Proc. of ISMIR*, pages 156–159, Victoria, Canada, 2006.

[11] K. Tanghe, M. Lesaffre, S. Degroeve, M. Leman, B. De Baets, and J. Martens. Collecting ground truth annotations for drum detection in polyphonic music. In *Proc. of ISMIR*, pages 50–57, London, UK, 2005.

[12] V. Sandvold. Percussion descriptors. A semantic approach to music information retrieval. Master's thesis, University of Oslo, 2004.

[13] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, CUIDADO project, 2004.

[14] M. Haro. Detecting and describing percussive events in polyphonic music. Master's thesis, Universitat Pompeu Fabra, 2008.

[15] P. Brossier. *Automatic annotation of musical audio for interactive systems*. PhD thesis, Queen Mary University of London, 2006.

[16] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proc. 17th International Conf. on Machine Learning*, pages 359–366, San Francisco, USA, 2000.

[17] O. Celma, P. Herrera, and X. Serra. Bridging the music semantic gap. In *Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, volume 187, Budva, Montenegro, 2006. CEUR.

[18] P. Herrera, V. Sandvold, and F. Gouyon. Percussion-related semantic descriptors of music audio files. In *Proc. of AES 25th*, London, UK, 2004.

[19] S. Streich and P. Herrera. Detrended fluctuation analysis of music signals danceability estimation and further semantic characterization. In *Proc. AES 118th*, Barcelona, Spain, 2005.

[20] P. H. Kvam and B. Vidakovic. *Nonparametric Statistics with Applications to Science and Engineering*. Wiley, 2007.

[21] E. Gómez, M. Haro, and P. Herrera. Music and geography: Content description of musical audio from different parts of the world. In *Proc. of ISMIR*, Kobe, Japan, 2009.