PERCUSSION CLASSIFICATION IN POLYPHONIC AUDIO RECORDINGS USING LOCALIZED SOUND MODELS

Vegard Sandvold
University of Oslo
Oslo, Norway

Fabien Gouyon
Universitat Pompeu Fabra
Barcelona, Spain

Perfecto Herrera
Universitat Pompeu Fabra
Barcelona, Spain

ABSTRACT

This paper deals with automatic percussion classification in polyphonic audio recordings, focusing on kick, snare and cymbal sounds. We present a feature-based sound modeling approach that combines general, prior knowledge about the sound characteristics of percussion instrument families (general models) with on-the-fly acquired knowledge of recording-specific sounds (localized models). This way, high classification accuracy can be obtained with remarkably simple sound models. The accuracy is on average around 20% higher than with general models alone.

1. INTRODUCTION

This paper deals with automatic symbolic transcription of percussion mixed in polyphonic audio signals. That is, given a multi-timbral audio signal, the goal is twofold: to automatically *classify* its percussion sounds and to automatically determine their *positions* on the time axis.

Snare drum sounds, for instance, can show large variations in timbral characteristics. In automatic *isolated* sound classification [8], this is typically dealt with from a machine learning perspective: a sound model (roughly, thresholds for specific relevant signal features) is built from a large, diverse collection of labeled snare drum sounds. This model is subsequently used to assign labels to unknown instances.

However, in our framework, the temporal boundaries of the sounds to classify are unknown. A list of *potential* percussion sound occurrences must be first extracted from the audio recording. Different rationales have been proposed to solve this issue. For instance, one may assume that percussion sounds are bound to occur in fixed-length regions around specific time-points, either sharp onsets [3, 4] or beats at the tatum level [5, 11, 1].

Dealing with polyphonic recordings raises an additional issue: percussion sounds are superposed with, and surrounded by, a high level of "noise", i.e. other instruments as e.g. voice, piano, guitars etc. Even worse, simultaneous occurrences of several classes of percussion instruments (e.g. kick + hihat or snare + hihat) may be encountered. To deal with this issue, existing literature advocates diverse research directions. Some advocate source separation techniques as Independent Subspace Analysis [1, 2] or signal models as 'Sinusoidal + Residual' (assuming that drums are in the residual component) [7, 11]. Noise reduction techniques, as RASTA [10], are also thinkable. Another option is to build sound models from a large collection of labeled "noisy" percussion instrument sounds extracted from polyphonic audio recordings [9]. The main assumption in this method is that, in average on the training database, the noise shows considerably higher variability than the drum sounds.

The approach by [4, 13] also assumes that percussion sound characteristics show less variability than surrounding noise; however, this assumption is not made at the scope of a training database, but rather at the smaller scope of *individual audio recordings*. They design very simple, general sound templates for each percussive sound (actual *waveform* templates [4, 13], at the difference with the sound models previously mentioned) and find the several sub-segments in the audio recording at hand that best match those templates (by means of a correlation function). This process is iterated several times and sound templates are gradually refined by time-domain averaging of the best-matching segment in the very audio recording at hand.

Our approach is to *combine* general, prior knowledge about the sound characteristics of percussion instrument families with on-the-fly acquired knowledge of recording-specific sounds. Instead of pursuing universally valid sound models and features [8, 9], unique, localized sound models are built for every recording using features that are *locally* noise-independent and give good class separation. Instead of actually synthesizing new waveform templates from the audio signal [4, 13], we tailor (in a gradual fashion) feature spaces to the percussion sounds of each recording.

Therefore, an onset detector yields N potential drum sound occurrences that are subsequently processed as follows:

1. Classification using general drum sound models
2. Ranking and selection of the $M < N$ most reliably

classified instances

3. Feature selection and design of localized models using those M instances

4. Classification of the N segments using the localized models

As it turns out, our attempts at the automatic ranking and selection has not yet provided satisfactory results. Therefore, we corrected manually the output of steps 1 and 2 and provided only correct percussion sound instances to the feature selection step. Consequently, in this paper, we present a more focused evaluation of the localized sound model design, as well as a proper comparison between general sound model and localized sound model performances. Using the corrected instance subsets, we investigate how the performance of the localized models evolve as increasingly smaller proportions of the data is used for feature selection and training.

Automatic techniques for instance ranking are currently being evaluated. Together with the evaluation of the fully automatic system they are the object of a forthcoming paper.

2. METHOD

2.1. Data and features

The training data set for general model building consists of 1136 instances (100 ms long): 1061 onset regions taken from 25 CD-quality polyphonic audio recordings and 75 isolated drum samples. These were then manually annotated, assigning category labels for *kick*, *snare*, *cymbal*, *kick+cymbal*, *snare+cymbal* and *not-percussion*. Other percussion instruments like toms and Latin percussions were left out. Cymbals denote hi-hats, rides and crashes.

Annotated test data consists of seventeen 20-second excerpts taken from 17 different CD-quality audio recordings (independent from the training data set). The total number of manually annotated onsets in all the excerpts is 1419, average of 83 per excerpt.

Training and test data are characterized by *115* spectral features (averages and variances of frame values) and temporal features (computed on the whole regions), see [9] and [12] for details.

The experiments described in the remainder of this paper were conducted with Weka . [1]

2.2. Classification with general models

In order to design general drum sound models, we first propose to reduce the dimensionality of the feature space by applying a *Correlation-based Feature Selection* (CFS) algorithm (Section 2.4) on the training data. From the total of 115, an average of 24.67 features are selected for each model.

This data is then used to induce a collection of C4.5 decision trees using the AdaBoost meta-learning scheme. Bagging or boosting approaches has turned out to yield better results when compared to other more traditional machine learning algorithms [9].

[1] *http://www.cs.waikato.ac.nz/ml/weka/*

2.3. Instance ranking and selection

The instances classified by the general models must be parsed in order to derive the most likely correctly classified subset. Several rationales are possible. For instance, we can use instance probability estimates assigned by some machine learning algorithms as indicators of correct classification likelihood. Another option is to use clustering techniques. Instances of the same percussion instrument, which we are looking for, would form the most populated and compact clusters while other drum sounds and non-percussive instances would be outliers.

However, as mentioned above, we went for a "safe" option: manually parsing the output of the general classification schemes. Using the corrected output, we investigated how the performance of the localized models evolved as increasingly smaller proportions of the instances selected from a recording were used to classify the remaining sound instances of the recording. Since dependence between training and test data sets is known to yield overly optimistic results, these test were performed by doing randomized, mutually exclusive splits on the complete collection of instances for each recording.

2.4. Feature selection

A collection of correctly classified instances from a recording are then used to build new, localized sound models.

Relevant features for the localized sound models are selected using a *Correlation-based Feature Selection* (CFS) algorithm that evaluates attribute subsets on the basis of both the predictive abilities of each feature and feature inter-correlations [6]. This method yields a set of features with recording-specific good class separability and noise independence. The localized models may differ from the general models in two respects: they may be based on 1) a different feature subset (feature space) and 2) different threshold values (decision boundaries) for specific features. As a general comment, the features showed significantly better class-homogeneity for localized models than for the general models.

2.5. Classification with localized models

Finally, the remaining instances must be classified with the recording-specific (localized) models.

For this final step, we propose to use instance-based classifiers, such as 1-NN (k-Nearest Neighbors, with $k = 1$). Instance-based classifiers are usually quite reliable and give good classification accuracies. However, usual criticisms are that they are not robust to noisy data, they are memory consuming and they lack generalization capabilities. Nevertheless, in our framework, these are not issues we should be worried about: by the very nature of our

Model	General		Localized	
	# feat.	Accuracy	# feat.	Accuracy
Kick	19	80.27	5.73	95.06
Snare	33	70.9	10.41	93.1
Cymbal	22	66.31	10.94	89.17

Table 1. Average number of features used and accuracy (%) for kick, snare and cymbal sound classification in polyphonic audio recordings, using both general and localized models.

method, instances are reliable, there are few of them and we explicitly seek localized (i.e. not general) models.

3. EXPERIMENTS, RESULTS AND DISCUSSION

Table 1 shows a comparison of the performance of the general and localized models applied to the polyphonic audio excerpts. The number of selected features is constant for the general models, but individual to each recording for the localized models. The classification accuracies of the localized models are determined using 10-fold cross-validation.

The number of features selected for the localized models is significantly less than for the general models. At the same time, the performance of the former is clearly superior. Maybe not surprisingly, this is resulting from the lesser variability of percussion sounds within a specific recording, which gives clearer decision boundaries in the feature space between instances of the different categories.

Doing feature selection on all sound instances of a recording (100%) should give what we consider the "ideal" feature subset, which should give optimal performance (noise-independence and class separation) on the individual recordings. Figure 1 shows the average classification accuracy of the kick, snare and cymbal models, using the optimal feature subsets for each localized model. The training-test data splits are repeated 30 times for each reported training data set percentage.

We see from the figure that the accuracy never drops down below that of the general sound models (marked by dotted lines). It seems like the performance makes a significant drop around 20% – 30%, indicating a sort of threshold on the minimum number of instances needed to permit successful classification. This proportion corresponds to about 17 – 25 samples. Further studies have to be done to establish whether it is the relative percentage or the approximate number of samples that is significant for the performance of the localized models.

In practice it is not possible to know the optimal feature subsets, as feature selection must be performed on a reduced data set. Table 2 shows average classification accuracies together with the average number of selected features for kick, snare and cymbal models, using truly localized features.

There is a slight loss of performance from localized

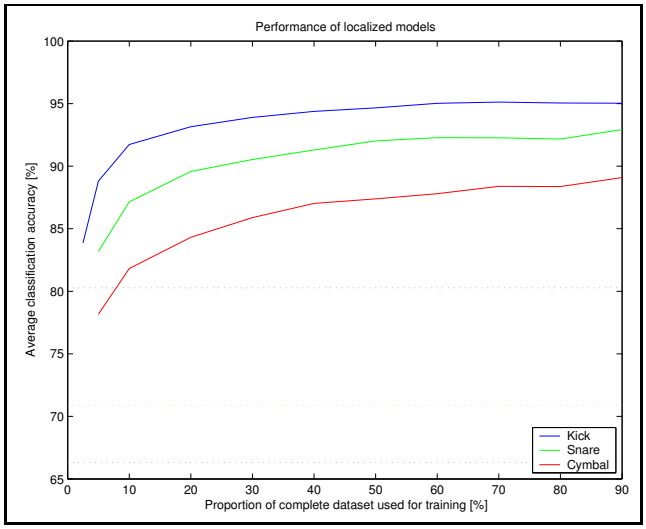

Figure 1. Accuracy for kick, snare and cymbal sound classification using the optimal feature subsets and decreasing proportions of correct instances to create the localized models. The dotted lines mark the accuracies obtained with general models.

models with optimal feature subsets (Figure 1). Using 30% of the instances, the accuracy decreases 7.3% for kicks, 7.57% for snares and 1.17% for cymbals. We observe that a reduction in the amount of training instances greatly effects the feature selection. Besides a general decrease in number of selected features, the variation in types of features selected for each recording can be high.

What is not evident from the tables, is the variability of the performance among individual recordings. At one extreme 96.72% accuracy is obtained using only 1 feature and 10% of the complete data set. When comparing to classification with general models, it appears that recordings having the least successful localized models are also least favorable for classification with general models.

Also, it is important to notice that relevant features for localized models usually differ from one recording to the other, which justifies the proposed method. Let us focus, for instance, on single-feature based kick models. Depending on the specific recording at hand, some used e.g. the mean of the frame energy values in the 1st Bark band, others the mean of the 3rd spectral moment in successive frames, or other features. Snare models used e.g. the mean of the frame energy values in the 11th Bark band or the mean of the 4th MFCC in successive frames, etc. Cymbals models used e.g. the mean of 9th MFCC in successive frames or the mean of frame spectral flatness values, etc.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose to design feature-based percussion instrument sound models specialized on individual polyphonic audio recordings. Initial classification with general sound models and parsing of their output provides a reduced set of correctly classified sound instances from

Percentage	Kick		Snare		Cymbal	
	# features	Accuracy	# features	Accuracy	# features	Accuracy
50%	5	89.9	11.86	90.84	6.67	86.73
40%	5.1	88.42	8.71	87.88	7.13	85.35
30%	3	86.6	5.71	82.96	3.57	84.72
20%	2.5	85.51	4	77.22	3.4	79.4
10%	1	77.92	1.71	73.34	1.27	71.53

Table 2. Average number of features used and accuracy (%) for kick, snare and cymbal sound classification using decreasing proportions of correct instances to select relevant features and perform 1-NN classification.

a single recording. By applying a a feature selection algorithm to the reduced instance set we obtain the reduced feature sets required to design recording-specific, localized sound models. The localized models achieved an average classification accuracy (and feature dimensionality) of 95.06% (5.73) for kicks, 93.1% (10.41) for snares and 89.17% (10.94) for cymbals, which represents improvements of respectively 14.79%, 22.2% and 22.86% over general model classification accuracies. We also showed that the choice of relevant features for percussion model design should be dependent, at some extent, on individual audio recordings.

Part of future work is to implement a semi-automatic percussion transcription tool based on the approach presented in this paper. Our results are encouraging, but we need to process more and longer recordings to claim that the method is general and scales up well. More effort has to be put into determination of reliable estimators of general model classifications. We must also consider the influence of noisy data in localized model design. Another direction for future work is to explore whether ISA, RASTA or 'Sinusoidal + Residual' pre-processing can improve the classification performance.

5. REFERENCES

[1] Dittmar C. and Uhle C. "Further Steps towards Drum Transcription of Polyphonic Music" *Proc. AES 116th Convention*, Berlin, 2004.

[2] FitzGerald D., Coyle E. and Lawlor B. "Sub-band Independent Subspace Analysis for Drum Transcription" *Proc. 5th International Conference on Digital Audio Effects*, Hamburg, 2002.

[3] Goto M., Tabuchi M. and Muraoka Y. "An Automatic Transcription System for Percussion Instruments" *Proc. 46th Annual Convention IPS Japan*, 1993

[4] Gouyon F., Pachet F. and Delerue O. "On the use of zero-crossing rate for an application of classification of percussive sounds" *Proc. 3d International Conference on Digital Audio Effects*, Verona, 2000.

[5] Gouyon F. and Herrera P. "Exploration of techniques for automatic labelling of audio drum tracks' instruments" *Proc. MOSART*, Barcelona, 2001.

[6] Hall M. A., "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning", *Proc. of the Seventeenth International Conference on Machine Learning*, 2000.

[7] Heittola T. and A. Klapuri. *Locating Segments with Drums in Music Signals*. Technical Report, Tampere University of Technology, 2002.

[8] Herrera P., Peeters G. and Dubnov S. "Automatic Classification of Musical Instrument Sounds" *Journal of New Music Research Vol.32 .1*, 2003

[9] Herrera P., V. Sandvold and F. Gouyon. "Percussion-related Semantic Descriptors of Music Audio Files", *Proc. AES 25th International Conference*, London, 2004.

[10] Klapuri, Virtanen, Eronen, Seppnen. "Automatic transcription of musical recordings", *Proc. Consistent & Reliable Acoustic Cues Workshop*, Aalborg, 2001.

[11] Paulus J. and Klapuri A. "Model-Based Event Labeling in the Transcription of Percussive Audio Signals" *Proc. 6th International Conference on Digital Audio Effects*, London, 2003.

[12] Peeters G., *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. CUIDADO I.S.T. Project Report, 2004.

[13] Zils A., Pachet F., Delerue O. and Gouyon F., "Automatic Extraction of Drum Tracks from Polyphonic Music Signals" *Proc. 2nd International Conference on Web Delivering of Music*, Darmstadt, 2002.