AUTOMATIC CYMBAL CLASSIFICATION USING NON-NEGATIVE MATRIX FACTORIZATION

Sofia Cavaco[*] *and Hugo Almeida*[+]

CITI, Departamento de Informática
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
2829-516 Caparica, Portugal
scavaco@fct.unl.pt, hrcalmeida@gmail.com

ABSTRACT

Several musical instrument classifiers have been proposed. While many approaches in sound-feature extraction and in sound classification have been successfully used, most focus on distinguishing different harmonic instruments such as the violin and the flute, whose sounds have very different characteristics. On the other hand, much less attention has been given to percussion instruments, especially if we consider the discrimination of instruments of the same type, like the cymbals in a drum kit.

Here, we propose a classifier that is able to distinguish this latter type of instruments. The classifier is able to distinguish sounds with very similar properties, like sounds produced by instruments with similar geometry that differ in material or size. In particular it is able to distinguish sounds from the cymbals in a drum kit. Instead of using a set of predefined features, the classifier learns spectral features from the data using non-negative matrix factorization. This work is important to fill the gap on percussion instrument classification and transcription (since most music transcribers focus on harmonic instruments).

Index Terms— sound classification, percussion instruments, indefinite pitch, acoustic signal processing, non-negative matrix factorization

1. INTRODUCTION

Most proposed classifiers of musical instruments deal with string and wind harmonic instruments, while much less attention has been given to percussion instruments with non-perceptible pitch, that is, with indefinite pitch. Still, there have been a few studies that focus on the automatic classification of this latter type of instrument. A few of these studies focus on the recognition of different types of strokes in a single instrument, like the snare drum and conga drums [1, 2, 3]. Yet, most of these studies focus on distinguishing different instruments in the drum kit (such as bass drum, snare drum, hi-hat, toms and cymbals) [4, 5, 6, 7, 8, 9, 10, 11, 12]. Nonetheless, while some of the proposed classifiers can distinguish cymbals from other instruments in the drum kit, they cannot discriminate between the cymbals, that is, sounds from any of the cymbals in the drum kit are all assigned to the same class.

Here we propose a classifier of indefinite pitched percussion instruments of the same type, such as the cymbals in a drum kit. That is, we are interested in discriminating sounds from instruments of the same type, such as instruments with similar geometries that differ in size, material, or other subtle properties. This is a significantly more difficult problem than differentiating a bass drum from a cymbal, because while bass and cymbal sounds have very different characteristics, sounds from different cymbals are more alike. To the best of our knowledge, there has been no previous work on classification of indefinite pitched percussion instruments of the same type.

Sound classifiers are characterized by a stage of sound features extraction and another stage of classification. Many low and high level temporal, spectral and short-time features have been tried to characterize indefinite pitch percussion instruments, but due to the difficulty on deciding which are the most appropriate features to characterize the data, many classifiers use a combination of several features to achieve good classification rates [1, 2, 3, 7, 9, 11, 13]. For instance, starting with a set of about fifty features of the attack and decay sections of the sound, of the energy in the sub-bands, mel-frequency cepstral coefficients and variances, Herrera, Yetarian, and Gouyon were able to determine the twenty most relevant features to distinguish sounds from the five main instruments in the drum kit [8].

Whereas most sound classifiers use a set of pre-defined features, there are also some classifiers that learn the features using a decomposition method such as independent component analysis, independent sub-space analysis, sub-band independent subspace analysis and non-negative matrix factorization (NMF) [4, 5, 10, 12]. While exploring the techniques proposed by FitzGerald and colleagues on the classification of

[*] This work was supported in part by grant UTA-Exp/MAI/0025/2009 from *Fundação para a Ciência e Tecnologia* to S.C.

[+] Both authors contributed equally to this work.

cymbals, we were able to conclude that some of those techniques are appropriate to distinguish cymbals from other elements in the drum kit (namely from the bass, snare, hi-hat, and toms), but they are not suitable to discriminate the individual cymbals. We then investigated the use of other statistical sound source separation techniques and were able to develop a classifier that can discriminate cymbals.

While our cymbal classifier uses spectral features, it does not use a set of pre-defined features. Instead, it learns them from the data using non-negative matrix factorization [14, 15]. Using these features on a 1-nearest neighbor (1-NN) algorithm, we obtained very high classification rates: 95% when distinguishing between two cymbals and 86% when distinguishing between three cymbals.

2. THE CLASSIFIER

As mentioned before, our classifier learns the features from the data: it starts by representing the sounds with magnitude spectrograms[1], and then it extracts spectral features from these spectrograms using NMF. Here we prove that the features learned in this manner can successfully separate drum kit cymbal sounds, more specifically sounds from hi-hats, and china, splash, crash and ride cymbals.

Here we consider that the spectrogram is a sequence of frames, which are the columns of the spectrogram. Each frame is the power spectrum at a given time interval. In other words, the data runs over frequency: the frames are initially represented in an F-dimensional space, which we call the *frequency space*, with one dimension for each frequency bin (where the frequency bins are the rows of the spectrogram).

Each spectrogram, $\mathbf{S}_n$, can be represented as a product of a matrix, $\mathbf{\Theta}$, which has a spectrum per column, with a matrix, $\mathbf{P}_n$, which has a temporal envelope (i.e., time-varying gain function) per line: $\mathbf{S}_n = \mathbf{\Theta}\mathbf{P}_n$. We estimate $\mathbf{\Theta}$ and $\mathbf{P}_n$ with NMF, as explained below.

In the training phase, the classifier performs NMF[2] on the concatenation of the spectrograms, $(\mathbf{S}_1, \ldots, \mathbf{S}_\mathrm{N})$[3], where $\mathbf{S}_n$ is the spectrogram of a training sample, and N is the number of samples. As a result it learns a set of *spectral basis functions*, that consist of the spectra in matrix $\mathbf{\Theta}$, that describe the spectral regularities in the frames in $(\mathbf{S}_1, \ldots, \mathbf{S}_\mathrm{N})$.To put it in another way, the data is now represented in a new space, whose axes are spectral basis functions. Fig. 1 shows the basis functions learned by NMF of the concatenation of the spectrograms from samples from a china and a crash cymbal.

The basis functions in $\mathbf{\Theta}$ are the spectral features that later will be used in the classification stage. But we still need the values of those features: we have seen that each frame in the data set is now represented in the space defined by the basis functions in $\mathbf{\Theta}$, but we still need to know by which coefficients. These coefficients are the values of the features, that is, these are the values that will be used in the 1-NN algorithm. Along with matrix $\mathbf{\Theta}$, the NMF of the concatenated spectrograms also produces a matrix $\mathbf{P}$ that contains the representation of the frames in the new space, that is, the coefficients that we are looking for.

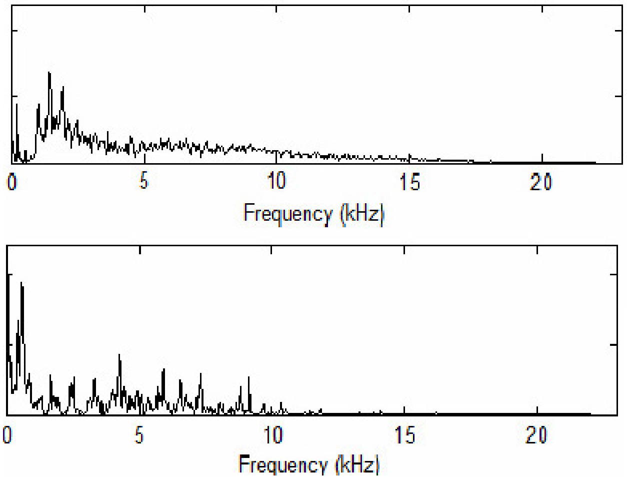


Fig. 1. Two spectral basis functions learned by NMF of $(\mathbf{S}_{s_1}, \ldots, \mathbf{S}_{s_6}, \mathbf{S}_{c_1}, \ldots, \mathbf{S}_{c_6})$, where $\mathbf{S}_{s_i}$ is the spectrogram of a sample from a stroke on the edge of a 16 inch china cymbal and $\mathbf{S}_{c_i}$ is the spectrogram of a sample from a stroke on the edge of a 16 inch crash cymbal, with $1 \leqslant i \leqslant 6$.

Each column of $\mathbf{P}$ consists of the representation of one frame in the new space (that is, it contains the coefficients associated to one frame) and $\mathbf{P}$ contains as many lines as the number of basis functions. Now we can define matrix $\mathbf{P}_n$ which contains the coefficients associated to spectrogram $\mathbf{S}_n$ ($\mathbf{P}_n$ contains as many columns as the number of frames in $\mathbf{S}_n$). Therefore, the data set $(\mathbf{S}_1, \ldots, \mathbf{S}_\mathrm{N})$ can be expressed as:

$$(\mathbf{S}_1, \ldots, \mathbf{S}_\mathrm{N}) = \mathbf{\Theta}\,(\mathbf{P}_1, \ldots, \mathbf{P}_\mathrm{N}) \quad . \qquad (1)$$

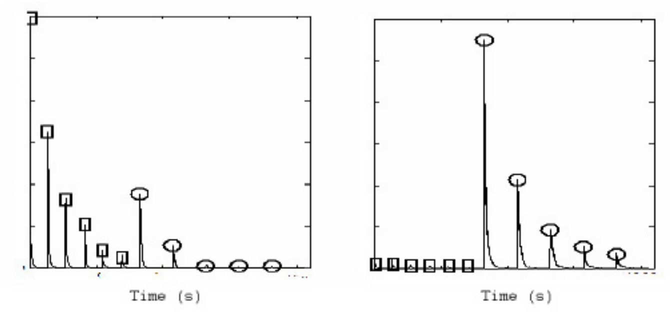
The ith row of every matrix $\mathbf{P}_n$ is associated to the ith basis function in $\mathbf{\Theta}$, that is, the ith column of $\mathbf{\Theta}$, and $\mathbf{P} = (\mathbf{P}_1, \ldots, \mathbf{P}_\mathrm{N})$. Each line in $\mathbf{P}_n$ is a sequence of coefficients associated to one basis function, which here we call *temporal envelopes*. In this way, a spectrogram is represented by a set of temporal envelopes (one temporal envelope per spectral basis function).

Fig. 2 shows the temporal envelopes related to the basis functions from Fig. 1 (each graph consists of one line from $\mathbf{P}$). The temporal envelopes in Fig. 2a are related to the top basis function in Fig. 1 and the temporal envelopes in Fig. 2b are related to the bottom basis function in Fig. 1. Note that each graph contains twelve envelopes. Each envelope is associated to one of the twelve spectrograms used in this example: six spectrograms from strokes on the edge of a 16 inch china

[1] We do not use the spectrograms phase information. For simplicity, in the remaining text we refer to the magnitude spectrogram as spectrogram.

[2] We used an NMF software package by Virtanen [16].

[3] $(\mathbf{A}, \mathbf{B})$ represents two concatenated matrices.



(a) (b)

Fig. 2. Temporal envelopes obtained by NMF of $(\mathbf{S}_{s_1}, \ldots, \mathbf{S}_{s_6}, \mathbf{S}_{c_1}, \ldots, \mathbf{S}_{c_6})$, where $\mathbf{S}_{s_i}$ is the spectrogram of a sample from a stroke on the edge of a 16 inch china cymbal and $\mathbf{S}_{c_i}$ is the spectrogram of a sample from a stroke on the edge of a 16 inch crash cymbal, with $1 \leqslant i \leqslant 6$. The peaks of the temporal envelopes are marked with rectangles (for samples from the china cymbal) and circles (for samples from the crash cymbal). (The envelopes of the sixth crash are not noticeable because of the sample's very low energy.)

cymbal and other six from strokes on the edge of a 16 inch crash cymbal.

In particular, note how the highest peaks in the temporal envelopes from the crash cymbal (marked with a circle) are much higher than the peaks in the temporal envelopes from the china cymbal in Fig. 2b. This suggests that the corresponding basis function (from the bottom Fig. 1) describes a property of the sounds from the crash cymbal.

In the example given in Fig. 1 and 2, $\mathbf{\Theta}$ is a matrix with 2 columns, where the first column contains the basis functions in the top Fig. 1 and the second column contains the basis function in the bottom Fig. 1. $\mathbf{P}_n$ (with $n \in \{s_1, \ldots s_6, c_1 \ldots, c_6\}$) is a matrix with 2 lines, with the nth envelopes from Fig. 2a and b.

3. RESULT ANALYSIS

The data used to test and train the classifier was a set of strokes on six different cymbals with different diameters and of different classes: 16 inch crash cymbal, 14 inch crash cymbal, 16 inch china cymbal, 9 inch splash cymbal, 20 inch ride cymbal, and 14 inch hi-hats. For each cymbal we chose to analyze samples from the areas that are most commonly used by drummers: the edge (for china, splash and crashes) and the bow (for ride and closed hi-hat). We used two wooden drumsticks made of *pau-santo*. The recorded samples have different loudness levels.

The sounds were digitized using a sampling frequency of 44100 Hz. The spectrograms were computed with the fast Fourier transform (FFT) using a sliding Hanning window of 2048 samples and 50% overlap between successive frames. The length of the FFT was the same as the size of window.

Cymbal combination	A	B	C	D	E	F	G
Correctly classified samples	11	11	11	12	12	15	16

Table 1. Number of correctly classified samples out of the 12 samples in the tests with two cymbals (A to E) and out of the 18 samples in the tests with three cymbals (F and G). We chose cymbal combinations commonly played by drummers: splash and china (A), 14 inch crash and 16 inch crash (B), splash and 16 inch crash (C), china and 16 inch crash (D), closed hi-hat bow and ride bow (E), china, 16 inch crash and 14 inch crash (F), and splash, 16 inch crash and 14 inch crash (G). (The stroke area is edge unless otherwise specified.)

The training and test sets were built with samples from two or three cymbals and contained six samples from each of the cymbals in the set. The samples were spread along the various levels of amplitude for each set (from loud to soft strokes). Obviously, the samples used in the training sets were different from the ones in the test sets.

A training sample, which is a column of $\mathbf{P}$, is an M dimensional vector of coefficients from M temporal envelopes from one sound, where M is the number of basis functions in $\mathbf{\Theta}$. More specifically, the training sample $t_{i,n}$ is a vector that contains the coefficients associated to the ith frame of spectrogram $\mathbf{S}_n$: $t_{i,n} = (p_{1,i,n}, \ldots, p_{M,i,n})$, where $p_{r,c,n}$ is the coefficient in the rth row and cth column of $\mathbf{P}_n$.

A test sample consists also of an M dimensional vector of coefficients extracted from M temporal envelopes. Yet, while the temporal envelopes in the training set are learned by NMF of the spectrograms, the temporal envelopes in the test set are determined by the following equation:

$$\mathbf{P}_{\text{test sound}} = \mathbf{\Theta}^{-1}\mathbf{S}_{\text{test sound}} \quad , \qquad (2)$$

where $\mathbf{\Theta}^{-1}$ is the pseudo-inverse of $\mathbf{\Theta}$.

Table 1 shows the results obtained for two cymbal combinations (combinations A to E). Out of the 5 tests performed, only 3 samples were misclassified: one sample from the china cymbal (in combination A), one from the 16 inch crash cymbal (in combination B) and one from the splash cymbal (in combination C). The overall classification rate was 95%. The most surprising results came from the combination of the ride with the hi-hat (combination E), and of the two crashes (combination B), given these cymbals have very similar characteristics. The crashes are of the same class of cymbals, while the ride bow and closed hi-hat bow, both have very low energy and a fast decay.

The overall classification rate obtained for three cymbal combinations was 86% (Table 1, combinations F and G). Unsurprisingly, out of the 5 samples misclassified, 3 were from the 14 inch crash, which were misclassified as being from the 16 inch crash.

4. CONCLUSIONS

A classifier for drum kit cymbals has been proposed. While most drum kit classifiers focus on distinguishing different instruments in the drum kit (like bass drum, snare drum, hihat, toms and cymbals) little attention has been given to the classification of different cymbals. This is a harder problem than distinguishing the different instruments in the drum kit, because while the sounds from those instruments have very different characteristics, sounds from different cymbals are much more alike. In spite of those similarities, here we have proven that these sounds can be correctly classified.

The proposed classifier uses spectral features learned by NMF of the spectrograms, to train a 1-NN algorithm that is then used to classify new sounds. It achieves very high classification rates: 95% for training and test sets composed of samples from two cymbals and 86% for sets composed of samples from three cymbals.

There have been other studies that use NMF of the spectrograms of drum sounds [10, 12]. Yet, the techniques used differ from ours. While we use the learned spectral basis functions directly as features that are feed into the k-NN algorithm, Moreau and Flexer extract other pre-defined features from the basis functions. On the other hand, Paulus and Virtanen use NMF of the spectrograms to estimate the spectra of each instrument: they separately analyze different instruments and obtain several spectral basis functions that characterize the samples from each instrument. The basis functions from each instrument are then averaged to obtain a spectrum that characterizes the instrument.

Whereas we only explored spectral features, temporal features can also be useful for distinguishing cymbal sounds. In particular, they may be useful for distinguishing the china cymbal from the other cymbals, because while the shape of the relative decay envelopes of the different frequency frames of most cymbals behave in a similar fashion, the relative decay envelopes of the china show a different behavior.

5. REFERENCES

[1] A. Tindale, A. Kapur, and I. Fujinaga, "Towards timbre recognition of percussive sounds," in *Proceedings of the International Computer Music Conference*, 2004, pp. 592–595.

[2] W. Schloss, *On the automatic transcription of percussive music -from acoustic signal to high-level analysis*, Ph.D. thesis, CCRMA, Department of Music, Stanford University, 1985.

[3] J. Bilmes, "Timing is of the essence: Perceptual and computational techniques for representing, learning and reproducing expressive timing in percussive rhythm," M.S. thesis, Massachussetts Institute of Technology, Media Laboratory, 1993.

[4] D. FitzGerald, E. Coyle, and B. Lawlor, "Sub-band independent subspace analysis for drum transcription," in *Proceedings of the Digital Audio Effects Conference (DAFX02)*, 2002, pp. 65–69.

[5] D. FitzGerald, *Automatic Drum Transcription and Source Separation*, Ph.D. thesis, Dublin Institute of Technology, 2004.

[6] D. FitzGerald and J. Paulus, "Unpitched percussion transcription," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds., pp. 131–162. Springer, 2006.

[7] J. Sillanpää, "Drum stroke recognition," Tech. Rep., Tampere University of Technology, 2002.

[8] P. Herrera, A. Yetarian, and F. Gouyon, "Automatic classification of drum sounds: A comparison of feature selection and classification techniques," in *Proceedings of the International Conference on Music and Artificial Intelligence*, 2002, pp. 79–91.

[9] F. Gouyon and P. Herrera, "Exploration of techniques for automatic labeling of audio drum tracks' instruments," in *Proceedings of MOSART: Workshop on Current Directions in Computer Music*, 2001.

[10] J. Paulus and T. Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *Proceedings of the European Signal Processing Conference*, 2005, pp. 4–8.

[11] J. Paulus, "Acoustic modelling of drum sounds with hidden markov models for music transcription," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, vol. 5.

[12] A. Moreau, "Drum transcription in polyphonic music using non-negative matrix factorization," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2007.

[13] I. Kaminskyj, "Multi-feature musical instrument sound classifier," in *Proceedings of Australasian Computer Music Conference*, 2001.

[14] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[15] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," *Neural Information Processing Systems*, pp. 556–562, 2001.

[16] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, 2007.