ON AUTOMATIC DRUM TRANSCRIPTION USING NON-NEGATIVE MATRIX DECONVOLUTION AND ITAKURA SAITO DIVERGENCE

Axel Roebel, Jordi Pons, Marco Liuni[*]

UMR 9912 - IRCAM/UPMC/CNRS
Paris, France

Mathieu Lagrange[†]

IRCCYN
CNRS Ecole Centrale de Nantes, France

ABSTRACT

This paper presents an investigation into the detection and classification of drum sounds in polyphonic music and drum loops using non-negative matrix deconvolution (NMD) and the Itakura Saito divergence. The Itakura Saito divergence has recently been proposed as especially appropriate for decomposing audio spectra due to the fact that it is scale invariant, but it has not yet been widely adopted. The article studies new contributions for audio event detection methods using the Itakura Saito divergence that improve efficiency and numerical stability, and simplify the generation of target pattern sets. A new approach for handling background sounds is proposed and moreover, a new detection criteria based on estimating the perceptual presence of the target class sources is introduced. Experimental results obtained for drum detection in polyphonic music and drum soli demonstrate the beneficial effects of the proposed extensions.

Index Terms— Source separation, music information retrieval, audio event detection, non-negative matrix deconvolution, drum transcription.

1. INTRODUCTION

Drum transcription is part of a larger application, called music transcription, that deals with the automatic description of a music piece in form of a symbolic score. The automatic transcription of music is a very active area of research [1] but nevertheless todays methods are still far from being robust enough to allow resynthesis of high quality music from the derived scores. Methods based on non-negative factorisation have shown good potential to improve results of automatic music transcription methods, and the present article deals with the application of non-negative matrix factorisation methods to the drum transcription problem.

Drum transcription methods proposed in the literature can be divided into three groups: *segment and classify* [2] [3], *match and adapt* [4] and *separate and detect* [5]. The method based on non negative factorisation that will be developed in the present paper belongs to the last approach. The factorisation methods assume that the observed spectrogram $\mathbf{V} \in \Re^{\geq 0, M \times N}$ results from the superposition of K source spectrograms $\mathbf{Y}_k \in \Re^{\geq 0, M \times N}$. Where K is the number of sources, M is the number of frequency bins and N is the number of time frames of the segment under analysis. Each of the components $\mathbf{Y}_k$ is represented by the outer product of K basis ($\vec{W}_k$ of size M) with a corresponding activation ($\vec{H}_k$ of length N).

NMD is an extension of NMF which is capable to use templates with a temporal structure. This makes the system capable to exploit the time-frequency "signature" of each source (patterns from now on), formulated as follows:

$$\mathbf{V} \approx \hat{\mathbf{V}} = \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \mathbf{H}^{t\rightharpoonup} \tag{1}$$

where $(\cdot)^{t\rightharpoonup}$ is a column shift operator described in [6], $\hat{\mathbf{V}}$ approximates $\mathbf{V} \in \Re^{\geq 0, M \times N}$, in $\mathbf{W}_t \in \Re^{\geq 0, M \times K}$ are represented the patterns that model the K sources, T is the number of frames available for each pattern and $\mathbf{H} \in \Re^{\geq 0, K \times N}$.

The results that are obtained with the algorithms based on factorisation depends critically on the objective function that is used. In most studies the Kullback-Leibler (KL) divergence is favoured [5, 7]. Alternatively the generalised beta divergence has been used [8] and especially the Itakura Saito divergence (IS) [9, 10]. A special advantage of the IS is the scale invariance that allows taking into account spectral features with low energy. Therefore, we have selected to use the Itakura Saito divergence for the following study.

The main contributions of this paper can be summarised as follows: A new method for the representation of background (non target) sounds (see 2.4). A new method to stabilise the factorisation algorithm against numerical problems due to very small amplitudes in the observed spectra or bases (see 2.2). A new detection criterion based on an estimate of the ratio between target and non target energy (see 2.6). A new method to construct the templates for representing a class of events from a given set of examples (see 2.3).

This paper is organised as follows: Section 2 introduces the basic ideas of the proposed drum transcription system. Section 3 describes the experimental results that have been obtained and Section 4 summarises the conclusions and future work. The notation to be used is as follows: Matrices are in bold: ($\mathbf{H}$); vectors are denoted with an arrow ($\vec{H}$); and scalar values are denoted as italic text (k).

2. DESCRIPTION OF THE TRANSCRIPTION SYSTEM

The drum detection algorithm to be described in the following is based on the non negative tensor deconvolution (NTD) audio event detection algorithm described in [11]. In the present study the audio signals have a single channel so that the NTD algorithm simplifies into NMD.

In this preliminary study we have focussed on drum events generated by bass drum (bd), hi-hat (hh) both open and closed, and snare drum (sd). The restriction to these three drum sound classes follows common practice in the state of the art [5, 4, 12]. Prior to drum detection we run an onset detection algorithm [13], that allows us to limit the decomposition to potentially most interesting signal segments, which leads to a strong reduction in false positives, and also a reduction in runtime.

[*]This work has been funded partly by the FP7 project 3DTVS.
[†]This work has been funded partly by the ANR project HOULE.

In the following the time frequency representation of the signal is denoted as $\mathbf{V}$. $\mathbf{V}$ is constructed from the power 2 of the short time Fourier transform (STFT) of the audio signal using a Hanning window of size M, with analysis step M/4 and fft-size N. For decomposition and detection the power spectrogram is converted into a mel spectrogram using 40 frequency bands that are distributed equally in mel scale over the full signal band. The mel band signals are obtained using overlapping triangular filters [14].

Note that the probabilistic interpretation of the IS-NMF given in [15] will no longer hold for the proposed MEL representation. The probabilistic interpretation could be preserved if one would sum complex spectral coefficients. Experimental evaluation has shown, however, that this leads to decreasing performance. Comparing the performance of the latter MEL representation with the performance of the full power spectrogram, no consistent advantages were observed. However, the run time of the algorithm using MEL representation is about 25 times shorter when using the full spectrogram. Therefore, the MEL representation has been selected for the following experiments.

2.1. NMD update rules

The update rules are obtained for minimising the cost function given by the Itakuro Saito divergence (IS)

$$d_{IS}(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{ik} \frac{v_{ik}}{\hat{v}_{ik}} - \log \frac{v_{ik}}{\hat{v}_{ik}} - 1 \tag{2}$$

where the sum goes over all time frequency samples v_{ik} of the target $\mathbf{V}$ and reconstructed $\hat{\mathbf{V}}$ MEL spectrograms. As shown in [10] the following multiplicative update rules are guaranteed to reduce the objective function in each step:

$$\mathbf{W}^t \leftarrow \mathbf{W}^t \circledast \frac{\left(\mathbf{V} \circledast \hat{\mathbf{V}}^{\circledast(-2)}\right) \circ \mathbf{H}^{t\rightharpoonup}}{\hat{\mathbf{V}}^{\circledast(-1)} \circ \mathbf{H}^{t\rightharpoonup}} \tag{3}$$

$$\mathbf{H} \leftarrow \mathbf{H} \circledast \frac{\sum_t \left(\mathbf{V} \circledast \hat{\mathbf{V}}^{\circledast(-2)}\right)^T \circ \mathbf{W}^{t\rightharpoonup}}{\sum_t \left(\hat{\mathbf{V}}^{\circledast(-1)}\right)^T \circ \mathbf{W}^{t\rightharpoonup}} \tag{4}$$

The $\circ$ symbol denotes the outer product, while $\circledast$ is the Hadamard product and powers of matrices indicated with $\circledast(\cdot)$ are elementwise. After each update the patterns are normalised to ensure the energy can be obtained from $\mathbf{H}$.

The motivation to use the IS cost function in the present study is due to its scale invariance, which is an interesting property for decomposing audio signals [15]. This cost however, leads to problems with low energy noise as discussed in the following section.

2.2. Controlling the Noise Robustness of IS-NMD

The IS divergence cost function is a ratio based measure and therefore, all amplitude levels of the target spectrum are taken into account with the same importance. The advantage is that perceptually relevant properties in lower amplitude levels are fully taken into account. On the other hand, ambient noise or band stop filters in very high frequency bands that are perceptually not relevant, will also have a strong impact on the decomposition. Moreover, the numerical stability of the non-negative decomposition may be affected by individual very small values or even zeros that might be present in the reconstructed matrix $\hat{\mathbf{V}}$.

In [8, 7] this problem has been addressed by means of adjusting the β parameter of the Bregman divergence. But this solution does not allow an intuitive control. Moreover, this solution does not adress the numerical problems with zeros in the bases $\mathbf{W}$ and by consequence in $\hat{\mathbf{V}}$. Note that limiting $\hat{\mathbf{V}}$ from below is no appropriate solution because it introduces a systematic bias into the estimate which can have very strong impact on the decomposition notably due to the scale invariance property.

Here a new approach is proposed that consists of adding a small constant value to the observed MEL spectrogram $\mathbf{V}$ and at the same time adding a fixed constant basis with fixed and matching activation to the decomposition. A properly selected offset masks the effect of low level noise avoiding the undesirable impact of irrelevant effects on the decomposition preserving nevertheless the positive impact of the ratio based measure for signal components above the noise floor. The modification follows coherently the structure of the non-negative representation and ensures a few desirable properties: first, for the case that a perfect reconstruction is possible the global optimum is not modified. If a perfect reconstruction is not possible then the impact of the amplitude levels below the noise floor will be limited. Second, seeing the noise floor as an additional component we conclude that the probabilistic interpretation proposed in [15] still applies (if the MEL representation is not used). The question that arises here is the question of how the constant should be selected. The answer to this question depends notably on the SNR between background signals and target signals. The experimental investigation will demonstrate that a proper selection of the noise floor can positively effect the performance of the detection system (see 3).

2.3. Composing the target pattern set

For each of the target sound events - in this preliminary study bass drum (bd), open and closed hi-hat (hh), snare drum (sd) are used - a dictionary of patterns has to be learned from a collection of target sounds. For our experiments we used drum sounds from Vienna symphonic database, ENST drums database [16] and the training example provided in the IDMT-SMT-Drum database [17]. The database of target patterns for each class is constructed such that it allows to represent the complete target class with a small error. To simplify interpretation of the activations in the detection phase each pattern should itself be a valid sound of the target class.

The set of target patterns is grown iteratively. It is initialised by means of selecting the target pattern having the minimum average distance (as measured by the Itakura Saito divergence) to all the other elements in the target class training database. For extending the set of target patterns first the NMD algorithm is used to represent the full set of events of the target class, however with adapting only the activations, and then the target event with maximum representation cost is selected to extend the training database. In contrast to vector quantisation this procedure allows to fully exploit the mixing of the patterns in the target set. For selecting the target bases a noise floor should be used as described in section 2.2 to avoid adding new patterns due to irrelevant variations of the noise floor.

2.4. Representation of the background

To limit the impact of the other instruments on the performance of the drum detection results the system should include a background representation. In [11] this was done by means of extending the dictionary of fixed target patterns with a small number of patterns not belonging to the target classes. These patterns were adapted such that they could represent any non target signals. A disadvantage of this approach is the fact that the performance of the background strongly depends on the number of background bases and their random initialisation. Depending on the number of background

bases, and the complexity of the background sound the background bases may or may not be used to fit the target events, and there are no means to select the appropriate number of background patterns. Another approach was used in [18] where a set of noise bases was trained on 6h of noise data. Given the nearly unlimited amount of variations in background noise this approach seems unrealistic even for the special case when the background noise is known to be music.

In this study a new approach for handling the background scene is proposed allowing for a more precise control of the impact of the background sounds. The background model is setup such that it is guaranteed to be able represent the entire audio signal. Here this means that the decomposition is performed in segments that are exact multiples of the NMD bases time length T. For a segment of length lT there will be l non overlapping background bases, and all but l activations are zero. Given that all the mel spectrum templates $\mathbf{W_j}$ are normalised in energy the representation of the complete signal has only one perfect solution. The control of the amount of energy represented by the background bases is provided through ℓ_1-norm regularisation of the activation of the background bases only. The objective function eq. 2 becomes

$$\min_{W_j,H_j} d_{IS}(\mathbf{V}|\hat{\mathbf{V}}(\mathbf{W},\mathbf{H})) + \lambda \sum_{k_b \in backgrd,i} h_{k_b,i}. \quad (5)$$

Here $h_{k_b,i}$ is the activation of the background bases k_b for time i and λ is the regularisation factor that penalises activations of the background.

The key problem here is the determination of λ. For obtaining a first idea of the effect of the regularisation one can study a toy system with a single background base $\vec{W}$ with elements w_i a scalar activation h and a observed mel power spectrum $\vec{V}$ with elements v_i. In this case the closed form solution for h can be obtained

$$h = \sum_i (\sqrt{4\lambda v_i/w_i + 1} - 1)/(2\lambda) \quad (6)$$

$$= \sum_i 2v_i/(w_i(\sqrt{4\lambda v_i/w_i + 1} + 1)) \quad (7)$$

where the second step has been achieved after multiplying the numerator and denominator with $(\sqrt{4\lambda v_i/w_i + 1} + 1)$. For $\lambda = 0$ this gives the known solution $h = \sum_i v_i/w_i$. And for $\lambda > 0$ one can see that the impact of λ depends on the ratio v_i/w_i, which means that for situations with only background components one can achieve corresponding regularisation effects if the regularisation constant is increased proportional to $1/v$. While the toy system is certainly not sufficient to create a full understanding of the impact and scaling behaviour of λ the present result is in agreement with the experimental results described in section 3.

2.5. Decomposing

Starting from the onset positions (specified by start and end time for each onset event) that are detected with the algorithm described in [13] and the same setup that was used in [19] sound segments are determined that are covering $3T$ spectral frames. The segment is formed such that it start exactly one analysis window before the start of the onset given by the onset detection algorithm and ends $3T$ analysis frames later.

The decomposition is done using all the target and background bases as described in sec. 2.4. Initialisation of the target pattern activations is derived from the energy contour of the signal. Background activations are treated the same besides that background activations are non zero only at 1 position over the segment for each of the individual background bases, and that their activation is penalised by the regularisation term given in eq. 5. Then NMD decomposition is performed until convergence.

2.6. Drum event detection

Once the decomposition of the onsets zones is done the final decisions have to be formed. Three criteria are used to decide whether at any given point in time a drum class is active. First, an activation based criteria is used: the activation vector of each drum (for obtaining the activation vector the activations of all the patterns representing the respective drum class are summed together) is convolved with the kernel $[0.9, 1, 0.9]$ to be able to properly take into account drum event positions between the analysis frames. The resulting activation for class c is denoted $H_c(n)$, it represents a time evolution of the activation of the class. Only the local maxima of these activations are retained and the median of these maxima $\hat{H}_c$is formed over the complete sound file. Then a threshold μ_P is introduced and only activations fulfilling $H_c(n) \geq \mu_p \hat{H}_c$ will be retained. The argument here is that each drum type will be active less than half the time and so the median value of the activations will represent a notion of the activations due to background events. Only activations that are above this background activation should be retained.
Second, a sum of all activations is calculated followed by smoothing with a rectangular smoothing window $F(n)$ of duration T and amplitude $1/T$. The result is a smoothed full activation $H_F(n)$, which due to the energy normalisation of all bases patterns approximates the signal energy. A second threshold factor μ_F is used and only class activations fulfilling $H_c(n) \geq \mu_F H_F(n)$ will be retained. This criterion establishes a minimum SNR for detected events.
For the third criterion the power spectrum related to the target class is re-synthesised and a band wise target to background ratio is calculated in form of a floating average of the energy the separated target class within each MEL channel as follows

$$P(k,n) = \frac{F(n) * v_c(k,n)^2/\hat{v}(k,n)}{F(n) * v_c(k,n)} \quad (8)$$

where $P(k,n)$ is the average relative energy in the MEL bin k at time position n, $F(n)$ is the same smoothing window described above, and $v_c(k,n)$ and $\hat{v}(k,n)$ are the power spectrum samples in MEL sub band k at time position n of the separated target spectrogram and the complete estimated spectrogram respectively. This value indicates the prominence of the target class in the observed spectrum weighted by the energy distribution in the synthesised target spectrum. An average of the three largest values of all sub bands is calculated $\hat{P}(n)$ and this value is compared with a third threshold μ_{PE}. Only target spectra leading to $\hat{P}(n) \geq \mu_{PE}$ are retained.

Locations that pass all threshold tests are accepted as drum events only if they are located within onset start and end times provided by the onset detection algorithm mentioned above. Additionally, the drum events are then filtered such that for each class only the strongest detected event for each onset zone is retained.

3. EXPERIMENTAL EVALUATION

For the following evaluation each of the three target classes (bd, hh, sd) are represented using 15 patterns that were derived from target training set using noise level parameters covering the noise level attenuation from -15 to -55dB compared to the maximum value of the target class. The experimental investigation aims to answer the questions that were posed in the previous sections: what noise floor should be used during decomposition, what regularisation parameter

DB	drum	F-meas μ_P	F-meas μ_{PE}	F-meas μ_P, μ_{PE}, μ_F
RWC	bd	0.79	0.63	0.81
	hh	0.79	0.75	0.79
	sd	0.78	0.64	0.88
ENST Drums	bd	0.88	0.81	0.96
	hh	0.89	0.86	0.89
	sd	0.57	0.66	0.74

Table 1. F-measures obtained using the median peak threshold (left), the perceptual dominance threshold (centre), all thresholds (right).

should be used for controlling the background activations, and how does the noise floor applied during the training of the target datasets impact the detection properties.

Two datasets were used: 1) 100s of excerpts of drum solos of the ENST drum database [16] and 2) 120s of polyphonic synthesised from midi provided by the music genre data of the RWC database [20] using the method described in [21]. The sample rate of the sound samples is 44.1kHz, and the window size of the initial STFT analysis is 1500 samples, step size 375 samples and FFT size 2048, and the number of analysis frames covered in each pattern $T = 20$. The evaluation measure is a standard F-measure calculated from recall and precision individually for the three traget classes. A drum event is considered correctly detected if the detection appears with the correct class within $\pm$30ms of the annotated drum event.

The three thresholds described in the section 2.6 were optimised on a set of training data separately for each of the target classes such that maximum F-measure is obtained. Given the very strong difference between playing technique and spectral distribution of the differente drums it appears normal that the thresholds μ_p, μ_{pe}, μ_F strongly depend on the drum type.

3.1. Results

A first results concerns the effect of the noise level on the detection performance for a given setup of target classes. This result is shown in Fig. 1. The result is obtained for the synthesised RWC samples, but similar results however with rather different convergence points have been obtained as well for the ENST drum examples. The results displayed is the F-measure for each drum type as a function of the noise level parameter discussed in 2.2.
It becomes clear that for the bass drum, the noise floor should be set higher than for hh and sd. This can be easily understood by the fact that the bass drum is generally rather dominant when it appears. For snare and hi-hat lowering the noise floor seems to reveal more relevant spectral details and detection performance increases until about 55dB attenuation relativ to the maximum amplitude of $\mathbf{V}$.

A second result is related to the regularisation control parameter λ. The results show that the average of the F-measures of the three drums is improved when compared to a situation with no background bases, or when using background bases trained with standard NMD configuration. Using the same regularisation control parameter for all drum detectors the optimum value is $\lambda = 25$ leading for the RWC sound examples to an average F-measure increasing from 0.78 to 0.83. However, the results are rather different for the different drums. The optimal regularisation parameters for the respective drum detectors are: for bd $\lambda_{bd} = 50$ with F-measure increasing from 0.79 to 0.82, for sd $\lambda_{sd} = 10$ with increase in F-measure from 0.78 to 0.88, and for hh $\lambda_{hh} = 100$ leading to an increase in F-measure from 0.78 to 0.8. Following the mathematical analysis described in sec. 2.4 this behaviour should be related to the signal energy present

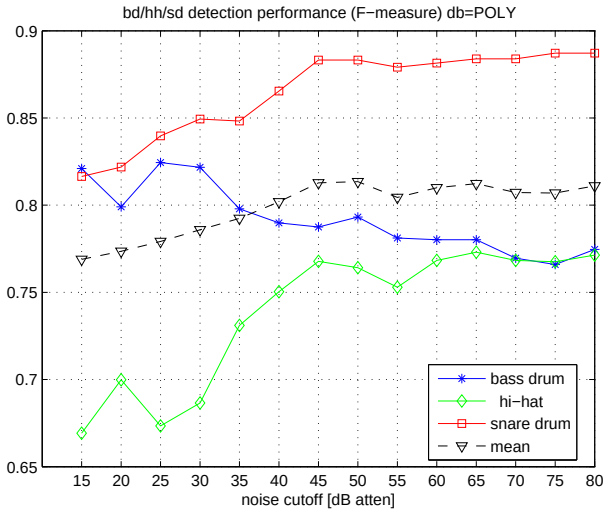

Fig. 1. Detection performance as a function of the noise level parameter in the decomposition phase for the RWC sound examples.

in the components related to the different drums in $\mathbf{V}$. hh has by far the smallest energy. The bd and sd components are relatively strong and therefore the background bases will not be able to represent these two drums even with small λ.

The overall performance of the algorithm using parameter settings that are optimal for the global F-measure comprising all detections (noise level for detection: -55dB, noise level for training the bases: -25dB, and $\lambda = 25$) is shown in Tab. 1. This table allows to see the effect of the individual detection criteria on the final F-measure for all instruments individually. If only the threshold acting on the activation relative to the median of the activation is used the performance is overall quite good, with the exception of the snare drum that is too low. Using the estimated perceptual impact of the estimated component is a bit less satisfying with F-measures below 0.7. Combining all the three criteria however we get a significant improvement of the performance such that all results, for the polyphonic and the drum loop data, are always above 0.74 F-measure.
Note that assuming that the subsequent drum detection phase will produce perfect results, the initial onset detection described above would result in an F-measure of 0.91 (bd), 0.95 (hh), and 0.94 (sd).

4. CONCLUSIONS AND FUTURE WORK

The present article has proposed a number of improvements of a NMD based audio event detection framework that was previously used for detection of audio events in film scenes, and that was here used for the detection and classification of drum events.

The proposed changes have been motivated and the experimental evaluation has shown that the changes have potential to significantly improve the algorithm. The overall system performs well, and it can be noted that there are no random initialisations to be performed. This means that the result is completely deterministic, which simplifies the use of the algorithm.

A main result of the present study is the fact that the detectors require different parameter settings (noise level and regularisation) for different drums. Therefor, one can expect that overall performance would be improved if events for the different drum types would be detected individually, each using a dedicated decomposition of the observed input spectrum. Further work is planned to study how the configuration of the detection framework can be optimised.

5. REFERENCES

[1] Anssi Klapuri, Manuel Davy, et al., Eds., *Signal processing methods for music transcription*, vol. 1, Springer, 2006.

[2] Jouni Paulus, "Drum transcription from polyphonic music with instrument-wise hidden Markov models," in *Proc. of the First Annual Music Information Retrieval Evaluation eXchange*, London, UK, Sept. 2005.

[3] Koen Tanghe, Sven Degroeve, and Bernard De Baets, "An algorithm for detecting and labelling drum events in polyphonic music," *Proceedings of the 1st Annual Music Information Retrieval Evaluation Exchange*, pp. 11–15, 2005.

[4] Kazuyoshi Yoshii, Masataka Goto, and Hiroshi G Okuno, "Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 333–345, 2007.

[5] Jouni Paulus and Tuomas Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *Proceedings of the 13th European Signal Processing Conference*, 2005, p. 4.

[6] P. Smaragdis, "Non-negative matrix factor deconvolution, extraction of multiple sound sources from monophonic inputs," *International Symposium on Independent Component Analysis and Blind Source Separation (ICA) 3195 (2004) 494*, 2004.

[7] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*, F. Nielsen and R. Bhatia, Eds., pp. 341–372. Springer, 2012.

[8] Emmanuel Vincent, Nancy Bertin, and Roland Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *ITASLP*, vol. 18, no. 3, pp. 528–537, 2010.

[9] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu, "Non-negative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[10] Cédric Févotte and Jérôme Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[11] Yuki Mitsufuji, Marco Liuni, Alex Baker, and Axel Roebel, "Online non-negative tensor deconvolution for source detection in 3dtv audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 3082–3086.

[12] Olivier Gillet and Gaël Richard., "Transcription and separation of drum signals from polyphonic music.," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 3, pp. 529–540, 2008.

[13] A. Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, 2003, pp. 344–349.

[14] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," Cuidado project report, IRCAM, 2004.

[15] Cédric Févotte, "Itakura-Saito nonnegative factorizations of the power spectrogram for music signal decomposition," *Machine Audition: Principles, Algorithms and Systems*, pp. 266–296, 2011.

[16] Olivier Gillet and Gaël Richard, "ENST-Drums: an extensive audio-visual database for drum signals processing.," in *ISMIR*, 2006, pp. 156–159.

[17] Christian Dittmar and Daniel Gärtner, "Real-time transcription and separation of drum recordings based on nmf decomposition," in *Proc Int. Conf on Digital Audio Effects (DAFx)*, 2014.

[18] Felix Weninger, Martin Wöllmer, Jürgen Geiger, Björn Schuller, Jort F. Gemmke, Antti Hurmalainen, and Tuomas Virtanen abd Gerhard Rigoll, "Non-negative matrix factorization for highly noise-robust asr: To enhance or to recognize?," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 4681–4684.

[19] Francois Rigaud, Mathieu Lagrange, Axel Roebel, and Geoffroy Peeters, "Drum extraction from polyphonic music based on a spectro-temporal model of percussive sounds," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 381 – 384.

[20] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "RWC music database: Popular, classical and jazz music databases.," in *Proc. of the Int. Soc. on Music Information Retrieval Conf. (ISMIR)*, 2002, vol. 2, pp. 287–288.

[21] C. Yeh, N. Bogaards, and A. Röbel, "Synthesized polyphonic music database with verifiable ground truth for multiple f0 estimation," in *Proc. of the Int. Soc. on Music Information Retrieval Conf. (ISMIR)*, 2007, pp. 393–398.