STUDY OF REGULARIZATIONS AND CONSTRAINTS IN NMF-BASED DRUMS MONAURAL SEPARATION

*Ricard Marxer, Jordi Janer**

Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
ricard.marxer@upf.edu, jordi.janer@upf.edu

ABSTRACT

Drums modelling is of special interest in musical source separation because of its widespread presence in western popular music. Current research has often focused on drums separation without specifically modelling the other sources present in the signal. This paper presents an extensive study of the use of regularizations and constraints to drive the factorization towards the separation between percussive and non-percussive music accompaniment. The proposed regularizations control the frequency smoothness of the basis components and the temporal sparseness of the gains. We also evaluated the use of temporal constraints on the gains to perform the separation, using both ground truth manual annotations (made publicly available) and automatically extracted transients. Objective evaluation of the results shows that, while optimal regularizations are highly dependent on the signal, drum event position contains enough information to achieve a high quality separation.

1. INTRODUCTION

Drums transcription has been regarded as an important task by the Music Information Retrieval (MIR) community and in the past decade there has been increasing interest in developing techniques for separating the drums track from music mixes. [1] derive a method based on synthetic drums sound pattern matching. The matching is performed using the correlation as the objective function. [2] computes the presence of percussive events based on the temporal derivative of the spectral magnitudes on the decibel scale. The separation is then performed by spectral modulation, weighting the spectral bins by the individual bin derivatives previously computed. [3] propose another method based on spectrotemporal features. However in this case both the temporal and frequency derivatives are taken into account. [4] decompose the signal into a basis of Exponentially Damped Sinusoids (EDS) using a noise subspace projection approach. This leads to a harmonic/noise decomposition that is used to extract the percussive sources. [5] use a template-based pattern matching technique to estimate and separate the drums spectra from the rest. The authors show several applications such as remixing, drum timbre modification and rhythmic sources equalization. [6] propose the use of Non-negative Matrix Factorization (NMF) and Support Vector Machine (SVM) classification to perform drum separation. The technique consists in performing an NMF decomposition of the spectrogram of the mixture and classifying the basis components of the factorization using Mel-Frequency Cepstrum Coefficients (MFCC) and an SVM trained using isolated drums and harmonic audio recordings. [7] propose a similar approach where Nonnegative Matrix Partial Co-Factorization (NMPCF) is used to avoid training harmonic components. In [8] the authors propose using the Flexible Audio Source Separation Toolbox (FASST) to perform an isolation of the drum components in a mixture. FASST is based on non-negative factorization of a complex spectrum model that contains templates for specific spectral and temporal patterns which are able to reconstruct harmonic and percussive components when combined.

The use of temporal constraints on NMF is not new and has proven useful in several scenarios. [9] use score-based temporal restrictions on the gains of an NMF decomposition to estimate piano notes attacks.

Here we address the separation of drums in polyphonic music mixtures, typically containing lead vocals. One approach to the separation of the singing voice of special interest to us is the SIMM (Smoothed Instantaneous Mixture Mode) by [10], which uses a source/filter decomposition based on NMF.

2. PROPOSED METHOD

We propose an extension to the SIMM method that includes an extra additive spectral component to represent percussive events. The proposed spectrum model can be defined as $\hat{\boldsymbol{V}} = \hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_{m'} + \hat{\boldsymbol{X}}_d$, where the additional component $\hat{\boldsymbol{X}}_d$ corresponds to the estimation of the drums. The lead vocals spectrum $\hat{\boldsymbol{X}}_v$ is decomposed in multiple factors representing a source-filter harmonic model, the other components are decomposed into two factors $\hat{\boldsymbol{X}}_{m'} = \boldsymbol{W}_{m'}\boldsymbol{H}_{m'}$ and $\hat{\boldsymbol{X}}_d = \boldsymbol{W}_d\boldsymbol{H}_d$. It is trivial to show that without any further modifications and with a specific ordering of the multiplicative updates, the proposed spectrum model is equivalent to SIMM with $\boldsymbol{W}_m = [\boldsymbol{W}_{m'}; \boldsymbol{W}_d]$ and $\boldsymbol{H}_m = [\boldsymbol{H}_{m'}, \boldsymbol{H}_d]$.

As in the original SIMM the actual separation is performed by Weiner filtering using the drums spectra estimation $\hat{\boldsymbol{X}}_d$. Thus the time-frequency mask becomes:

$$\boldsymbol{m}_d = \frac{\hat{\boldsymbol{X}}_d}{\hat{\boldsymbol{X}}_v + \hat{\boldsymbol{X}}_{m'} + \hat{\boldsymbol{X}}_d} \tag{1}$$

In the following sections we show different techniques in order to achieve the differentiation between the drums and the other musical accompaniment sources in $\hat{\boldsymbol{X}}_d$ and $\hat{\boldsymbol{X}}_{m'}$. First we present a method based on NMF regularizations and then one that uses information specific to the processed signal to apply constraints to the factorization.

* Authors thank Yamaha Corp. for their support. This research was partially funded by the PHENICX project (EU FP7 Nr. 601166).

2.1. Training

We study the use of semisupervised NMF in conjunction with the SIMM method for the separation of drums sources. In this scenario we first learn a set of basis components $\boldsymbol{W}^t_{drums}$ using recordings of drums in isolation and then use these components during the separation stage. The learned components will be used as constants and complemented with $N_{W\,d'}$ basis components that will be free and learned during the separation $\boldsymbol{W}_d = [\boldsymbol{W}^t_{drums}; \boldsymbol{W}_{d'}]$. We trained basis for two different types of percussive instruments: snare drums and cymbals. The bass drum was not used for training since preliminary results showed that by doing so, a large amount of low frequency content from other sources was assigned to the drums. The result is two sets of learned basis components $\boldsymbol{W}^t_{drums} = [\boldsymbol{W}^t_{snare}; \boldsymbol{W}^t_{cymbal}]$.

2.2. Regularizations

[11] proposed the use of temporal continuity and sparseness regularizations on the gains of an NMF process to isolate sustained harmonic sources. We extend these regularization terms to the the basis factor and integrate them into the proposed spectrum model based on SIMM.

In our proposed method we apply different regularizations to the factors $\hat{\boldsymbol{X}}_{m'}$ and $\hat{\boldsymbol{X}}_d$ in order to disambiguate between drums and other musical accompaniment. Drums are characterized by their wideband smooth spectral shape and their sparseness in the time axis, since they are often transient sounds with a short decay and a shorter attack. On the other hand we assume the spectral evolution of the other musical accompaniment to be smooth in time. We define two additional regularization terms to include this prior knowledge into the factorization. We propose a regularization on the basis that penalizes frequency domain discontinuities in the spectra. The term is similar to the one proposed by [11] that penalizes temporal discontinuities of the gains. In our case the smoothness is enforced on the frequency axis of the basis components. The resulting frequency continuity regularization is defined as:

$$\boldsymbol{J}^{fc}_{\boldsymbol{W}}(\boldsymbol{W}) = \sum_{w}^{N_W} \frac{1}{\sigma_w^{\omega\,2}} \sum_{\omega}^{N_\omega} \Big([\boldsymbol{W}]_{\omega,w} - [\boldsymbol{W}]_{\omega-1,w} \Big)^2 \quad (2)$$

where the standard deviation of the components is estimated as $\sigma^\omega_w = \sqrt{(1/N_\omega)\sum_\omega^{N_\omega}([\boldsymbol{W}]^2_{\omega,w})}$. The term w represent the basis index, ω the frequency index and t the time index (columns in $\boldsymbol{H}$).

The gradient of the regularization then becomes:

$$\Big[\varphi^{fc}_{\boldsymbol{W}}(\boldsymbol{W})\Big]_{\omega,w} = \quad 2N_\omega \frac{2[\boldsymbol{W}]_{\omega,w} - [\boldsymbol{W}]_{\omega-1,w} - [\boldsymbol{W}]_{\omega+1,w}}{\sum_i^{N_\omega}[\boldsymbol{W}]^2_{w,i}} - N_\omega \frac{2[\boldsymbol{W}]_{\omega,w} \sum_{i=2}^{N_\omega} \big([\boldsymbol{W}]_{i,w} - [\boldsymbol{W}]_{i-1,w}\big)^2}{\big(\sum_i^{N_\omega}[\boldsymbol{W}]^2_{i,w}\big)^2} \quad (3)$$

which can easily be expressed as an addition of positive and negative terms $\varphi^{fc^+}_{\boldsymbol{W}}$ and $\varphi^{fc^-}_{\boldsymbol{W}}$.

We also propose a regularization on the drums activation matrix $\boldsymbol{H}_d$ that penalizes gains that are non-sparse in time. The regularization is a simple variation on that proposed by [11].

$$\boldsymbol{J}^{ts}_{\boldsymbol{H}}(\boldsymbol{H}) = \sum_{t}^{N_T} \sum_{w}^{N_W} g([\boldsymbol{H}]_{w,t} / \sigma_t) \quad (4)$$

where $g(\cdot)$ is a function that penalizes non-zero gains, in our case $g(x) = |x|$. The only difference between the regularization term proposed in [11] and the one we propose is that the standardization is done with respect to each time frame instead of each basis. The gradient then becomes:

$$\big[\varphi^{ts}_{\boldsymbol{H}}(\boldsymbol{H})\big]_{w,t} = \quad \frac{1}{\sqrt{\frac{1}{N_W}\sum_i^{N_W}[\boldsymbol{H}]^2_{i,t}}} - \sqrt{N_W} \frac{[\boldsymbol{H}]_{w,t} \sum_i^{N_W}[\boldsymbol{H}]_{i,t}}{\big(\sum_i^{N_W}[\boldsymbol{H}]^2_{i,t}\big)^{3/2}} \quad (5)$$

Due to the additive nature of the spectrum model and regularizations, the derivation of the multiplicative update rules are quite straightforward. The multiplicative update rule for accompaniment $\boldsymbol{W}_{m'}$ remains the same as for $\boldsymbol{W}_m$ in the original SIMM method. The update rules for the $\boldsymbol{H}_{m'}$ $\boldsymbol{W}_d$ and $\boldsymbol{H}_d$ become:

$$\boldsymbol{H}_{m'} \leftarrow \boldsymbol{H}_{m'} \otimes \frac{\boldsymbol{W}^\top_{m'} \Big(\hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\Big) + \varphi^-_{\boldsymbol{H}_{m'}}}{\boldsymbol{W}^\top_{m'} \hat{\boldsymbol{V}}^{(\beta-1)} + \varphi^+_{\boldsymbol{H}_{m'}}} \quad (6)$$

$$\boldsymbol{H}_d \leftarrow \boldsymbol{H}_d \otimes \frac{\boldsymbol{W}^\top_d \Big(\hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\Big) + \varphi^-_{\boldsymbol{H}_d}}{\boldsymbol{W}^\top_d \hat{\boldsymbol{V}}^{(\beta-1)} + \varphi^+_{\boldsymbol{H}_d}} \quad (7)$$

$$\boldsymbol{W}_d \leftarrow \boldsymbol{W}_d \otimes \frac{\Big(\hat{\boldsymbol{V}}^{(\beta-2)} \otimes \boldsymbol{V}\Big) \boldsymbol{H}^\top_d + \varphi^-_{\boldsymbol{W}_d}}{\hat{\boldsymbol{V}}^{(\beta-1)} \boldsymbol{H}^\top_d + \varphi^+_{\boldsymbol{W}_d}} \quad (8)$$

where the gradient terms are defined as follows:

$$\begin{aligned} &\varphi^-_{\boldsymbol{H}_{m'}} = \alpha_{tc} \varphi^{tc^-}_{\boldsymbol{H}_{m'}}, \varphi^-_{\boldsymbol{H}_d} = \alpha_{ts} \varphi^{ts^-}_{\boldsymbol{H}_d}, \varphi^-_{\boldsymbol{W}_d} = \alpha_{fc} \varphi^{fc^-}_{\boldsymbol{W}_d} \\ &\varphi^+_{\boldsymbol{H}_{m'}} = \alpha_{tc} \varphi^{tc^+}_{\boldsymbol{H}_{m'}}, \varphi^+_{\boldsymbol{H}_d} = \alpha_{ts} \varphi^{ts^+}_{\boldsymbol{H}_d}, \varphi^+_{\boldsymbol{W}_d} = \alpha_{fc} \varphi^{fc^+}_{\boldsymbol{W}_d} \end{aligned} \quad (9)$$

and the parameters $\alpha_{tc} \in \Re^+$, $\alpha_{ts} \in \Re^+$ and $\alpha_{fc} \in \Re^+$ control the enforcement of the temporal continuity of the accompaniment gains $\boldsymbol{H}_{m'}$, the temporal sparseness of the drums gains $\boldsymbol{H}_d$ and the frequency continuity on the drums basis $\boldsymbol{W}_d$ respectively.

The regularizations can improve the separation between the musical accompaniment and the percussive components in the SIMM method. This separation is performed in an unsupervised manner since no signal-specific knowledge is needed. However the parameters controlling the regularizations may have a large influence on the results.

2.3. Constraints

Another extension proposed to the SIMM method for isolating the percussive instruments is the use of constraints. In this extension we assume the temporal positions of the drum events are known. This information is used to restrict the activation of the gains of the percussive components, reducing the degrees of freedom of the factorization problem. The constraints are performed in a manner similar to [9].

We consider a set of percussive sources $m_d \in [1, N_{M\,d}]$. We denote $t_e^{m_d}$ for $e \in [1, N_e]$ the frame indices of the attacks of the events of the percussive source m_d. The dictionary $\boldsymbol{W}_d$ is the set of basis components for all the percussive sources, with ${N_W}^s$ components assigned to each percussive source. The constraints

are set in the form of initializations to 0 in the corresponding gains matrix $\boldsymbol{H}_d$:

$$\boldsymbol{H}_d[w,t] = \begin{cases} \gamma, & \text{if } t_e^{m_d} - (1-\alpha)\tau < t < t_e^{m_d} + \alpha\tau \\ & \text{and } (m-1){N_W}^s < w < m{N_W}^s \quad \forall m_d, t_e \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $\gamma > 0$ is a random positive value, τ is a parameter that controls the size of the event region and α controls the position of the active region around the event position.

We examine two different ways of supplying the drum event positions $t_e^{m_d}$. We propose an unsupervised approach based on transient estimation and two scenarios with user-supplied annotations.

2.3.1. Transient Analysis

The transient analysis used to evaluate the constraint-based unsupervised method is the same one used in [12]. It is based on the work by [13] where the spectral peak center of gravitiy is used as a measure of transient quality. This measure is coupled with a band analysis and thresholding in order to extract a frame-level decision about the presence of a percussive event attack. This method for drum event estimation is quite straightforward and serves as a baseline for constraint-based blind drums separation methods. State of the art drum estimation techniques can acheive much better results, probably leading to improved separation.

2.3.2. Annotations

Two main scenarios for user-supplied annotations are considered. The first consists in creating different annotations sets for each of the drum sounds (bass drum, snare drum, closed hi-hat, open hi-hat,...). This implies having multiple drum sources $N_{Md_{ind}} > 1$ in our spectrum model. The second technique uses a single set of annotations, by merging all the drum sounds together $N_{Md_{join}} = 1$, in order to keep both approaches comparable, the number of basis components used in the second approach is ${N_W}^s_{join} = N_{Md_{ind}}{N_W}^s$. The annotations of the drum events were manually performed by an amateur experienced drum player using the SonicVisualiser software application [1]. The annotations were created using the isolated drum tracks in order to evaluate the near-optimal separation using a constraint-based method. The annotations dataset has been made publicly available online [2].

3. EXPERIMENTS

We used the same dataset of multitrack audio recordings with drums as in [12] to evaluate the proposed methods. A quantitative evaluation is done by using the perceptually motivated objective measures in the PEASS toolbox [14]: OPS (Overall Perceptual Score), TPS (Target-related Perceptual Score), IPS (Interference-related Perceptual Score), APS (Artifact-related Perceptual Score). For all the excerpts we have also computed the near-optimal time-frequency mask-based separation using the BSS Oracle ([15]) framework. The evaluation measures of the oracle versions of each excerpt were used as references to reduce dependance of the performance on the difficulty of each audio. Therefore the values shown are error values with respect to the near-optimal version.

[1] `http://www.sonicvisualizer.org`

[2] `http://mtg.upf.edu/download/datasets/dreanss`

We performed two series of experiments (regularization and constraints), which evaluate the performances of different methods.

The first set of tests consists of parameter explorations of the regularization-based methods (REG). In these experiments we tested the separation for multiple values of the time continuity regularization $\alpha_{tc} = 25$ (SM25), $\alpha_{tc} = 50$ (SM50), $\alpha_{tc} = 75$ (SM75), $\alpha_{tc} = 100$ (SM100) for the non-percussive accompaniment basis $\boldsymbol{W}_m$. We also evaluated the effect of employing a sparseness regularization $\alpha_{ts} = 10$ (SP10) on the drums gains. The regularizations for the frequency continuity of the non-percussive accompaniment has been kept to a fixed value $\alpha_{fc} = 1$. These tests were conducted in an unsupervised scenario (UNS) where all the drum basis components are learned during the separation and a semisupervised (SUP) case where the basis components are learned previously using training data with the drums in isolation.

In a second series of experiments we evaluated three constraint-based methods. We compared a blind transient analysis method (CON-TR) to two annotated methods: individual sources model (CON-AN-I), and joint sources model (CON-AN-J). We explored the influence of the main parameter ${N_W}^s$ on each method and the effect of using the SIMM lead voice model with an external annotated pitch (CON-TR-NP, CON-AN-I-NP, CON-AN-J-NP). Finally we performed a comparative evaluation with state of the art methods THPS-TIK (similar to [12]), HPSS [3] and FASST [8]. The best parameter combination resulting from the parameter exploration was used in the comparative tests.

4. DISCUSSION

4.1. Regularizations Experiments

In Figure 1 and Figure 2 we can observe the Overall Perceptual Score (OPS) error relative to Oracle, for the individual excerpts in the unsupervised and semi-supervised configurations. We can appreciate that in both scenarios the results are not conclusive, since the OPS error varies a lot with changes in the regularization parameters. For the unsupervised configuration, on average we observe an increase of the error with the amount of temporal continuity regularization applied to the accompaniment gains. The average result also shows that the application of the sparseness is detrimental since it increases the separation errors. The average results show very little variation for the semisupervised scenario.

However we do notice that for certain excerpts, such as for excerpt 0 in the unsupervised case, temporal continuity regularization causes a significant improvement. This improvement for individual excerpts is more visible still for the temporal sparseness regularization parameter of the drums gains $\boldsymbol{H}_d$.

In Figures 3 and 4 we plot a histogram of the improvements from adding sparseness regularization. This value is computed as the difference of OPS error obtained with the method using sparseness regularization and that obtained without using it. These values are computed for all the values of the temporal continuity regularization. The histograms show a large variance in improvement. In some cases the use of $\alpha_{ts} = 10$ creates a large improvement and in others the opposite.

These results suggest the utility of future investigation of the dependency of optimal regularization parameters on the data, and the potential for deriving methods to estimate optimal regularization for each excerpt to be analyzed.

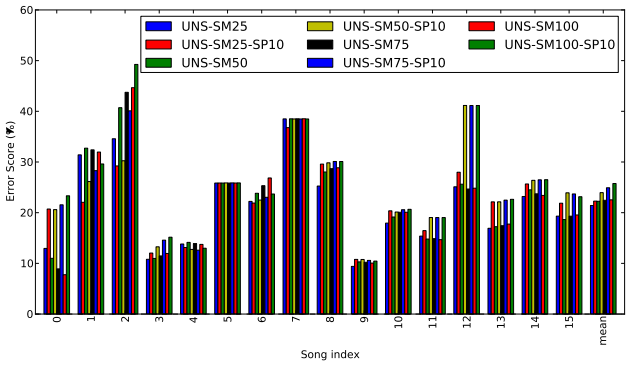

Figure 1: Individual OPS error (%) measures for the drums separation unsupervised scenario with relation to the regularizations applied.

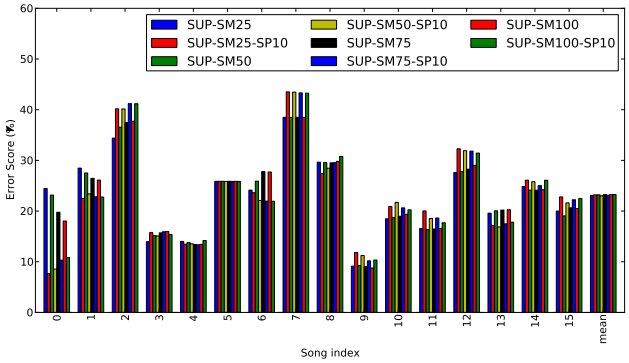

Figure 2: Individual OPS error (%) measures for the drums separation semisupervised scenario with relation to the regularizations applied.

Informal listening to the results confirms the findings that we show here. In some excerpts the regularization improves the separation while in others it is disadvantageous. We can also appreciate that the regularizations behave as expected, controlling the desired spectro-temporal qualities of the estimated sources. In general we also observe that semisupervised separation maintains the bass drum and snare sources better. Unsupervised separation tends to produce a filtered signal keeping only mid-high components. A drawback of the supervised version is the greater interference between lead vocals and the bass line.

4.2. Constraints Experiments

Figures 5 and 6 show the ${N_W}^s$ parameter exploration experiment for the constraint-based method that uses annotations of the individual drums sources (CON-AN-I). This is the method with the most prior information supplied about the mixture and serves as a maximum for our proposed contraint-based methods. The plot of the OPS and APS score errors shows that the results vary slightly depending on the number of basis components assigned to each drum source ${N_W}^s$. There are several local minima implying that there is no unique optimal value for all excerpts and drum sources.

In terms of TPS and IPS the number of basis components ${N_W}^s$ controls the tradeoff between target fidelity and interference. This

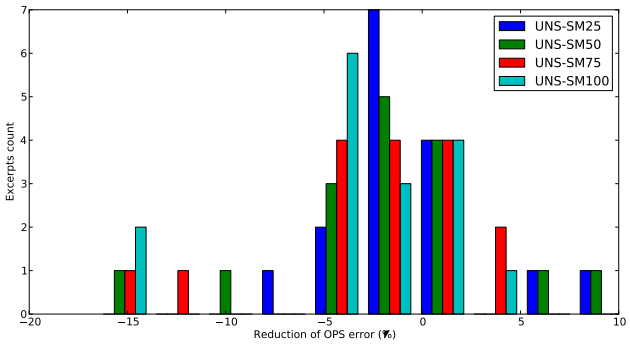

Figure 3: Histogram of the OPS improvement (%) by using the sparseness regularization (SP10) in the unsupervised scenario.

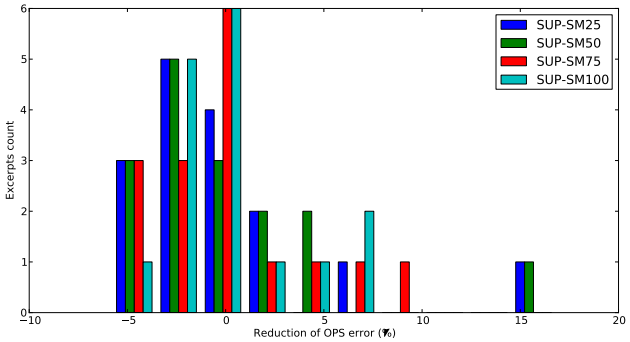

Figure 4: Histogram of the OPS improvement by using the sparseness regularization (SP10) in the supervised scenario.

is an expected result, since a large number of basis components to reconstruct drum components could lead to overfitting of the mixture spectra and therefore capturing other non-percussive components and increasing the interference while at the same time better reconstructing the target drums.

Figures 7 and 8 show a similar trend when the constraints are based on generic drums annotations $N_{Md_{join}}$, without making a distinction between drum sounds (CON-AN-J). In future work we should investigate optimizing the parameter for each drum type and its dependence on the number of occurrences in the excerpt.

In Figure 9 we show the effect of implementing such constraint-based methods as extensions of the SIMM approach, in contrast to not performing the lead voice estimation (NP). These results show a reduction of the OPS error (%) in all the constraint-based methods. This improvement is mainly due to a decrease in interference and informal listening to the results confirms this finding. The lead voice is often an energetic component and by specifically modelling it we significantly reduce the parts of it that are counted as drum sounds.

Figure 10 shows how these constraint-based methods relate to other state of the art drums separation approaches. The annotation-based informed source separation methods show a clear improvement in OPS over the blind techniques. This shows that the development of proper temporal estimation of the drum event positions could lead to significant improvements in blind drums separation. The difference between annotations of individual drum sources (CON-AN-I) and generic drum sources (CON-AN-J) is in-

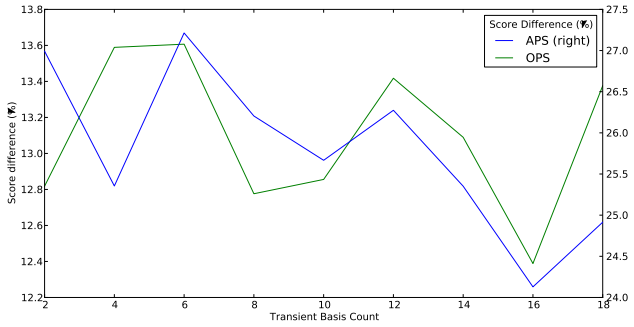

Figure 5: OPS and APS score errors (%) with relation to $N_W{}^s$ for the constraint-based individual annotation method (CON-AN-I).

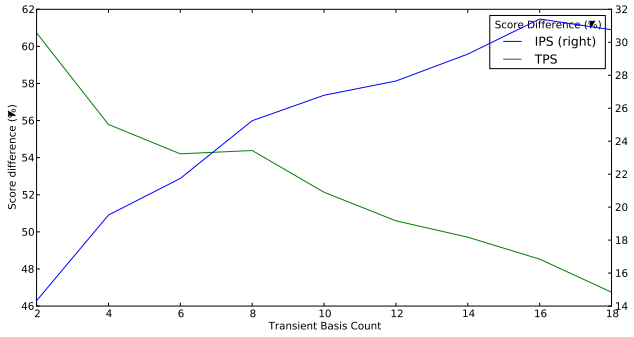

Figure 6: TPS and IPS score errors (%) with relation to $N_W{}^s$ for the constraint-based individual annotation method (CON-AN-I).

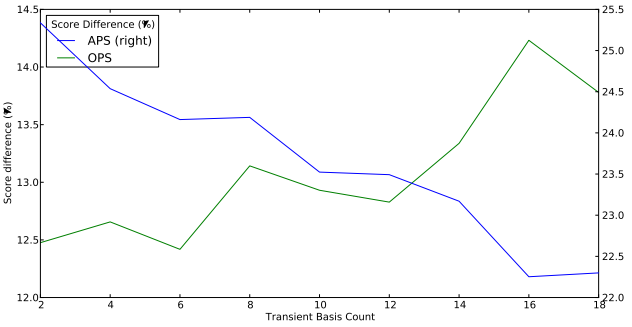

Figure 7: OPS and APS score errors (%) with relation to $N_W{}^s$ for the constraint-based joint annotation method (CON-AN-J).

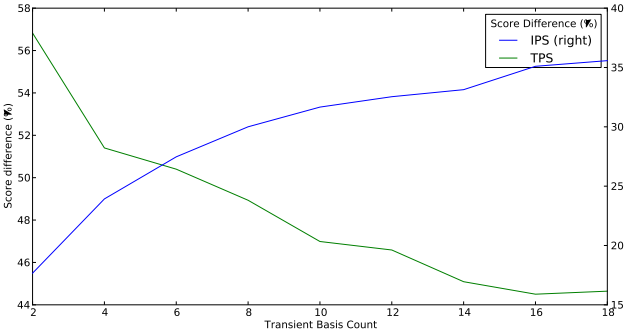

Figure 8: TPS and IPS score errors with relation to $N_W{}^s$ for the constraint-based joint annotation method (CON-AN-J).

significant, from which we can conclude that estimation of general drums events should be sufficient.

The artifact-related scores (APS) show unexpected results where the FASST method achieves better averages (negative score difference) than the Oracle version. This is probably due to the perceptual-inspired relations in the PEASS framework, since the non-perceptual related BSSEval results in Figure 11 do not present this behavior.

Finally we observe that the blind transient constraint-based method (CON-TR-J) does not achieve results comparable to other blind techniques. The transient detection method is not adapted to drums and thus is prone to false positives caused by other sources.

Subjective assessment by informal listening to the comparative study confirms the trend presented in Figure 10. The main shortcoming of the constraint-based methods is that the full decay of the drums is often not preserved. Increasing the parameter τ could help reduce this issue, however it would also increase the amount of noise in the learning process of the drums component basis during the factorization. In the future studying the relations between τ and $N_{W\,d}$ might be useful since together they influence the amount of overfitting and underfitting of the problem.

5. CONCLUSIONS

We proposed and studied an extension to the SIMM method to perform drums separation. The proposed extension makes use of regularizations and constraints to drive the factorization towards the separation between percussive and non-percussive music accompaniment. We proposed two new regularization terms that consist in small variations of the ones proposed by [11]. The proposed regularizations control the frequency smoothness of the basis components and the temporal sparseness of the gains. These regularizations were used together with the temporal continuity regularization of the gains to perform blind drums separation. We also studied the effect of using a set of pre-trained basis components for drums sources. The experiments showed that there was no optimal value for the strength of the regularizations and that these were highly dependent on the excerpt.

We evaluated the use of temporal constraints on the gains to perform drums separation. The technique consists of using the positions of the drums events in the mixture to limit the regions of activation of the drums basis. This technique was tested using both ground truth manual annotations from the isolated tracks and automatically extracted transients from the mixture. This allowed us to assess both a glass ceiling and a baseline for this approach. Results show however that a simple transient estimation technique is insufficient for this task, compared to the method with manual annotations or other state of the art methods. Additionally we tested how the number of basis components assigned to each drum source affects the quality of the separation. The results showed that the overall performance and the artifacts related score did not vary much with respect to this parameter. This parameter controlled the tradeoff between interference and target related scores.

We also observed that it may not be of much benefit to estimate the positions of the individual drum sounds (closed hi-hat, open hi-hat, snare drum,...) since this does not significantly improve the separation results. However it remains to be tested whether using

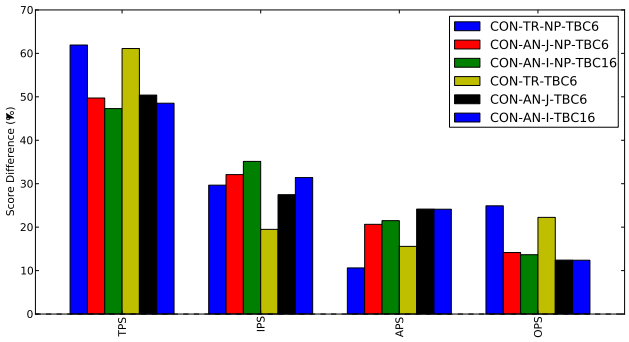

Figure 9: Effect of the lead voice estimation on the constraint-based methods, using ${N_W}^s = 6$.

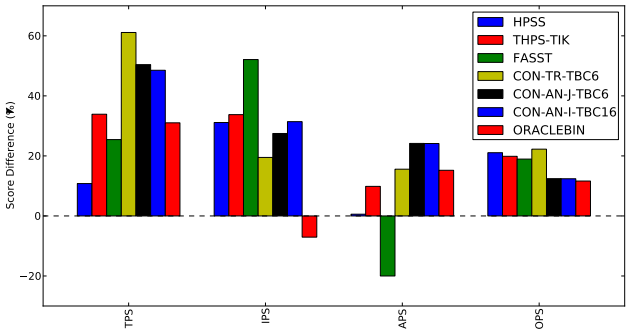

Figure 10: PEASS results of the comparative study of the constraint-based methods for drums separation.

different parameter values per type of drum sound enhances the results. Furthermore the use of frequency domain constraints specific to each drum type could also improve the separation. Another possible future direction could be to perform a two step strategy, where a subset of the drum positions are first used to estimate the basis components, and a second step in which the separation is done loosening the temporal constraints.

6. REFERENCES

[1] A. Zils, F. Pachet, O. Delerue, and F. Gouyon, "Automatic extraction of drum tracks from polyphonic music signals," in *Web Delivering of Music. WEDELMUSIC. Proc. Second Int. Conf. on*, 2002, pp. 179–183.

[2] D. Barry, "Drum source separation using percussive feature detection and spectral modulation," *IEEE Irish Signals and Systems Conf.*, pp. 13–17(4), 2005.

[3] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. EUSIPCO*, 2008.

[4] O. Gillet and G. Richard, "Extraction and remixing of drum tracks from polyphonic music signals," in *Applications of Signal Processing to Audio and Acoustics. IEEE Workshop on*, 2005, pp. 315–318.

[5] K. Yoshii, M. Goto, and H.G. Okuno, "INTER:D: a drum sound equalizer for controlling volume and timbre of drums," in *Integration of Knowledge, Semantics and Digital Media Technology, EWIMT. The 2nd European Workshop on the*, 2005, pp. 205–212.

[6] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proc. EUSIPCO*, 2005.

[7] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Acoustics Speech and Signal Processing (ICASSP), IEEE Int. Conf. on*, 2010, pp. 1942–1945.

[8] A. Ozerov, E. Vincent, and F. Bimbot, "A general modular framework for audio source separation," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Saint-Malo, France, Sept. 2010.

[9] S. Ewert and M. Müller, "Score-Informed Voice Separation For Piano Recordings.," in *ISMIR*, Anssi Klapuri and Colby Leider, Eds. 2011, pp. 245–250, University of Miami.

[10] J-L Durrieu, G. Richard, B. David, and C. Févotte, "Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 18, no. 3, pp. 564–575, Mar. 2010.

[11] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[12] J. Janer, R. Marxer, and K. Arimoto, "Combining a harmonic-based NMF decomposition with transient analysis for instantaneous percussion separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 281–284.

[13] A. Röbel, "Transient detection and preservation in the phase vocoder," in *Proc. Int. Computer Music Conf. (ICMC)*, 2003, pp. 247–250.

[14] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and Objective Quality Assessment of Audio Source Separation.," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[15] E. Vincent, R. Gribonval, and M.D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.

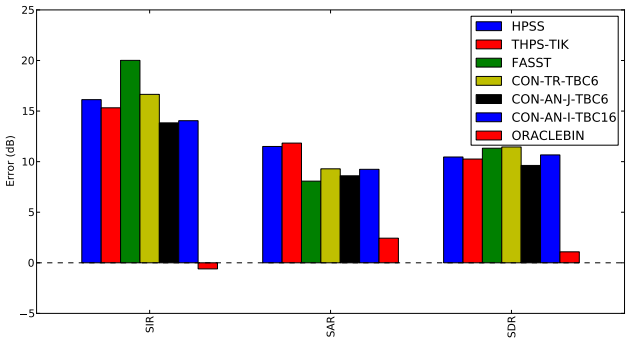

Figure 11: BSSEval results of the comparative study of the constraint-based methods for drums separation.