Response to the reviewers' comments

Clément Laroche, Matthieu Kowalski, Hélène Papadopoulos, Gaël Richard

This document presents the responses to the comments that have been made by the reviewers during a first review of this article, originally entitled "Structured Projective Non Negative Matrix Factorization with drum dictionaries for Harmonic/Percussive Source Separation". This paper has been accepted after major changes which has led to modify its content regarding several aspects. In particular, we reorganised the introduction and the section II. We also provide more details on some technical contributions (notably the MM algorithm). Finally, the minor comments (regarding some references and notations) have been taken into account.
Reviewer 1' comments are written in green and presented in Section 1. Reviewer 2' comments are written in blue and presented in Section 2. Reviewer 3' comments are written in red and presented in Section 3. At the end of this document is enclosed the original paper submission, in order to make the references to the different sections clearer.
We would like to thank the reviewers for their insightful remarks, that resulted in the present version of the paper, that we hope largely enhanced.

1 Reviewer 1 (decision: Review Again After Major Changes)

The paper is generally well written and motivated. The results presented do represent an improvement over the state of the art, but there are a number of things that I would like to see addressed in the paper before it is published. Firstly, it needs a very thorough proof reading, there are a relatively large numbers of typos in the paper which need to be sorted.
Secondly, the figures are quite difficult to read, particularly with regards to the line representing the median, this is missing in a number of places, and it is very hard to see due to the choice of colours. This needs to be amended.
Most of the figures have been updated in order to improve the readability.
Thirdly, I have a couple of concerns/questions regarding the testing. The authors learned an NMF dictionary from the "ENST mixture" files for each drummer, but then used the spectrograms from the separate drum audio files to create the STFT dictionary. Why was the NMF dictionary not learned from the separate drum audio files? Secondly, the NMF dictionary size was fixed on an ad-hoc basis, I would like to see how changing the size of the dictionary changes the results obtained.
The NMF dictionary have been learned of the "ENST mixture files". However, these files only contain drum sounds with some post processing (reverb, compression). In order to make it clearer, we insisted on this point in the article. Secondly, we reworked extensively this part of the journal paper. It is clear that the dictionary learning is a crucial part of our algorithm. We decided to make an in-depth analysis on how the size of the dictionary changes the outcome of our method. As the dictionary is fixed, it is important to find a good compromise with the size and the quantity of information contained. Two types of audio recordings are available, the first type contains drum hits of the independents element of the drum kit the other type of audio files are drum phrases. We created three signals of different length (6 min, 12 min, 30 min) for the three drummers for a total of 9 audio signal. Finally, we compute the STFT of these 12 signals and we execute a NMF on each of the spectrograms to obtain various dictionaries. The rank of the decomposition is chosen as $6 \leq k \leq 500$. In total, 108 dictionaries are created.
This allowed us to provide a much more in-depth discussion on how to create an appropriate drum dictionary.
Further, the choice of the factorization rank is a source of confusion for me, a rank of 100 is chosen, but inspection of figure 3 clearly shows that a rank of 150 is a much better choice for both KL and Euc. I would like to see the sisec tests repeated with this rank, as it may make a considerable difference in the performance of KL and EUC.
This is a valid remark and we repeated the experiment with a rank of 150 but there was no noticeable improvement for EUC and KL.
On another note, why use 50% overlap, it is generally the case that 75% overlap usually increases the performance of most separation methods?

After looking at the literature, it seems like there is no consensus on using 75% over 50% overlap. We repeated some of our experiments with 75% overlap but we did not found any noticeable improvement.
Finally, in section V-C, the authors state "Our explanation is that all methods rely on Wiener filtering for phase reconstruction (see Equation (8)). As the percussive instruments have flat spectra, the percussive mask is a non sparse matrix and small estimation errors drastically decrease the results of the percussive instruments. This tendency is not visible on small scale tests (see Figure 6)." I am not entirely sure what point the authors are trying to make here and would like the statement reworded and clarified.
After discussion, we found that this statement was not relevant and we decided to remove it from the article.

2 Reviewer 2 (decision: Review Again After Major Changes)

The authors present a method for separating percussive from harmonic instruments. The method is based on a combination of standard pretrained NMF for the percussive elements, and a variant of NMF (Projective NMF published in [27]) for the harmonic parts. The idea behind projective NMF is (similar to incoherent dictionary learning in sparse coding) to constrain the learnt templates in the NMF dictionary to be orthogonal to each other. Since broadband percussive templates are typically "less" orthogonal to each other (due to the spectral overlap) than harmonic templates for different pitches, the projective NMF tends to represent the harmonic parts. The paper is an extended version of paper [15], i.e. the method itself is already published. Therefore, the main contributions are: 1.) additional experiments showing the influence of some parameters / design choices 2.) experiments using an increased data set 3.) Update rules for two more divergence/distance measures The other two contributions mentioned in the introduction, i.e. the introduction of a method and the comparison with state of the art methods, were already part of the original paper.
The journal paper is the continuity of the paper [15]. For real audio signals, the original sources are used to sort the components of the decomposition. The method presented in [15] is not capable to perform a blind HPSS.
The contribution of this article is that we improve the method [15] into a blind method that does not require the information of the original sources to perform a HPSS.
More precisely, similar to other methods (e.g. [11]), the method employs an NMF split into two parts: one for the percussive and one for the harmonic part. Instead of using regularizers encouraging certain behaviour, the percussive part uses a pretrained dictionary. For the harmonic part, again instead of using regularizers leading to specific behaviour, the method relies on an implicit tendency of the PNMF method to produce fewer templates with broadband properties. Therefore, the main novelty could be summarized with replacing one part of a method (regularized NMF) with another previously published one (projective NMF). Overall, this is a nice idea and worth trying out. It is just somewhat limited. The argument also extends to the scientific depth: The main argument being made in the paper is that projective NMF is better for the harmonic part than regularized NMF (or diffusion NMF or the median filter approach as used in the other approaches). This is only tested in terms of experiments involving entire systems, which probably contain various different design choices and hence the argument that one module might be better than another one is mostly hidden by various other parameters.
The goal of HPSS is to model the harmonic and percussive components in two different layers in order to separate them. We based our method on the observation that harmonic instruments are tonal sounds very localized in frequency that can be modelled by a sum of sinusoidal waves and percussive instruments are wide band transient signals.
The HPSS algorithm usually decompose the spectrogram as

$$V \approx V = V_H + V_P.$$

The article on constrained NMF [13] uses constraints of spectral smoothness and temporal sparseness to extract the percussive part and spectral sparseness and temporal smoothness to extract the harmonic part. We decided to use a dictionary to constrain the percussive part and a PNMF decomposition in order to obtain a similar decomposition. The philosophy of our work was to buy a robust method that could be easy to tune (i.e., few hyper parameters).
The structure of the existing methods, however, would allow for replacing parts of in single system and comparing then systematically what choices for different components make sense (after carefully setting the overall parameters for the resulting variant); i.e. replacing the proposed PNMF with the existing regualarized NMF approach (or Ono's diffusion NMF or Fitzgeralds median approach), to

see how the proposed system really behaves. Comparing methods this way would, resulting from the much more direct comparability, drastically increase the scientific depth of the entire study and help support any more general conclusions.
This is an interesting question and we decided to add this part to the article:
"In order to highlight the advantages of using the PNMF for the harmonic part, we also designed a regularized method. Let V be the magnitude spectrogram of the input data, the model is given by

$$V \approx \tilde{V} = W_H H_H + W_P H_P. \quad (1)$$

The optimization problem is then:

$$\min_{W_H, W_P, H_P} D(V|\tilde{V}) + k_{TSM} TSM + k_{SSP} SSP. \quad (2)$$

The constraints TSM and SSP are from the constraints of temporal smoothness and spectral sparseness from the article of constrained NMF [13]. The hyperparameters k_{TSM} and k_{SSP} control the amount of temporal smoothness and spectral sparseness. This model replace the PNMF components by a regularized NMF (RegNMF) to extract the harmonic part. This requires prior tuning of the variables k_{SSP} and k_{SSP}. "
Also, there are various hard to justify or inprecise statements that further weaken the scientific depth. For example: P2: "However, on a small scale test conducted by [11], this simple approach does not give the best separation results." What does 'best' mean? Such statements should be rephrased.
We changed this sentence for "However, this method does not perform well on non stationary signal such as tremolos or vibratos as these get divided between the harmonic and percussive part."
1.) There are various claims throughout the paper, which are only weakly or not backed up by previous research. To give an example: P2: "This method gives good results compared to other state-of-the-art methods but the hyperparameters tuning is a tedious process and results depend heavily on the training database."
This statement was clarified " This method gives good results compared to other state-of-the-art methods but this article will show that it is not robust to the wide variety of audio signal as the results depend heavily on how the parameters are optimized during the training."
P7: "As the dictionary is fixed, it is important to have a large dictionary to be able to extract a large type of percussive instruments." (This is not a logical conclusion as the large dictionary could have side effects)
We clarified this sentence as "As the dictionary is fixed, it is important to find a good compromise with the size and the quantity of information contained."
2.) Various statements not clear enough. For example: P2: "Numerous other methods exist but in this work, we will focus on three methods that are particularly relevant to our benchmark." : Why are these more relevant than the others? Relevant with respect to what?
We decided to compare our method to the MF method [9] as it is the most used HPSS method for preprocessing.
The constrained NMF [13] is a good benchmark for our method as it represent a state of the art constrained HPSS method in the NMF framework.
Finally, the NMPCF [27] is a recent method that uses drum dictionaries in the NMF framework.
These three methods are recent state of the art method that will provide a good benchmark to our method. Also it provides a comparison between supervised/unsupervised and constrained/unconstrained methods.
3.) Many statements are repeated several times leading to a high redundancy in the paper. Examples: P8: Most of the entire section V.E is redundant P1: Top of second column. First few sentences.

Fixed
4.) Various typos and grammar issues. Examples: P1: "As the data matrix V is usually redundant, the product WH can be thought as a" P1: "The downside of this approach is that it requires a well organized prior information that..." P4: "(Algorithme 1)"
Fixed
* P1: it is not clear why the approaches in paragraph are referred to as parametric, physical models. In which way are they more 'physical' than others?
- P1: "These inherent properties are particularly interesting for audio source separation as shown in [11], [14]." It is not clear why these references were chosen.
Fixed
- P2: "However, as shown in [15] this direct approach obtains limited performances in a blind source separation" : In what way is this approach more direct?

The method in [15] does not use a drum dictionary but it cannot perform a blind source separation.
- P2 Col2: First paragraph. If I remember correctly, none of these methods actually employs HPSS methods, and therefore it remains unclear how these citations actually back up the claim that HPSS is useful in this context.
The citation have been changed and the paragraph have been modified as follow
"HPSS has numerous applications as a preprocessing step. For example most multi-pitch estimation models [3], instruments recognition [4] and melody extraction [5], [6] algorithms are much more efficient on harmonic instruments alone. Indeed, these algorithms often rely on the analysis of the harmonic structures that are blurred by the percussive instruments. Similarly, drum transcription algorithms [7], [8] are more accurate if the harmonic instruments are not part of the signal. Finally, using the Median Filtering (MF) algorithm [9] as a preprocessing step increases the performance for singing pitch extraction and voice separation [10]."
- P2: "A well known unsupervised method to extract the harmonic/ percussive components consists in applying a median filtering on the spectrogram of the audio signal [23], [25]." This is incorrect: [25] is not based on median filtering but on a diffusion process.
Fixed
- P3: "Contrary to the NMF, the sparsity of the PNMF is an inherent features of the decomposition." Originally, NMF was mentioned as a highly sparse method due to its non-negativity constrained (when compared to PCA factorization). Sparsity is relative not absolute.
We modified this part and we removed this statement.
- There should be more information, where the implementations for the state of the art methods were obtained from (or whether they were implemented by the authors). Also, how the method were parametrized, and any other relevant details.
Constrained NMF and NMPCF are re-implemented in this paper and we used the optimal parameters recommended by the author in there respective articles. The MF implementation is taken from [9] and we used the standard parameters for a HPSS task.
- P9: Last paragraph in V.E should be mentioned earlier.
Fixed

3 Reviewer 3 (decision: Review Again After Major Changes)

This paper describes an NMF-based method that separates a music signal into a harmonic part and a percussive part. The harmonic spectrogram is approximated as a low-rank matrix using projective NMF that tend to learn sparse basis spectra. Similarly, the percussive spectrogram is modeled using standard NMF whose basis spectra are learned from a drum database in advance. The experimental results show that the proposed method outperformed the state-o-the-art methods.
The key idea of the proposed method (combination of two types of NMF) is interesting and justified well. However, the paper is not well organized and no examples are provided, so it is hard to understand some parts of the paper. The appendix is not necessary (the content of the appendix should be placed in the many body). It is necessary to reorganize the whole paper and add many figures and sample outputs.
We removed all the figures from the appendix and placed them in the main body. The equation for the algorithm are still in the appendix.
This part is very redundant as an introduction of the paper on HPSS. Please extract important information related to the proposed method and move other information into Section 2 (Related Work). You should start with introduction to HPSS, not with introduction to NMF. For example, a basic flow of this section is like:
1) Explain why HPSS is important.
2) Introduce some relevant conventional methods of HPSS and the problems of those methods.
3) Explain why combination of SNMF and standard NMF can solve those problems.
Thank you this great remark. We put a lot of time reorganising the introduction and the 2 section and we organised this 1st section as you advised. We also tried to make this part less redundant in order to make the paper easier to read.
You do not need to introduce ONMF in Section 1 because it is much less relevant to the proposed method than PNMF. Note that in Section 3-A, introducing ONMF for comparison is a good idea to understand the behavior of PNMF.
We moved the section 3-A to the Section 2 with the Related work on NMF.
Please explain four constraints proposed by Canadas et al.

We presented 2 of the four constraints, the Temporal Smoothness (TSM) is defined as follow

$$TSM = \frac{F}{R_h} \sum_{r_h=1}^{R_h} \frac{1}{\sigma^2_{H_{H_{r_h}}}} \sum_{t=2}^{T} (H_{H_{r_h,t-1}} - H_{H_{r_h,t}})^2 \tag{3}$$

The value $\sigma_{H_H} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} H^2_{H_{r_h,t}}}$ is a normalization term to make the global objective function independent to the norm of the signal. Spectral Sparseness (SSP) is designed using the same constrain as in [23]:

$$SSP = \frac{T}{R_h} \sum_{r_h=1}^{R_h} \sum_{f=1}^{F} | \frac{W_{H_{f,r_h}}}{\sigma_{W_{H_{r_h}}}} | \tag{4}$$

The other two constraints are identical but they are applied to the in reverse order in order to favour spectral smoothness and temporal sparseness.
The title should be "Related Work and be divided into two subsections on NMF and HPSS.
Done.
A method based on median filtering [23] does need tuning of horizontal and vertical filter sizes.
This subsection should be moved into Section 2 (merged into subsection on NMF).
Done
Please provide some sample outputs (spectrograms and variables obtained by PNMF and ONMF.
Please explain more carefully why PNMF can find a sparse solution, using figures.
Section 3-C: The title should be Optimization Algorithm."
Done
It must be easy to derive a unified multiplicative update algorithm including those for EU-, KL-, and IS-SPNMF as its special cases because the EU distance and the KL and IS divergences are special cases of the beta divergence. See some papers using beta-divergence NMF.
Section 3-D: Each drummer uses Each drummer used"
Done
I cannot understand why there are tree long drum audio signals, the rank is set to k=12, and each element is represented as two basis functions.
Please enumerate here the tree" methods of dictionary learning. The third method should not be introduced in Section 4-E.
We extensively modified this part of the article and these statement are not in the paper any more.
I cannot understand the second method at all.
"another types" -¿ "other types" Done
Section 3-E: Please clarify the value of alpha.
We fixed the value $\alpha = 2$
, where -¿ where" Done
$\alpha \in (1, 2)$ -¿ $1 \leq \alpha \leq 2$? Done
Sections 4 and 5: Please clearly describe the average values of SDR, SIR, and SAR in each experiment.
All the figures should be placed on the top. Done
Please provide a sufficient number of examples (spectrograms and decomposition results). It is very hard to understand the discussions in the current form.
Please do not paint the boxplots for readability. For example, use the standard style of of matlabs boxplot.
All the old and new boxplot have been updated.
Figures 4 and 5 should be put in other in the same page and each figure should be updated for readability, i.e., the horizontal axis indicates Harmonic, Percussive, or Mean," each of which has three bars corresponding to Euc, KL," and IS."
Done
Similarly, Figures 6, 11, 12 should be put in order in the same page and each figure should be updated for readability, i.e., the horizontal axis indicates Harmonic, Percussive, or Mean," each of which has three bars corresponding to STFT," NMF," and "Concatenated."
The figure have been update and is no longer part of the article.
Similarly, Figures 10, 13, 14 should be put in order in the same page and each figure should be updated for readability, i.e., the horizontal axis indicates Harmonic, Percussive, or Mean," each of which has four bars corresponding to SPNMF, HPSS, NMPCF," CoNMF."
The 3 figure have been condensed into one big figure.
Section 4-A: The term straight is not appropriate. We remove the inappropriate word.
Algorithme -¿ Algorithm" Done

The symbols k, k, and e are not defined in the many body. We rephrased that part to clarify "for the SPNMF the rank of the harmonic part is $k_H = 2$ and the rank of the percussive part is $k_P = 2$."
Figure 1 has no color bar (not compatible with Figure 2). We added the colorbar.
Section 4-E, we test a third dictionary is made by -¿ we test a third dictionary made by?
Done
Section 5-B: The name HPSS cannot be recommended for referring to [23] because it has a very general meaning.
In the paper, we now refer to the method [23] by Median Filtering (MF).
It is necessary to compare the proposed method with a conventional version without dictionary training [15].
Section 5-E, "compare to -¿ , compared to?
Done