Adapted NMFD update procedure for removing double hits in drum mixture decompositions

Len Vande Veire[1], Cedric De Boom, and Tijl De Bie

Ghent University, Belgium
Contact: len.vandeveire@ugent.be

Abstract. Non-negative matrix factor deconvolution (NMFD) can be used to decompose a drum solo recording into K time-varying spectral templates (the constituent sounds) with corresponding activation functions. Unfortunately, choosing the template length, an important hyperparameter, is hard: it must be long enough to capture drum hits with a long decay, but when chosen too large, the algorithm often captures multiple drum hits within the same template. We propose to detect the emergence of such 'double hits' during optimization, and to replace them with an exponentially decaying extrapolation of the preceding template frames. Experiments demonstrate the effectiveness of this approach.

Keywords: Non-negative matrix factor deconvolution · Automated drum transcription

1 Introduction

The non-negative matrix factor deconvolution (NMFD) algorithm [8] decomposes a spectrogram matrix $X \in \mathbb{R}_{\geq 0}^{N \times T}$ with N frequency bins and T time frames into a dictionary of K time-varying spectral *templates* $W^{(k)} \in \mathbb{R}_{\geq 0}^{N \times L_\tau}$, and an *activation matrix* $H \in \mathbb{R}_{\geq 0}^{K \times T}$. The spectrogram is modeled as the convolution of the templates with the activation matrix:

$$X_{n,t} \approx \hat{X}_{n,t} = \sum_{k=1}^{K} \sum_{\tau=1}^{L_\tau} W_{n,\tau}^{(k)} H_{k,t-\tau} \tag{1}$$

where $H_{k,t-\tau}$ is zero when $t < \tau$. $W^{(k)}$ and H are updated iteratively using multiplicative updates in order to minimize a divergence measure $\mathcal{L}(X, \hat{X})$. In this paper, we use the KL divergence, $\mathcal{L}_{KL}$, and the corresponding update rules for $W^{(k)}$ and H [7]:

$$\mathcal{L}_{KL}(X, \hat{X}) = \sum_{n,t} X_{n,t} \log \frac{X_{n,t}}{\hat{X}_{n,t}} - X_{n,t} + \hat{X}_{n,t}, \tag{2}$$

$$W_{n,\tau}^{(k)} \leftarrow W_{n,\tau}^{(k)} \frac{\sum_t H_{k,t-\tau} \left(X_{n,t} / \hat{X}_{n,t} \right)}{\sum_t H_{k,t-\tau}}, \tag{3}$$

$$H_{k,t} \leftarrow H_{k,t} \frac{\sum_\tau \sum_n W_{n,\tau}^{(k)} \left(X_{n,t+\tau} / \hat{X}_{n,t+\tau} \right)}{\sum_\tau \sum_n W_{n,\tau}^{(k)}}. \tag{4}$$

The templates $W^{(k)}$ can be interpreted as short spectrograms of length L_τ that model the constituent sounds of the mixture. Ideally, each $W^{(k)}$ would capture an individual drum hit of a particular instrument, e.g. $W^{(0)}$ captures a single kick drum hit, $W^{(1)}$ captures a single snare drum hit and so on. The corresponding activations H_k then describe where in the mixture these sounds occur. NMFD has already been applied successfully for automated drum transcription and drum separation tasks [1,2,4,5,6,9]. These works only consider constrained settings, though, e.g. only optimizing for H and keeping the dictionary W fixed. We note the absence in literature of a successful application of NMFD where both W and H are optimized jointly.

The template length L_τ is an important hyper-parameter in NMFD. Percussive mixtures often contain some instrument(s) with a long decay, e.g. a kick drum; therefore, L_τ needs to be large enough to adequately capture a single drum hit of these instruments. However, percussive mixtures also often contain hits that follow each other in rapid succession, e.g. the hi-hats. In this case, NMFD often captures multiple drum hits within one template, as has been noted before in the context of drum mixture decomposition using NMFD [5]. This is problematic: the discovered templates then no longer contain single drum hits, or they can even contain drum hits of multiple instruments, so that the resulting activations no longer reflect the onsets of the individual instruments, making the decomposition less interpretable and useful. Figure 1(b) illustrates this problem.

2 Detecting emerging double hits during optimization

We propose to solve the 'double-hit' problem by checking after each update of $W^{(k)}$ whether a second onset can be detected in the template. If this is the case, then $W^{(k)}$ is modified by overwriting this second onset with an exponentially decaying extension of the preceding template frames. This will initially lead to a worse approximation of the spectrogram, as important information for the decomposition was removed. However, the expected effect of this modification is that, in the next update of the activations H, some activation value(s) will increase to compensate for the removal of the secondary onset in the template; eventually, after a few updates, each $W^{(k)}$ will ideally only contain a single drum hit, and all onsets will be captured in H_k.
The adapted update procedure for $W^{(k)}$ is as follows:

1. Calculate the updated version of $W^{(k)}$, as in Eqn. (3).
2. Calculate the log-envelope $a^{(k)}[\tau]$ of each updated template $W^{(k)}$:

$$\tilde{a}^{(k)}[\tau] = \sum_n \log\left(W^{(k)}_{n,\tau} + \epsilon\right), \tag{5}$$

$$a^{(k)}[\tau] = \tilde{a}^{(k)}[\tau] - \min_\tau\left(\tilde{a}^{(k)}[\tau]\right). \tag{6}$$

3. Calculate $\Delta a^{(k)}[\tau] = a^{(k)}[\tau + \tau_u] - a^{(k)}[\tau]$. When $\Delta a^{(k)}[\tau]$ is large for some τ, then there is an onset at time τ in the template.

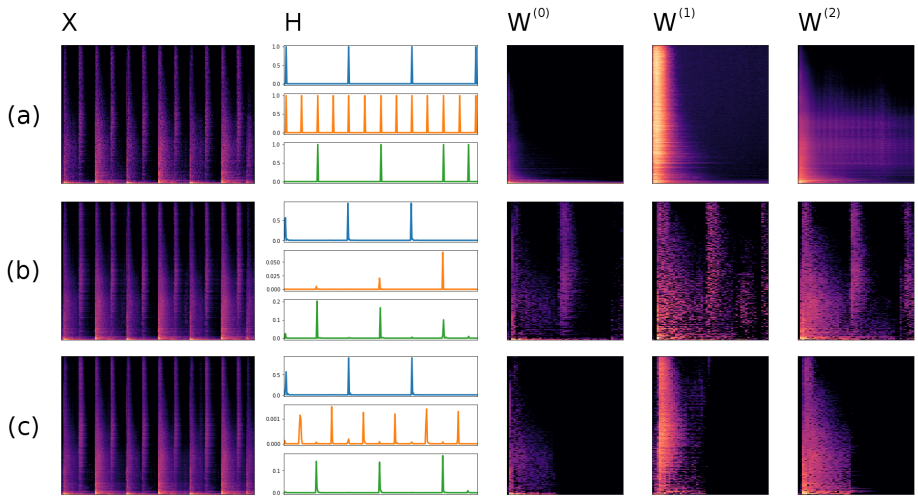

Fig. 1. Illlustration of the decomposition of a short drum loop: (a) ground-truth decomposition; (b) decomposition with NMFD; (c) decomposition with the modified NMFD algorithm. Columns: X, the spectrogram; H, the activations; $W^{(0)}$, the first template, capturing the kick drum; $W^{(1)}$, capturing the hi-hats; $W^{(2)}$, capturing the snare drum.

4. Set $a^{(k)}_{\max} = \max\left(a^{(k)}[\tau]\right)$. Detect onsets in $W^{(k)}$ by determining whether there is an onset larger than some threshold θ_{thr}, $\Delta a^{(k)}[\tau] \geq (\theta_{\text{thr}}\, a^{(k)}_{\max})$, for some $\tau \geq \tau_{\text{thr}}$. Only peaks that lie past the shift threshold τ_{thr} are considered, in order to not erroneously correct the first (and correct) hit in the template.
5. If there is a second onset in the template at $\tau_{\text{err}} \geq \tau_{\text{thr}}$, then all the frames after this onset are replaced by an exponentially decaying extension of the template frames preceding it:

$$W^{(k)}_{n,\tau} \leftarrow W^{(k)}_{n,\tau_{err}-\tau_u} \exp\left(-\gamma(\tau - \tau_{\text{err}})\right), \tau = \tau_{\text{err}} \ldots L_\tau. \tag{7}$$

In our experiments, we use the following settings for the hyper-parameters of this procedure: $\tau_u = 3$, $\theta_{\text{thr}} = 0.05$, $\tau_{\text{thr}} = 10$, $\gamma = 1$, $L_\tau = 50$, $\epsilon = 10^{-18}$, which were empirically found to lead to good results. The STFT spectrogram is calculated with a hop size of 512, and the audio sampling rate is 44.1 kHz.

3 Case study: decomposing a drum loop

As an example, we consider the drum loop in Figure 1(a)[1]. It contains three instruments: a kick drum, a snare drum and a hi-hat. The kick drum decays over approximately 50 frames; hence, we set $L_\tau = 50$. We note, however, that the hi-hats occur in rapid succession, i.e. approximately every 25 frames.

[1] This drum loop is a 4 second extract of a solo drum recording from the ENST dataset [3], *"062_phrase_rock_simple_medium_sticks.wav"*.

When decomposed with the original NMFD algorithm, shown in Figure 1(b), the templates $W^{(k)}$ capture not the individual drum hits, but rather repeating *sub-sequences* of drum hits. The activations consequently are very sparse and are not informative to determine the onset locations of the individual instruments.

When decomposed with NMFD using the proposed modifications, the templates each capture only a single drum hit, as shown in Figure 1(c). Note that the extracted templates very much resemble their ground-truth counterpart, see Figure 1(a). The activations also match the ground-truth onsets quite well; for the hi-hat, i.e. the second component, there is some discrepancy, as only every other onset is clearly captured. The other activations are 'absorbed' into the kick drum and snare drum components. This is a consequence of the fact that NMFD cannot distinguish a single-instrument hit from such a consistent layering of multiple instantaneous drum hits (i.e. in this example, each kick/snare drum hit always coincides with a hi-hat hit); an additional mechanism to disentangle such sounds is beyond the scope of this paper.

4 Evaluation on the ENST dataset

We evaluate our approach on all *fast simple* phrases from the ENST dataset [3]. We run the original NMFD algorithm and our adaptation on these extracts, and quantify how many excess drum hits can be detected in each template by counting the number of peaks in $\Delta a^{(k)}[\tau]$, see Section 2. We furthermore measure the spectrogram reconstruction quality using the Mean Absolute Error between X and $\hat{X}$: $\text{MAE}(X, \hat{X}) = \frac{1}{NT} \sum_{n,t} |X_{n,t} - \hat{X}_{n,t}|$.

For each decomposed mixture, the MAE for the decomposition with the original algorithm and the MAE for the adapted version are nearly identical; furthermore, all spectrograms are approximated well (mean MAE $5.6 \cdot 10^{-5}$ for both the original and the adapted algorithm, stdev. $3.0 \cdot 10^{-5}$ and $3.1 \cdot 10^{-5}$ resp.). The average number of excess peaks detected in $\Delta a^{(k)}[\tau]$ is 2.2 (stdev. 1.0) for default NMFD, and 0 for the adapted procedure[2]. Visual inspection[3] of the results shows that in the decompositions with unmodified NMFD, double hits are often present, while these are removed with the proposed procedure.

5 Conclusion

We conclude that the proposed adaptation maintains the same spectrogram reconstruction quality, with the added advantage that NMFD now captures only one drum hit per template. This allows to choose the template length long enough to fully capture drum hits with a long decay, while maintaining a clear and interpretable decomposition even in the presence of rapid successive drum hits.

[2] Which is an expected result, of course, as we report on the metric that is used in the adapted algorithm to detect double hits in the templates.

[3] See the accompanying website for examples: `https://users.ugent.be/~levdveir/2020MML`

Acknowledgements

Len Vande Veire is supported by a PhD fellowship of the Research Foundation Flanders (FWO).

References

1. Dittmar, C., Müller, M.: Towards transient restoration in score-informed audio decomposition. In: Proc. Int. Conf. Digital Audio Effects. pp. 145–152 (2015)
2. Dittmar, C., Müller, M.: Reverse engineering the amen break: score-informed separation and restoration applied to drum recordings. IEEE/ACM Transactions on Audio, Speech, and Language Processing **24**(9), 1535–1547 (2016)
3. Gillet, O., Richard, G.: Enst-drums: an extensive audio-visual database for drum signals processing. In: Proc of 7th International Conference on Music Information Retrieval, ISMIR 2006 (2006)
4. Laroche, C., Papadopoulos, H., Kowalski, M., Richard, G.: Drum extraction in single channel audio signals using multi-layer non negative matrix factor deconvolution. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 46–50. IEEE (2017)
5. Lindsay-Smith, H., McDonald, S., Sandler, M.: Drumkit transcription via convolutive nmf. In: International Conference on Digital Audio Effects (DAFx-12), York, UK (2012)
6. Roebel, A., Pons, J., Liuni, M., Lagrangey, M.: On automatic drum transcription using non-negative matrix deconvolution and itakura saito divergence. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 414–418. IEEE (2015)
7. Schmidt, M.N., Mørup, M.: Nonnegative matrix factor 2-d deconvolution for blind single channel source separation. In: International Conference on Independent Component Analysis and Signal Separation. pp. 700–707. Springer (2006)
8. Smaragdis, P.: Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In: Int. Conf. on Independent Component Analysis and Signal Separation. pp. 494–499. Springer (2004)
9. Ueda, S., Shibata, K., Wada, Y., Nishikimi, R., Nakamura, E., Yoshii, K.: Bayesian drum transcription based on nonnegative matrix factor decomposition with a deep score prior. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 456–460. IEEE (2019)