Source Separation Based on Non-Negative Matrix Factorization of the Synchrosqueezing Transform

janvier 2021

Plan

Plan

- We consider the problem of single channel blind source separation using non-negative matrix factorization on the TFR of the mixture
- We propose to replace the spectrogram by the modulus of the synchrosqueezing transform.
- We introduce two methods for reconstructing the sources:
  1. One based on the direct reconstruction from the synchrosqueezed representation
  2. Another based on both the short-time Fourier and the synchrosqueezing transforms (for long signal).

Plan

For a given signal f of length L and a window $w \in [0 : N-1]$, the *short-time Fourier transform* (STFT) is defined as

$$\mathbf{S}_f^w[n,k] = \sum_{l \in \mathbb{Z}} f[l] w[l-nH] e^{-i2\pi \frac{k(l-nH)}{N}} \tag{1}$$

where $k \in [0 : N-1]$ is the frequency index, $H \leq N$ the hop-size, N the frequency resolution, and $f[l] = f(\frac{l}{L})$. Assuming the length of w is smaller than N, the signal is traditionally reconstructed through overlap-add (OLA) :

$$f[l] = \frac{\sum_{n \in \mathbb{Z}} w[l-nH] \left(\frac{1}{N} \sum_{k=0}^{N-1} \mathbf{S}_f^w[n,k] e^{i2\pi \frac{k(l-nH)}{N}} \right)}{\sum_{n \in \mathbb{Z}} w[l-nH]^2}. \tag{2}$$

Synchrosqueezing Transform

To obtain a sharper *time-frequency representation* (TFR), one can alternatively consider the *synchrosqueezing transform* (SST):

$$\hat{\mathbf{m}}_f[n,k] = \frac{mL}{N} - \Im\left(\frac{\mathbf{S}_f^{w'}[n,k]}{\mathbf{S}_f^{w}[n,k]}\right) \tag{3}$$

where w' represents the derivative of window w.

$$\hat{\mathbf{S}}_f^{w}[n,k] = \sum_{k' \in \mathbb{Z}} \mathbf{S}_f^{w}[n,k'] \delta_{k', \lfloor \hat{\mathbf{m}}_f[n,k] \frac{N}{L} \rceil}. \tag{4}$$

Signal reconstruction is then carried out as:

$$f[n] = \frac{1}{w[0]N} \sum_{k=0}^{N-1} \hat{\mathbf{S}}_f^{w}[n,k]. \tag{5}$$

Plan

NMF decomposes a given non negative data matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$ into two non negative matrices, the *dictionary matrix* $\mathbf{W} \in \mathbb{R}^{N \times R}$ and the *activation matrix* $\mathbf{H} \in \mathbb{R}^{R \times L}$ such that $\mathbf{X} \approx \mathbf{WH}$. The decomposition is based on minimizing the reconstruction error of $\mathbf{X}$ through $\mathbf{WH}$, which can be formulated as

$$\min_{\mathbf{W},\mathbf{H}} \quad D(\mathbf{X}|\mathbf{WH}) \ \text{ subject to } \ \mathbf{W} \geq 0, \mathbf{H} \geq 0. \tag{6}$$

The most popular cost functions D are the Euclidean distance, Kullback-Leibler (KL) divergence and Itakura-Saito (IS) distance, which are special cases of β-divergence.

Plan

- We first propose here to apply NMF to SST modulus rather than STFT modulus.
- We then define TF masks used for source separation exploiting the fact that SST is invertible.

This can be formally described by:

$$\mathbf{S}_k = \mathbf{M}_k \odot \mathbf{S}, \tag{7}$$

where $\mathbf{S}_k$ represents the estimated STFT or SST of the k^{th} source, the soft mask $\mathbf{M}_k$ being defined by:

$$\mathbf{M}_k = \frac{\mathbf{X}_k}{\sum_{r=1}^{R} \mathbf{X}_r}, \tag{8}$$

where $\mathbf{X}_k$ corresponds to the k^{th} source in NMF decomposition, namely $\mathbf{W}_{:,k}\mathbf{H}_{k,:}$.

- Having defined $\mathbf{S}_k$, one reconstructs the modes using by replacing $\mathbf{S}_f^w$ by $\mathbf{S}_k$ in reconstruction formula.
- In what follows, we call STFT-NMF and SST-NMF the reconstruction processes based on NMF applied to STFT and SST respectively.

Plan

- Reconstruction with SST is not tractable when hop-size larger than 1
- To circumvent this limitation, we here introduce a two-step approach

We first apply NMF to SST and only keep the corresponding activation matrix **H**, and then recompute a dictionary matrix **W** from the spectrogram. With these new matrices, we are able to build soft masks, and then proceed with source separation:

$$\begin{aligned}(\mathbf{W},\mathbf{H}) &= \mathrm{NMF}(|\hat{\mathbf{S}}_f^w|)\\ \mathbf{W}' &= \mathrm{NMF}'(|\mathbf{S}_f^w|,\mathbf{H})\\ \mathbf{X}_k &= \mathbf{W}'_{:,k}\mathbf{H}_{k,:},\end{aligned}$$

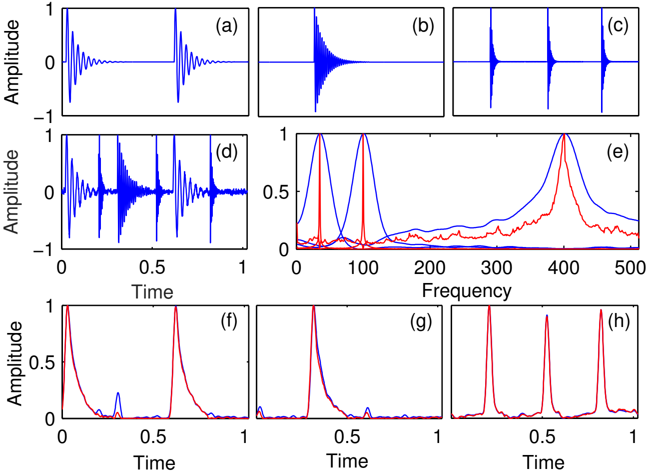
- Because SST is sparser than STFT, it should enable better source separation (more accurate activation and dictionary matrices).
- Cannot be directly used to recover the sources since SST is no longer invertible in that case.
- We recompute a dictionary from NMF applied to STFT modulus, and in which the activation matrix corresponds to that of NMF based on SST.
- The last step of the algorithm is a simple constrained least-square problem, which should converge faster than NMF and is convex if the divergence is convex.

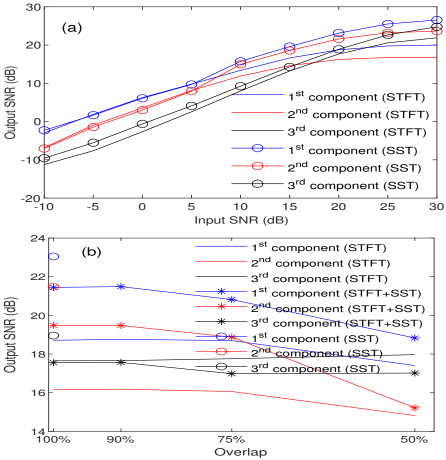
The technique is denoted by STFT+SST-NMF.

Plan

Application to Synthetic Drum Sound Signals

We investigate the behavior of the source separation procedures introduced above when these are applied to either synthetic or real signals.




(a)
Output SNR (dB)
Input SNR (dB)
1st component (STFT)
2nd component (STFT)
3rd component (STFT)
1st component (SST)
2nd component (SST)
3rd component (SST)
(b)
Output SNR (dB)
Overlap
100%
90%
75%
50%
1st component (STFT+SST)
2nd component (STFT+SST)
3rd component (STFT+SST)

Performance Evaluation on Drum Source Separation

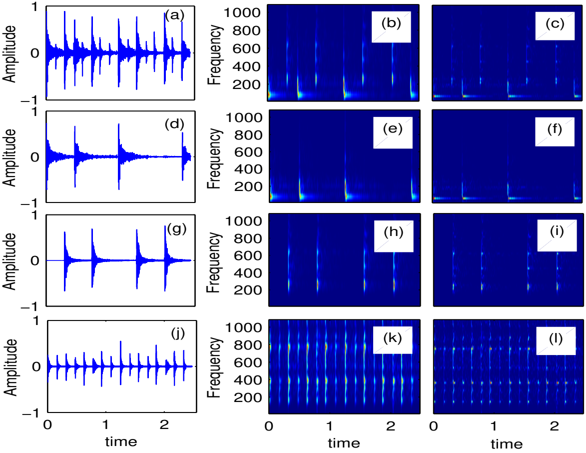


Figure: first column: Mixture signal of KD, SD and HH components (a) with KD, SD and HH components displayed in (d),(g) and (h); second column : STFT of the signals of the first column; third column: SST of the signals of the first column

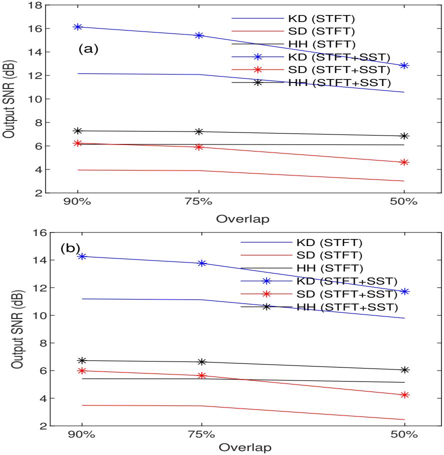


Figure: Output SNR corresponding to the reconstruction of KD, SD and HH components averaged over 60 drum loops with varying percentage overlaps for the analysis window; (a) without noise; (b) for input SNR 20 dB; and when the percentage of overlap varies, the results are averaged over 20 realizations