Drum transcription from multichannel recordings with non-negative matrix factorization

David S. Alves\textsuperscript{1,2}, Jouni Paulus\textsuperscript{1}, and José Fonseca\textsuperscript{2}

\textsuperscript{1}Department of Signal Processing, Tampere University of Technology, Finland
\textsuperscript{2}Department of Electrical Engineering, New University of Lisbon, Portugal

**Introduction**

- Drum transcription: from audio input
  - determine temporal locations of drum sound events, and
  - recognise the played instruments.
- Earlier methods operate mainly on single-channel (or stereo) signals.
- In studios, multichannel recordings are available.
- Extend an existing method to multichannel signals.

**Signal model**

- Observed magnitude spectrogram $X$ is a sum of $N$ source signals:
  \[ X = \sum_{n=1}^{N} X_n + \epsilon. \]
- Each source is assumed to be a product of two basis vectors (gain over time and magnitude on each frequency):
  \[ X_n = s_n a_n^T. \]
- As a matrix product: $X \approx S A$, where $S = [s_1, s_2, \cdots, s_N]$ and $A = [a_1, a_2, \cdots, a_N]^T$.
- Inverse problem: solve $S$ and $A$ minimising reconstruction error given $X$.
- Non-negative matrix factorization (NMF) restricts all elements to be non-negative.
- An example factorization of a drum loop to three sources ($X$ is a mel-frequency spectrogram):

**Baseline method**

- Calculate spectral templates $S$ for each target drum (training phase).
- Solve time-varying gains $A$ from input $X$ keeping $S$ fixed.
- Detect onsets from the gains $A$.

**Multichannel extension**

- Stack spectrograms $X_c$ from $C$ channels $c \in 1 \cdots C$ to $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_C \end{bmatrix}$.
- Spectral template stacking:
  \[ S = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_C \end{bmatrix} = \begin{bmatrix} s_{1,1}, s_{1,2}, \cdots, s_{1,N} \\ s_{2,1}, s_{2,2}, \cdots, s_{2,N} \\ \vdots \\ s_{C,1}, s_{C,2}, \cdots, s_{C,N} \end{bmatrix} = \begin{bmatrix} \tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_N \end{bmatrix}. \]

**Results**

- Evaluations with ENST drums data set
  - 3 drummers and drum kits (differing microphone setups with 7–8 mics), 64 tracks, average duration 55 s (30–75 s)
  - Transcribe bass drum (BD), snare drum (SD), and hi-hat (HH).
- Comparison to
  - a single-channel version operating on a mix-down, and
  - a naive onset detection based multichannel method (assuming each drum to have a close microphone).

**Conclusions:**

- Extend a drum transcription method using spectral templates to accept multichannel inputs.
- Performance increase from single-channel method $\rightarrow$ channel information helps.
- Performance increase from naive onset detection method $\rightarrow$ spectral information helps (and no dependency on having close microphones on all targets).