Drum transcription from multichannel recordings with non-negative matrix factorization

David S. Alves[1,2], *Jouni Paulus*[1], and José Fonseca[2]

[1]Department of Signal Processing, Tampere University of Technology, Finland

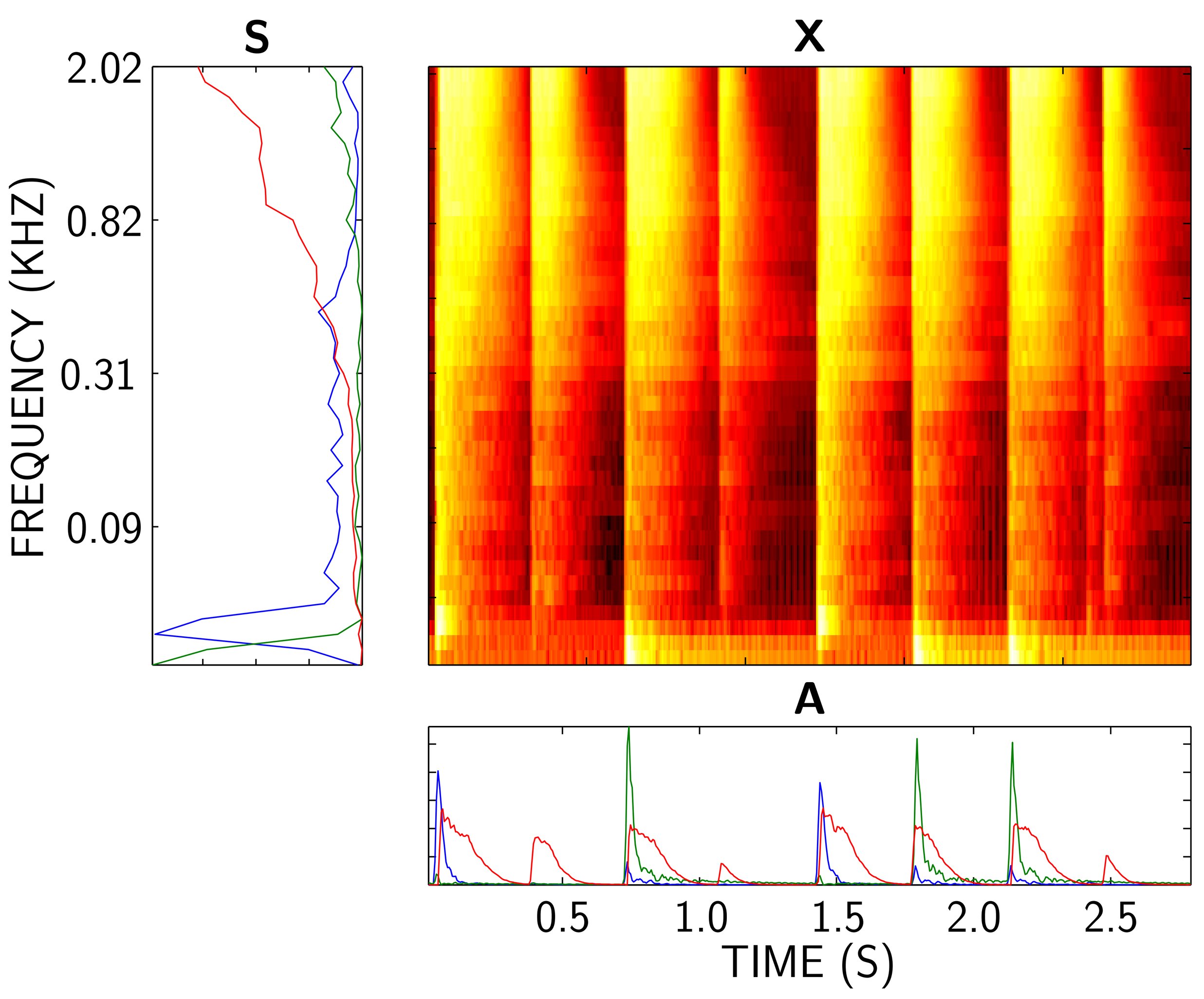
[2]Department of Electrical Engineering, New University of Lisbon, Portugal

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 $\mathbf{X}$ is a sum of N source signals: $\mathbf{X} = \sum_{n=1}^{N} \mathbf{X}_n + \epsilon$.
- Each source is assumed to be a product of two basis vectors (gain over time and magnitude on each frequency): $\mathbf{X}_n = \mathbf{s}_n \mathbf{a}_n^{\mathsf{T}}$.
- As a matrix product: $\mathbf{X} \approx \mathbf{SA}$, where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N]$ and $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N]^{\mathsf{T}}$.
- Inverse problem: solve $\mathbf{S}$ and $\mathbf{A}$ minimising reconstruction error given $\mathbf{X}$.
- Non-negative matrix factorization (NMF) restricts all elements to be non-negative.
- An example factorization of a drum loop to three sources ($\mathbf{X}$ is a mel-frequency spectrogram):



Baseline method

- Template-based NMF method from Paulus & Virtanen *"Drum transcription with non-negative spectrogram factorisation"*, EUSIPCO2005.
- Calculate spectral templates $\mathbf{S}$ for each target drum (training phase).
- Solve time-varying gains $\mathbf{A}$ from input $\mathbf{X}$ keeping $\mathbf{S}$ fixed.
- Detect onsets from the gains $\mathbf{A}$.

Multichannel extension

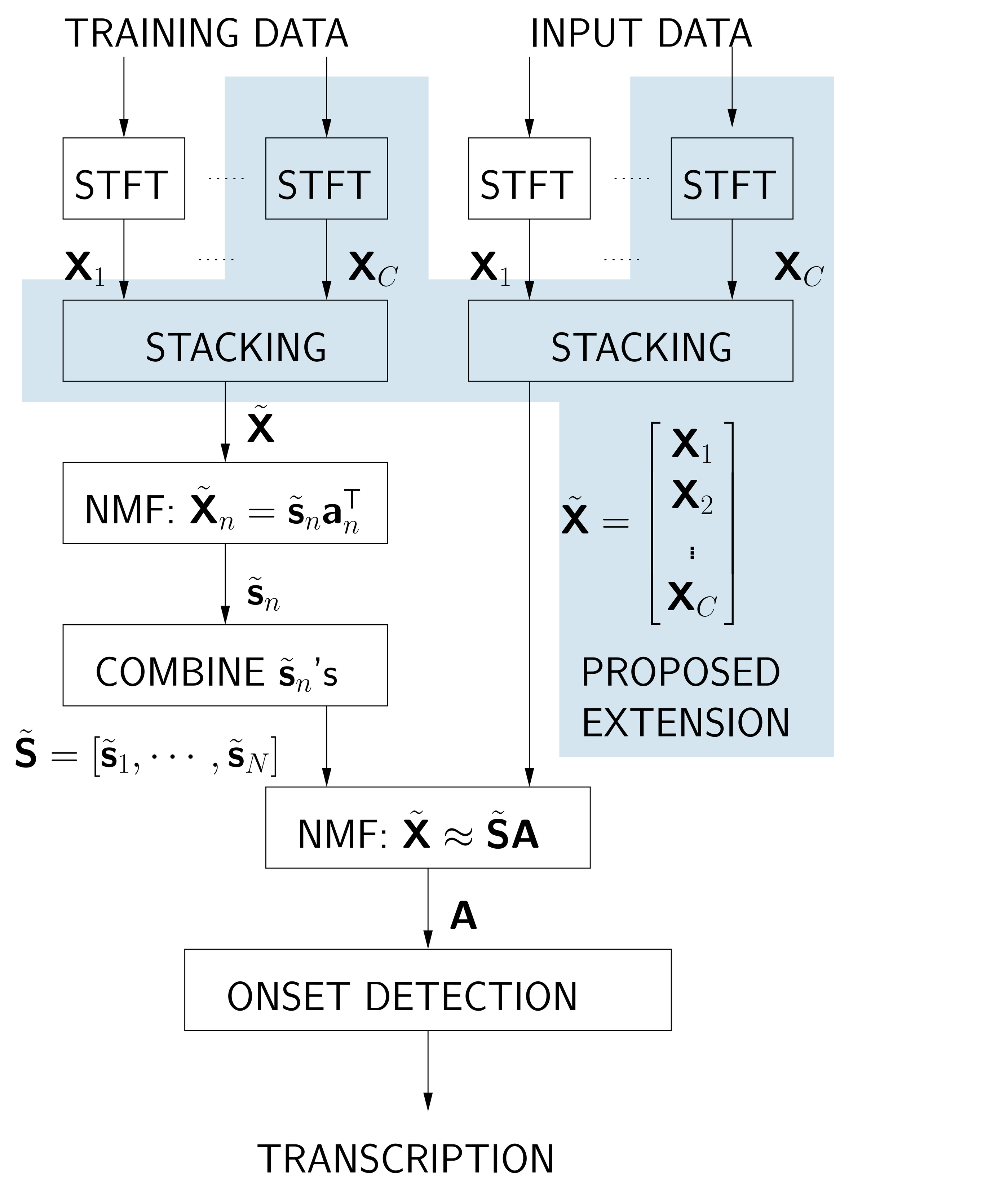
- Stack spectrograms $\mathbf{X}_c$ from C channels $c \in 1 \ldots C$ to

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_C \end{bmatrix}.$$

- Spectral template stacking:

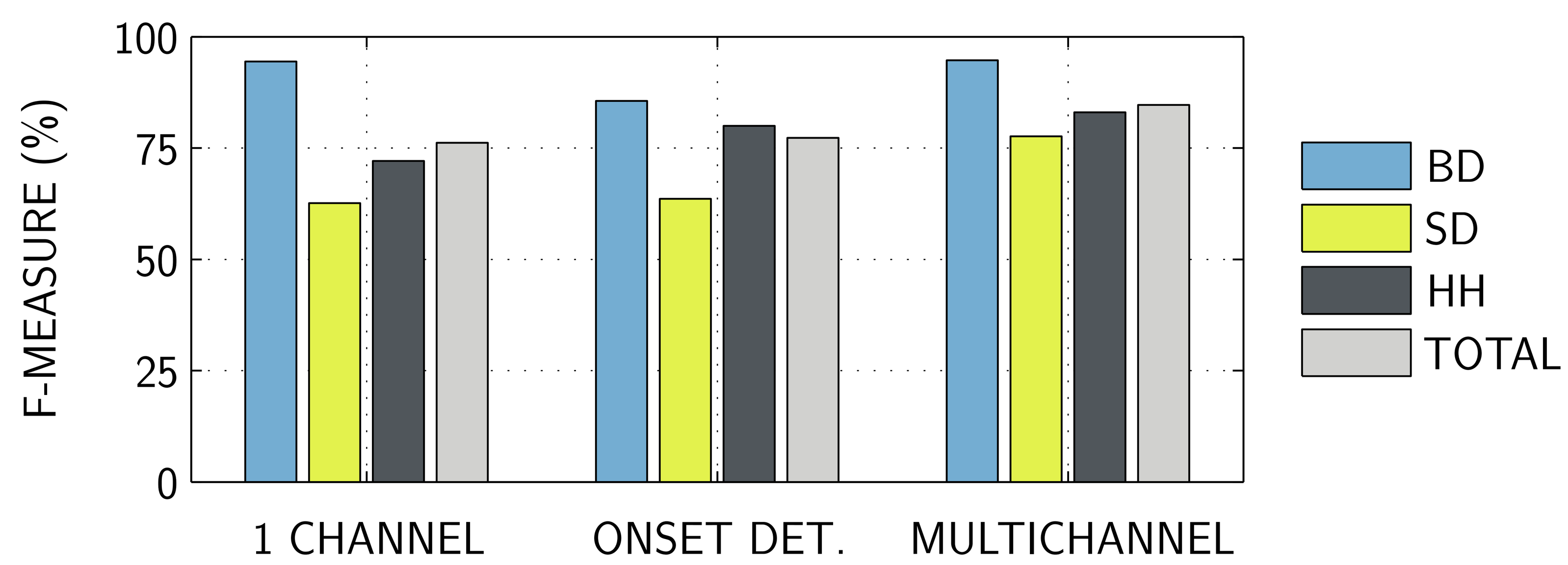
$$\tilde{\mathbf{S}} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_C \end{bmatrix} = \begin{bmatrix} \mathbf{s}_{1,1}, \mathbf{s}_{1,2}, \cdots, \mathbf{s}_{1,N} \\ \mathbf{s}_{2,1}, \mathbf{s}_{2,2}, \cdots, \mathbf{s}_{2,N} \\ \vdots \cdots \cdots \vdots \\ \mathbf{s}_{C,1}, \mathbf{s}_{C,2}, \cdots, \mathbf{s}_{C,N} \end{bmatrix} = [\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2, \cdots, \tilde{\mathbf{s}}_N].$$

- $\mathbf{s}_{n,c}$ spectrum basis of n^{th} drum on c^{th} channel. $\tilde{\mathbf{s}}_n$ spectrum basis of n^{th} drum across channels.
- Solve gains $\mathbf{A}$ from $\tilde{\mathbf{X}} \approx \tilde{\mathbf{S}}\mathbf{A}$.



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 naïve 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 → channel information helps.
- Performance increase from naïve onset detection method → spectral information helps (and no dependency on having close microphones on all targets).