

# REAL-TIME POLYPHONIC MUSIC TRANSCRIPTION WITH NON-NEGATIVE MATRIX FACTORIZATION AND BETA-DIVERGENCE

ISMIR 2010 - 11th International Society for Music Information Retrieval Conference

Arnaud Desein, Arshia Cont, Guillaume Lemaitre

IRCAM - CNRS UMR 9912, Paris, France

{desein, cont, lemaitre}@ircam.fr



## Introduction

- Context and motivations:
  - Real-time system for polyphonic music transcription.
  - Non-negative matrix factorization and  $\beta$ -divergence.
  - Front-end for musical interactions in live performances.
- Contributions:
  - Non-negative decomposition scheme tailored to real-time.
  - Intuition in understanding the relevancy of the  $\beta$ -divergence.
  - Comparative evaluation with off-line algorithms at the state-of-the-art.

## Standard non-negative matrix factorization

- Problem formulation and multiplicative updates:
  - Given  $\mathbf{V} \in \mathbb{R}_+^{n \times m}$  and  $r < \min(n, m)$ ,  $\mathbf{V}$  is modeled as follows:

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \quad \text{with } \mathbf{W} \in \mathbb{R}_+^{n \times r} \text{ and } \mathbf{H} \in \mathbb{R}_+^{r \times m} \quad (1)$$

- In the standard formulation,  $\mathbf{W}$  and  $\mathbf{H}$  are found by minimizing:

$$\frac{1}{2} \|\mathbf{V} - \mathbf{W}\mathbf{H}\|_F^2 = \frac{1}{2} \sum_j \|\mathbf{v}_j - \mathbf{W}\mathbf{h}_j\|_2^2 \quad (2)$$

- A popular scheme alternates between two multiplicative updates:

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T \mathbf{V}}{\mathbf{W}^T \mathbf{W} \mathbf{H}} \quad \mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\mathbf{V} \mathbf{H}^T}{\mathbf{W} \mathbf{H} \mathbf{H}^T} \quad (3)$$

- Applications in sound recognition:
  - $\mathbf{V}$  is a time-frequency representation of the sound to analyze.
  - The basis vectors  $\mathbf{w}_i$  contain spectral templates, while the decomposition coefficients  $h_{ij}$  represent their successive activations.
  - Numerous off-line applications to polyphonic music transcription.
  - Some on-line applications by employing non-negative decomposition.

## Non-negative decomposition with the beta-divergence

- Beta-divergence:
  - For  $\beta \in \mathbb{R}$  and  $x, y \in \mathbb{R}_{++}$ , the  $\beta$ -divergence from  $x$  to  $y$  is defined by:

$$d_\beta(x|y) = \frac{1}{\beta(\beta-1)} (x^\beta + (\beta-1)y^\beta - \beta xy^{\beta-1}) \quad (4)$$

- The scaling property is relevant to polyphonic music transcription:

$$d_\beta(\lambda x | \lambda y) = \lambda^\beta d_\beta(x|y) \quad \text{for any } \lambda \in \mathbb{R}_{++} \quad (5)$$

- $\beta$  can help to reduce octave and harmonic errors since it controls the trade-off between the fundamental, the first, and the higher partials.

- Problem formulation and multiplicative update:
  - The incoming signal  $\mathbf{v} \in \mathbb{R}_+^n$  is decomposed onto a dictionary of spectral templates  $\mathbf{W} \in \mathbb{R}_+^{n \times r}$  which is kept fixed:

$$\mathbf{v} \approx \mathbf{W}\mathbf{h} \quad \text{with } \mathbf{h} \in \mathbb{R}_+^r \quad (6)$$

- Using the  $\beta$ -divergence as a cost function,  $\mathbf{h}$  is found by minimizing:

$$\mathcal{D}_\beta(\mathbf{v}|\mathbf{W}\mathbf{h}) = \sum_i d_\beta(v_i | [\mathbf{W}\mathbf{h}]_i) \quad (7)$$

- A multiplicative update tailored to real-time is applied iteratively:

$$\mathbf{h} \leftarrow \mathbf{h} \otimes \frac{(\mathbf{W} \otimes (\mathbf{v}\mathbf{e}^T))^T (\mathbf{W}\mathbf{h})^{\beta-2}}{\mathbf{W}^T (\mathbf{W}\mathbf{h})^{\beta-1}} \quad (8)$$

## General architecture of the system

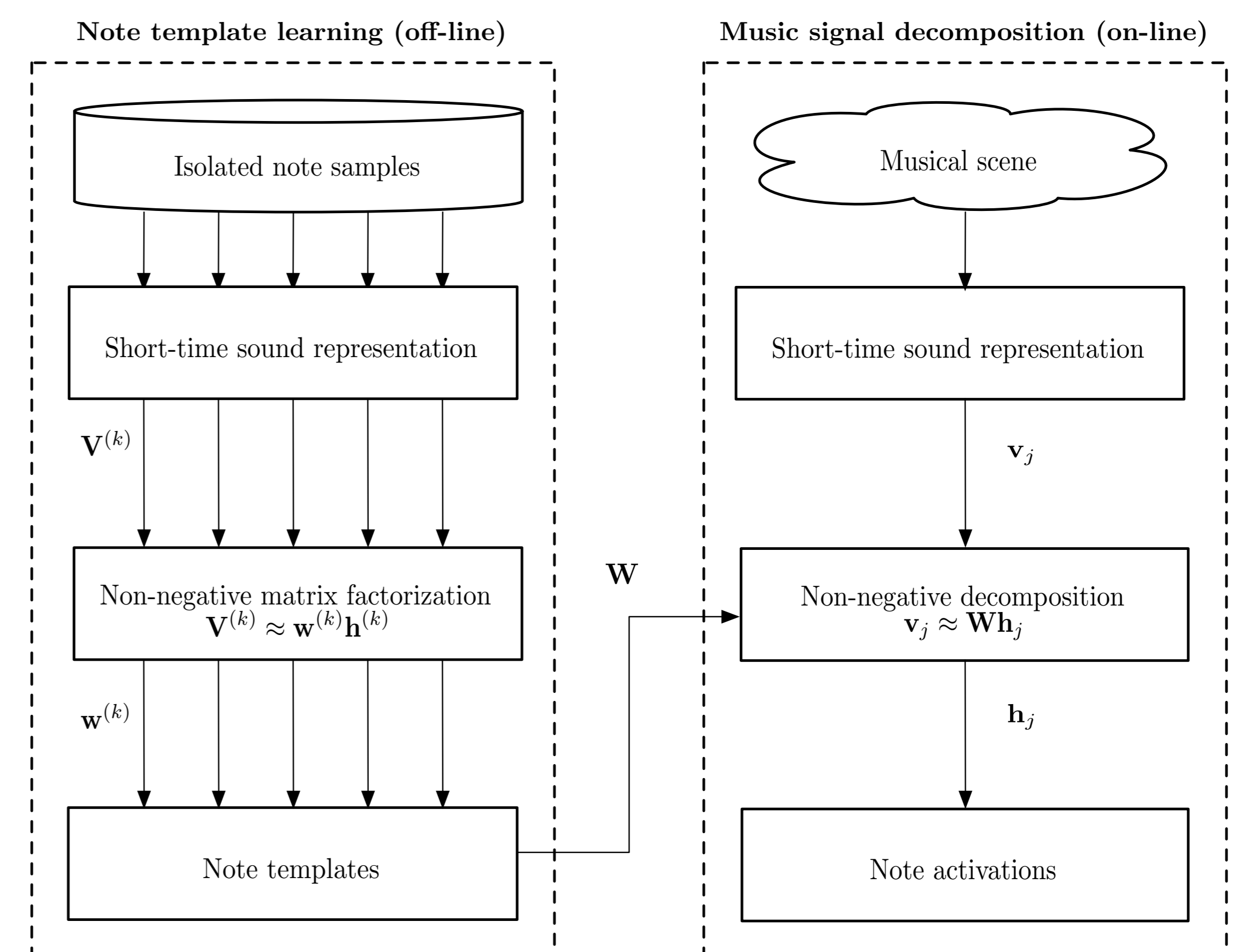


Figure 1: Schematic view of the general architecture.

## Evaluation and results

- Subjective evaluation.

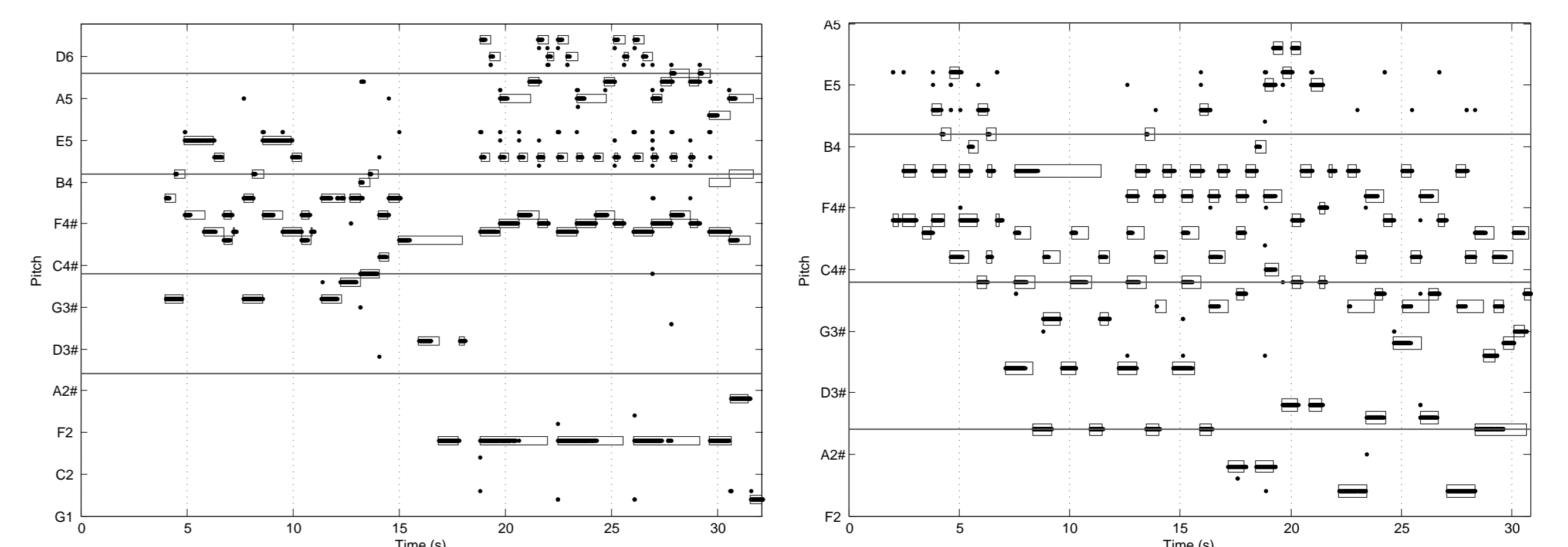


Figure 2: Transcription of two piano excerpts.

- Objective evaluation.

Alg.	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{A}$	$\mathcal{E}_{tot}$	$\mathcal{E}_{subs}$	$\mathcal{E}_{miss}$	$\mathcal{E}_{fa}$	$\mathcal{P}_1$	$\mathcal{R}_1$	$\mathcal{F}_1$	$\mathcal{A}_1$	$\mathcal{M}_1$	$\mathcal{P}_2$	$\mathcal{R}_2$	$\mathcal{F}_2$	$\mathcal{A}_2$
BND	63.9	67.3	65.5	48.7	58.9	11.9	20.8	26.2	75.5	67.1	71.1	55.1	56.7	30.0	26.6	28.2	16.4
END	55.3	58.6	56.9	39.8	71.4	17.3	24.1	29.9	57.9	58.2	58.1	40.9	53.9	21.4	21.6	21.5	12.0
Hoy.	58.5	55.2	56.8	39.7	67.1	16.8	28.0	22.3	57.2	56.3	56.8	39.6	54.1	21.0	20.7	20.8	11.6
Vin.	61.0	66.7	63.7	46.8	65.6	10.4	22.9	32.3	58.1	73.7	65.0	48.1	57.7	20.7	26.3	23.2	13.1
Yeh	60.0	70.8	65.0	48.1	60.0	16.3	12.8	30.8	33.0	58.8	42.3	26.8	55.1	11.6	20.7	14.9	8.0

Table 1: Comparative results for frame and note-level transcription.

## Conclusion

- The proposed system can outperform off-line approaches.
- Perspectives:
  - Employ multi-channel information.
  - Improve the template learning scheme.
  - Consider adaptive and non-stationary templates.
- Additional resources:
  - [http://imtr.ircam.fr/imtr/Arnaud\\_Desein](http://imtr.ircam.fr/imtr/Arnaud_Desein)
  - [http://imtr.ircam.fr/imtr/Realtime\\_Transcription](http://imtr.ircam.fr/imtr/Realtime_Transcription)