# Polyphonic Piano Note Transcription with Non-negative Matrix Factorization of Differential Spectrogram

Lufei Gao*, Li Su†, Yi-Hsuan Yang†, Tan Lee*

*Department of Electronic Engineering, The Chinese University of Hong Kong
†Academia Sinica, Taiwan

## Highlights

Note-level music transcription of pitched percussive instruments

- Investigate the idea of highlighting local energy increase in the TF representation;
- Propose algorithms based on existing NMF based methods;
- Validate the advantages of the differential spectrogram.

## Baseline Methods

- Standard NMF (NMF)
  - Approximate the STFT $X_{ft}$ as the product of two non-negative matrices:

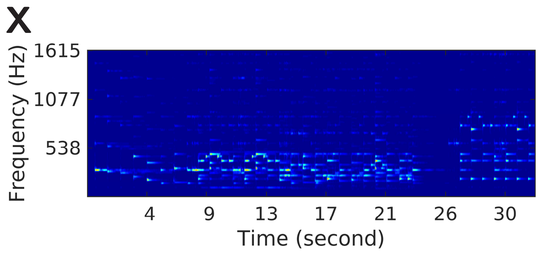
$$X_{ft} \simeq V_{ft} = \sum_{k=1}^{K} W_{fk} H_{kt}$$

  $\mathbf{W}$ is the template of single-note spectra; $\mathbf{H}$ is the time-varying activation.
  - The distortion $D(\mathbf{X}|\mathbf{V})$ is measured by the $\beta$-divergence.
- Attack/Decay Convolutional NMF (CNMF-AD)
  - $X_{ft}$ is assumed to be the summation of the attack phase and the decay phase. The model is defined as

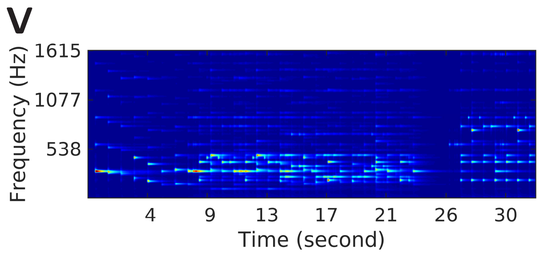
$$V_{ft} = \sum_{k=1}^{K} W^a_{fk} \sum_{\tau=t-T_t}^{t+T_t} H_{k\tau} P(t-\tau) + \sum_{k=1}^{K} W^d_{fk} \sum_{\tau=1}^{t} H_{k\tau} e^{-(t-\tau)\alpha_k}$$

  $\mathbf{W}^a$ is the percussive template for the attack phase; $\mathbf{W}^d$ is the harmonic template for the decay phase; $\mathbf{P}$ and $\alpha_k$ are the transient pattern and the exponential decay rate, respectively.

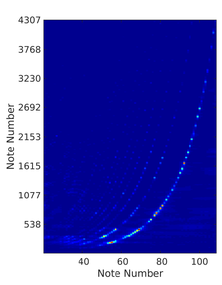
- $\mathbf{X}$



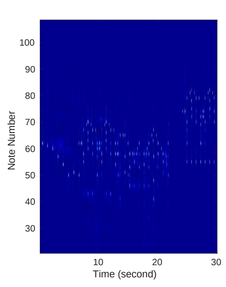

- $\mathbf{V}$



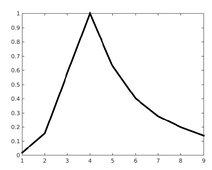

- $\mathbf{W}^a$



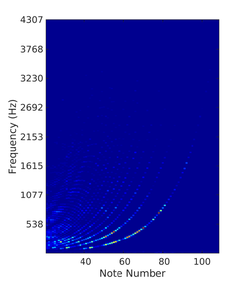

- $\mathbf{H}$



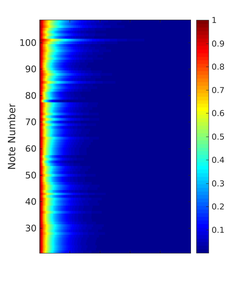

- $\mathbf{P}$



- $\mathbf{W}^d$




- $e^{-t\alpha_k}$

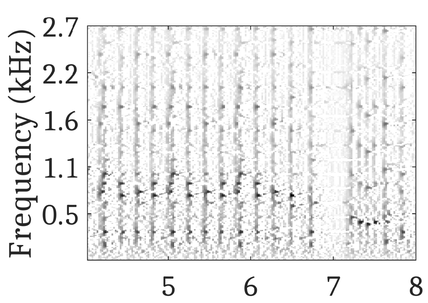



## Differential Spectrogram

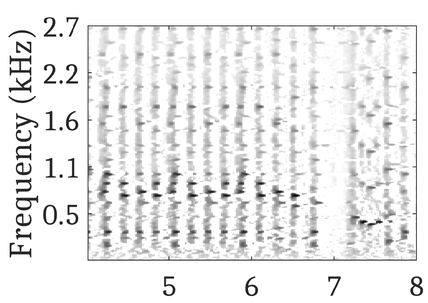
- Assuming that the intrument exhibits harmonics with locally stable frqeuencies, the differential spectrogram $\widehat{X}_L(f,t)$ is defined as:

$$\widehat{X}_L(f,t) = \mathrm{HWR}(|X(f,t+L)| - |X(f,t)|)$$

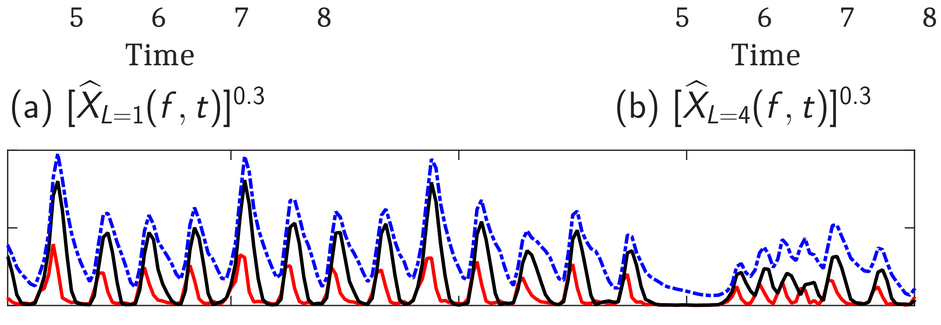
HWR stands for the half-wave rectification ($\mathrm{HWR}(x) = \frac{x+|x|}{2}$).




(a) $[\widehat{X}_{L=1}(f,t)]^{0.3}$




(b) $[\widehat{X}_{L=4}(f,t)]^{0.3}$




(c) Red: $\mathrm{SF}_{L=1}(t)$; black: $\mathrm{SF}_{L=4}(t)$; blue: $\sum_f |X(f,t)|$.

## Model Adaptation with Differential Spectrogram

- Standard NMF adaptation (NMF-$\Delta$): The following feature is used to replace $\mathbf{X}$ in the model:

$$\widetilde{X}_L(f,t) = c_1 X(f,t) + c_2 \widehat{X}_L(f,t),$$

## Model Adaptation with Differential Spectrogram

- Convolutional NMF adaptation (CNMF-$\Delta$): The following model is utilized to estimate the note activation.

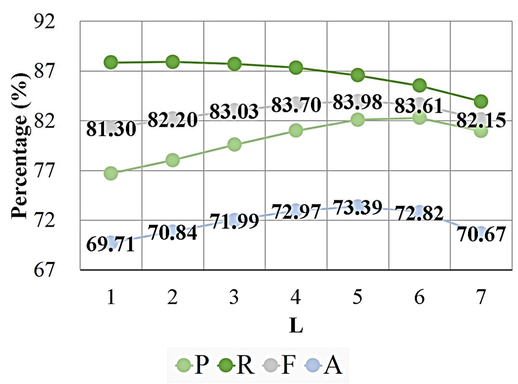
$$\widehat{X}_L(f,t) \simeq \widehat{V}_{ft} = \sum_{k=1}^{K} \widehat{W}_{fk} \sum_{\tau=t-T_t}^{t+T_t} \widehat{H}_{k\tau} \widehat{P}(t-\tau).$$

  Convolving $\widehat{\mathbf{H}}$ with $\widehat{\mathbf{P}}$ yields the attack activation denoted by $\widehat{\mathbf{H}}^a$.
- Model initialization (CNMF-AD-$\Delta$): Initialize $\widehat{\mathbf{H}}$ by $\mathbf{H}$ whihc is estimated using CNMF-AD before the estimation with differential spectrogram.

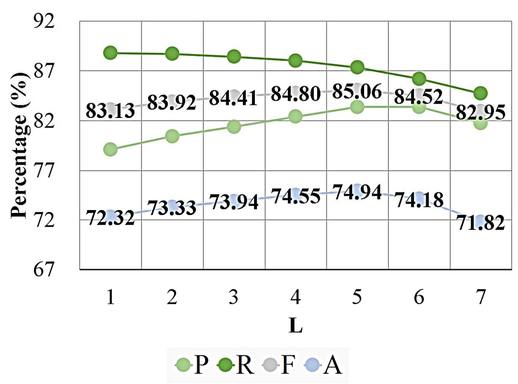
## Experiment setting

- Databases
  - For dictionary construction
    - The 88 forte isolated note recordings in MAPS-ENSTDkCl.
  - For testing tasks
    - The first 30-second excerpt of the 30 music pieces from MAPS-ENSTDkCl.
- Metric: Precision $\mathcal{P}$, Recall $\mathcal{R}$, F-measure $\mathcal{F}$, Accuracy $\mathcal{A}$.

## Results: System settings




(a) CNMF-$\Delta$




(b) CNMF-AD-$\Delta$

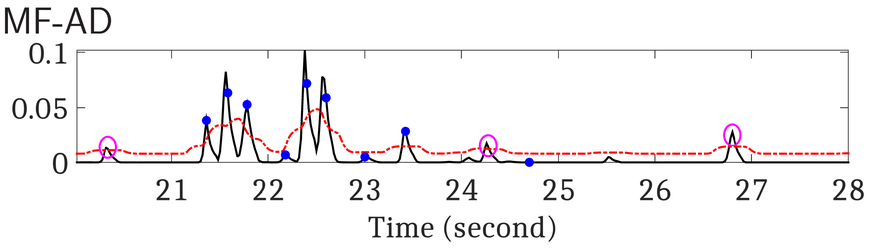
## Results: Comparison

Table: Performance comparison on "ENSTDkCl"

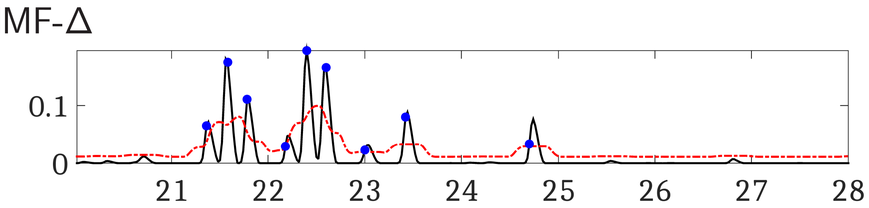
| Method | P | R | F | A |
|---|---|---|---|---|
| NMF ($\beta = 0.5$) | 59.70 | 34.51 | 41.81 | 27.24 |
| NMF-$\Delta$ ($\beta = 0.5$) | 71.04 | 42.48 | 50.70 | 35.13 |
| NMF ($\beta = 2$) | 51.67 | 43.11 | 46.34 | 30.54 |
| NMF-$\Delta$ ($\beta = 2$) | 67.83 | 58.21 | 61.76 | 45.10 |
| CNMF-$\Delta$ | 82.11 | 86.57 | 83.98 | 73.39 |
| CNMF-AD-$\Delta$ | 83.38 | **87.34** | **85.06** | **74.94** |
| CNMF-AD | **89.22** | 78.35 | 82.91 | 71.55 |
| Böck | – | – | – | 68.70 |
| Berg-Kirkpatrick | 78.10 | 74.70 | 76.40 | – |

## Result Analysis

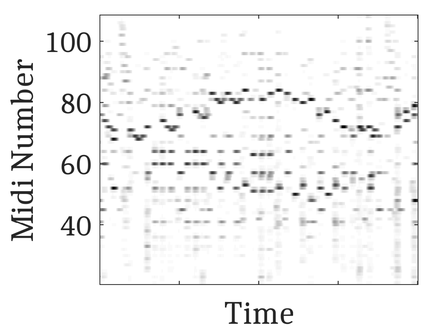
- The attack activations of note E4 in one test file.
  - $\mathbf{H}^a$ of CNMF-AD



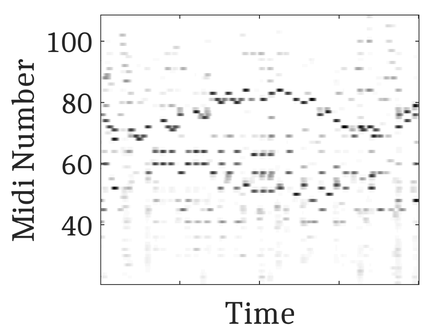

  - $\widehat{\mathbf{H}}^a$ of CNMF-$\Delta$




- The piano rolls of attack activations of our methods




$(\widehat{\mathbf{H}}^a)^{0.3}$ of CNMF-$\Delta$




$(\widehat{\mathbf{H}}^a)^{0.3}$ of CNMF-AD-$\Delta$