ISMIR 2015 MALAGA
16th International Society for Music Information Retrieval Conference - October 26-30, 2015 Málaga, Spain

A Hierarchical Bayesian Framework for Score-Informed Source Separation of Piano Music Signals

Wai Man SZETO
Office of University General Education
The Chinese University of Hong Kong
wmszeto@cuhk.edu.hk

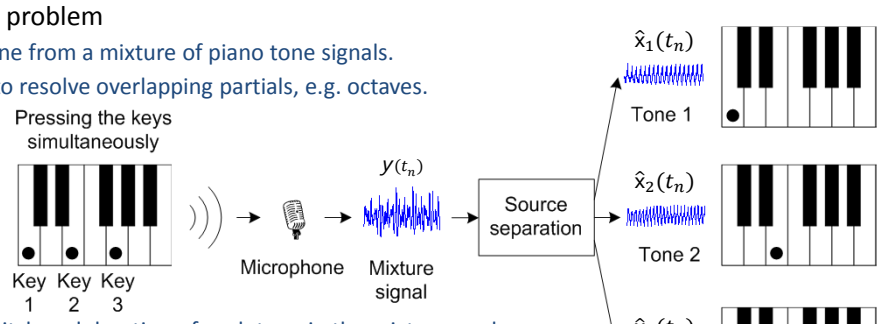
Kin Hong WONG
Department of Computer Science and Engineering
The Chinese University of Hong Kong
khwong@cse.cuhk.edu.hk

Abstract

Here we propose a score-informed monaural source separation system to extract every tone from a mixture of piano tone signals. Two sinusoidal models in our earlier work are employed in the above-mentioned system to represent piano tones: the General Model and the Piano Model. The General Model, a variant of sinusoidal modeling, can represent a single tone with high modeling quality, yet it fails to separate mixtures of tones due to the overlapping partials. The Piano Model, on the other hand, is an instrument-specific model tailored for piano. Its modeling quality is lower but it can learn from training data (consisting entirely of isolated tones), resolve the overlapping partials and thus separate the mixtures. We formulate a new hierarchical Bayesian framework to run both Models in the source separation process so that the mixtures with overlapping partials can be separated with high quality. The results show that our proposed system gives robust and accurate separation of piano tone signal mixtures (including octaves) while achieving significantly better quality than those reported in related work done previously.

Problem definition

- Source separation problem
  - To extract every tone from a mixture of piano tone signals.
  - Major difficulty is to resolve overlapping partials, e.g. octaves.



- Given that
  - *Scored-informed*. Pitch and duration of each tone in the mixture are known.
  - *Training data*. The pitches in the mixture should reappear as isolated tones in the target recording. These isolated tones serve as the training data.
  - *Without pedaling*. Both mixture and training data are performed without pedaling.

General Model (GM)

- Frame-wise sinusoidal model to represent piano tones [SW13a].
- The estimated kth tone at rth frame $\hat{x}_{k,r}[l] = \sum_{m=1}^{M_k} w[l]\left(\alpha_{k,m,r}\cos(2\pi f_{k,m}t_l) + \beta_{k,m,r}\sin(2\pi f_{k,m}t_l)\right)$.
- The estimated mixture at rth frame $\hat{y}_r[l] = \sum_{k=1}^{K}\hat{x}_{k,r}[l]$.
- The estimated mixture in the matrix form $\hat{\mathbf{Y}} = \mathbf{HG}$ where $\mathbf{H}$ is the frequency matrix, and $\mathbf{G}$ is the amplitude matrix.
- Goal: To estimate the GM parameters $\boldsymbol{\Theta}_y = \{\mathbf{H}, \mathbf{G}\}$ of the mixture $\mathbf{y}$.
- The overlap-and-add method is used to reconstruct the entire signal from GM.
- Pro: High modeling quality.
- Con: Unable to resolve overlapping partials ($\mathbf{H}$ is rank-deficient).

Piano Model (PM)

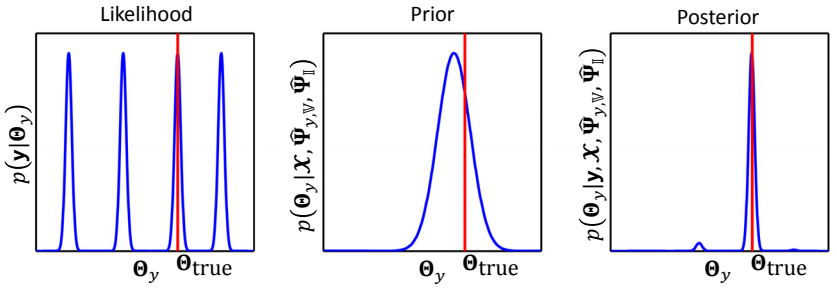
- Sinusoidal model to represent piano tones in an entire duration [SW13b]. The estimated kth tone $\hat{x}_k(t_n) = \sum_{m=1}^{M_k} a(t_n; c_k, \boldsymbol{\varphi}_{k,m})\cdot\cos(2\pi f_{k,m}t_n + \phi_{k,m})$ where $a(t_n; c_k, \boldsymbol{\varphi}_{k,m}) = b_{k,m}(c_k)^{d_{k,m}}\zeta_{k,m}(\exp\{-\lambda_{k,m}t_n\} - \exp\{-\gamma_{k,m}t_n\})$ and $\boldsymbol{\varphi}_{k,m} = \{b_{k,m}, d_{k,m}, \lambda_{k,m}\}$.
- The estimated mixture $\hat{y}(t_n) = \sum_{k=1}^{K}\hat{x}_k(t_n - \tau_k)$.
- Goal: To estimate the PM parameters $\boldsymbol{\Psi}_{\mathbb{I}}$ and $\boldsymbol{\Psi}_{y,\mathbb{V}}$.
  - Invariant PM parameters $\boldsymbol{\Psi}_{\mathbb{I}}$ are invariant to instances of the same pitch.
  - Varying PM parameters $\boldsymbol{\Psi}_{y,\mathbb{V}}$ may vary across instances.
- Pro: Able to resolve overlapping partials.
- Con: Medium modeling quality.

		Training	Source separation
Invariant PM parameters $\boldsymbol{\Psi}_{\mathbb{I}}$	Envelope parameters $\boldsymbol{\varphi}_{k,m}$	To be estimated	Given
	Frequencies $f_{k,m}$		
	Phases $\phi_{k,m}$		
Varying PM parameters $\boldsymbol{\Psi}_{y,\mathbb{V}}$	Intensity c_k	Given	To be estimated
	Time shift τ_k		

The Bayesian framework

- The hierarchical Bayesian framework first runs PM and then GM in the source separation process so that the mixtures with overlapping partials can be separated with high quality.
- If overlapping partials are present, the frequency matrix $\mathbf{H}$ in GM is rank-deficient and there are many peaks in the likelihood function.
- Given the estimated PM parameters and the training data, we can set the prior distributions of the GM parameters to favor the proper regions of values.

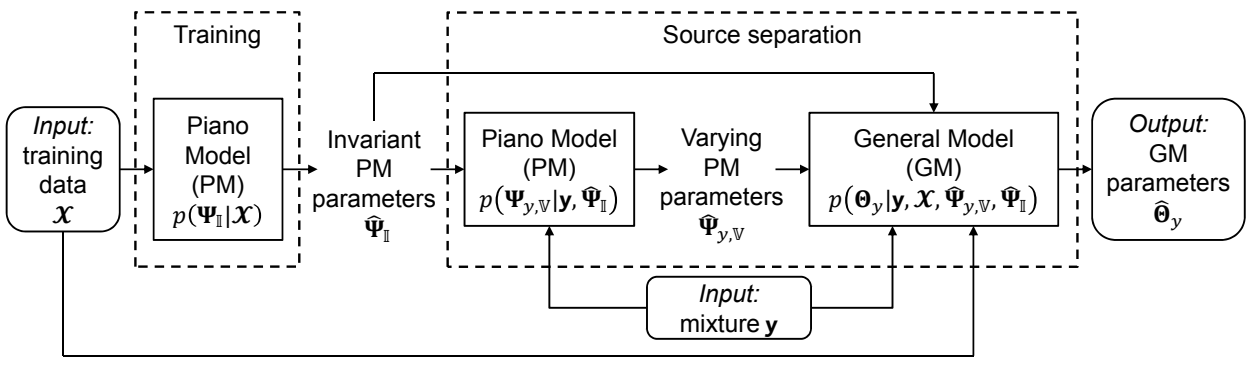
$$\underbrace{p(\boldsymbol{\Theta}_y|\mathbf{y}, \boldsymbol{\mathcal{X}}, \hat{\boldsymbol{\Psi}}_{y,\mathbb{V}}, \hat{\boldsymbol{\Psi}}_{\mathbb{I}})}_{\text{posterior}} \propto \underbrace{p(\mathbf{y}|\boldsymbol{\Theta}_y)}_{\text{likelihood}}\ \underbrace{p(\boldsymbol{\Theta}_y|\boldsymbol{\mathcal{X}}, \hat{\boldsymbol{\Psi}}_{y,\mathbb{V}}, \hat{\boldsymbol{\Psi}}_{\mathbb{I}})}_{\text{prior}}$$



This schematic diagram shows that an appropriate prior gives the desirable MAP solution.

Training and source separation

- *Training stage*. Given the training data $\boldsymbol{\mathcal{X}}$, find the most probable value of the invariant PM parameters $\hat{\boldsymbol{\Psi}}_{\mathbb{I}}$ of $p(\boldsymbol{\Psi}_{\mathbb{I}}|\boldsymbol{\mathcal{X}})$.
- *Source separation stage*. Given the mixture $\mathbf{y}$, the training data $\boldsymbol{\mathcal{X}}$ and the invariant PM parameters $\hat{\boldsymbol{\Psi}}_{\mathbb{I}}$, source separation functions in two steps:
  - *Source separation with PM*. Given $\mathbf{y}$ and $\boldsymbol{\Psi}_{\mathbb{I}}$, find the most probable value of the varying PM parameters $\hat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ of $p(\boldsymbol{\Psi}_{y,\mathbb{V}}|\mathbf{y}, \hat{\boldsymbol{\Psi}}_{\mathbb{I}})$.
  - *Source separation with GM*. Given $\mathbf{y}$, $\boldsymbol{\mathcal{X}}$, $\hat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ and $\boldsymbol{\Psi}_{\mathbb{I}}$, estimate the prior distribution $p(\boldsymbol{\Theta}_y|\boldsymbol{\mathcal{X}}, \hat{\boldsymbol{\Psi}}_{y,\mathbb{V}}, \hat{\boldsymbol{\Psi}}_{\mathbb{I}})$ and find the MAP solution of the GM parameters $\hat{\boldsymbol{\Theta}}_y$ of $p(\boldsymbol{\Theta}_y|\mathbf{y}, \boldsymbol{\mathcal{X}}, \hat{\boldsymbol{\Psi}}_{y,\mathbb{V}}, \hat{\boldsymbol{\Psi}}_{\mathbb{I}})$. Estimating the GM parameters under the Bayesian framework is called *Bayesian General Model (Bayes-GM)*.



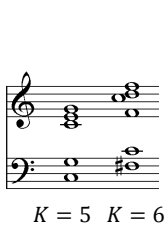
Data set and experimental setup

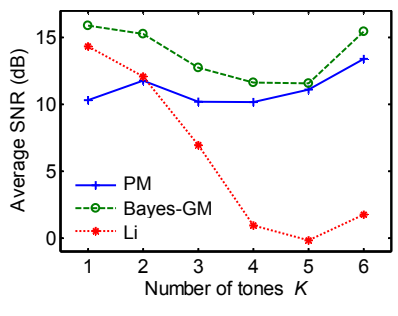
- The data set contains 25 mixtures. Each mixture was generated by mixing the isolated tones in the recorded piano databases (RWC and [SW13b]), taken from 4 different pianos.
- The pitches in each mixture correspond to a chord randomly selected from 11 piano pieces in the RWC database.
- The number of tones (represented by K) in the mixtures ranges from 1 to 6.
  - 1 tone (8 mixtures), 2 tones (6), 3 tones (5), 4 tones (4), 5 tones (1) and 6 tones (1).
- These 25 mixtures consist of 62 tones. 7 mixtures contain one pair of octaves, 2 ($K = 5$ and $K = 6$) contain 2 pairs of octaves.
- For the training data, two instances of each pitch are available.
- The first 0.5 seconds of the mixtures and the training data were used.
- SNR $= 10\log_{10}\left(\sum_n x_k(t_n)^2 / \sum_n (x_k(t_n) - \hat{x}_k(t_n))^2\right)$
- Demo: http://www.cse.cuhk.edu.hk/~khwong/www2/conference/ismir2015/ismir2015.html

Results

- Evaluation on modeling quality
  - The quality to represent an isolated tone before mixing.
  - Average SNRs of PM and Bayes-GM: 11.15 dB and 17.38 dB. Average SNR: Bayes-GM is much higher than PM.
- Comparing with other systems for separation quality
  - Li's system [LWW09] assumes that the amplitude envelope of each partial from the same note tends to be similar (known as common amplitude modulation (CAM)).
  - Both PM and Bayes-GM outperform Li's system.
  - A significant improvement is in the octave cases.

	Average SNR (dB)		
	PM	Bayes-GM	Li
All mixtures	10.88	13.51	6.63
$2 < K < 6$	10.97	13.15	5.40
$K = 5$	11.08	11.57	-0.17
$K = 6$	13.39	15.43	1.74
Upper tones in octaves	10.95	12.77	1.57





Selected bibliography

- [LWW09] Y. Li, J. Woodruff, & D. Wang. Monaural musical sound separation based on pitch and common amplitude modulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1361–1371, 2009.
- [SW13a] W. M. Szeto & K. H. Wong. Sinusoidal modeling for piano tones. In *2013 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC 2013)*, Kunming, China, Aug 2013.
- [SW13b] W. M. Szeto & K. H. Wong. Source separation and analysis of piano music signals using instrument-specific sinusoidal model. In *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland, Sep 2013.

ISMIR